P9120 - Statistical Learning and Data Mining

Lecture 2 - Linear Methods in Regression

Min Qian

Department of Biostatistics, Columbia University

September 12, 2024

Outline

1. Linear Regression
2. Ridge Regression
3. Principal Component Analysis
4. Stepwise Selection
5. Best Subset Selection (Mallow's C_p, AIC, BIC)

Linear Algebra

1. Basic vector matrix operations, norms, positive-definite matrix, etc.
2. Properties of matrix trace
3. Differentiation w.r.t. a vector
4. Singular value decomposition

Regression

- Input $X \in \mathbb{R}^p$, output $Y \in \mathbb{R}$, $(X, Y) \sim \mathcal{D}$ unknown.
- Optimal prediction model:

$$f^*(X) = \arg\min_f E[Y - f(X)]^2 = E(Y|X).$$

- Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ i.i.d. $\sim \mathcal{D}$
- Least squares (Empirical Risk Minimization)

$$\min_f \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2$$

Ordinary Least Squares (OLS)

- Linear Model: $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$.
- OLS estimate $\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator.

$$\begin{aligned}\hat{\boldsymbol{\beta}} \triangleq & (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) \\ = & \arg\min_{\beta_j' s} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \right]^2 \\ = & \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned}$$

where $\mathbf{y}$ is n-vector of response values, $\mathbf{X}$ is $n \times (p+1)$ matrix of regressors, and $\boldsymbol{\beta}$ is a $(p+1)$-vector of parameters.

Is an unbiased estimator always better than a biased estimator?

Parameter Estimation / Prediction Accuracy

- Parameter estimation accuracy can be measured by

$$E_{\mathcal{T}}(\hat{\beta}_j - \beta_j)^2 = (E_{\mathcal{T}}\hat{\beta}_j - \beta_j)^2 + E_{\mathcal{T}}(\hat{\beta}_j - E_{\mathcal{T}}\hat{\beta}_j)^2 = \text{Bias}^2(\hat{\beta}_j) + \text{Var}(\hat{\beta}_j)$$

- The prediction accuracy of $\hat{f}$ at a point $\mathbf{x}^*$ is measured by the mean squared error (MSE)

$$\begin{aligned} MSE(\hat{f}(\mathbf{x}^*)) =& E_{\mathcal{T}}\left[\hat{f}(\mathbf{x}^*) - f(\mathbf{x}^*)\right]^2 \\ =& \left(E_{\mathcal{T}}\left[\hat{f}(\mathbf{x}^*)\right] - f(\mathbf{x}^*)\right)^2 + Var_{\mathcal{T}}\left[\hat{f}(\mathbf{x}^*)\right] \\ =& (\text{Bias})^2 + \text{Variance}, \end{aligned}$$

  where $f(\mathbf{x}^*)$ is the true response and $\hat{f}(\mathbf{x}^*)$ is an estimate of $f(\mathbf{x}^*)$.

- Choosing estimators often involves a bias-variance trade-off.

Playing with Simulated Data: Case I

- Fix the sample size $n = 100$
- Draw a sample of size $n = 100$ of (x_1, x_2, x_3, x_4) from $N(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & 0 & 0.97 & 0 \\ 0 & 1 & 0 & -0.97 \\ 0.97 & 0 & 1 & 0 \\ 0 & -0.97 & 0 & 1 \end{pmatrix}$$

- Draw a corresponding sample of size $n = 100$ of $\epsilon \sim N(0, 1)$
- Calculate the 100 output y values using the generative model:

$$y = 1 + x_1 + x_2 + \epsilon,$$

- So far you have played the role of Nature - now become a human and analyze the data.

