Stat 577 "Statistical Learning Theory"

T2: Non-regularized regression methods

Xiongzhi Chen

Washington State University

Overview and Some Questions

Overview

Topics to cover

1. Regression with multivariate response
2. Regularized linear models
3. Other regularization techniques

Some questions: I

1. Let $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$. Let tr be the trace operator for matrices. Do you know that

$$\text{tr}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) = n - (p+1) \tag{1}$$

when $\text{rank}(\mathbf{X}) = p + 1$? Note $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is often called the "hat matrix" in statistics.

2. What is your understanding of the term "degrees of freedom"? What is the degrees of freedom for

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ is the fitted value for y_i?

Some questions: II

1. What is the purpose of regularization? What properties does an optimizer of the following regularization

$$R_\lambda(f) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx$$

   have?
2. Can you explain what properties of f each of these controls: $\int [f''(x)]^2 dx$, $\int [f'(x)]^2 dx$ and $\int [f(x)]^2 dx$?
3. What is a general principle for variable/model selection?

Some questions: III

1. What is the key difficulty for Subset Selection?
2. Why cannot Backward Stepwise selection be applied when $N < p$, i.e., when the sample size is smaller than the number of predictors?
3. Among Best Subset selection, Forward-stepwise selection, and backward-stepwise selection, which produces a sequence of nested model?

Some questions: IV

1. Why in Figure 3.6 the mean-squared errors of the estimated coefficient are close for Best Subset, Forward Stepwise and Backward Stepwise selection? The text book claims that "Their performance is very similar, as is often the case." Is there an example where their performances are not similar?
2. What are your comments about a variable selection method that includes a predictor if it is the most correlated with the current residual among all remaining predictors to be selected?
3. What are 3 principles of variable selection that were advocated by Dr. Jianqing Fan's 2001 paper "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties"? How to achieve these via a penalty?

Linear Models

Scalar response

- Predictor $X = \left(X_1, ..., X_p\right)^T \in \mathbb{R}^p$ and response $Y \in \mathbb{R}$
- Coefficient vector $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^{p+1}$ and population model

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon,$$

  where $\varepsilon \in \mathbb{R}$ is the random error term
- Given observation vector $\mathbf{y} = \left(y_1, \ldots, y_N\right)^T$ for Y and those $\mathbf{x}_i$ for X, we define design matrix

$$\mathbf{X} = \begin{pmatrix} \tilde{\mathbf{x}}_1^T \\ \vdots \\ \tilde{\mathbf{x}}_N^T \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{pmatrix} \in \mathbb{R}^{N \times (p+1)}$$

Scalar response

- Data version of model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_N)^T \in \mathbb{R}^N$ and each $\varepsilon_i \sim \varepsilon$
- Minimizing "residual sums of squares (RSS)"

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

gives "least squares estimate (LSE)" of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{X}^T\mathbf{y},$$

where $\mathbf{A}^-$ is a generalized inverse of a matrix $\mathbf{A}$

- Least squares estimate of $\mathbf{y}$ as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{X}^T\mathbf{y},$$

i.e., $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$

- $\hat{\mathbf{y}}$ is unique but $\hat{\boldsymbol{\beta}}$ not always

Distributional properties of LSE

- Recall the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and LSE $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$
- Assume $\boldsymbol{\varepsilon} \sim \text{Normal}\left(0, \sigma^2\mathbf{I}\right)$. Then we have unbiasedness

$$E\left[\hat{\boldsymbol{\beta}}\right] = E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right)\right] = \boldsymbol{\beta}$$

- Further, set $\mathbf{S} = \mathbf{X}^T\mathbf{X}$. Then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{S}^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$ and

$$\begin{aligned}
\text{Var}\left(\hat{\boldsymbol{\beta}}\right) &= E\left\{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\right\} = E\left\{\mathbf{S}^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\left(\mathbf{S}^{-1}\mathbf{X}^T\right)^T\right\} \\
&= \left(\mathbf{S}^{-1}\mathbf{X}^T\right)E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)\left(\mathbf{S}^{-1}\mathbf{X}^T\right)^T \\
&= \sigma^2\mathbf{S}^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{S}^{-1}\right)^T = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}
\end{aligned}$$

Estimating variance of error term

- Recall the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, LSE $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \left(\hat{y}_1, ..., \hat{y}_N\right)^T$
- Assume $\boldsymbol{\varepsilon} \sim \text{Normal}\left(0, \sigma^2\mathbf{I}\right)$. Then an estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{N-p-1},$$

and

$$\left(N-p-1\right)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1}$$

- Further, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are statistically independent

Linear model: recap

- Recall the model

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon \text{ with } E(\varepsilon) = 0$$

- Equivalent model: $E(Y|X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$
- Caution: if $E(Y|X)$ is restricted in, e.g., $[0,1]$ but $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ is allowed, then the above model is insensible
- In general, we model

$$g[E(Y|X)] = \beta_0 + \sum_{j=1}^{p} X_j \beta_j,$$

where g is a suitable "link function"

Derivation of LSE: I

- Recall

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}) &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \mathbf{y}^T\mathbf{y} \underbrace{-(\mathbf{X}\boldsymbol{\beta})^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta}}_{=2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}} + \boldsymbol{\beta}^T \underbrace{\mathbf{X}^T\mathbf{X}}_{=\mathbf{S}} \boldsymbol{\beta} \end{aligned} \tag{2}$$

