Decomposing PE

- Expected prediction error/ Mean squared error

$$\text{MSE} = \text{E(PE)} = \text{E} \| Y_{\text{new}} - \hat{Y}_{\text{new}} \|^2$$

- We have

$$\begin{aligned}\text{MSE} &= \| \text{E}(Y_{\text{new}}) - \text{E}(\hat{Y}_{\text{new}}) \|^2 + \text{tr}\{\text{var}(Y_{\text{new}} - \hat{Y}_{\text{new}})\} \\ &= \text{bias}^2 + \text{variance}\end{aligned}$$

  - $\hat{Y}_{\text{new}}$ is from old (training) data.
  - Y_{new} is from new data.
    - When independent, variance $= \text{tr}\{\text{var}(\epsilon_{\text{new}}) + \text{var}(\hat{Y}_{\text{new}})\}$
    - $\text{tr}\{\text{var}(\epsilon_{\text{new}})\}$ is the irreducible variane while $\text{tr}\{\text{var}(\hat{Y}_{\text{new}})\}$ depends on model.

Bias-variance trade-off

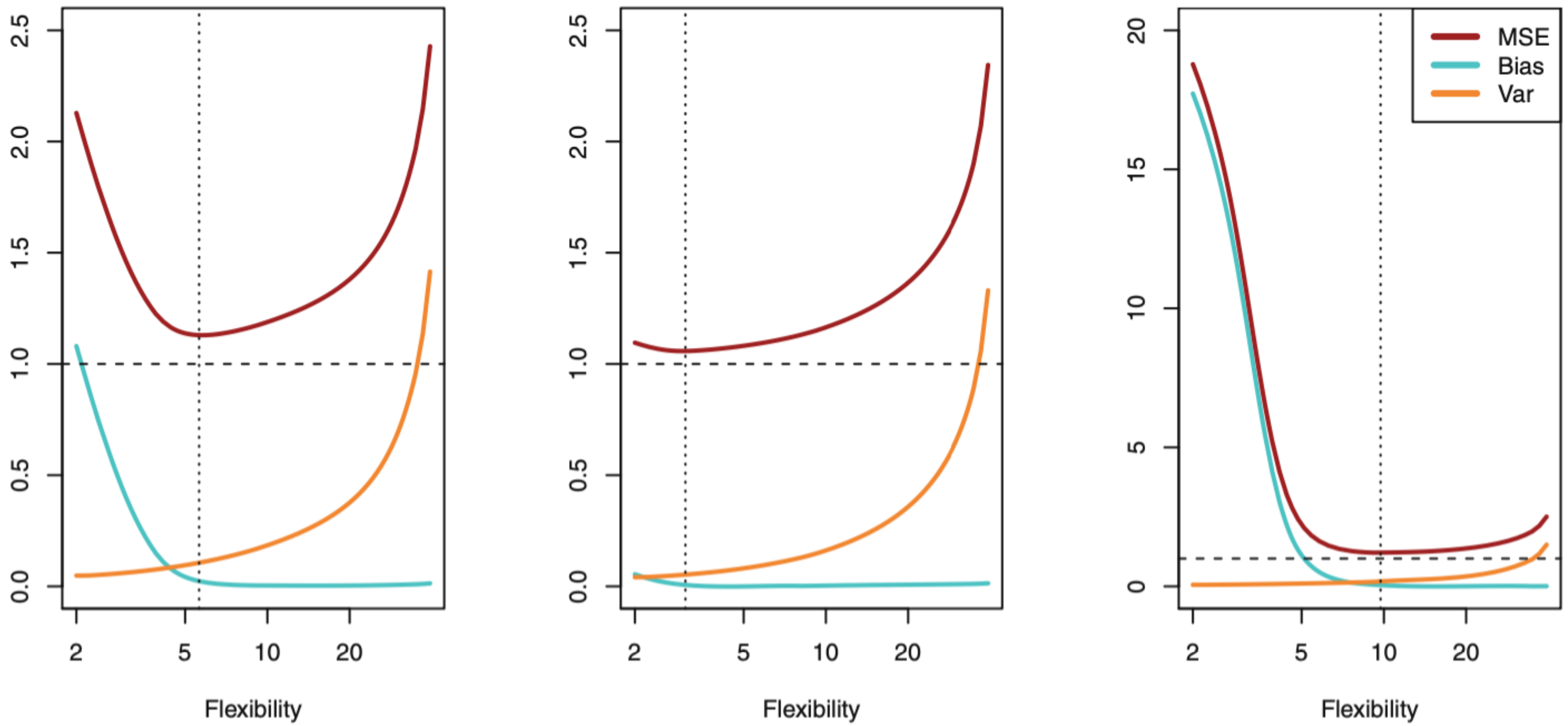


Figure 1: Figure from "An Introduction to Statistical Learning".

- It is possible to find a model with lower MSE than an unbiased model!
- Bias-variance trade-off is "generic" in statistics: almost always introducing some bias yields a decrease in MSE.

Bias-variance trade-off

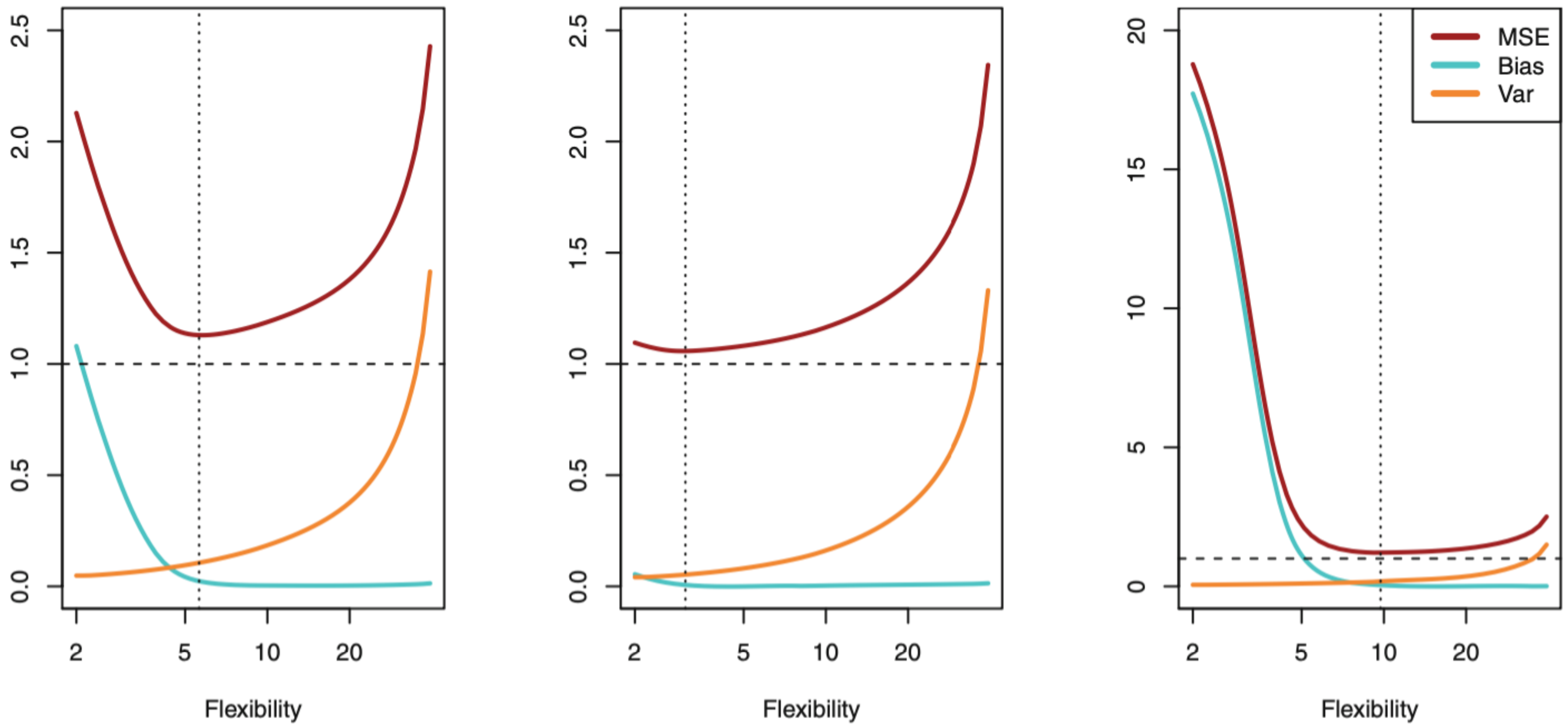


Figure 1: Figure from "An Introduction to Statistical Learning".