Results from Different Simulated Data Sets

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98007    0.08754  11.196  < 2e-16 ***
X1           0.93444    0.39437   2.369  0.01984 *
X2           1.03672    0.34872   2.973  0.00374 **
X3           0.14423    0.39077   0.369  0.71288
X4          -0.06307    0.36102  -0.175  0.86169
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.93093    0.10596   8.786 6.49e-14 ***
X1           2.07995    0.45148   4.607 1.27e-05 ***
X2           0.96422    0.42687   2.259   0.0262 *
X3          -1.03637    0.44524  -2.328   0.0221 *
X4           0.05816    0.43071   0.135   0.8929
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0999     0.1020  10.785   <2e-16 ***
X1            0.6769     0.4151   1.631    0.106
X2            0.1498     0.5072   0.295    0.768
X3            0.3601     0.4099   0.879    0.382
X4           -0.7713     0.4994  -1.545    0.126
```

Result varies from training set to training set.

Collinearity

- Suppose there is a perfect linear relation among the predictors:
  - $\mathbf{X}^T\mathbf{X}$ is not invertible.
  - OLS estimate is not unique.

- Collinearity refers to the case when there are very high correlations among predictors
  - Can't tell just by looking at simple correlations (why?)
  - OLS estimate has large variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{S_{jj}}\left(\frac{1}{1-R_j^2}\right),$$

where $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{\mathbf{x}}_j)^2$ and R_j^2 is the multiple R^2 for regressing $\mathbf{x}_j$ on other predictors.

  - $1/(1-R_j^2)$ is called the j-th variance inflation factor (VIF).

Shrinkage Methods

- Ridge Regression
- Principal Components Regression (PCR)
- Lasso and others (next week)

Ridge Regression

Penalizing the square of the coefficients (l_2 penalty)

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Used in the presence of collinearity where the usual OLS estimates are unstable.
- $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage.
  - What happens when $\lambda \to 0$?
  - What happens when $\lambda \to \infty$?
- There exist several criteria for choosing λ, e.g. cross-validation.

Ridge Regression in Practice

- The ridge estimate is not equivariant under scaling of the predictors.
  - Often standardize $\mathbf{x}_j$'s first.

- Common practice is to center and standardize the $\mathbf{x}_j$'s and center the $\mathbf{y}$ before applying the minimization procedure.

- Centering of $\mathbf{x}_j$'s and $\mathbf{y}$ implies that we don't need an intercept term any more ($\hat{\beta}_0^{\text{ridge}} = \bar{\mathbf{y}}$, $\hat{\beta}_j^{\text{ridge}}$s are the coefficients of standardized $\mathbf{x}_j$'s.).

- This means, we can work with a design matrix with no column of 1's – the dimension of $\mathbf{X}$ is now $n \times p$, as opposed to $n \times (p+1)$.

Solution for Ridge Regression