- Vector calculus to derive $\hat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \mathrm{RSS}(\boldsymbol{\beta})$
- Let $\mathbf{e}_j \in \mathbb{R}^{p+1}$ have its jth entry as 1 and the rest as 0. Set $\mathbf{a} = \mathbf{y}^T\mathbf{X} = (a_1, ..., a_{p+1})$. Then

$$\partial_{\beta_j}(\mathbf{a}\boldsymbol{\beta}) = \lim_{t\to 0} \frac{\mathbf{a}(\boldsymbol{\beta}+t\mathbf{e}_j) - \mathbf{a}\boldsymbol{\beta}}{t} = \mathbf{a}\mathbf{e}_j = a_j,$$

and the gradient $\nabla_{\boldsymbol{\beta}}$, always a column vector, is

$$\nabla_{\boldsymbol{\beta}}\left(\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}\right) = \nabla_{\boldsymbol{\beta}}(\mathbf{a}\boldsymbol{\beta}) = \mathbf{a}^T = \mathbf{X}^T\mathbf{y} \tag{3}$$

Derivation of LSE: II

- Obtain the gradient $\nabla_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} \right)$ as follows:

$$\begin{aligned}
\partial_{\beta_j} \left(\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} \right) &= \lim_{t \to 0} t^{-1} \left[\left(\boldsymbol{\beta} + t\mathbf{e}_j \right)^T \mathbf{S} \left(\boldsymbol{\beta} + t\mathbf{e}_j \right) - \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} \right] \\
&= \lim_{t \to 0} t^{-1} \left[t\boldsymbol{\beta}^T \mathbf{S} \mathbf{e}_j + t\mathbf{e}_j^T \mathbf{S} \boldsymbol{\beta} + t^2 \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j \right] \\
&= \lim_{t \to 0} t^{-1} \left[2t\mathbf{e}_j^T \mathbf{S} \boldsymbol{\beta} + t^2 \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j \right] \quad \text{since } \mathbf{S} = \mathbf{S}^T \\
&= 2\mathbf{e}_j^T \mathbf{S} \boldsymbol{\beta} = 2 \left(\mathbf{S} \boldsymbol{\beta} \right)_j = \text{the } j\text{th entry of } \mathbf{S} \boldsymbol{\beta}
\end{aligned}$$

and

$$\nabla_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} \right) = 2\mathbf{S} \boldsymbol{\beta} = 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \tag{4}$$

- Note: $\mathbf{e}_j^T \mathbf{b}$ gives the jth entry of $\mathbf{b}$, and $\mathbf{e}_j^T \mathbf{A}$ the jth *row* of $\mathbf{A}$, and $\mathbf{A}\mathbf{e}_j$ the jth *column* of $\mathbf{A}$

Derivation of LSE: III

- Combining (2), (3) and (4) gives

$$\nabla_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

- The critical condition for $\hat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \mathrm{RSS}(\boldsymbol{\beta})$ forces $\nabla_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta})\big|_{\hat{\boldsymbol{\beta}}} = 0$. This gives the "normal equation" for $\hat{\boldsymbol{\beta}}$ as

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0 \Leftrightarrow \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}, \tag{5}$$

  where the symbol $\Leftrightarrow$ denotes "equivalence" or "if and only if"
- So, LSE $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y}$

Vector response

- Linear model with a scalar response: $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$ and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- Linear model with vector response:
  - Let $Y = (Y_1, ..., Y_m)$ and $X = (1, X_1, ..., X_p)$
  - Let $\mathbf{B} \in \mathbb{R}^{(p+1)\times m}$ be the unknown coefficient matrix, and $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ the random error vector
  - Models:

$$\begin{cases} \text{population version:} & Y = X\mathbf{B} + \boldsymbol{\varepsilon} \\ \text{data version:} & \mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \\ \text{data version (rowwise):} & \mathbf{y}_i = \mathbf{x}_i\mathbf{B} + \boldsymbol{\varepsilon}_i \end{cases},$$

where the ith row $\mathbf{y}_i$ of $\mathbf{Y}$ contains the ith observation for Y, the ith row $\mathbf{x}_i$ of $\mathbf{X}$ the ith observation on X, and ith row $\boldsymbol{\varepsilon}_i$ of $\mathbf{E}$ the ith realization of $\boldsymbol{\varepsilon}$ for $1 \leq i \leq N$

LSE of coefficient matrix

- Recall model $\mathbf{y}_i = \mathbf{x}_i\mathbf{B} + \boldsymbol{\varepsilon}_i$ or $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$ and

$$\mathrm{RSS}(\mathbf{B}) = \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{x}_i\mathbf{B}\|^2$$

- Then minimizing $\mathrm{RSS}(\mathbf{B})$ for $\mathbf{B} \in \mathbb{R}^{(p+1)\times m}$ gives the LSE

$$\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-}\mathbf{X}^T\mathbf{Y},$$

where $\left(\mathbf{X}^T\mathbf{X}\right)^{-}$ is a general inverse of $\mathbf{S} = \mathbf{X}^T\mathbf{X}$, and is $\mathbf{S}^{-1}$ if $\mathbf{S}$ is invertible

Derivation of LSE: overview

- Frobenius norm and trace operator
- Special matrices and basis
- Matrix calculus

Trace operator and Frobenius norm

Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times m}$.

