—— WORKSHOP ——

Applied Classical and Modern Multivariate Statistical Analysis

Module 7: Classification and Clustering

Weixing Song, Juan Du

Department of Statistics
Kansas State University

August 13, 2018

Part I: Discrimination and Classification

Outline

Discrimination and classification are multivariate techniques concerned with separating different sets of objects and with allocating new objects to previously defined groups.

Discriminant analysis is rather exploratory, and classification is less exploratory, and often requires more problem structure.

Goals of discrimination and classification:

- Discrimination: Finding the features that separate known groups in a multivariate sample.
- Classification: Developing a rule to allocate a new object into one of a number of known groups.

Connections: A classification rule is based on the features that separate the groups, so the goals overlap.

Outline

Notations and Concepts

Let take two populations/classes as the illustration example.

Notations:

- Classes: π_1, π_2.;
- Measurements: $\mathbf{X} = [X_1, X_2, \ldots, X_p]$;
- Sample Space: Ω;
- Class Sample Spaces: R_1, R_2, $\Omega = R_1 \cup R_2$.
- Class Density Functions: $f_1(\mathbf{x})$, $f_2(\mathbf{x})$;

Concepts:

Denote $P(\mathbf{X} \in R_j|\pi_i)$ the probability of classifying an object as π_j when it is from π_i.

- Misclassification Rate:

$$P(2|1) = P(\mathbf{X} \in \mathbb{R}_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x},$$

$$P(1|2) = P(\mathbf{X} \in \mathbb{R}_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}.$$

- Bayesian Misclassification Rate
  Let p_i be the prior probability of π_i, $i = 1, 2$.

$$P(\text{Observation is misclassified as } \pi_1) = P(\mathbf{X} \in \mathbb{R}_1|\pi_2)P(\pi_2) = P(1|2)p_2,$$
$$P(\text{Observation is misclassified as } \pi_2) = P(\mathbf{X} \in \mathbb{R}_2|\pi_1)P(\pi_1) = P(2|1)p_1.$$

- Sometimes, making a wrong assignment comes with certain cost. The costs of misclassification can be defined by a cost matrix

		Classify as	
		π_1	π_2
True Populations:	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

- Expected Cost of Misclassification (ECM):

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2.$$

A reasonable classification rule should have an ECM as small as possible.

Classification Rule Based on Minimizing ECM

The regions R_1 and R_2 that minimize the ECM are defined by the values of $\mathbf{x}$ for which the following inequalities hold:

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Special cases include:

- $p_1 = p_2$;
- $c(1|2) = c(2|1)$: The classification rule in this case is equivalent to the classification rule based on the total probability of misclassification (TPM).
- $p_2/p_1 = c(1|2)/c(2|1)$.

Outline

Recall the density function of $MVN(\boldsymbol{\mu}, \Sigma)$:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

Classification of Normal Populations When $\Sigma_1 = \Sigma_2$

The regions R_1 and R_2 that minimize the ECM are defined by the values of $\mathbf{x}$ for which the following inequalities hold:

$$R_1: \quad \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \geq \frac{c(1|2)p_2}{c(2|1)p_1}$$

$$R_2: \quad \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] < \frac{c(1|2)p_2}{c(2|1)p_1}$$

Given the regions R_1 and R_2, we can construct the following classification rule.

Population Classification Rule: Equal Covariances

Let the populations π_1 and π_2 be described by the MVN with mean $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and with the same covariance matrix Σ. The allocation rule that minimizes the ECM is to allocate $\mathbf{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right],$$

and allocate $\mathbf{x}_0$ to π_2 otherwise.

In most practical situations, the population quantities $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Σ are unknown.

To implement the above classification rules, the population quantities are replaced by the sample analogues.

Suppose we have n_1 observations of $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ from π_1 and n_2 measurements of this quantity from π_2, with $n_1 + n_2 - 2 \geq p$. The respective data matrices are

$$\mathbf{X}_1 = \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{1n_1} \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \vdots \\ \mathbf{x}'_{2n_2} \end{bmatrix}.$$

Let $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ be the sample mean vectors, and $\mathbf{S}_1$, $\mathbf{S}_2$ be the sample covariance matrices.

Define the pooled sample covariance matrix as

$$\mathbf{S}_{\text{pooled}} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}.$$

Sample Classification Rule: Equal Covariances

The sample classification rule allocates $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right],$$

and allocate $\mathbf{x}_0$ to π_2 otherwise.