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}} \left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}\right] \\ &= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned}$$

- Even if $\mathbf{X}^T\mathbf{X}$ is not of full-rank, $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p$ is invertible if $\lambda > 0$.
- $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ is linear in $\mathbf{y}$.
- $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ is biased.
- $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ has smaller variance than the OLS (proved under fixed design), thus may have smaller mean squared error (MSE).

Shrinkage in Ridge

Suppose orthonormal design (i.e. $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$). Then

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{ols}} =& \mathbf{X}^T\mathbf{y},\\ \hat{\boldsymbol{\beta}}^{\text{ridge}} =& \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{1+\lambda}\mathbf{X}^T\mathbf{y} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{\text{ols}}.\end{aligned}$$

- Shrink OLS towards zero by a positive constant less than 1.
- $Var(\hat{\boldsymbol{\beta}}^{\text{ridge}}) = \frac{1}{(1+\lambda)^2} Var(\hat{\boldsymbol{\beta}}^{\text{ols}})$.
- $\lambda \uparrow$, shrinkage $\uparrow$, bias $\uparrow$, variance $\downarrow$.
- $\lambda \downarrow$, shrinkage $\downarrow$, bias $\downarrow$, variance $\uparrow$.

Solution Path for Ridge in Prostate Cancer Example

Examine the risk factors for prostate cancer, based on clinical and demographic variables.

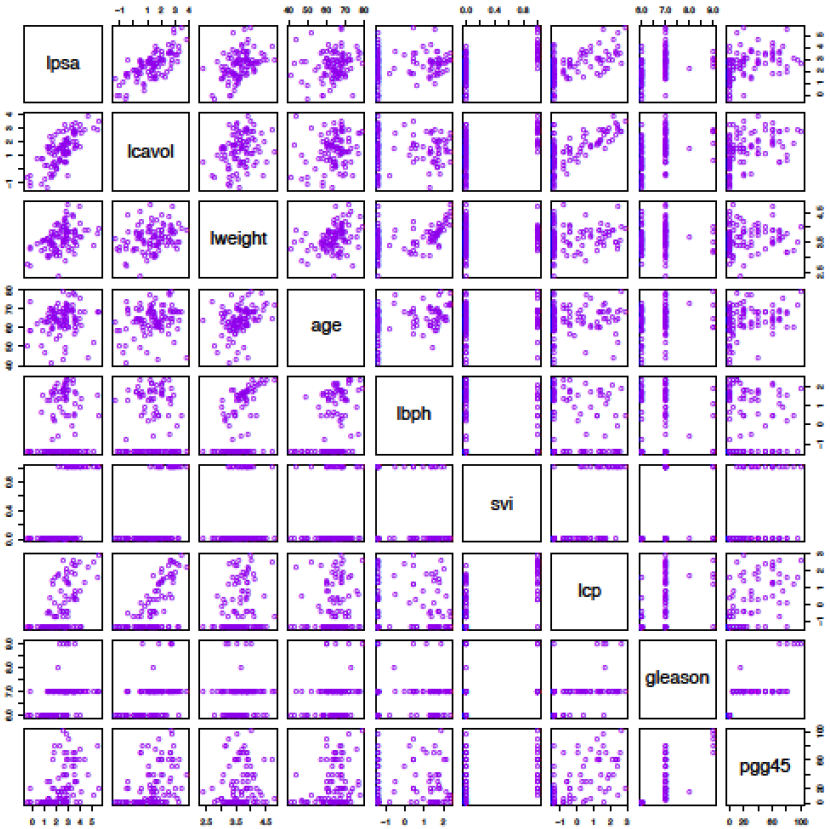


FIGURE 1.1. *Scatterplot matrix of the prostate cancer data.*

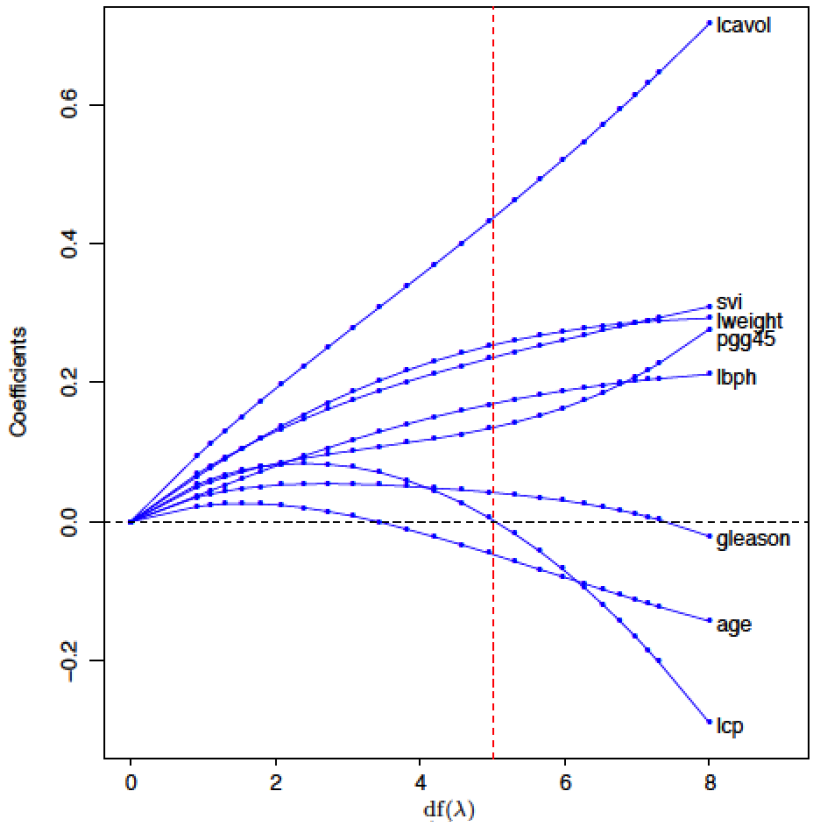


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

Singular Value Decomposition (SVD)

The SVD of any $n \times p$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

- $\mathbf{U}_{n\times p}$ has orthogonal columns (i.e. $\mathbf{U}^T\mathbf{U} = I_p$).
- $\mathbf{D}_{p\times p}$ is diagonal with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$, called singular values of $\mathbf{X}$.
- $\mathbf{V}_{p\times p}$ is an orthogonal matrix (i.e. $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = I_p$).
- span{columns of $\mathbf{U}$} = span {columns of $\mathbf{X}$}.

Understanding Ridge from SVD

Let $\mathbf{u}_j$'s be the columns of $\mathbf{U}$. The OLS fitted vector

$$\hat{\mathbf{y}}^{ols} = \mathbf{X}\hat{\boldsymbol{\beta}}^{ols} = \mathbf{U}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{p} \mathbf{u}_j\mathbf{u}_j^T\mathbf{y}$$

The Ridge fitted vector

$$\hat{\mathbf{y}}^{ridge} = \mathbf{X}\hat{\boldsymbol{\beta}}^{ridge} = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j\mathbf{u}_j^T\mathbf{y},$$

Regarding $\mathbf{U}$ as an orthonormal basis for the column space of $\mathbf{X}$.

- The coordinate onto $\mathbf{u}_j$ is shrunken by the factor $d_j^2/(d_j^2+\lambda)$.
- Smaller values of d_j^2 incur larger amount of shrinkage.

How to interpret d_j^2?