- Define its "Frobenius norm" as

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}$$

- When $m = n$, i.e., $\mathbf{A}$ is a square matrix, its "trace" is defined as

$$\text{trace}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

- For any two (compatible) matrices $\mathbf{A}$ and $\mathbf{C}$, we have

$$\begin{cases} \text{trace}(\mathbf{A}) = \text{trace}(\mathbf{A}^T); \text{trace}(\mathbf{AC}) = \text{trace}(\mathbf{CA}); \\ \text{trace}(\mathbf{A} + \mathbf{C}) = \text{trace}(\mathbf{C} + \mathbf{A}) \end{cases}$$

since trace operates on the diagonal entries of a matrix

Frobenius norm

- Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times m}$ and write

$$\mathbf{A} = \underbrace{\begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}}_{\text{row layout}} = \underbrace{\begin{pmatrix} \tilde{\mathbf{a}}_1 & \cdots & \tilde{\mathbf{a}}_m \end{pmatrix}}_{\text{column layout}}$$

- Recall the Frobenius norm $\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}$. Then

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n} \|\mathbf{a}_i\|^2} = \sqrt{\sum_{i=1}^{m} \|\tilde{\mathbf{a}}_j\|^2}$$

- Also, $\|\mathbf{A}\| = \|\text{vec}(\mathbf{A})\|$, where

$$\text{vec}(\mathbf{A}) = \underbrace{\left(\tilde{\mathbf{a}}_1^T, ..., \tilde{\mathbf{a}}_m^T\right)}_{\text{column concatenation}} \quad \text{or vec}(\mathbf{A}) = \underbrace{(\mathbf{a}_1, ..., \mathbf{a}_n)}_{\text{row concatenation}}$$

Trace operator and Frobenius norm

- Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times m}$ and write

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{a}}_1 & \cdots & \tilde{\mathbf{a}}_m \end{pmatrix}$$

- Its Frobenius norm: $\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n} \|\mathbf{a}_i\|^2} = \sqrt{\sum_{i=1}^{m} \|\tilde{\mathbf{a}}_j\|^2}$
- Since $\left(\mathbf{A}\mathbf{A}^T\right)(i,i) = \mathbf{a}_i \mathbf{a}_i^T = \|\mathbf{a}_i\|^2$ and $\left(\mathbf{A}^T\mathbf{A}\right)(j,j) = \|\tilde{\mathbf{a}}_j\|^2$, then

$$\begin{aligned} \|\mathbf{A}\|^2 &= \sum_{i=1}^{n} \left(\mathbf{A}\mathbf{A}^T\right)(i,i) = \text{trace}\left(\mathbf{A}\mathbf{A}^T\right) \\ &= \sum_{i=1}^{m} \left(\mathbf{A}^T\mathbf{A}\right)(j,j) = \text{trace}\left(\mathbf{A}^T\mathbf{A}\right) \end{aligned}$$

- Namely, $\|\mathbf{A}\| = \sqrt{\text{trace}\left(\mathbf{A}^T\mathbf{A}\right)} = \sqrt{\text{trace}\left(\mathbf{A}\mathbf{A}^T\right)}$

Frobenius norm and inner product

- Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times m}$ and recall its Frobenius norm

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2} = \sqrt{\text{trace}\left(\mathbf{A}^T\mathbf{A}\right)} = \sqrt{\text{trace}\left(\mathbf{A}\mathbf{A}^T\right)}$$

- For $\mathbf{C} \in \mathbb{R}^{n\times m}$, we can define "inner product"

$$\langle \mathbf{A}, \mathbf{C} \rangle = \text{trace}\left(\mathbf{A}^T\mathbf{C}\right) = \text{trace}\left(\mathbf{C}^T\mathbf{A}\right) \tag{6}$$

- When $\mathbf{A}$ and $\mathbf{C}$ are vectors in $\mathbb{R}^s$, $\langle \mathbf{A}, \mathbf{C} \rangle$ is just the Euclidean inner product
- With (6), we can endow a geometric structure on $\mathbb{R}^{n\times m}$, where we can discuss angle, length, etc and where essentially we regard $\mathbf{A}$ as vec $(\mathbf{A})$

Special matrices

- Let $\mathbf{P}_{ij} \in \mathbb{R}^{n\times m}$ have only its (i,j)th entry as 1 and the rest as 0
- The set $\left\{\mathbf{P}_{ij}\right\}_{1\le i\le n, 1\le j\le m}$ form a basis for $\mathbb{R}^{n\times m}$, i.e., any $\mathbf{A} = \left(a_{ij}\right) \in \mathbb{R}^{n\times m}$ is a linear combination of the $\mathbf{P}_{ij}$'s. In fact,

$$\mathbf{A} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\mathbf{P}_{ij} \Leftrightarrow \mathbf{A} = \left(a_{ij}\right)$$