Example 7.1 (Hemophilia A Carriers): To construct a procedure for detecting potential hemophilia A carriers, blood samples were assayed for two groups of women and measurements on the two variables

$$X_1 = \log_{10}(\text{AHF activity}), \quad X_2 = \log_{10}(\text{AHF-like antigen}).$$

The data set can be found in R package `rrcov`.

Remark:

- **Scaling:** The coefficient vector $\hat{\mathbf{a}} = \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ is frequently "scaled" or "normalized" to help the interpretation of its elements.

  Two commonly used scalings are

  - $\hat{\mathbf{a}}^* = \frac{\hat{\mathbf{a}}}{\sqrt{\hat{\mathbf{a}}'\hat{\mathbf{a}}}}$;
  - $\hat{\mathbf{a}}^* = \frac{\hat{\mathbf{a}}}{\hat{a}_1}$, where $\hat{a}_1$ is the first element in $\hat{\mathbf{a}}$.

  Scaling is recommended only if the X variables have been standardized.

- **Fisher's approach:** Using an entirely different argument, Fisher developed a dimension-reduction-type classification approach, which is equivalent to linear discriminant approach.

  Fisher's approach does not assume that the population is normal. However, it does assume that the population covariance matrices are equal.

Population Classification Rule: Unequal Covariances

Let the populations π_1 and π_2 be described by the MVN densities with mean vectors and covariance matrices $\boldsymbol{\mu}_1$, Σ_1 and $\boldsymbol{\mu}_2$, Σ_2, respectively. The allocation rule that minimizes the ECM is to allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x}_0 + (\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1})\mathbf{x}_0 - k \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right],$$

and allocate $\mathbf{x}_0$ to π_2 otherwise, where

$$k = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2).$$

Quadratic Classification Rule: Normal Population with Unequal Covariance Matrices

The sample classification rule allocates $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right],$$

and allocate $\mathbf{x}_0$ to π_2 otherwise, where

$$k = \frac{1}{2}\log\left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\right) + \frac{1}{2}(\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1}\bar{\mathbf{x}}_2).$$

Outline

The performance of any classification procedure is often measured by the error rates or misclassification probabilities.

- When the population density functions are known, the misclassification probabilities can be calculated.

  The optimum error rate (OER)

  $$\text{OER} = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x},$$

  where

  $$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \quad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}.$$

  Example: Derive an expression for the OER when $p_1 = p_2 = 1/2$, and $\pi_1 : MVN(\boldsymbol{\mu}_1, \Sigma)$, $\pi_2 : MVN(\boldsymbol{\mu}_2, \Sigma)$.

  We can show that, with $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$,

  $$\text{OER} = \text{minimum TPM} = \Phi(-\Delta/2)$$

- The performance of sample classification functions can be evaluated by the actual error rate (AER).

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x}) d\mathbf{x},$$

where $\hat{R}_1$ and $\hat{R}_2$ represent the classification regions determined by the sample. For LDA,

$$\hat{R}_1 : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \log \left[\frac{c(1|2) p_2}{c(2|1) p_1} \right]$$

$$\hat{R}_2 : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) < \log \left[\frac{c(1|2) p_2}{c(2|1) p_1} \right]$$

Unfortunately, AER cannot be calculated because it depends on the unknown density functions.

- Apparent Error Rate (APER): The fraction of observations in the training sample that are misclassified by the sample classification function.

  The APER can be easily calculated from the confusion matrix, which shows actual versus predicted group membership. For n_1 observations from π_1 and n_2 observations from π_2, the confusion matrix has the form

		Predicted membership	
		π_1	π_2
Actual Membership:	π_1	n_{1c}	$n_{1m} = n_1 - n_{1c}$
	π_2	$n_{2m} = n_2 - n_{2c}$	n_{2c}

  where

  - n_{1c} = number of π_1 items correctly classified as π_1 items
  - n_{1m} = number of π_1 items misclassified as π_2 items
  - n_{2c} = number of π_2 items correctly classified as π_2 items
  - n_{2m} = number of π_2 items misclassified as π_1 items

  The APER is defined as

$$\text{APER} = \frac{n_{1m} + n_{2m}}{n_1 + n_2}$$

Remark:

- The APER tends to underestimate the AER.
- A better procedure to estimate the AER is to split the total sample into a training sample and a validation sample. The training sample is used to construct the classification function, and the validation sample is used to evaluate it.