Principle Component Analysis (PCA)

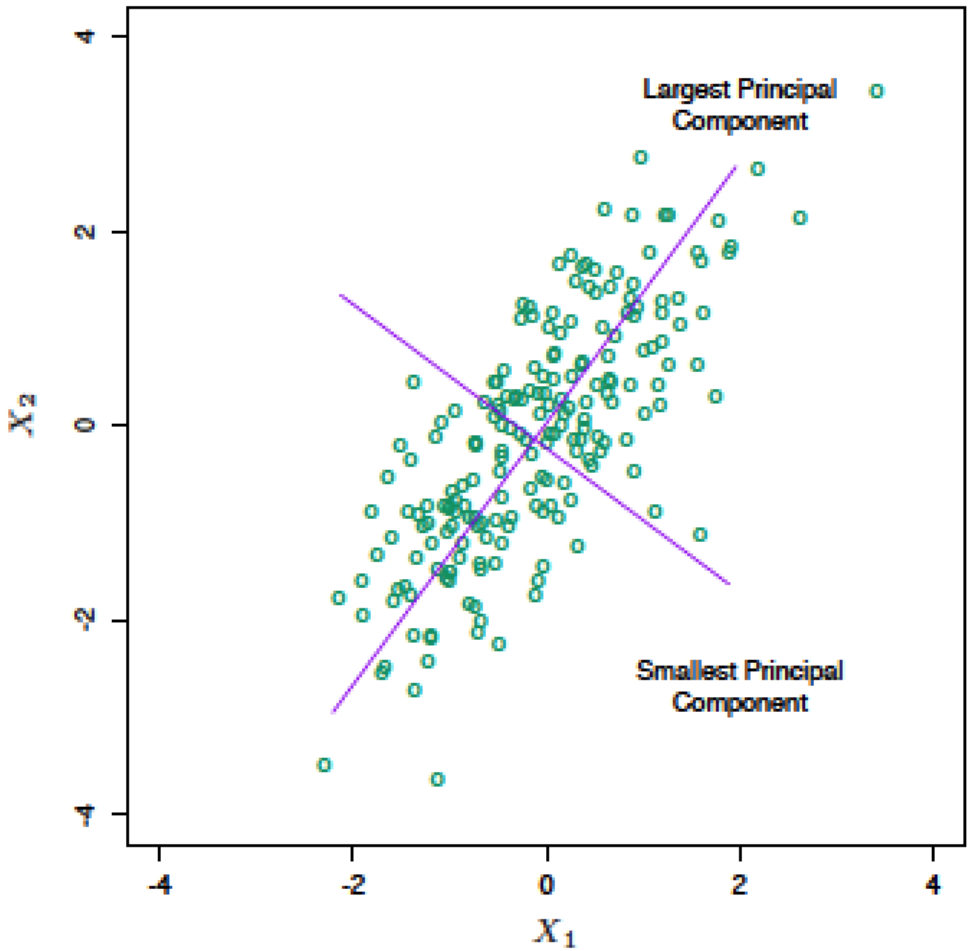


FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*

PCA

- Reduce the dimensionality of the data (Unsupervised Learning)
- Find linear combinations of predictors that explain most of the variation in the data
  - The 1st PC has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$.
  - The 2nd PC has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$ that are orthogonal to the 1st PC.
  - ...
- In regression, make predictors orthogonal to each other, and thus address the issue of collinearity (use of PCs in Supervised Learning).

Principal Component Analysis

- Center each input by its mean, resulting in the $\mathbf{X}_{n\times p}$ matrix.
- Find the $p \times 1$ vector $\mathbf{a}_1$ to maximize sample variance

$$Var(\mathbf{X}\mathbf{a}_1) \text{ s.t. } \mathbf{a}_1^T\mathbf{a}_1 = 1.$$

  Solution: $\mathbf{v}_1$, the first column of $\mathbf{V}$ in SVD.

- For each $j = 2, \ldots, p$, given $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$, find $\mathbf{a}_j$ to maximize

$$Var(\mathbf{X}\mathbf{a}_j) \text{ s.t. } \mathbf{a}_j^T\mathbf{a}_j = 1, \mathbf{a}_j^T\mathbf{v}_k = 0 \text{ for } k = 1, \ldots, j-1.$$

  Solution: $\mathbf{v}_j$, the j-th column of $\mathbf{V}$ in SVD.

- We obtain $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j$, the j-th principal component of $\mathbf{X}$ for $j = 1, \ldots, p$.

SVD and PCA

- When each x_j is centered, the sample variance-covariance matrix of $(x_1, \ldots, x_p)$ is

$$\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\mathbf{V}\mathbf{D}^2\mathbf{V}^T.$$

- The j-th PC of $\mathbf{X}$ is $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = d_j\mathbf{u}_j$ with $Var(\mathbf{z}_j) = d_j^2/n$, where $\mathbf{v}_j$ and $\mathbf{u}_j$ are the j-th column of $\mathbf{V}$ and $\mathbf{U}$, respectively.

- $\mathbf{v}_j$ is called the j-th PC direction.

- $\mathbf{u}_j$ is the normalized j-th PC of $\mathbf{X}$.

PAC in Handwritten Digits Example

Goal: Representing high-dimensional data by a small subset of features.

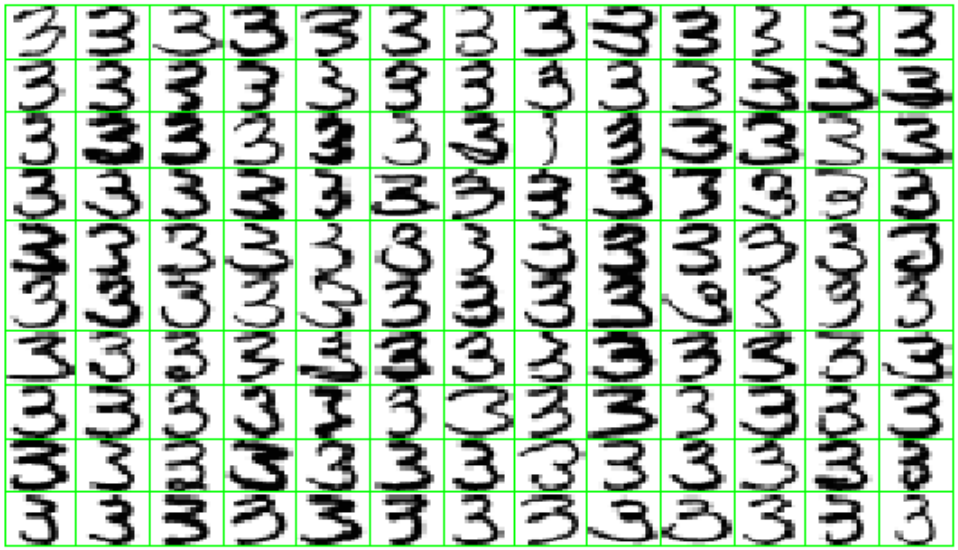


FIGURE 14.22. *A sample of 130 handwritten 3's shows a variety of writing styles.*

- Each 3 can be viewed as a data point of $\mathbf{x} \in \mathbb{R}^{256}$ (16×16 image).
- PCs of $\mathbf{X}$ can be computed via SVD.
- First 50 PCs account for 90% of the variation, and 12 PCs account for 63% variation.

Handwritten Digits with 2 PCs

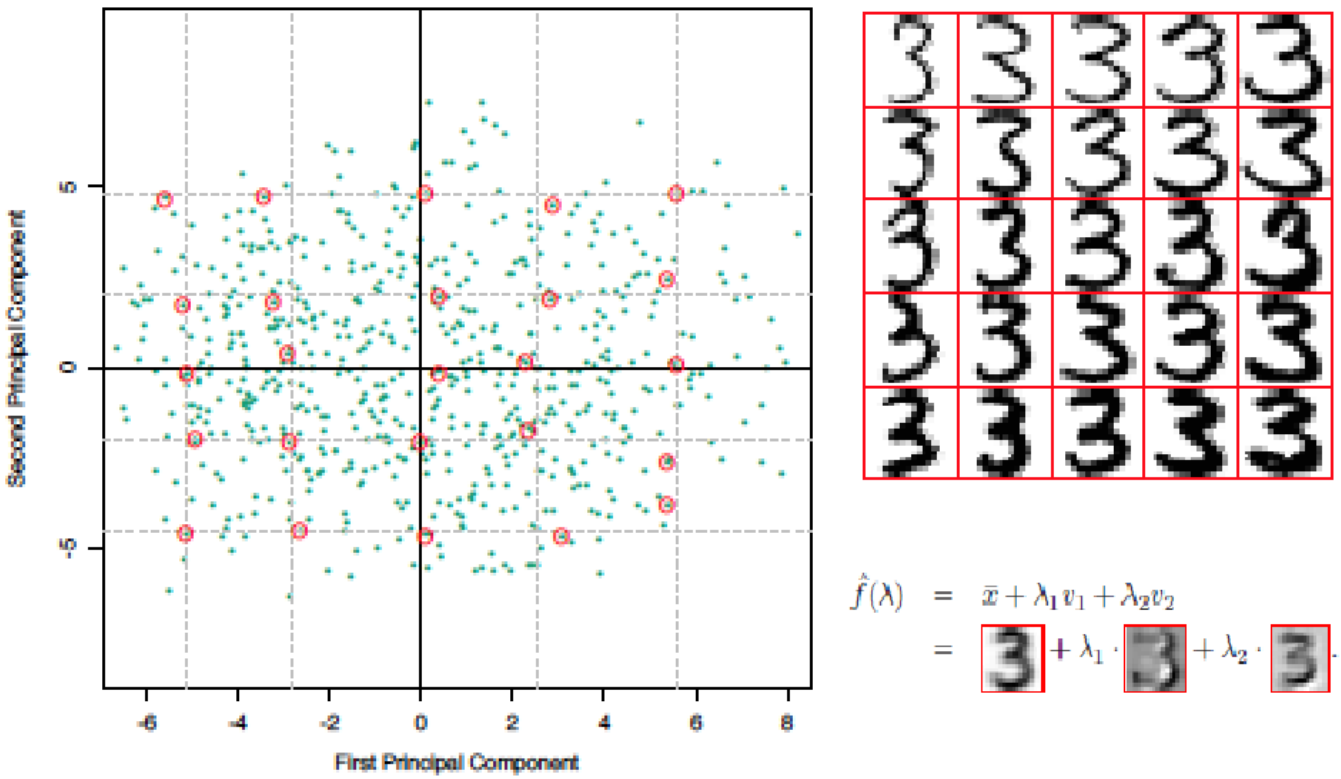


FIGURE 14.23. *(Left panel:) the first two principal components of the handwritten threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. (Right panel:) The images corresponding to the circled points. These show the nature of the first two principal components.*

Principal Component Regression

- PCR replaces the original regression model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_p \mathbf{x}_p + \boldsymbol{\epsilon}$$