- Note $\mathbf{P}_{ij}^T = \mathbf{P}_{ji}$. Let $\mathbf{P}_{ij}^T\mathbf{A} \in \mathbb{R}^{m\times m}$. Then

$$\mathbf{P}_{ij}^T\mathbf{A} = \begin{pmatrix} 0 \\ a_{i1}, a_{i2}, ..., a_{im} \\ 0 \end{pmatrix},$$

i.e., the jth row of $\mathbf{P}_{ij}^T\mathbf{A}$ is the ith row of $\mathbf{A} = \left(a_{ij}\right)$, and all other rows of $\mathbf{P}_{ij}^T\mathbf{A}$ are zero

Matrix calculus

- Recall $\mathbf{P}_{ij} \in \mathbb{R}^{n\times m}$ has only its (i,j)th entry as 1 and the rest as 0, and pick any $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times m}$
- Let $g : \mathbb{R}^{n\times m} \to \mathbb{R}$ and define

$$\frac{\partial g(\mathbf{A})}{\partial \mathbf{A}} = \left(\frac{\partial g(\mathbf{A})}{\partial a_{ij}}\right) \in \mathbb{R}^{n\times m},$$

where the partial derivative

$$\frac{\partial}{\partial a_{ij}} g(\mathbf{A}) = \lim_{t\to 0} \frac{1}{t}\left[g\left(\mathbf{A} + t\mathbf{P}_{ij}\right) - g(\mathbf{A})\right]$$

- An example g is

$$g(\mathbf{B}) = \mathrm{RSS}(\mathbf{B}) = \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{x}_i\mathbf{B}\|^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2$$
$$= \mathrm{trace}\left[(\mathbf{Y} - \mathbf{X}\mathbf{B})^T (\mathbf{Y} - \mathbf{X}\mathbf{B})\right]$$

Derivation of LSE: strategy

- Recall $\mathbf{B} = \left(b_{ij}\right)$ and

$$\mathrm{RSS}\left(\mathbf{B}\right) = g\left(\mathbf{B}\right) = \mathsf{trace}\left[\left(\mathbf{Y} - \mathbf{X}\mathbf{B}\right)^T \left(\mathbf{Y} - \mathbf{X}\mathbf{B}\right)\right]$$

and

$$\frac{\partial g\left(\mathbf{B}\right)}{\partial \mathbf{B}} = \left(\frac{\partial g\left(\mathbf{B}\right)}{\partial b_{ij}}\right) = \left(\lim_{t\to 0} \frac{1}{t}\left[g\left(\mathbf{B} + t\mathbf{P}_{ij}\right) - g\left(\mathbf{B}\right)\right]\right)$$

- The critical condition for

$$\hat{\mathbf{B}} \in \underset{\mathbf{B}\in\mathbb{R}^{(p+1)\times m}}{\operatorname{argmin}} \mathrm{RSS}\left(\boldsymbol{\beta}\right) \Longrightarrow \left.\frac{\partial g\left(\mathbf{B}\right)}{\partial \mathbf{B}}\right|_{\hat{\mathbf{B}}} = 0,$$

which induces the normal equation for $\hat{\mathbf{B}}$ as

$$-2\mathbf{X}^T\left(\mathbf{Y} - \mathbf{X}\mathbf{B}\right) = 0 \Leftrightarrow \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\mathbf{B} \Rightarrow \hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-}\mathbf{X}^T\mathbf{Y}$$

Derivation of LSE: I

- Step 1: compute difference as

$$
\begin{aligned}
& g\left(\mathbf{B}+t\mathbf{P}_{ij}\right)-g\left(\mathbf{B}\right) \\
&= \text{trace}\left\{\left[\mathbf{Y}-\mathbf{X}\left(\mathbf{B}+t\mathbf{P}_{ij}\right)\right]^{T}\left[\mathbf{Y}-\mathbf{X}\left(\mathbf{B}+t\mathbf{P}_{ij}\right)\right]\right\} \\
&\quad -\text{trace}\left[\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)^{T}\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)\right] \\
&= \text{trace}\left\{\left[\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)-t\mathbf{X}\mathbf{P}_{ij}\right]^{T}\left[\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)-t\mathbf{X}\mathbf{P}_{ij}\right]\right\} \\
&\quad -\text{trace}\left[\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)^{T}\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)\right] \\
&= \text{trace}\left\{-t\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)^{T}\mathbf{X}\mathbf{P}_{ij}-t\mathbf{P}_{ij}^{T}\mathbf{X}^{T}\left(\mathbf{Y}-\mathbf{X}\mathbf{B}\right)+t^{2}\mathbf{P}_{ij}^{T}\mathbf{X}^{T}\mathbf{X}\mathbf{P}_{ij}\right\}
\end{aligned}
$$

(note linear terms and quadratic term in t)