- It is possible to find a model with lower MSE than an unbiased model!
- Bias-variance trade-off is "generic" in statistics: almost always introducing some bias yields a decrease in MSE.

Bias-variance trade-off

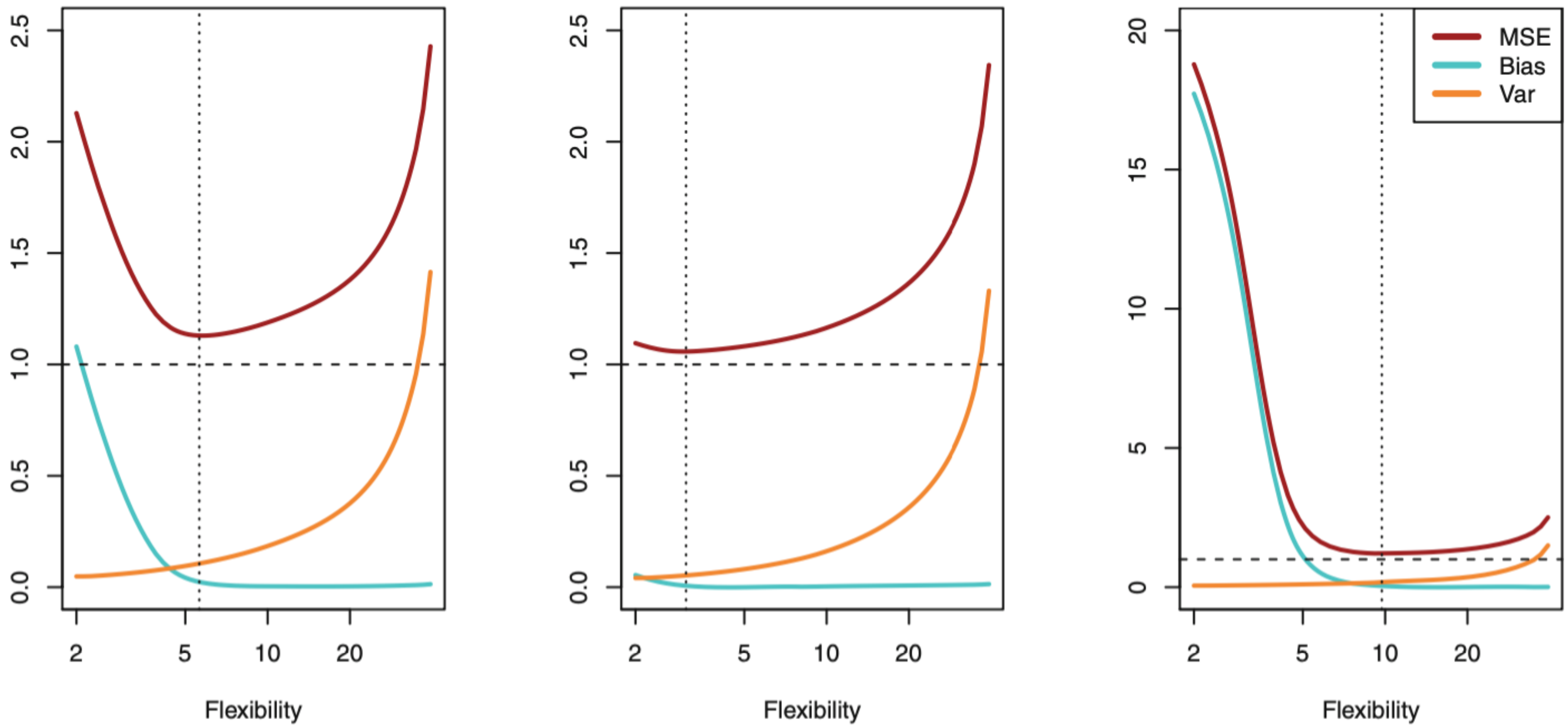


Figure 1: Figure from "An Introduction to Statistical Learning".

- It is possible to find a model with lower MSE than an unbiased model!
- Bias-variance trade-off is "generic" in statistics: almost always introducing some bias yields a decrease in MSE.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.
   - Unbiased estiamte.
   - But $\|\mathbf{Z}\|^2$ tends to be too large.
     - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
     - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.
3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte. $\Rightarrow E(Z) = \mu$
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Proof:

$$E\|z\|^2 = \sum_{i=1}^{p} E(z_i^2)$$

$$= \sum_{i=1}^{p} \left[\{E(z_i)\}^2 + var(z_i) \right]$$

$$= \sum_{i=1}^{p} (\mu_i^2 + \sigma^2)$$

$$= \|\mu\|^2 + p \times \sigma^2$$

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased. $E(cZ) \neq \mu$
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

$$
\begin{aligned}
&MSE(c) \\
&= E\left(\|cZ - \mu\|^2\right) \\
&= \|E(cZ-\mu)\|^2 + tr\{var(cZ-\mu)\} \quad \text{By } E\|a\|^2 = \|E(a)\|^2 + tr\{var(a)\} \\
&= \|cE(Z)-\mu\|^2 + tr\{var(cZ)\} \\
&= (c-1)^2\|\mu\|^2 + c^2\, tr(var(Z)) \\
&= (c-1)^2\|\mu\|^2 + c^2 \times tr(\sigma^2 I_p) = (c-1)^2\|\mu\|^2 + c^2 p \sigma^2
\end{aligned}
$$

Quadratic function in terms of c

$$
\frac{\partial MSE(c)}{\partial c} = 2(c-1)\|\mu\| + 2cp\sigma^2 = 0
$$

$$
\Rightarrow \hat{c} = \frac{\|\mu\|^2}{p\sigma^2 + \|\mu\|^2} \in (0,1) \Rightarrow \hat{c}Z \text{ achieves minimum of MSE}
$$

Shrinkage and Penalty

- Corresponds to

$$\text{minimize}_{\mu} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 + \lambda \times \|\boldsymbol{\mu}\|^2$$

- This is also Lagrange form of the "constrained" minimization.

$$\text{minimize}_{\mu} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 \qquad \text{subject to } \; \|\boldsymbol{\mu}\|^2 \leqslant C$$

  - For any λ, there is some C such that the solutions of two problems are the same, and vice versa.
  - Intuitively, constrains $\|$ minimizer $\|^2$ not too large.
    - If $C = \infty$ or $\lambda = 0$, solution is OLS.
    - As C gets smaller, λ gets larger, find solution subject to the constraint $\|\boldsymbol{\mu}\|^2 \leqslant C$.

minimize $L(\mu) = \| Z - \mu \|^2 + \lambda \| \mu \|^2$

Quadratic in μ.

$$\frac{\partial L(\mu)}{\partial \mu} = -2(Z-\mu) + 2\lambda\mu = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{1+\lambda} Z \quad \text{(A shrinked estimator)}$$

If choose λ such that $\frac{1}{1+\lambda} = \hat{c}$

then solution $\frac{1}{1+\lambda} Z = \hat{c} Z$

Shrinkage and Penalty

- Corresponds to

$$\text{minimize}_{\boldsymbol{\mu}} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 + \lambda \times \|\boldsymbol{\mu}\|^2$$

- This is also Lagrange form of the "constrained" minimization.

$$\text{minimize}_{\boldsymbol{\mu}} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 \qquad \text{subject to } \|\boldsymbol{\mu}\|^2 \leqslant C$$

  - For any λ, there is some C such that the solutions of two problems are the same, and vice versa.
  - Intuitively, constrains $\|$ minimizer $\|^2$ not too large.
    - If $C = \infty$ or $\lambda = 0$, solution is OLS.
    - As C gets smaller, λ gets larger, find solution subject to the constraint $\|\boldsymbol{\mu}\|^2 \leqslant C$.

Shrinkage and Penalty

- Corresponds to

$$\text{minimize}_{\boldsymbol{\mu}} \|\mathbf{Z} - \boldsymbol{\mu}\|^2 + \lambda \times \|\boldsymbol{\mu}\|^2$$

- This is also Lagrange form of the "constrained" minimization.

$$\text{minimize}_{\boldsymbol{\mu}} \|\mathbf{Z} - \boldsymbol{\mu}\|^2 \qquad \text{subject to } \|\boldsymbol{\mu}\|^2 \leqslant C$$

  - For any λ, there is some C such that the solutions of two problems are the same, and vice versa.
  - Intuitively, constrains $\|$ minimizer $\|^2$ not too large.
    - If $C = \infty$ or $\lambda = 0$, solution is OLS.
    - As C gets smaller, λ gets larger, find solution subject to the constraint $\|\boldsymbol{\mu}\|^2 \leqslant C$.