  It suffers from two main defects:

  - It requires large samples;
  - The function evaluated is not the function of interest.

- Lachenbruch's "holdout" procedure
  - Start with the π_1 group of observations. Omit the one observation from this group, and develop a classification function based on the remaining $n_1 - 1, n_2$ observations.
  - Classify the "holdout" observation, using the function construct in Step 1.
  - Repeat Step 1 and Step 2 until all of the π_1 observations are classified. Let $n_{1m}^{(H)}$ be the number of holdout (H) observations misclassified in this group.
  - Repeat Step 1 through Step 3 for the π_2 observations. Let $n_{1m}^{(H)}$ be the number of holdout observations misclassified in this group.

Then we can estimate $P(2|1)$ and $P(1|2)$ by

$$\hat{P}(2|1) = \frac{n_{1m}^{(H)}}{n_1}, \quad \hat{P}(1|2) = \frac{n_{2m}^{(H)}}{n_2}.$$

and estimate $E(AER)$ by

$$\hat{E}(AER) = \frac{n_{1m}^{(H)} + h_{2m}^{(H)}}{n_1 + n_2}.$$

Note: For moderate sample size, Lachenbruch's estimate is nearly unbiased.

Outline

In theory, the generalization of classification procedure from 2 to $g > 2$ groups is straightforward. However, not much is known about the properties of the corresponding sample classification functions, for example, their error rates have not been fully investigated.

Notation:

- π_i: the i-th population, $i = 1, 2, \ldots, g$;
- $f_i(\mathbf{x})$: the density associated with π_i;
- $c(k|i)$: the cost of allocating an item to π_k when it belongs to π_i;
- R_k: the set of $\mathbf{x}$-values belonging to π_k;
- $P(k|i)$: the probability of classifying item as π_k when it belongs to π_i.

Expected Cost of Misclassifying (ECM)

- The conditional ECM of a item from π_i into other populations is

$$\mathrm{ECM}(i) = \sum_{k \neq i} P(k|i)c(k|i).$$

- The overall ECM is defined as

$$\mathrm{ECM} = \sum_{i=1}^{g} p_i \left(\sum_{k \neq i} P(k|i)c(k|i) \right).$$

Classification Based on Minimizing ECM

The classification regions that minimizes the overall ECM are defined by allocating $\mathbf{x}$ to that population π_k for which

$$\sum_{i \neq k} p_i f_i(\mathbf{x})c(k|i)$$

is smallest, $k = 1, 2, \ldots, g$. If a tie occurs, $\mathbf{x}$ can be assigned to any of the tied populations.

Minimum ECM Classification Rule with Equal Misclassification Costs

Allocate $\mathbf{x}_0$ to π_k if $p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x})$ for all $i \neq k$.

Remark:

- The above classification rule is identical to the one that maximizes the "posterior" probability $P(\pi_k|\mathbf{x})$. Note that

$$P(\pi_k|\mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{\sum_{i=1}^{g} \pi_i f_i(\mathbf{x})}, \quad k = 1, 2, \ldots, g.$$

- To implement the above classification rule, we must specify: prior probabilities, misclassification costs, and population densities.

Classification with Normal Populations

Recall the MVN density function.

Based on the classification rule, $\mathbf{x}$ will be allocated to π_k if

$$\log p_k f_k(\mathbf{x}) = \log p_k - \frac{p\log(2\pi)}{2} - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)$$

is the largest among all $\log \pi_i f_i(\mathbf{x})$, $i = 1, 2, \ldots, g$.

Define the quadratic discrimination score (QDS) for the i-th population as

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log p_i.$$

Minimum TPM Rule for MVN: Unequal Σ_i

Allocate $\mathbf{x}$ to π_k if

$$d_k^Q(\mathbf{x}) = \text{the largest of } d_1^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x}).$$

In practice, the $\boldsymbol{\mu}_i$ and Σ_i are unknown. Sample analogs will be used.

Let n_i, $\bar{\mathbf{x}}_i$, $\mathbf{S}_i$ denote the sample size, sample mean vector and covariance matrix of the sample from the i-th population.