Derivation of LSE: II

- Step 2: compute partial derivative as

$$\begin{aligned}
\frac{\partial}{\partial b_{ij}} g(\mathbf{B}) &= \lim_{t\to 0} \frac{1}{t}\left[g\left(\mathbf{B}+t\mathbf{P}_{ij}\right)-g(\mathbf{B})\right] \\
&= \lim_{t\to 0} \frac{1}{t}\text{trace}\left\{-t\left(\mathbf{Y}-\mathbf{XB}\right)^T\mathbf{XP}_{ij}-t\mathbf{P}_{ij}^T\mathbf{X}^T\left(\mathbf{Y}-\mathbf{XB}\right)\right\} \\
&\qquad + \lim_{t\to 0} \frac{1}{t}\text{trace}\left\{t^2\mathbf{P}_{ij}^T\mathbf{X}^T\mathbf{XP}_{ij}\right\} \\
&= \text{trace}\left\{-\left(\mathbf{Y}-\mathbf{XB}\right)^T\mathbf{XP}_{ij}-\mathbf{P}_{ij}^T\mathbf{X}^T\left(\mathbf{Y}-\mathbf{XB}\right)\right\} \\
&= -2\text{trace}\left\{\mathbf{P}_{ij}^T\mathbf{X}^T\left(\mathbf{Y}-\mathbf{XB}\right)\right\},
\end{aligned}$$

since $\text{trace}(\mathbf{C}) = \text{trace}\left(\mathbf{C}^T\right)$ with $\mathbf{C} = \mathbf{P}_{ij}^T\mathbf{X}^T\left(\mathbf{Y}-\mathbf{XB}\right)$

Derivation of LSE: III

- Step 3: obtain matrix derivative. Set

$$\mathbf{A} = \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\mathbf{B}) = (a_{ij}) \quad \text{and} \quad \mathbf{Q} = \mathbf{P}_{ij}^T \mathbf{A}.$$

  Then

$$\mathbf{Q} = \mathbf{P}_{ij}^T \mathbf{A} = \begin{pmatrix} 0 \\ (a_{i1}, ..., a_{ij}, ...a_{im}) \\ 0 \end{pmatrix} (j\text{th row}),$$

  i.e., the jth row of $\mathbf{Q}$ is the ith row of $\mathbf{A}$ and the rest are 0

- So,

$$\frac{\partial g(\mathbf{B})}{\partial b_{ij}} = -2\text{trace}\left\{\mathbf{P}_{ij}^T \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\mathbf{B})\right\} = -2a_{ij}$$

  i.e., $\frac{\partial}{\partial b_{ij}} g(\mathbf{B})$ is the (i,j)th entry of $\mathbf{A}$

- By the definition of $\nabla_{\mathbf{B}}$ as $\nabla_{\mathbf{B}} = \left(\frac{\partial}{\partial b_{ij}}\right)$, we see

$$\nabla_{\mathbf{B}} g(\mathbf{B}) = -2\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\mathbf{B})$$

Derivation of LSE: V

- Obtain normal equation: since

$$\nabla_{\mathbf{B}} g(\mathbf{B}) = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{B}),$$

setting $\nabla_{\mathbf{B}} g(\mathbf{B}) = 0$ gives

$$-2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{B}) = 0 \Leftrightarrow \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\mathbf{B}$$

- When $\mathbf{X}^T\mathbf{X}$ is invertible, the unique solution is

$$\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{7}$$

Dependence adjustment: settings and LSE

- Recall model:

$$\begin{cases} \text{data version:} & \mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \\ \text{data version (rowwise):} & \mathbf{y}_i = \mathbf{x}_i\mathbf{B} + \boldsymbol{\varepsilon}_i \end{cases}$$

- Assume: $\boldsymbol{\varepsilon}_i \sim \mathrm{G}(0, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| \neq 0$ and $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exitsts
- Then $\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ minimizes

$$h(\mathbf{B}) = \sum_{i=1}^{N} \underbrace{(\mathbf{y}_i - \mathbf{x}_i\mathbf{B})\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{y}_i - \mathbf{x}_i\mathbf{B})^T}_{=r_i(\boldsymbol{\Sigma})=\|\mathbf{r}_i(\boldsymbol{\Sigma})\|^2} = \left\|(\mathbf{Y} - \mathbf{X}\mathbf{B})\,\boldsymbol{\Sigma}^{-1/2}\right\|^2,$$

- Note: residual vector $\mathbf{r}_i = \mathbf{y}_i - \mathbf{x}_i\mathbf{B}$ is adjusted by $\boldsymbol{\Sigma}^{-1/2}$ into

$$\mathbf{r}_i(\boldsymbol{\Sigma}) = (\mathbf{y}_i - \mathbf{x}_i\mathbf{B})\,\boldsymbol{\Sigma}^{-1/2}$$

LSE based on adjusted residuals

- "Spectral decomposition": $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{D}$ is a diagonal matrix. Then $\mathbf{\Sigma}^{-1/2} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{U}^T$
- Let $\tilde{\mathbf{y}}_i = \mathbf{y}_i\mathbf{\Sigma}^{-1/2}$ and $\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{\Sigma}^{-1/2}$. Then the covariance of $\tilde{\mathbf{y}}_i$ is $\mathbf{I}$
- Set $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Sigma}^{-1/2}$. Then

$$h\left(\mathbf{B}\right) = \sum_{i=1}^{N}\left(\mathbf{y}_i - \mathbf{x}_i\mathbf{B}\right)\mathbf{\Sigma}^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\mathbf{B}\right)^T$$
$$\Longleftrightarrow h\left(\tilde{\mathbf{B}}\right) = \sum_{i=1}^{N}\left(\tilde{\mathbf{y}}_i - \mathbf{x}_i\tilde{\mathbf{B}}\right)\left(\tilde{\mathbf{y}}_i - \mathbf{x}_i\tilde{\mathbf{B}}\right)^T$$