Shrinkage and Penalty

- Corresponds to

$$\text{minimize}_{\mu} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 + \lambda \times \|\boldsymbol{\mu}\|^2$$

- This is also Lagrange form of the "constrained" minimization.

$$\text{minimize}_{\mu} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 \qquad \text{subject to } \|\boldsymbol{\mu}\|^2 \leqslant C$$

  - For any λ, there is some C such that the solutions of two problems are the same, and vice versa.
  - Intuitively, constrains $\|$ minimizer $\|^2$ not too large.
    - If $C = \infty$ or $\lambda = 0$, solution is OLS.
    - As C gets smaller, λ gets larger, find solution subject to the constraint $\|\mu\|^2 \leqslant C$.

Shrinkage and Penalty

- Corresponds to

$$\text{minimize}_{\mu} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 + \lambda \times \|\boldsymbol{\mu}\|^2$$

- This is also Lagrange form of the "constrained" minimization.

$$\text{minimize}_{\mu} \, \|\mathbf{Z} - \boldsymbol{\mu}\|^2 \qquad \text{subject to } \|\boldsymbol{\mu}\|^2 \leqslant C$$

  - For any λ, there is some C such that the solutions of two problems are the same, and vice versa.
  - Intuitively, constrains $\|$ minimizer $\|^2$ not too large.
    - If $C = \infty$ or $\lambda = 0$, solution is OLS.
    - As C gets smaller, λ gets larger, find solution subject to the constraint $\|\mu\|^2 \leqslant C$.

Shrinkage and Penalty

- Corresponds to

$$\text{minimize}_{\boldsymbol{\mu}} \|\mathbf{Z} - \boldsymbol{\mu}\|^2 + \lambda \times \|\boldsymbol{\mu}\|^2$$

- This is also Lagrange form of the "constrained" minimization.

$$\text{minimize}_{\boldsymbol{\mu}} \|\mathbf{Z} - \boldsymbol{\mu}\|^2 \qquad \text{subject to } \|\boldsymbol{\mu}\|^2 \leqslant C$$

  - For any λ, there is some C such that the solutions of two problems are the same, and vice versa.
  - Intuitively, constrains $\|$ minimizer $\|^2$ not too large.
    - If $C = \infty$ or $\lambda = 0$, solution is OLS.
    - As C gets smaller, λ gets larger, find solution subject to the constraint $\|\boldsymbol{\mu}\|^2 \leqslant C$.

[MS 6] Shrinkage Method for Model Selection

Ridge Regression

Motivation: Suppose $\mathbf{X}^\top \mathbf{X} = n\mathbf{I}_p$.

- $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2 \mathbf{I}_p)$.
- $\hat{\beta}$ has a shrinkaged version $\tau\hat{\beta}$ with smaller MSE.
- Ridge Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$$

  - Also corresponds to an $\|\beta\|^2$ constrained optimization.
  - Solution: $\hat{\beta}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}$

Ridge Regression

Motivation: Suppose $\mathbf{X}^\top\mathbf{X} = n\mathbf{I}_p$.

- $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2\mathbf{I}_p)$.
- $\hat{\beta}$ has a shrinkaged version $\tau\hat{\beta}$ with smaller MSE.
- Ridge Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$$

  - Also corresponds to an $\|\beta\|^2$ constrained optimization.
  - Solution: $\hat{\beta}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}$

Ridge Regression

Motivation: Suppose $\mathbf{X}^\top \mathbf{X} = n\mathbf{I}_p$.

- $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2 \mathbf{I}_p)$.
- $\hat{\beta}$ has a shrinkaged version $\tau\hat{\beta}$ with smaller MSE.
- Ridge Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$$

  - Also corresponds to an $\|\beta\|^2$ constrained optimization.
  - Solution: $\hat{\beta}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}$

Ridge Regression

Motivation: Suppose $\mathbf{X}^\top \mathbf{X} = n\mathbf{I}_p$.

- $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2 \mathbf{I}_p)$.
- $\hat{\beta}$ has a shrinkaged version $\tau\hat{\beta}$ with smaller MSE.
- Ridge Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$$

  - Also corresponds to an $\|\beta\|^2$ constrained optimization.
  - Solution: $\hat{\beta}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}$

Ridge Regression

Motivation: Suppose $\mathbf{X}^\top \mathbf{X} = n\mathbf{I}_p$.

- $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2 \mathbf{I}_p)$.
- $\hat{\beta}$ has a shrinkaged version $\tau\hat{\beta}$ with smaller MSE.
- Ridge Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$$

  - Also corresponds to an $\|\beta\|^2$ constrained optimization.
  - Solution: $\hat{\beta}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^\top \mathbf{Y}$

$$L(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|^2 \quad \text{(Quadratic in } \beta)$$

$$\frac{\partial L(\beta)}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta$$

$$= -2\left\{X^TY - (X^TX + \lambda I_p)\beta\right\} = 0$$

$$\Rightarrow \text{Solution:} \quad \hat{\beta}_\lambda = (X^TX + \lambda I_p)^{-1}X^TY$$

If $X^TX = I_p$

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY = X^TY$$

$$\hat{\beta}_\lambda = \left\{(1+\lambda)I_p\right\}^{-1}X^TY$$

$$= \frac{1}{1+\lambda}X^TY = \frac{1}{1+\lambda}\hat{\beta}_{OLS}$$

Shrink each $\hat{\beta}_{i,OLS}$ by $\frac{1}{1+\lambda} \in (0,1)$ with $\lambda > 0$

- Ridge Regression will include all p predictors in the final model.
- The penalty $\lambda ||\beta||^2$
  - will shrink all of the coefficients towards zero
  - but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - may not be a problem for prediction accuracy
  - can create a challenge in model interpretation if p is too large

- Ridge Regression will include all p predictors in the final model.
- The penalty $\lambda\|\beta\|^2$
  - will shrink all of the coefficients towards zero
  - but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - may not be a problem for prediction accuracy
  - can create a challenge in model interpretation if p is too large

- Ridge Regression will include all p predictors in the final model.
- The penalty $\lambda\|\beta\|^2$
  - will shrink all of the coefficients towards zero
  - but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - may not be a problem for prediction accuracy
  - can create a challenge in model interpretation if p is too large

- Ridge Regression will include all *p* predictors in the final model.
- The penalty $\lambda \|\beta\|^2$
  - will shrink all of the coefficients towards zero
  - but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - may not be a problem for prediction accuracy
  - can create a challenge in model interpretation if *p* is too large

- Ridge Regression will include all p predictors in the final model.
- The penalty $\lambda\|\beta\|^2$
  - will shrink all of the coefficients towards zero
  - but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - may not be a problem for prediction accuracy
  - can create a challenge in model interpretation if p is too large

Lasso Regression

Lasso Regression:

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$
- Also corresponds to an $\|\beta\|_1$ constrained optimization.
- Lasso can zero some coefficients.
  - If $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ and $\lambda = 2\gamma$, lasso solution

$$\tilde{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j) \times (|\hat{\beta}_j| - \gamma), & \gamma \leq |\hat{\beta}_j|, \\ 0, & \text{otherwise} \end{cases}$$

Lasso Regression

Lasso Regression:

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$
- Also corresponds to an $\|\beta\|_1$ constrained optimization.
- Lasso can zero some coefficients.
  - If $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ and $\lambda = 2\gamma$, lasso solution

$$\tilde{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j) \times (|\hat{\beta}_j| - \gamma), & \gamma \leq |\hat{\beta}_j|, \\ 0, & \text{otherwise} \end{cases}$$

Lasso Regression

Lasso Regression:

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$
- Also corresponds to an $\|\boldsymbol{\beta}\|_1$ constrained optimization.
- Lasso can zero some coefficients.
  - If $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ and $\lambda = 2\gamma$, lasso solution

$$\tilde{\beta}_j = \begin{cases} \operatorname{sign}(\hat{\beta}_j) \times (|\hat{\beta}_j| - \gamma), & \gamma \leq |\hat{\beta}_j|, \\ 0, & \text{otherwise} \end{cases}$$

Lasso Regression

Lasso Regression:

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$
- Also corresponds to an $\|\boldsymbol{\beta}\|_1$ constrained optimization.
- Lasso can zero some coefficients.
  - If $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$ and $\lambda = 2\gamma$, lasso solution

$$\tilde{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j) \times (|\hat{\beta}_j| - \gamma), & \gamma \leq |\hat{\beta}_j|, \\ 0, & \text{otherwise} \end{cases}$$

Lasso Regression

Lasso Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$
- Also corresponds to an $\|\boldsymbol{\beta}\|_1$ constrained optimization.
- Lasso can zero some coefficients.
  - If $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ and $\lambda = 2\gamma$, lasso solution

$\hat{\beta}$ denotes OLS

$$\tilde{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j) \times (|\hat{\beta}_j| - \gamma), & \gamma \le |\hat{\beta}_j|, \\ 0, & \text{otherwise} \end{cases}$$

Therefore, as λ gets larger but not ∞, $\tilde{\beta}_i$ can be 0

When $X^TX = I_p$. $\quad \hat{\beta}_{OLS} = (X^TX)^{-1}X^TY = X^TY$

Then $\min\limits_{\beta} \|Y - X\beta\|^2 + \lambda\|\beta\|_1$

$$\iff \min_{\beta} \underbrace{Y^TY}_{\text{doesn't involve } \beta} - 2\underbrace{Y^TX}_{\hat{\beta}_{OLS}^T}\beta + \beta^T\underbrace{X^TX}_{X^TX = I_p}\beta + \lambda\|\beta\|_1$$

$$\iff \min_{\beta} -2\hat{\beta}_{OLS}^T\beta + \beta^T\beta + \lambda\|\beta\|_1$$

$$\iff \min_{\beta} \sum_{i=1}^{p} \left(\beta_i^2 - 2\hat{\beta}_{i,OLS}\beta_i + \lambda|\beta_i|\right)$$

$\iff$ For each $i = 1 \cdots p$

$$\min_{\beta_i \geq 0} \beta_i^2 - 2\hat{\beta}_{i,OLS}\beta_i + \lambda\beta_i \qquad \circledast$$

$$\min_{\beta_i \leq 0} \beta_i^2 - 2\hat{\beta}_{i,OLS}\beta_i - \lambda\beta_i$$

Step 1:

Claim: If $\hat{\beta}_{i,OLS} > 0$, to minimize above objective $\circledast$

then solution $\tilde{\beta}_i \geq 0$ (non-negative)

If $\hat{\beta}_{i.OLS} < 0$, then $\tilde{\beta}_i \leq 0$.

Proof: Suppose solution $\tilde{\beta}_i < 0$,
$$\begin{cases} \hat{\beta}_{i.OLS}\tilde{\beta}_i < \hat{\beta}_{i.OLS}(-\tilde{\beta}_i) & (\hat{\beta}_{i.OLS} > 0) \\ \lambda\tilde{\beta}_i < -\lambda\tilde{\beta}_i & (\lambda > 0) \end{cases}$$

Thus
$$\tilde{\beta}_i^2 - 2\hat{\beta}_{i.OLS}\tilde{\beta}_i - \lambda\tilde{\beta}_i > \tilde{\beta}_i^2 - 2\hat{\beta}_{i.OLS}(-\tilde{\beta}_i) + \lambda\tilde{\beta}_i$$

showing $-\tilde{\beta}_i$ would achieve smaller value in $\circledast$

contradicts with $\tilde{\beta}_i$ is the solution.

Therefore $\tilde{\beta}_i$ shouldn't be negative.

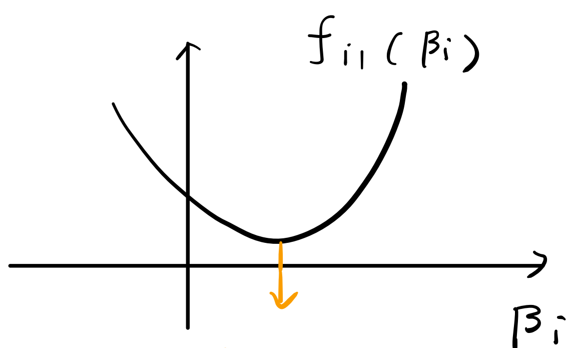
Step 2: Minimizer in the domain $\beta_i \geq 0$

Let $f_{i1}(\beta_i) := \beta_i^2 - (2\hat{\beta}_{i,OLS} - \lambda)\beta_i$

$$= \left\{\beta_i - \left(\hat{\beta}_{i,OLS} - \frac{\lambda}{2}\right)\right\}^2 - \underbrace{\left(\hat{\beta}_{i,OLS} - \frac{\lambda}{2}\right)^2}_{\text{not change with } \beta}$$

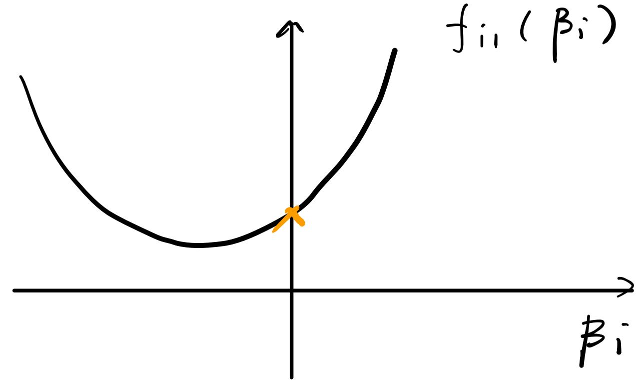
$$= \left\{\beta_i - (\hat{\beta}_{i,OLS} - r)\right\}^2 + \text{fixed terms} \qquad (\lambda = 2r)$$

① If $\hat{\beta}_{i,OLS} - r \geq 0$



minimizer $\hat{\beta}_{i,OLS} - r$

② If $\hat{\beta}_{i,OLS} - r \leq 0$



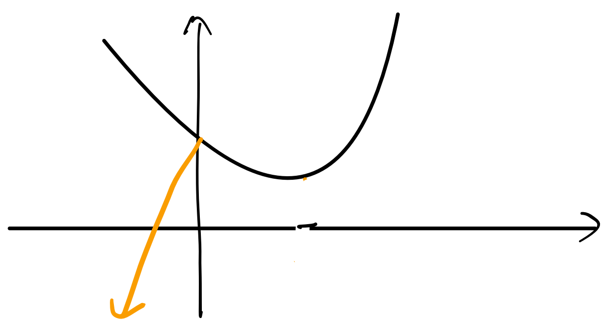
minimizer 0

Step 3: minimizer in the domain $\beta_i \le 0$

Let $f_{i2}(\beta_i) := \beta_i^2 - (2\hat{\beta}_{i,OLS} + \lambda)\beta_i$

$$= \{\beta_i - (\hat{\beta}_{i,OLS} + r)\}^2 + \text{fixed terms}$$

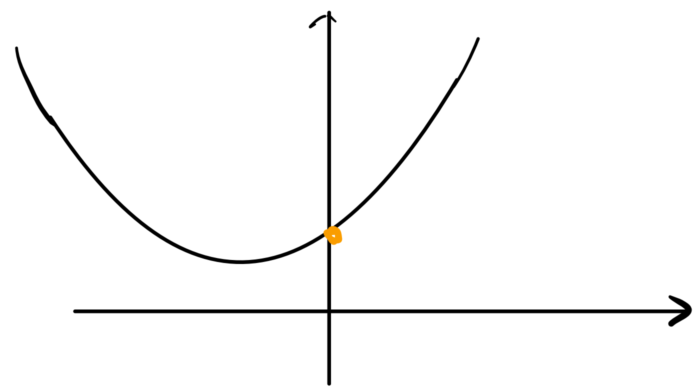
① If $\hat{\beta}_{i,OLS} + r \ge 0$



minimizer 0

(if $\beta_i \le 0$)

② If $\hat{\beta}_{i,OLS} + r < 0$



minimizer $\hat{\beta}_{i,OLS} + r$

(if $\beta_i \le 0$)

Step 4: In summary

① If $\hat{\beta}_{i.OLS} \geq 0$, solution over $\beta_i \geq 0$ gives

$$\widehat{\beta}_i = \begin{cases} \hat{\beta}_{i.OLS} - r & \text{if } \hat{\beta}_{i.OLS} - r \geq 0; \\ 0 & \text{if } \hat{\beta}_{i.OLS} - r < 0. \end{cases}$$

In this case

$\hat{\beta}_{i,OLS} - r = |\hat{\beta}_{i.OLS}| - r$

② If $\hat{\beta}_{i,OLS} < 0$, solution over $\beta_i < 0$ gives

$$\widehat{\beta}_i = \begin{cases} -(\hat{\beta}_{i.OLS} + r) & \text{if } \hat{\beta}_{i.OLS} + r < 0; \\ 0 & \text{if } \hat{\beta}_{i.OLS} + r \geq 0. \end{cases}$$

In this case,

$\hat{\beta}_{i.OLS} + r = -(|\hat{\beta}_{i.OLS}| - r)$

Thus,

$$\widehat{\beta}_i = \begin{cases} \text{sign}(\hat{\beta}_{i.OLS}) \times (|\hat{\beta}_{i.OLS}| - r) & \text{if } |\hat{\beta}_{i.OLS}| - r \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

Graph Illustration

- Consider $p = 2$.
- The solid blue areas are the constraint regions $|\beta_1|^2 + |\beta_2|^2 \leqslant C$ and $|\beta_1| + |\beta_2| \leqslant C$
- The red ellipses given regions of constant RSS.