The sample QDS is defined as

$$\hat{d}_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{S}}_i) + \log p_i.$$

Minimum TPM Rule for MVN: Unequal Σ_i

Allocate $\mathbf{x}$ to π_k if

$$\hat{d}_k^Q(\mathbf{x}) = \text{the largest of } \hat{d}_1^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x}).$$

Note: If all Σ_i are equal, the population QDS becomes

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_i'\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i + \log p_i.$$

Define the linear discriminant score (LDS) as

$$d_i(\mathbf{x}) = \boldsymbol{\mu}_i'\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i + \log p_i, \quad i = 1, 2, \ldots, g,$$

and an estimated LDS is given by

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \frac{1}{2}\bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_i + \log p_i, \quad i = 1, 2, \ldots, g,$$

where $\mathbf{S}_{\text{pooled}}$ is defined as

$$\mathbf{S}_{\text{pooled}} = \frac{1}{n_1 + n_2 + \cdots + n_g - g}[(n_1 - 1)\mathbf{S}_1 + \cdots + (n_g - 1)\mathbf{S}_g].$$

Estimated Minimum TPM Rule for MVN: Equal Σ_i

Allocate $\mathbf{x}$ to π_k if

$$\hat{d}_k(\mathbf{x}) = \text{the largest of } \hat{d}_1(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x}).$$

Fisher's Method

Motivation: To obtain a reasonable representation of the populations that involve only a few linear combinations of the observations, such as $\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x}, \ldots$.

Benefits:

- Convenient representations of the g populations that reduce the dimension from a very large number of characteristics to a relatively few linear combinations.
- Plotting of the means of the first two or three linear combinations (discriminants). This helps display the relationships and possible grouping of the populations.
- Scatter plots of the sample values of the fist two discriminants, which can indicate outliers or other abnormalities in the data.

Note: Fisher's method does not assume normality. However, the population covariance matrices are assumed to be equal and of full rank.

Theory:

- $\boldsymbol{\mu}_i$: mean vectors, $i = 1, 2, \ldots, g$;
- Σ : mean covariance matrix;
- $\bar{\boldsymbol{\mu}}$: combined mean vectors.

Define

$$\mathbf{B}_{\boldsymbol{\mu}} = \sum_{i=1}^{g} (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})', \quad \bar{\boldsymbol{\mu}} = \frac{1}{g} \sum_{i=1}^{g} \boldsymbol{\mu}_i.$$

For any linear combination $Y = \mathbf{a}'\mathbf{X}$, we have, for population π_i,

$$E(Y) = \mathbf{a}'\boldsymbol{\mu}_i, \quad \mathrm{Var}(Y) = \mathbf{a}'\Sigma\mathbf{a}.$$

Define

$$\frac{\substack{\text{SS distances from populations} \\ \text{to overall mean of } Y}}{\mathrm{Var}(Y)} = \frac{\sum_{i=1}^{g}(\mathbf{a}'\boldsymbol{\mu}_i - \mathbf{a}'\bar{\boldsymbol{\mu}})^2}{\mathbf{a}'\Sigma\mathbf{a}} = \frac{\mathbf{a}'\mathbf{B}_{\boldsymbol{\mu}}\mathbf{a}}{\mathbf{a}'\Sigma\mathbf{a}}.$$

This ratio measures the variability between the groups of Y-values relative to the common variability within groups.

We can select $\mathbf{a}$ to maximize this ratio.

$\boldsymbol{\mu}_i$ and Σ are unavailable in general, and sample analogs will be used.

The sample between and within groups matrices are defined by

$$\mathbf{B} = \sum_{i=1}^{g} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \quad \mathbf{W} = \sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$$

Fisher's Sample Linear Discriminants

Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_s > 0$ denote the $s \leq \min(g-1, p)$ nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_s$ be the corresponding eigenvectors (scaled so that $\hat{\mathbf{e}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{e}} = 1$). Then the vector of coefficients $\hat{\mathbf{a}}$ that maximizes the ratio $\hat{\mathbf{a}}'\mathbf{B}\hat{\mathbf{a}}/\hat{\mathbf{a}}'\mathbf{W}\hat{\mathbf{a}}$ is given by $\hat{\mathbf{a}}_1 = \hat{\mathbf{e}}_1$. The linear combination $\hat{\mathbf{a}}_1\mathbf{x}$ is called the sample first discriminant. The choice $\hat{\mathbf{a}}_k = \hat{\mathbf{e}}_k$ produces the sample k-th discriminant, $k \leq s$.

Fisher's Discriminant: Classification

Fisher's discriminants were derived for the purpose of obtaining a low-dimensional representation of the data that separates the populations as much as possible.

Although they were derived from considerations of separation, the discriminant also provide the basis for a classification rule.