with a model with fewer derived inputs ($q < p$)

$$\mathbf{y} = \beta_0' + \beta_1' \mathbf{z}_1 + \ldots + \beta_q' \mathbf{z}_q + \boldsymbol{\epsilon}'$$

- Since each $\mathbf{z}_j$ is centered, and the $\mathbf{z}_j$'s are orthogonal, the above regression is just a sum of univariable regressions:

$$\hat{\beta}_0' = \bar{\mathbf{y}} \text{ and } \hat{\beta}_j' = \frac{\mathbf{z}_j^T \mathbf{y}}{\mathbf{z}_j^T \mathbf{z}_j} \text{ for } j = 1, \ldots, q.$$

- As with Ridge regression, PCs depend on the scaling of the inputs, so typically they are standardized first

How do we pick the number of PCs to use?

- This is a tuning parameter of the procedure
- Typically most variation in $\mathbf{X}$ can be represented by a few principal components.
- One can choose q to explain certain percent of variation (method from the pre-computing era), e.g. pick first q PCs so that

$$\sum_{j=1}^{q} d_j^2 \geq (1-\alpha) \sum_{j=1}^{p} d_j^2.$$

- Alternatively, one can use cross-validation (more common in the machine learning area)

Ridge vs PCR for Shrinkage

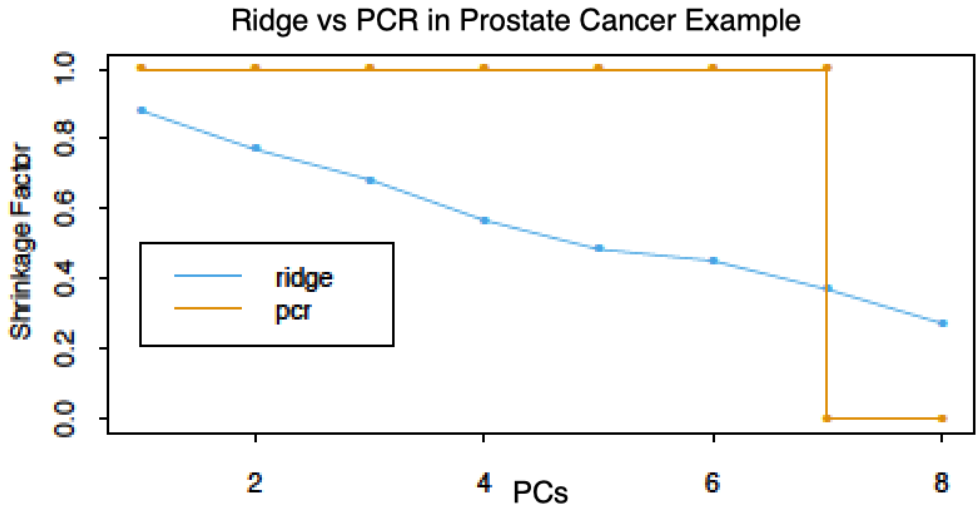
Assume $\mathbf{X}$ and $\mathbf{y}$ are both centered (so no intercept).

- SVD: $\mathbf{X} = (\mathbf{u}_1, \ldots, \mathbf{u}_p) \times \text{diag}(d_1, \ldots, d_p) \times (\mathbf{v}_1, \ldots, \mathbf{v}_p)^T$
- j-th PC of $\mathbf{X}$ is $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = d_j \mathbf{u}_j$ and $var(\mathbf{z}_j) = d_j^2/n$.

Fitted $\hat{\mathbf{y}}$ can be written as a linear combination of PCs:

$$\hat{\mathbf{y}}^{ols} = \sum_{j=1}^{p} \mathbf{u}_j (\mathbf{u}_j^T \mathbf{y}),$$

$$\hat{\mathbf{y}}^{ridge} = \sum_{j=1}^{p} \mathbf{u}_j \left(\frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j^T \mathbf{y} \right)$$



$$\hat{\mathbf{y}}^{pcr} = \sum_{j=1}^{q} \frac{\mathbf{z}_j^T \mathbf{y}}{\mathbf{z}_j^T \mathbf{z}_j} \mathbf{z}_j = \sum_{j=1}^{q} \mathbf{u}_j (\mathbf{u}_j^T \mathbf{y})$$

Summary

- OLS: BLUE; but may have high variance.
- Ridge: biased but has smaller variance. Shrink coefficients on each PC. More shrinkage on PCs with smaller variance.
- PCR: use the first q PCs; discard $(p - q)$ Pcs.

The final estimates use all of the p predictors.

Variable Selection

- Often we encounter situations with too many potential predictors of a certain outcome
- Parsimony: We would like to determine a smaller subset of predictors that exhibit the strongest effects
- Selecting the active set of variables is important:
  - Estimates and predictions based on models involving only the active terms are usually more precise
  - Collinearity causes instability in the estimates
  - Simplicity of interpretation
  - Can save time and money by not collecting data on irrelevant variables

Forward, Backward, Stepwise Regression

Forward selection:

- Starts with the intercept.
- Sequentially adds the predictor that most improves the fit
- Stop when meet some criterion (e.g. p-value > 0.05).