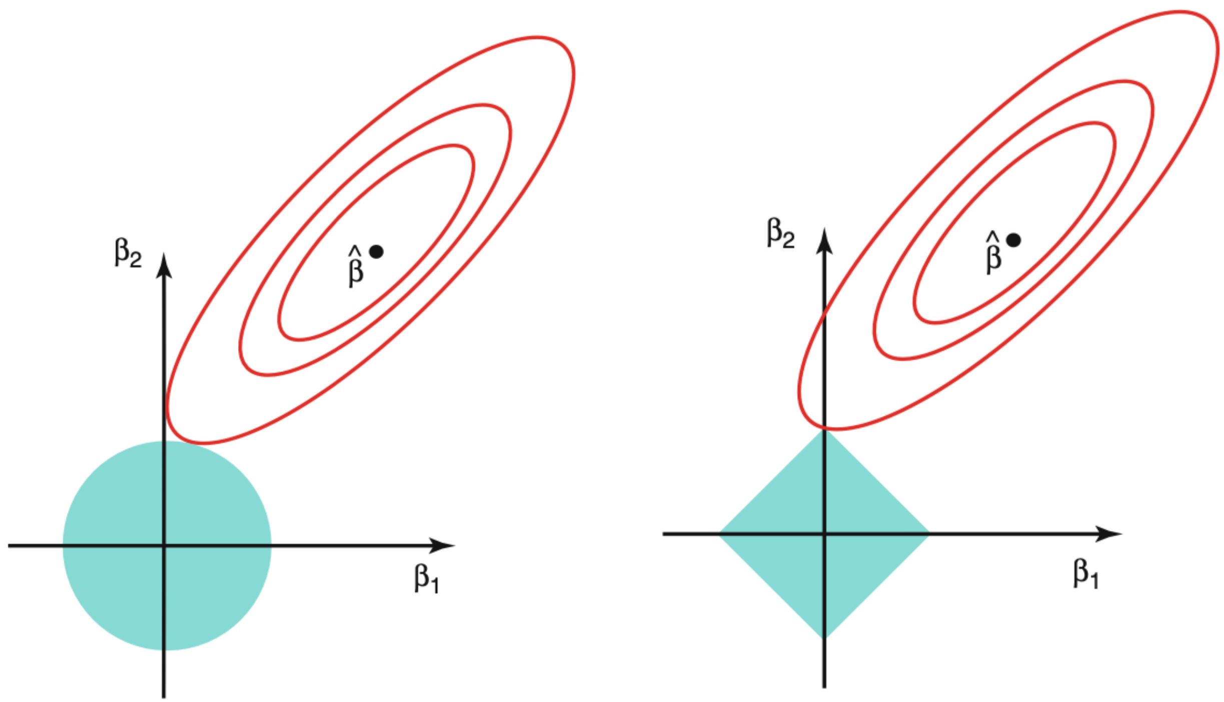


Figure 2: From "An Introduction to Statistical Learning".

Graph Illustration

- Consider $p = 2$.
- The solid blue areas are the constraint regions $|\beta_1|^2 + |\beta_2|^2 \leqslant C$ and $|\beta_1| + |\beta_2| \leqslant C$
- The red ellipses given regions of constant RSS.

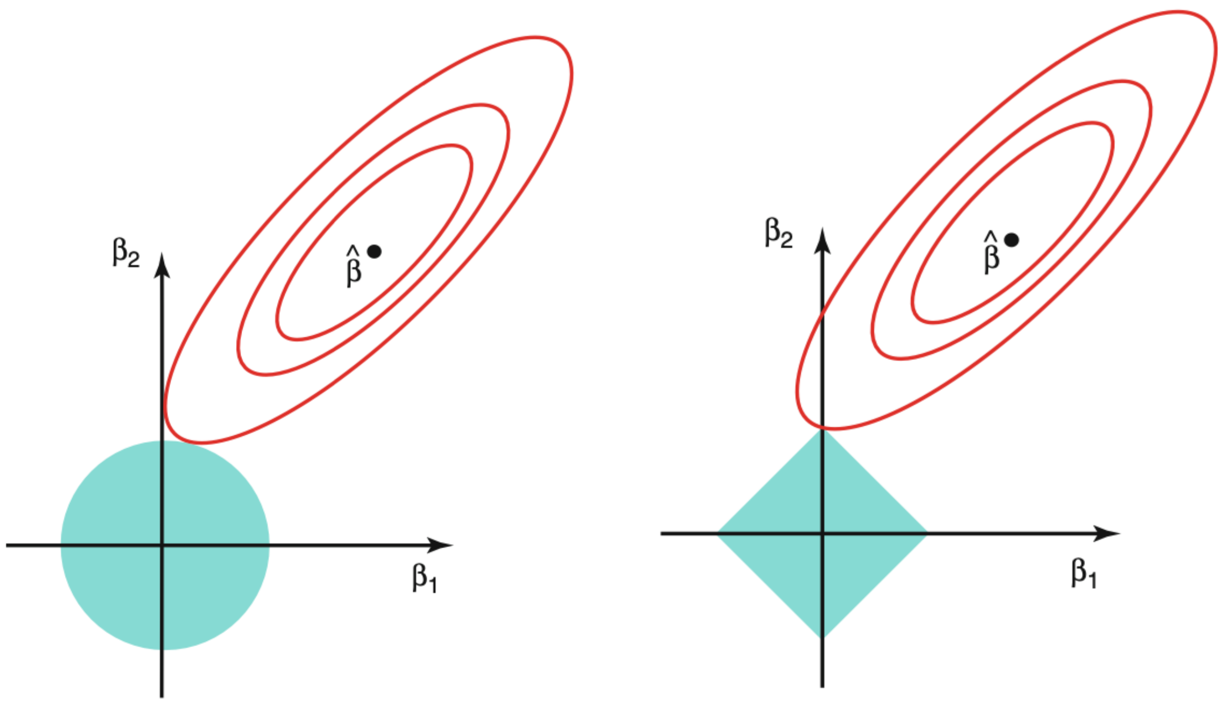


Figure 2: From "An Introduction to Statistical Learning".

