What to do today ?

Part I. Introduction and Preparation
Part II. Inference under Multivariate Normal Distribution

Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-9, 11-12)

III.1. Discrimination and Classification (Chp 11)

III.1.1 Introduction
III.1.2 Two-group discriminant analysis
III.1.3 Classification with two populations

III.2. Principal Component Analysis (Chp 8)

III.3. Factor Analysis (Chp 9)

III.4. Clustering (Chp 12)

Two groups of observations: $\mathbf{x}_{A1}, \ldots, \mathbf{x}_{Am}$; $\mathbf{x}_{B1}, \ldots, \mathbf{x}_{Bn}$

The (standardized) difference between their projections on $\mathbf{b}$

$$diff_{AB}(\mathbf{b}) = \left[\frac{\bar{z}_A(\mathbf{b}) - \bar{z}_B(\mathbf{b})}{s_z}\right]^2 = \frac{\left[\mathbf{b}'(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)\right]^2}{\mathbf{b}'\mathbf{S}_{pooled}\mathbf{b}}$$

is maximized on the direction $\mathbf{a} = \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)$.

The largest difference is

$$diff_{AB}(\mathbf{a}) = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B) \propto T^2\text{-test statistic.}$$

Discriminant function: $\mathbf{a}'\mathbf{x} = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)'\mathbf{S}_{pooled}^{-1}\mathbf{x}$.

III.1.3A Classification with two populations: with known distn

Two populations with sample space Ω: Population π_A with $f_A(\cdot)$; Population π_B with $f_B(\cdot)$. Provided the costs of two misclassifications $c(A|B)$ and $c(B|A)$,

Optimal Classification Rule: $\Omega = R_A \bigcup R_B$ R_A minimizes ECM if

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{c(A|B)p_B}{c(B|A)p_A}\}.$$

Special cases:

- (a) If $p_A/p_B = 1$ (a subject from the two populations with the same probability),

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{c(A|B)}{c(B|A)}\}.$$

- (b) If $c(A|B)/c(B|A) = 1$ (the costs of the two types of misclassification are equal),

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{p_B}{p_A}\}.$$

- (c) If $c(A|B)/c(B|A) = p_A/p_B = 1$,

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq 1\}.$$

III.1.3B Classification with two populations: examples

Two populations with sample space Ω: Population π_A with $MN(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$; Population π_B with $MN(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$;

- Case 1: $\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}$
  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}$ are known,

$$R_A = \{\mathbf{x} : (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)' \boldsymbol{\Sigma}^{-1}[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B)] \geq \log\left(\frac{c(A|B)p_B}{c(B|A)p_A}\right)\}.$$

  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}$ are unknown,

$$R_A = \{\mathbf{x} : (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)' \mathbf{S}_{pooled}^{-1}[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_A + \bar{\mathbf{x}}_B)] \geq \log\left(\frac{c(A|B)p_B}{c(B|A)p_A}\right)\}.$$

- Case 2: $\boldsymbol{\Sigma}_A \neq \boldsymbol{\Sigma}_B$
  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_B$ are known,

$$R_A = \{\mathbf{x} : (\boldsymbol{\mu}_A'\boldsymbol{\Sigma}_A^{-1} - \boldsymbol{\mu}_B'\boldsymbol{\Sigma}_B^{-1})\mathbf{x} - \frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_A^{-1} - \boldsymbol{\Sigma}_B^{-1})\mathbf{x} - k \geq \log\left(\frac{c(A|B)p_B}{c(B|A)p_A}\right)\}$$

$$k = \tfrac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_A|}{|\boldsymbol{\Sigma}_B|}\right) + \tfrac{1}{2}(\boldsymbol{\mu}_A'\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\mu}_A - \boldsymbol{\mu}_B'\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B)$$

  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_B$ are unknown, use $\bar{\mathbf{x}}_A, \bar{\mathbf{x}}_B$ and $\mathbf{S}_A, \mathbf{S}_B$ to approximate them in R_A above.