- The minimizer of $h\left(\tilde{\mathbf{B}}\right)$ is

$$\mathbf{B}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{Y}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{\Sigma}^{-1/2},$$

and $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Sigma}^{-1/2}$ gives $\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ as the minimizer of $h\left(\mathbf{B}\right)$

LSE fails to incorporate dependence adjustment

- $\hat{\mathbf{B}} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the minimizer of the "weighted RSS"

$$h(\mathbf{B}) = \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{x}_i\mathbf{B})\,\mathbf{\Sigma}^{-1}\,(\mathbf{y}_i - \mathbf{x}_i\mathbf{B})^T = \left\|(\mathbf{Y} - \mathbf{X}\mathbf{B})\,\mathbf{\Sigma}^{-1/2}\right\|^2,$$

  where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{x}_i\mathbf{B}$ is adjusted by $\mathbf{\Sigma}^{-1/2}$ (encoding dependence)

- However, $\hat{\mathbf{B}} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is also the minimizer of

$$g(\mathbf{B}) = \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{x}_i\mathbf{B})\,(\mathbf{y}_i - \mathbf{x}_i\mathbf{B})^T = \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2$$

- Namely, for a linear model, least squares method cannot take into account the dependence structure of random error vector

Weighted least squares: settings and LSE

- Model with $\tilde{\varepsilon}_j \sim \mathrm{G}(0, \mathbf{\Sigma})$,

$$\begin{cases} \text{data version:} & \mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \\ \text{data version (columnwise):} & \tilde{\mathbf{y}}_j = \mathbf{X}\tilde{\mathbf{b}}_j + \tilde{\varepsilon}_j \end{cases}$$

- Consider the weighted RSS

$$u(\mathbf{B}) = \sum_{j=1}^{N} \left[\mathbf{\Sigma}^{-1/2}\left(\tilde{\mathbf{y}}_j - \mathbf{X}\tilde{\mathbf{b}}_j\right)\right]^T \left[\mathbf{\Sigma}^{-1/2}\left(\tilde{\mathbf{y}}_j - \mathbf{X}\tilde{\mathbf{b}}_j\right)\right]$$

Then we have the "weighted LSE"

$$\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{Y}$$

Derivation of weighted LSE

- "Spectral decomposition": $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{D}$ is a diagonal matrix. Then $\mathbf{\Sigma}^{-1/2} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{U}^T$ and

$$u(\mathbf{B}) = \text{trace}\left\{\left[\mathbf{\Sigma}^{-1/2}(\mathbf{Y} - \mathbf{X}\mathbf{B})\right]^T \left[\mathbf{\Sigma}^{-1/2}(\mathbf{Y} - \mathbf{X}\mathbf{B})\right]\right\}$$

- Set $\tilde{\mathbf{X}} = \mathbf{\Sigma}^{-1/2}\mathbf{X}$ and $\tilde{\mathbf{Y}} = \mathbf{\Sigma}^{-1/2}\mathbf{Y}$. Then

$$u(\mathbf{B}) = \text{trace}\left\{(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\mathbf{B})^T (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\mathbf{B})\right\}$$

- The minimizer $\hat{\mathbf{B}}$ for $u(\mathbf{B}), \mathbf{B} \in \mathbb{R}^{(p+1)\times m}$ is

$$\begin{aligned}\hat{\mathbf{B}} &= \left[\left(\mathbf{\Sigma}^{-1/2}\mathbf{X}\right)^T \mathbf{\Sigma}^{-1/2}\mathbf{X}\right]^{-1} \left(\mathbf{\Sigma}^{-1/2}\mathbf{X}\right)^T \mathbf{\Sigma}^{-1/2}\mathbf{Y} \\ &= \left(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{Y}\end{aligned}$$

Comparisons: LSE and weighted LSE

- For model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, i.e.,

$$\text{rowwise: } \mathbf{y}_i = \mathbf{x}_i\mathbf{B} + \boldsymbol{\varepsilon}_i \text{ with } \boldsymbol{\varepsilon}_i \sim \mathrm{G}(0, \boldsymbol{\Sigma}),$$

the LSE $\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ is the minimizer of both

$$\begin{cases} g(\mathbf{B}) = \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{x}_i\mathbf{B})(\mathbf{y}_i - \mathbf{x}_i\mathbf{B})^T \\ h(\mathbf{B}) = \sum_{i=1}^{N} \left[(\mathbf{y}_i - \mathbf{x}_i\mathbf{B})\boldsymbol{\Sigma}^{-1/2}\right]\left((\mathbf{y}_i - \mathbf{x}_i\mathbf{B})\boldsymbol{\Sigma}^{-1/2}\right)^T \end{cases}$$

- For model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, i.e.,

$$\text{columnwise: } \tilde{\mathbf{y}}_j = \mathbf{X}\tilde{\mathbf{b}}_j + \tilde{\boldsymbol{\varepsilon}}_j \text{ with } \tilde{\boldsymbol{\varepsilon}}_j \sim \mathrm{G}(0, \boldsymbol{\Sigma}),$$

the LSE $\hat{\mathbf{B}} = \left(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y}$ is the minimizer of

$$u(\mathbf{B}) = \sum_{j=1}^{N} \left[\boldsymbol{\Sigma}^{-1/2}\left(\tilde{\mathbf{y}}_j - \mathbf{X}\tilde{\mathbf{b}}_j\right)\right]^T \left[\boldsymbol{\Sigma}^{-1/2}\left(\tilde{\mathbf{y}}_j - \mathbf{X}\tilde{\mathbf{b}}_j\right)\right]$$