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Example

- Hitters Data: Records and salaries for baseball players.

```
Hitters=na.omit(Hitters)
head(Hitters,2)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns
## -Alan Ashby         315   81     7   24  38    39    14   3449   835     69   321
## -Alvin Davis        479  130    18   66  72    76     3   1624   457     63   224
##                   CRBI CWalks League Division PutOuts Assists Errors Salary
## -Alan Ashby        414    375      N        W     632      43     10    475
## -Alvin Davis       266    263      A        W     880      82     14    480
##                   NewLeague
## -Alan Ashby               N
## -Alvin Davis              A
```

```
x=model.matrix(Salary ~ ., Hitters)[,-1]
y=Hitters$Salary
```

- In glmnet() function: alpha option determines the model type.
  - alpha = 0 ridge; alpha = 1 lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

```
ridge.mod$lambda[60] #//beta//^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

- In glmnet() function: alpha option determines the model type.
  - alpha $= 0$ ridge; alpha $= 1$ lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

```
ridge.mod$lambda[60] #//beta//^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

- In glmnet() function: alpha option determines the model type.
  - alpha $= 0$ ridge; alpha $= 1$ lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

```
ridge.mod$lambda[60] #||beta||^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

- In glmnet() function: alpha option determines the model type.
  - alpha $=0$ ridge; alpha $=1$ lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

```
ridge.mod$lambda[60] #//beta//^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

- In glmnet() function: alpha option determines the model type.
  - alpha $= 0$ ridge; alpha $= 1$ lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

```
ridge.mod$lambda[60] #//beta//^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

- In glmnet() function: alpha option determines the model type.
  - alpha = 0 ridge; alpha = 1 lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

```
ridge.mod$lambda[60] #||beta||^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

- In glmnet() function: alpha option determines the model type.
  - alpha $= 0$ ridge; alpha $= 1$ lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

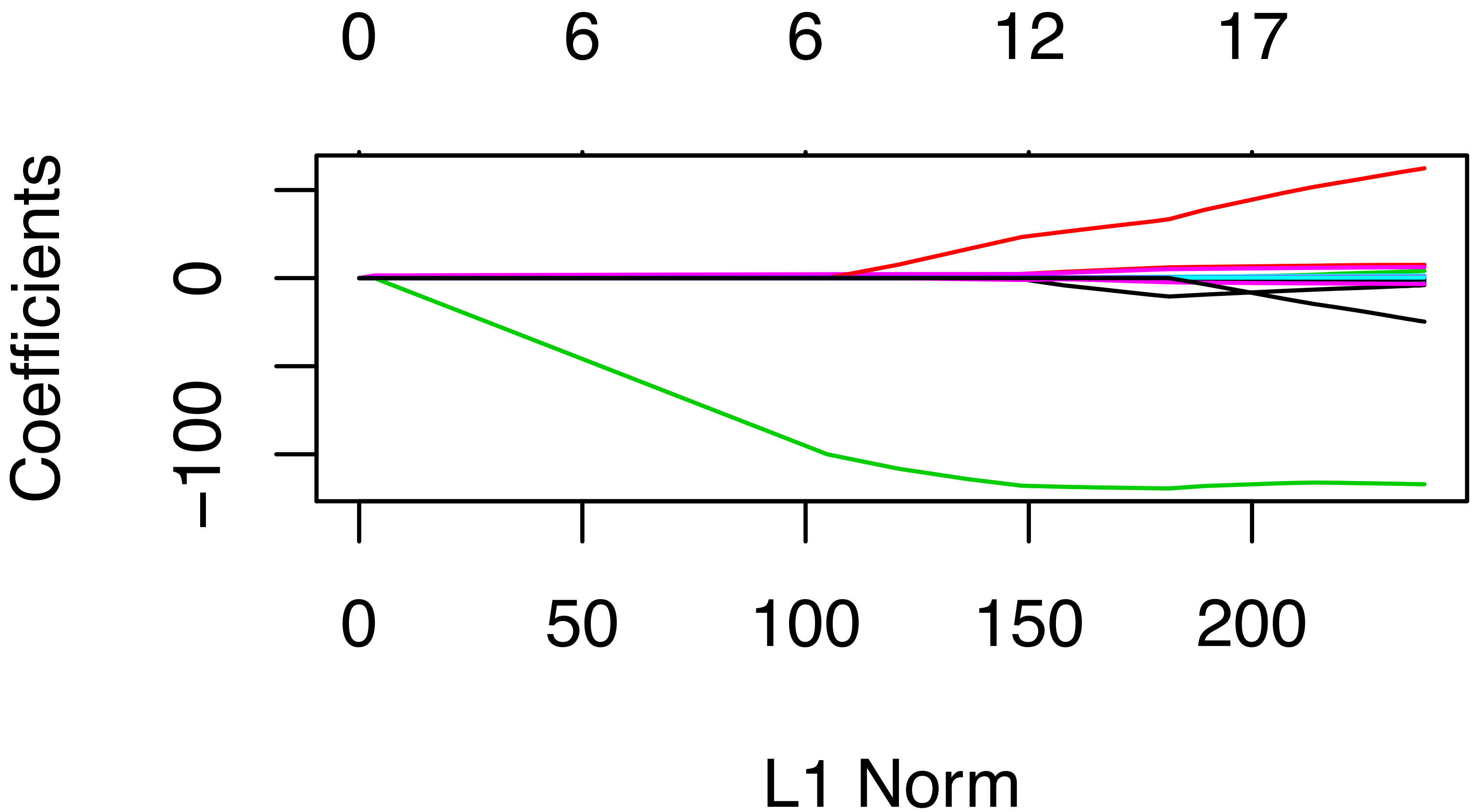
```
ridge.mod$lambda[60] #||beta||^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

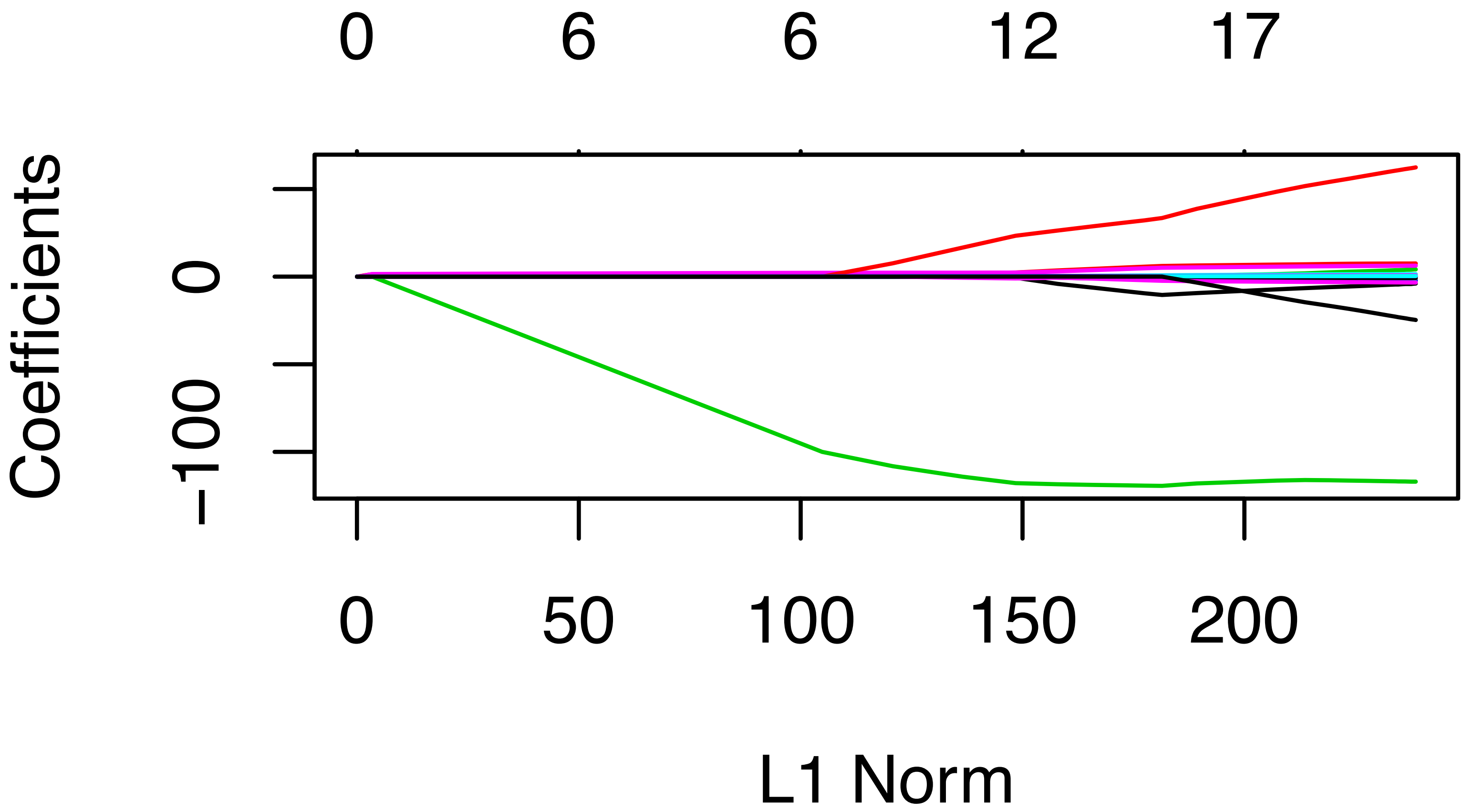
```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

```r
lasso.mod=glmnet(x,y,alpha=1,lambda=grid)
plot(lasso.mod)
```