Backward elimination:

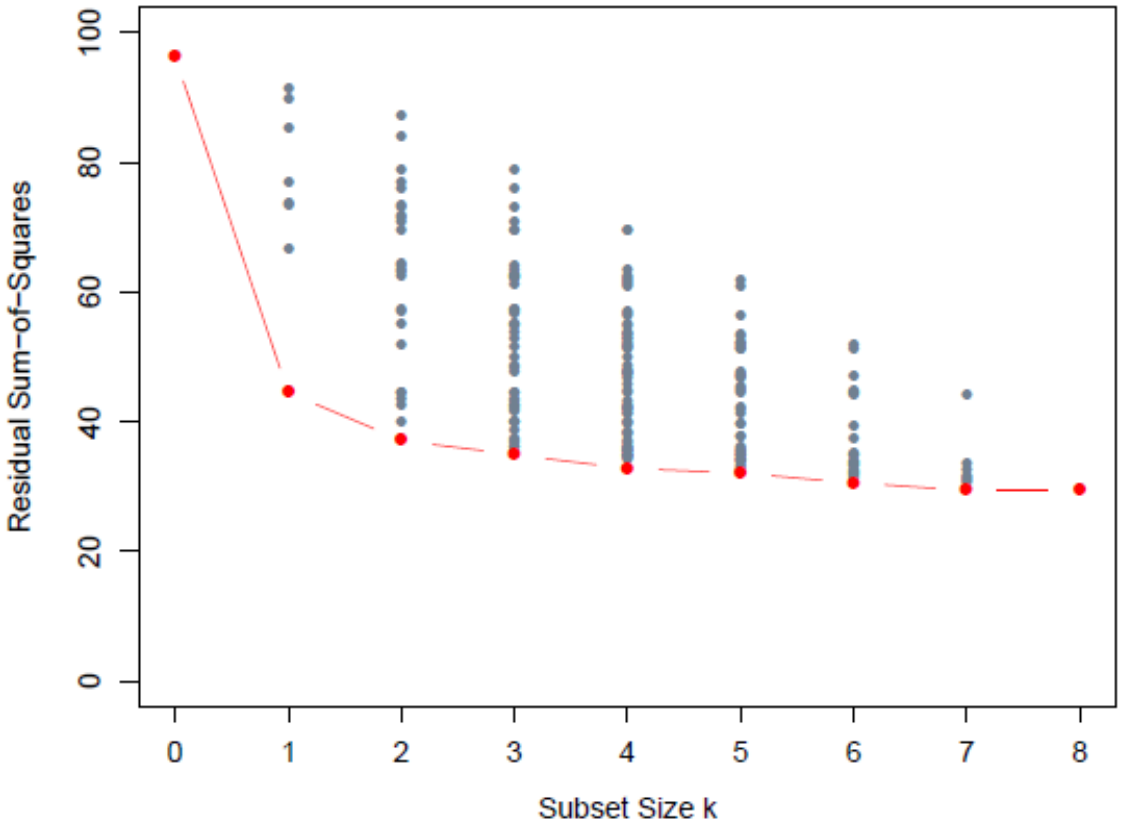
- Starts with the full model.
- Sequentially deletes the predictor that has the least impact.
- Stop when meet some criterion (e.g. p-value ≤ 0.05).

Forward-Backward Stepwise Regression: Modify forward selection by – Each time a new variable is added to the model, we test for significance of variables currently in the model, and remove the one with the largest insignificant p-value.

Forward-Backward Stepwise Regression: Caution

- Invalid p-values since it do not take into account the variable selection process.
- Inflated type-I error
- In R, only AIC is provided in stepwise selection (not p-value)
- The final model is not guaranteed to be optimal in any specified sense. It may include some unimportant predictors or exclude some important predictors.

Best-Subset Selection



- Consider all 2^p models.
- For models with the same dimension $k \in \{0, \ldots, p\}$, find the model that minimizes the residual sum of squares (i.e., OLS).
- How to select k?

Supervised Learning (Prediction)

- Training Sample: $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} = (\mathbf{X}, \mathbf{y})$
- Model class $\mathcal{F}$, e.g. $\mathcal{F} = \{f(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}^T \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}$
- Loss function $L(Y, f(X))$: measures errors between Y and $f(X)$

  e.g. Square error loss: $L(Y, f(X)) = [Y - f(X)]^2$
- Estimated Prediction model $\hat{f}(X)$

  e.g. ERM: $\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i))$

  How to evaluate the prediction performance of $\hat{f}$?

Prediction Error and Training Error

Assume each $(\mathbf{x}_i, y_i)$ is randomly drawn from the population.

- Prediction Error (also called Test Error or Generalization Error):

$$\mathrm{PE}(\hat{f}) = \mathrm{E}_{Y,X} L\big(Y, \hat{f}(X)\big) = \mathrm{E}\big[L\big(Y, \hat{f}(X)\big)|\mathcal{T}\big]$$

- Expected Prediction Error:

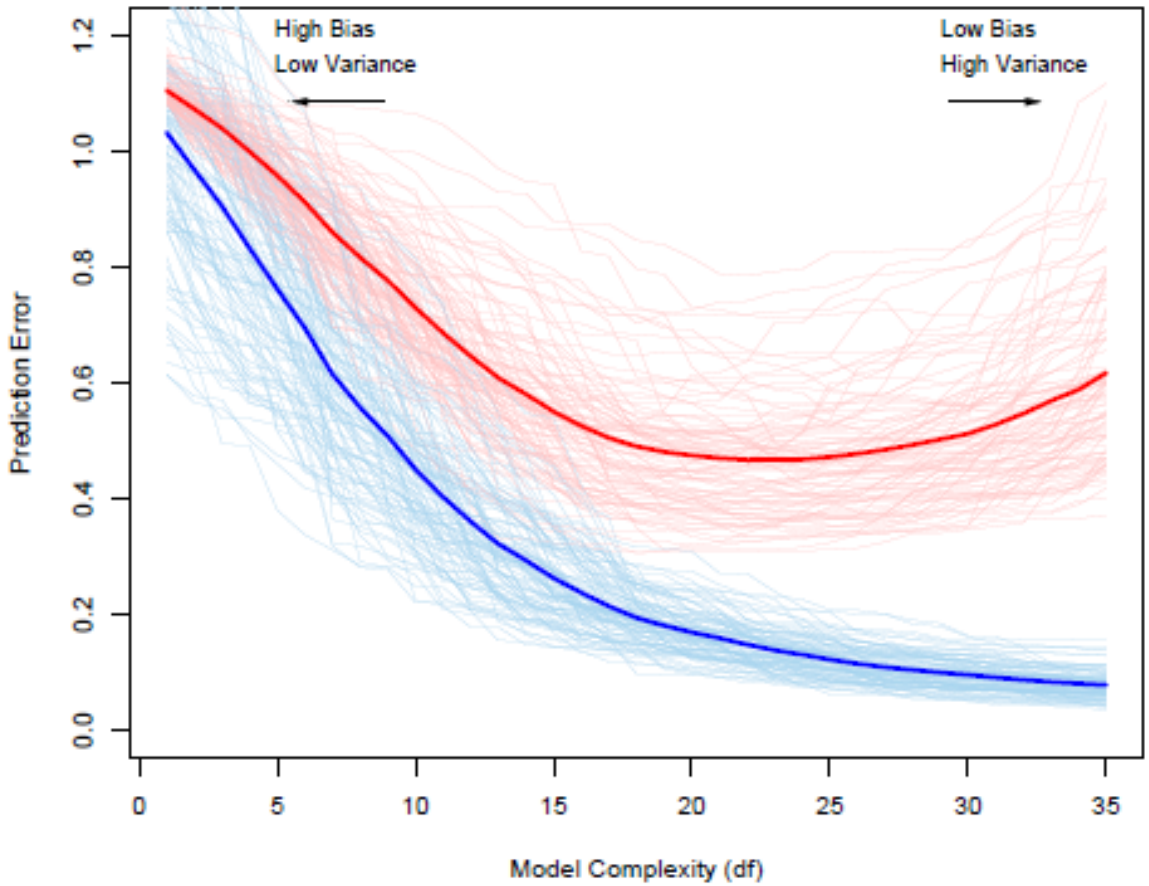
$$\mathrm{EPE} = \mathrm{E}[\mathrm{PE}(\hat{f})] = \mathrm{E}_{\mathcal{T}}\mathrm{E}_{Y,X} L\big(Y, \hat{f}(X)\big)$$

- Training Error:

$$\mathrm{TE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i))$$

Is the training error a good estimate of the prediction error?

Errors v.s. Model Complexity



Light blue/light red curves: TE/PE for 100 training sets of size n=50 each.

Solid curves: ETE/EPE

Overfitting: a predictive model may well describe the relationship between predictors and outcome in the training data, but may subsequently fail to be generalized to the target population.

Model Selection

Goal: choose a model (estimator) that minimize the (expected) prediction error.

In a data-rich situation:

- Training set: fit each model
- Validation set: (estimate PE of each model) and choose model.
- Test set: assess the final chosen model

A typical split might be 50% for training, 25% for validation and testing.