Traditional variable/model selection

Linear model with scalar response

- Predictor $X = \left(X_1, ..., X_p\right)^T \in \mathbb{R}^p$ and response $Y \in \mathbb{R}$
- Coefficient vector $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^{p+1}$ and population model

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon, \tag{8}$$

  where $\varepsilon \in \mathbb{R}$ is the random error term $E\left(\varepsilon\right) = 0$ and $\text{var}\left(\varepsilon\right) = \sigma^2$
- Data version: set $\boldsymbol{\varepsilon} = \left(\varepsilon_1, ..., \varepsilon_N\right)^T \in \mathbb{R}^N$ and each $\varepsilon_i \sim \varepsilon$, then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Variable/model selection for (8) is the same as retaining X_j whose $\beta_j \neq 0$

Motivations for variable selection

- "Simple model with good predictive performance" is desired
- There is need for variable/model selection
  - when sample size $N \geq p+1$ to obtain simpler models
  - when $N < p+1$ to obtain more stable models
  - when structural assumptions are reasonable to obtain more efficient models
- Usually variable/model selection is performed in
  - "Non-interpolation regime" for, i.e., small and non-zero training error, for "small models" such as generalized additive models
  - "Interpolation regime", i.e., zero training error, for "big models" such as neural networks
- Variable/model selection requires criteria

Selection criteria

- Always keep β_0 in $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$
- Let d be the number of predictors retained in a model and $\hat{\sigma}^2$ an estimate of σ^2
- Criterion on this model:

Mallow's C_p:	$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$
Akaike information criterion (AIC):	$\mathrm{AIC} \simeq \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$
Bayesian information criterion (BIC):	$\mathrm{BIC} \simeq \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + d\hat{\sigma}^2 \log n\right)$
Adjusted R-square:	$R^2_{\mathrm{Adj}} = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$
Cross-validation:	no explicit formula

Best-subset selection (BSS)

- Always keep β_0 in $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$
- There are 2^p submodels to check

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward-stepwise selection

- Always keep β_0 in $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$
- There are $1 + \sum_{k=0}^{p-1} (p-k) = 1 + 2^{-1} p\,(p+1)$ submodels to check

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward-stepwise selection

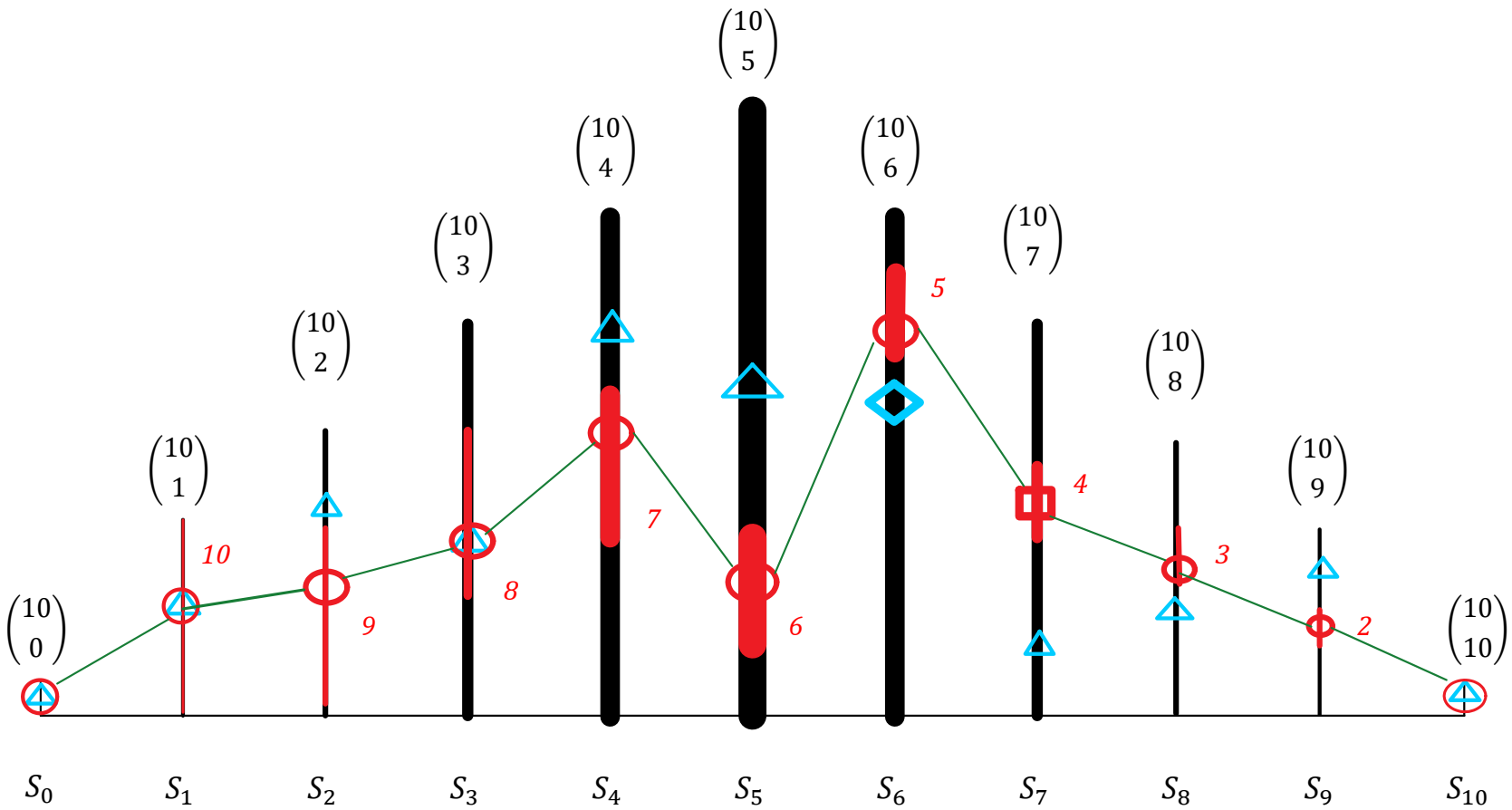
- Always keep β_0 in $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$
- There are $1 + \sum_{k=0}^{p-1} (p-k) = 1 + 2^{-1} p\,(p+1)$ submodels to check