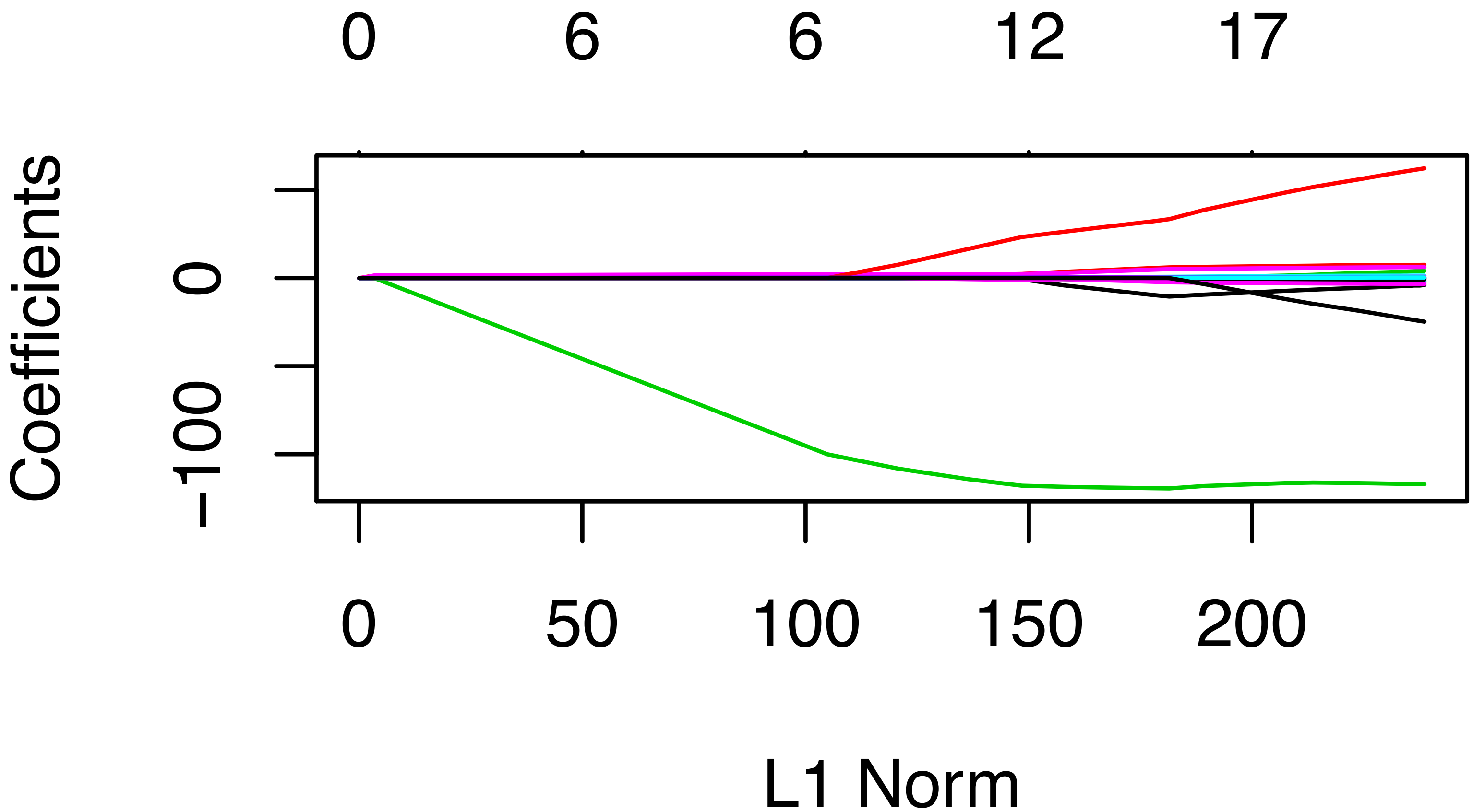
- Each curve corresponds to a variable.
- It shows the path of its coefficient against the $\|\hat{\beta}\|_1$.
- The axis above indicates # of nonzero coefficients at the current λ.

```r
lasso.mod=glmnet(x,y,alpha=1,lambda=grid)
plot(lasso.mod)
```



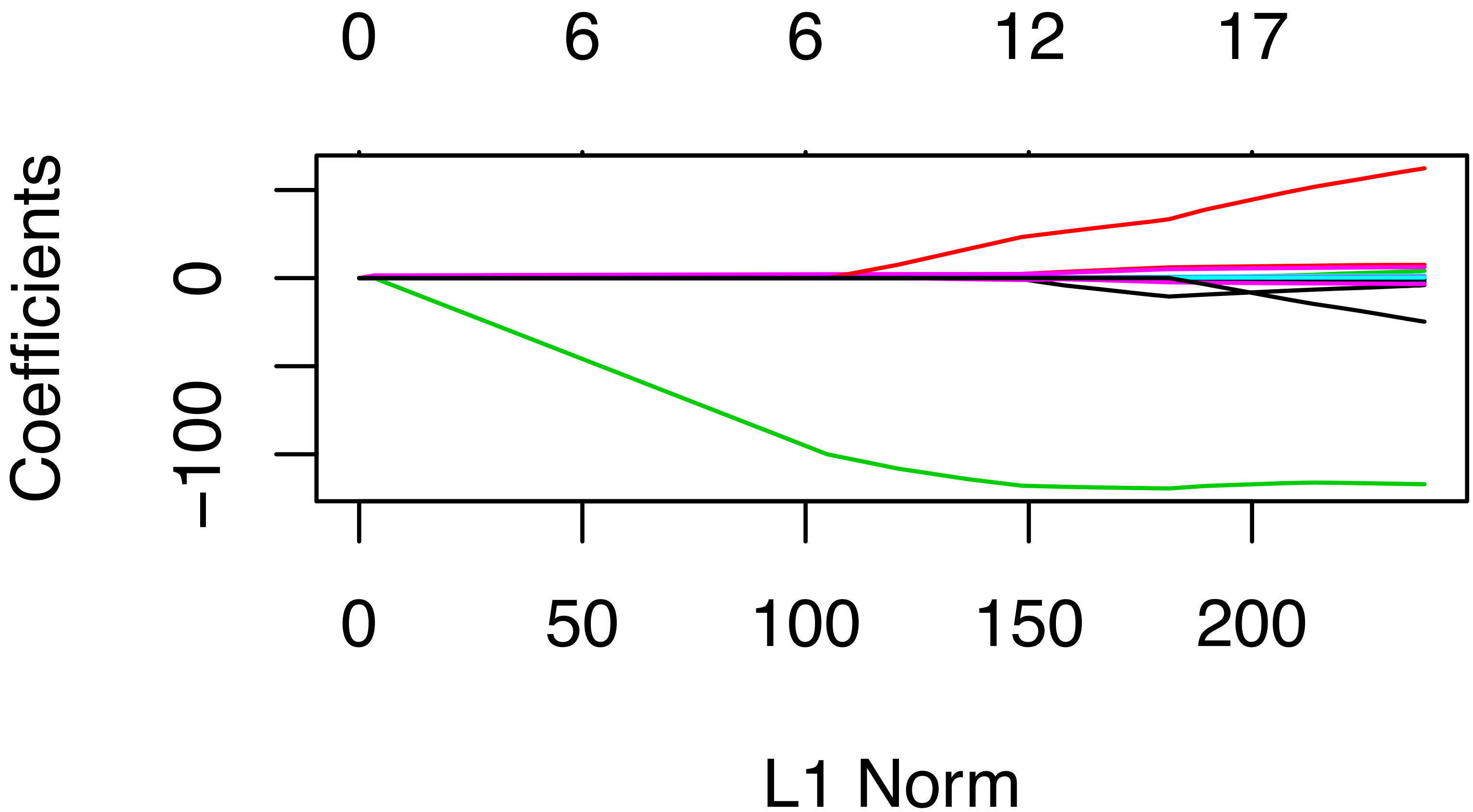
- Each curve corresponds to a variable.
- It shows the path of its coefficient against the $\|\hat{\beta}\|_1$.
- The axis above indicates # of nonzero coefficients at the current λ.

```r
lasso.mod=glmnet(x,y,alpha=1,lambda=grid)
plot(lasso.mod)
```



- Each curve corresponds to a variable.
- It shows the path of its coefficient against the $\|\hat{\beta}\|_1$.
- The axis above indicates # of nonzero coefficients at the current λ.