What Can We Do with Insufficient Data?

How to deal with validation (model selection) step?

- Approximate the validation step analytically (e.g. C_p, AIC, BIC, etc).
- Use resampling approach to estimate the (expected) prediction error (e.g. Cross-validation, Bootstrap, etc).
- Other methods...

Optimism of The Training Error

Typically, training error $>\overset{?}{=}<$ prediction error

- Same data is being used to fit the model and assess its error
- Training error will be an overly optimistic estimate of the prediction error.

Can we estimate the discrepancy between training error and prediction error?

- Very difficult in general. Doable in some situations.

Prediction Error under Fixed Design

When the design matrix $\mathbf{X}$ is random,

$$\begin{aligned}\mathrm{PE}(\hat{f}) &= \mathrm{E}_{Y,X} L\big(Y, \hat{f}(X)\big) = \mathrm{E}_{Y^N,X^N}[L\big(Y^N, \hat{f}(X^N)\big)|\mathcal{T}] \\ \mathrm{EPE} &= \mathrm{E}_{\mathcal{T}}[\mathrm{PE}(\hat{f})] = \mathrm{E}_{\mathbf{y},\mathbf{X}}[\mathrm{PE}(\hat{f})]\end{aligned}$$

When the design matrix $\mathbf{X}$ is fixed,

$$\begin{aligned}\mathrm{PE}_{in}(\hat{f}) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{Y_i^N}\left[L\big(Y_i^N, \hat{f}(\mathbf{x}_i)\big)|\mathcal{T}\right] \quad \longrightarrow \text{“in sample” error} \\ \mathrm{EPE}_{in} &= \mathrm{E}_{\mathcal{T}}[\mathrm{PE}_{in}(\hat{f})] = \mathrm{E}_{\mathbf{y}}[\mathrm{PE}_{in}(\hat{f})]\end{aligned}$$

Goal: use $\mathrm{TE}(\hat{f})$ to estimate $\mathrm{PE}_{in}(\hat{f})$ or EPE_{in}.

Estimating EPE_{in} using Training Sample

The *optimism* of $\text{TE}(\hat{f})$ is defined as

$$\text{op} \triangleq \text{PE}_{in}(\hat{f}) - \text{TE}(\hat{f})$$

The expectation of the optimism over training sets is

$$\text{E}_{\mathbf{y}}(\text{op}) \triangleq \text{EPE}_{in}(\hat{f}) - \text{E}_{\mathbf{y}}[\text{TE}(\hat{f})]$$

If $\text{E}_{\mathbf{y}}(\text{op})$ is computable from the data, then

$$\text{E}_{\mathbf{y}}(\text{op}) + \text{TE}(\hat{f})$$

is an unbiased estimate of $\text{EPE}_{in}(\hat{f})$.

Optimism

Claim: For squared error, 0-1, and some other loss functions,

$$\mathrm{E}_{\mathbf{y}}(\mathrm{op}) = \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}_{\mathbf{y}}(y_i, \hat{f}(\mathbf{x}_i)),$$

where the expectation is over the training set outcome values, $\mathbf{y}$.

If $\mathbf{y}$ arise from an additive error model $Y = f(X) + \epsilon$ with $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, then for any linear estimate $\hat{f} = \mathbf{S}\mathbf{y}$,

$$\sum_{i=1}^{n}\mathrm{Cov}_{\mathbf{y}}(y_i, \hat{f}(\mathbf{x}_i)) = \mathrm{trace}(\mathbf{S})\sigma^2,$$

where $\mathbf{S}$ is an $n \times n$ matrix depending on $\mathbf{X}$ but not on $\mathbf{y}$.

Proof (square error loss)

Mallows' C_p (Mallows 1973)

Assume

- $y_i = f(\mathbf{x}_i) + \epsilon_i$ with iid $\epsilon_i \sim (0, \sigma^2)$.
- $\hat{f}$ is the OLS estimate with p inputs.