Example. Classifying alaskan and Canadian salmon (textbook Example 11.8) Typically, the rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon. The study data include observations on female/male, X_1=1st year freshwater growth, X_2=1st year marine growth.

Assume equal prior probabilities, equal costs and equal covariance structure.

$$R_A = \left\{ \mathbf{x} : -0.131x_1 + 0.047x_2 \geq 3.261 \right\}$$

III.1.3C Classification with two populations: evaluating classification

Actual error rate (AER). Given a sample classification rule R_A,

$$AER = p_B \int_{R_A} f_B(\mathbf{x})d\mathbf{x} + p_A \int_{R_B} f_A(\mathbf{x})d\mathbf{x} +$$

Problem: $f_A(\cdot)$ and $f_B(\cdot)$ are unknown.

Apparent error rate (APER): $\frac{n_{AM}+n_{BM}}{n_A+n_B}$
n_A, n_B: num of subjects in Populations A,B; n_{AM}, n_{BM}: num of subjects in Populations A,B and misclassified.

- Pros: easy to obtain and no need of parametric assumptions;
- Cons: may underestimate AER – the data used to build the classification function are also used to evaluate it, bias is smaller for larger training samples.
- Example.

III.1.3C Classification with two populations: evaluating classification

Cross validation:

- (a) Randomly split data in a sample into g groups
- (b) Set aside one of the g groups as a validation sample
  - build the classification rule from the other $g-1$ groups (the training sample)
  - classify the data in the validation sample and record the results
- (c) Repeat the previous step g times

Holdout procedure (jackknife procedure)

- Cross-validation with $g = n$ groups, each with one obs.
- Estimate AER by $\frac{n_{AM}^{(H)}+n_{BM}^{(H)}}{n_A+n_B}$: $n_{AM}^{(H)}, n_{BM}^{(H)}$ are the number of holdout subjects in Populations A, B that are misclassified to Populations B,A.

III.1.3D Classification with two populations: logistic regression

Consider a logistic regression model:
$\pi(\mathbf{x}) = P(\text{obs from Population A}|\mathbf{x})$,

$$\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \boldsymbol{\beta}' \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix}$$

With *Trainining Sample*: $\mathbf{x}_{A1}, \ldots, \mathbf{x}_{Am}$; $\mathbf{x}_{B1}, \ldots, \mathbf{x}_{Bn}$,

- Introduce $Y_i = 1$ or 0 for the study's observation i from population A or B and denote the obs as $\mathbf{x}_i$, and $Y_i \sim \mathit{Binomial}(1, \pi(\mathbf{x}_i))$.
- Maximize the likelihood function $L(\boldsymbol{\beta}) = \prod_{i=1}^{m+n} \pi(\mathbf{x}_i)^{Y_i}(1-\pi(\mathbf{x}_i))^{1-Y_i}$, and obtain the MLE $\hat{\boldsymbol{\beta}}$.
- Use $\hat{\pi}(\mathbf{x}) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_p x_p)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_p x_p)}$ to classify a new data point.

III.1.3D Classification with two populations: logistic regression

Remarks of logistic regression

- Directly estimates the probability for an observation (a set of the covariate values) to be in "success" population.
- Does not assume a joint normal distribution.
- may give better results than the linear discriminant analysis when some variables are discrete.
- When the number of populations is larger than 2, can be extended to the multinomial logistic model.

What will we study next?

- *Part I. Introduction and Preparation*
- *Part II. Inference under Multivariate Normal Distribution (Chp 4-7)*
- **Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-9; 11-12)**
  - **III.1. Discriminant Analysis and Classification (Chp 11)**
    - **III.1.4 Discrimination and classification with several populations**
  - *III.2. Principal Component Analysis (Chp 8)*
  - *III.3. Factor Analysis (Chp 9)*
  - *III.4. Clustering (Chp 12)*