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Trajectory of selected models

- "Triangle": trajectory for BSS with "Diamond" as best model
- "Circle": trajectory for FSS with "Rectangle" as best model



An illustration

- Always keep β_0 in $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$

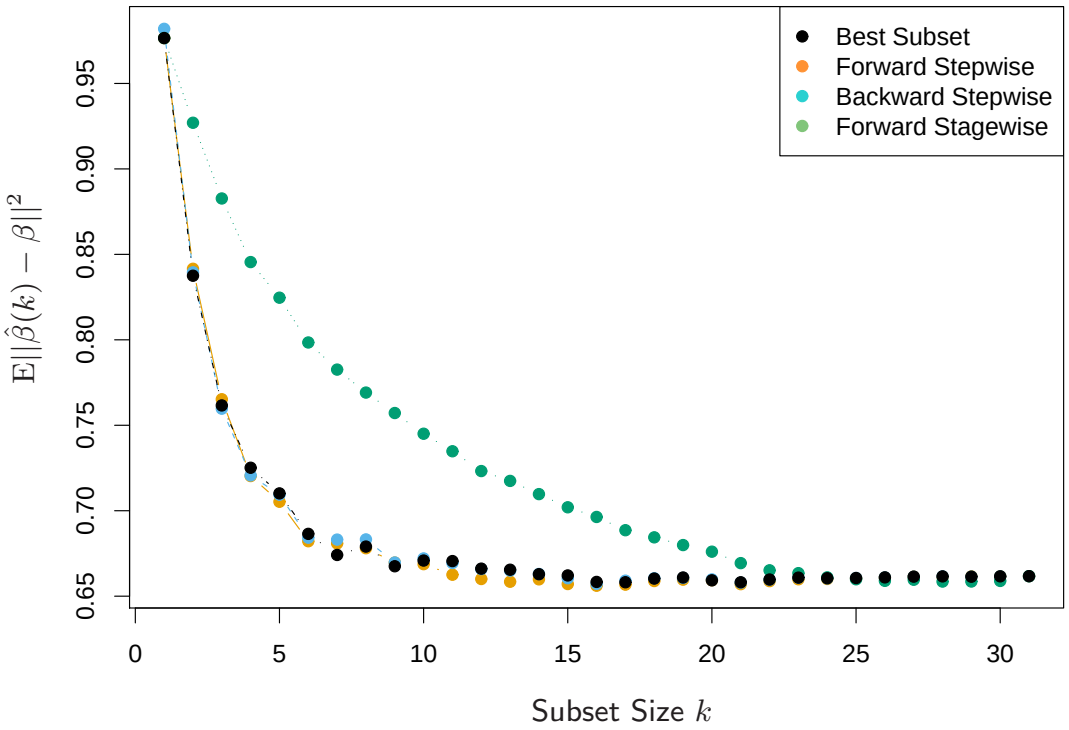


FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem* $Y = X^T\beta + \varepsilon$. *There are* $N = 300$ *observations on* $p = 31$ *standard Gaussian variables, with pairwise correlations all equal to* 0.85. *For* 10 *of the variables, the coefficients are drawn at random from a* $N(0, 0.4)$ *distribution; the rest are zero. The noise* $\varepsilon \sim N(0, 6.25)$, *resulting in a signal-to-noise ratio of* 0.64. *Results are averaged over* 50 *simulations. Shown is the mean-squared error of the estimated coefficient* $\hat{\beta}(k)$ *at each step from the true* β.

Recent progresses on BSS

- Best subset selection (BSS) is NP-hard but Mixed Integer Optimization (MIO) can help
- "Best subset selection via a modern optimization lens" by Dimitris Bertsimas, Angela King and Rahul Mazumder, The Annals of Statistics, 2016
- "Sparse high-dimensional regression: Exact scalable algorithms and phase transitions" by Dimitris Bertsimas and Bart Van Parys, The Annals of Statistics, 2020