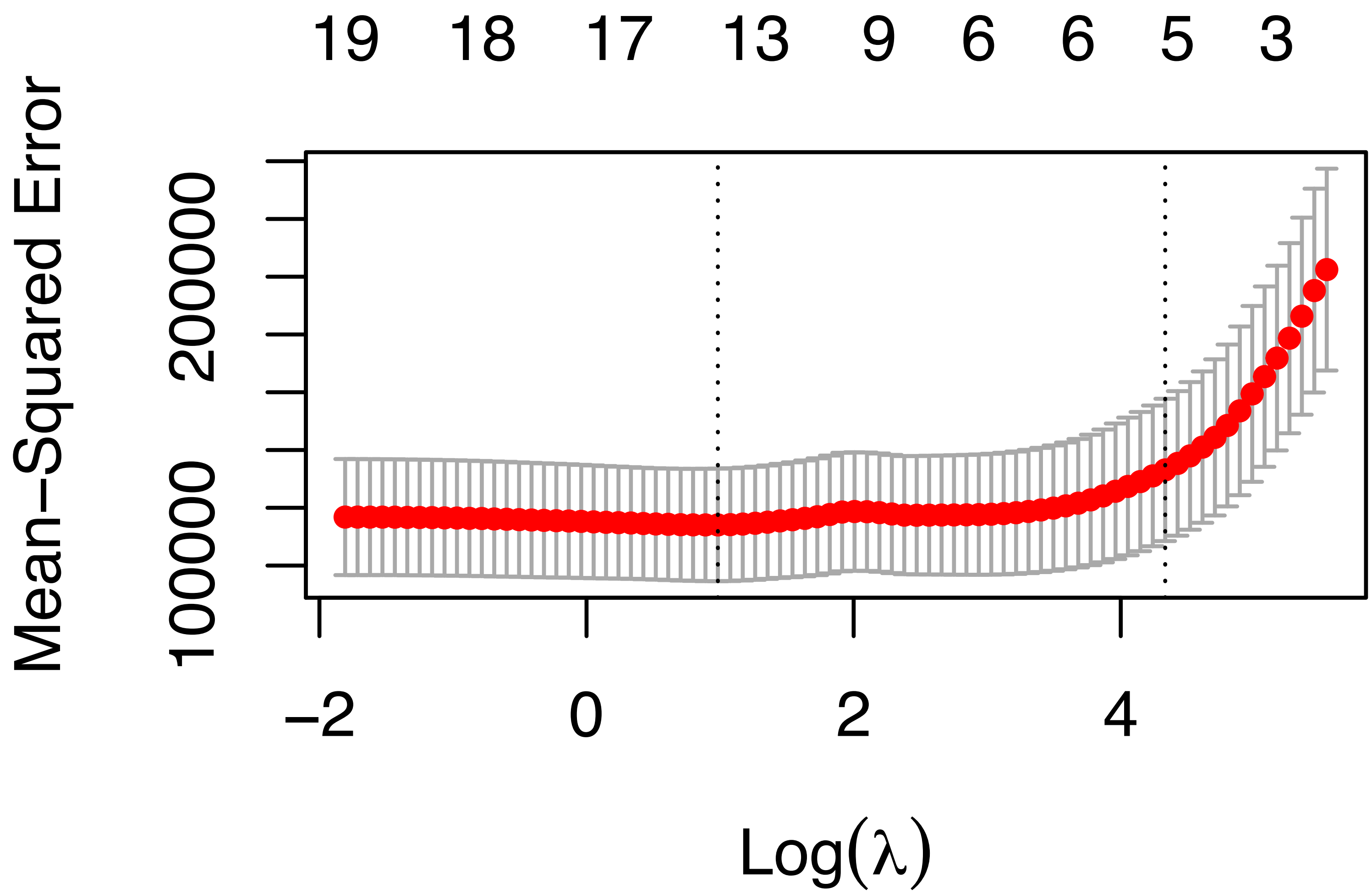
```
lasso.mod=glmnet(x,y,alpha=1,lambda=grid)
plot(lasso.mod)
```



- Each curve corresponds to a variable.
- It shows the path of its coefficient against the $\|\hat{\beta}\|_1$.
- The axis above indicates $\#$ of nonzero coefficients at the current λ.

Cross validaiton

```
cv.out <- cv.glmnet(x, y, alpha=1) #default # of folds is 10
plot(cv.out)
```

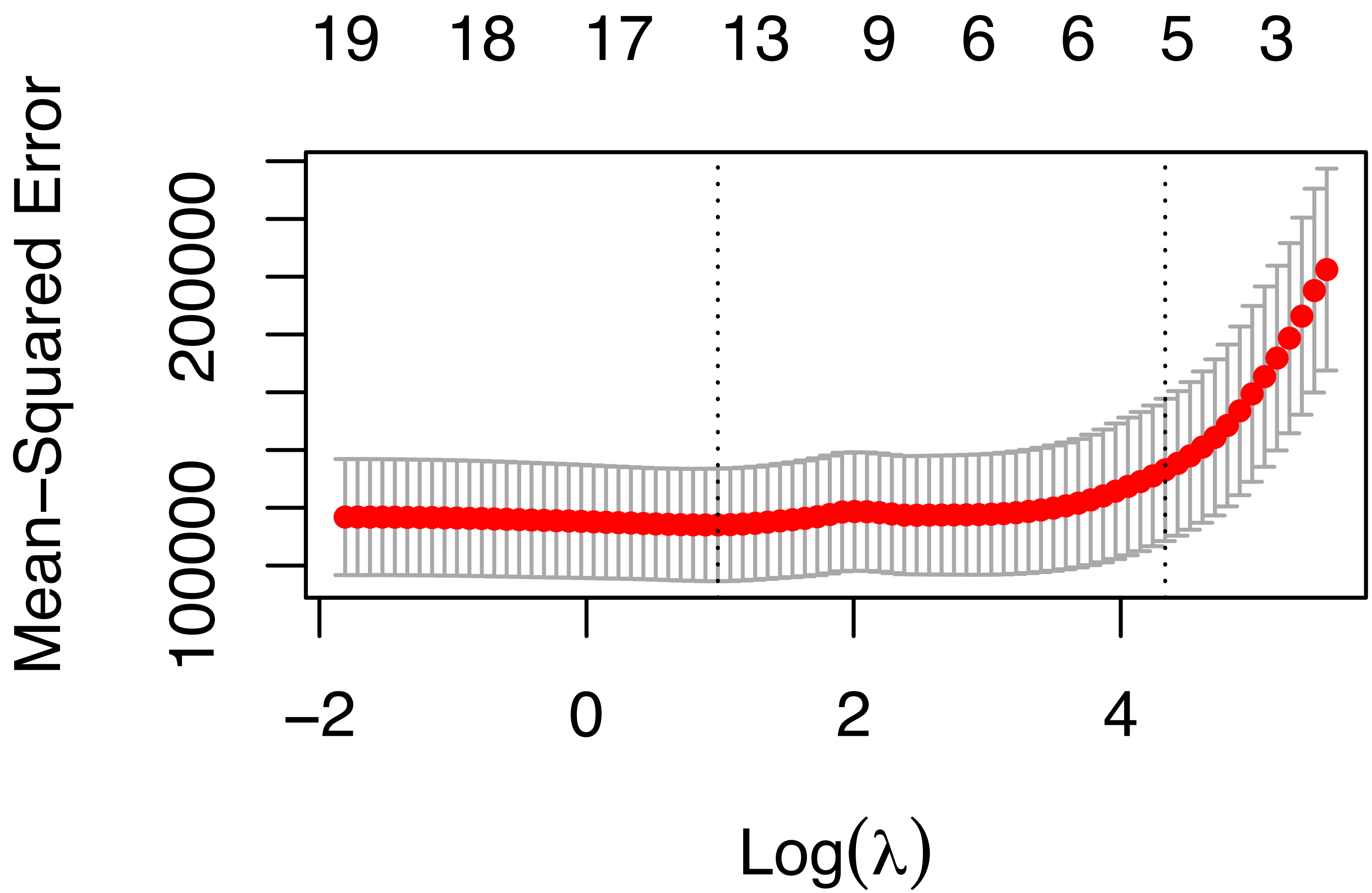



```
cv.out$lambda.min
```

```
## [1] 2.674375
```

Cross validaiton

```
cv.out <- cv.glmnet(x, y, alpha=1) #default # of folds is 10
plot(cv.out)
```




```
cv.out$lambda.min
```

```
## [1] 2.674375
```