Then $\hat{f}(\mathbf{X}) = \mathbf{S}\mathbf{y}$ with $\text{trace}(\mathbf{S}) = ?$

$$\text{EPE}_{in} = \text{E}_{\mathbf{y}}[\text{PE}_{in}(\hat{f})] = \text{E}_{\mathbf{y}}[\text{TE}(\hat{f})] + \frac{2}{n}\text{trace}(\mathbf{S})\sigma^2$$

Mallows' C_p Criterion:

$$C_p(\hat{f}) = \text{TE}(\hat{f}) +$$

- $C_p(\hat{f})$ is an unbiased estimator of EPE_{in}, i.e. $\text{E}_{\mathbf{y}}[C_p(\hat{f})] = \text{EPE}_{in}$.
- If the noise variance σ^2 is unknown, replace it by the MSE of a low-bias model.

AIC (Akaike, 1974)

AIC is the generalization of C_p where a log-likelihood loss is used.

Assume

- y_i, $i = 1, \ldots, n$ are iid from prob. density $f(Y; \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^p$.
- $\hat{\boldsymbol{\beta}}$ is the MLE.

By Taylor Expansion,

$$\mathrm{E}_{\mathbf{y}} \mathrm{E}_{Y^N} \log f(Y^N; \hat{\boldsymbol{\beta}}) \approx \mathrm{E}_{\mathbf{y}} \left[\frac{1}{n} \sum_{i=1}^{n} \log f(y_i; \hat{\boldsymbol{\beta}}) \right] - \frac{p}{n}.$$

$$AIC(\hat{\boldsymbol{\beta}}) = -\frac{2}{n} \sum_{i=1}^{n} \log f(y_i; \hat{\boldsymbol{\beta}}) + \frac{2p}{n} = -\frac{2}{n} \mathrm{loglik}(\hat{\boldsymbol{\beta}}) + \frac{2p}{n}.$$

When will AIC be equivalent to C_p?

BIC (Schwarz 1978)

$$BIC(\hat{\boldsymbol{\beta}}) = -\frac{2}{n}\text{loglik}(\hat{\boldsymbol{\beta}}) + \frac{(\log n)p}{n},$$

where $\hat{\boldsymbol{\beta}}$ is the MLE.

- M models for the outcome variable Y.
  Each model $\mathcal{M}_m$ is parameterized by $\boldsymbol{\beta}_m$.
- Goal: choose the model with largest posterior prob. $Pr(\mathcal{M}_m|\text{Data})$.
- Posterior odds

$$\frac{Pr(\mathcal{M}_m|\text{Data})}{Pr(\mathcal{M}_k|\text{Data})} = \frac{Pr(\mathcal{M}_m)}{Pr(\mathcal{M}_k)} \cdot \frac{Pr(\text{Data}|\mathcal{M}_m)}{Pr(\text{Data}|\mathcal{M}_k)}.$$

  The rightmost quantity $Pr(\text{Data}|\mathcal{M}_m)/Pr(\text{Data}|\mathcal{M}_k)$ is called the Bayes factor.

BIC (Schwarz 1978), Continued

- Assume uniform prior over models,

$$\frac{Pr(\mathcal{M}_m|\text{Data})}{Pr(\mathcal{M}_k|\text{Data})} = \frac{Pr(\text{Data}|\mathcal{M}_m)}{Pr(\text{Data}|\mathcal{M}_k)}$$

- Goal: choose the model with largest $Pr(\text{Data}|\mathcal{M}_m)$.
- It can be shown that

$$\log Pr(\text{Data}|\mathcal{M}_m) \approx \log Pr(\text{Data}|\hat{\boldsymbol{\beta}}_m, \mathcal{M}_m) - \frac{(\log n)p_m}{2},$$

where $\hat{\boldsymbol{\beta}}_m$ is the MLE from model $\mathcal{M}_m$.

What is the BIC under Gaussian model?

Summary of C_p, AIC and BIC

- Target:
- Loss function:
- Form of the criteria:
- Under Gaussian model (with known variance):
- Property of BIC:

Reminders

- Install Python and JupyterLab on your computer (before Oct 3rd).
- Lecture 2 quiz has been posted. Please remember to complete it before Monday (9/16) 9pm EST.
- TA office hour:
  In person office hour: 11:30-12:30pm on Tuesdays, room TBA
  Zoom office hour: 8am-9am on Fridays
  https://columbiacuimc.zoom.us/j/4698005235