Let

$$Y_k = \mathbf{a}_k'\mathbf{X} = k\text{-th deiscriminant}, \quad k \leq s.$$

Under population π_i, $\mathbf{Y} = (Y_1, \ldots, Y_s)'$ has mean vector $\boldsymbol{\mu}_{iY} = (\boldsymbol{\mu}_{iY_1}, \ldots, \boldsymbol{\mu}_{iY_s})' = (\mathbf{a}_1'\boldsymbol{\mu}_i, \ldots, \mathbf{a}_s'\boldsymbol{\mu}_i)$ and covariance matrix $\mathbf{I}$.

If only r of the discriminants are used for allocation, the rule is to allocate $\mathbf{x}$ to π_k if

$$\sum_{j=1}^{r}(y_j - \mu_{kY_j})^2 = \sum_{j=1}^{r}[\mathbf{a}_j'(\mathbf{x} - \boldsymbol{\mu}_k)]^2 \leq \sum_{j=1}^{r}[\mathbf{a}_j'(\mathbf{x} - \boldsymbol{\mu}_i)]^2 \quad \text{for all } i \neq k.$$

Relationship to the "normal theory" discriminant scores

Connection between Fisher's and Normal-Theory Based Classification Rules

Let $\mathbf{a}_j = \Sigma^{-1/2}\mathbf{e}_j$ and $\mathbf{e}_j$ is an eigenvector of $\Sigma^{-1/2}\mathbf{B}_{\boldsymbol{\mu}}\Sigma^{-1/2}$. Then

$$\sum_{j=1}^{p}(y_j - \mu_{iY_j}) = (\mathbf{x} - \boldsymbol{\mu}_i)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) = -2d_i(\mathbf{x}) + \mathbf{x}'\Sigma^{-1}\mathbf{x} + 2\log(p_i).$$

If $\lambda_1 \geq \cdots \geq \lambda_s > 0 = \lambda_{s+1} = \cdots = \lambda_p$, $\sum_{j=s+1}^{p}(y_j - \mu_{iY_j})$ is constant for all populations $i = 1, 2, \ldots, g$, so only the first s discriminants contribute to the classification.

Remark: If the prior probabilities are the same, Fisher's rule with $r = s$ is equivalent to the population version of the minimum TPM rule.

Fisher's Classification Procedure Based on Sample Discriminants

Allocate $\mathbf{x}$ to π_k if

$$\sum_{j=1}^{r}(\hat{y}_j - \bar{y}_{kj})^2 = \sum_{j=1}^{r}[\hat{\mathbf{a}}_j(bx - \bar{\mathbf{x}}_k)]^2 \leq \sum_{j=1}^{r}[\hat{\mathbf{a}}_j(bx - \bar{\mathbf{x}}_i)]^2 \quad \text{for all } i \neq k$$

where $\hat{\mathbf{a}}_j$ is defined on page 35, $\bar{y}_{kj} = \hat{\mathbf{a}}_j'\bar{\mathbf{x}}_k$ and $r \leq s$.

Remark:

- If the prior probabilities are the same and $r = s$, Fisher's rule is equivalent to the sample version of the minimum TPM rule.
- Why the fist few discriminants are more important than the last few. In fact, we can show

$$\begin{aligned}
\Delta_S^2 &= \sum_{i=1}^{g} (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})' \Sigma^{-1} (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}) \\
&= \lambda_1 + \cdots + \lambda_p \\
&= \sum_{i=1}^{g} (\boldsymbol{\mu}_{iY} - \bar{\boldsymbol{\mu}}_Y)' (\boldsymbol{\mu}_{iY} - \bar{\boldsymbol{\mu}}_Y) \\
&= \sum_{i=1}^{g} (\mu_{iY_1} - \bar{\mu}_{Y_1})^2 + \cdots + \sum_{i=1}^{g} (\mu_{iY_p} - \bar{\mu}_{Y_p})^2
\end{aligned}$$

It follows that the first discriminant makes the largest contribution λ_1 to the separative measure Δ_S^2.

Outline

Statistical Model:

- Y : bi-valued response variable. Often coded as 0 and 1;
- $\mathbf{z} = (1, z_1, z_2, \ldots, z_r)'$: predictors;
- $E(Y|\mathbf{z}) = P(Y = 1|\mathbf{z}) = p(\mathbf{z})$ and

$$\log\left(\frac{p(\mathbf{z})}{1 - p(\mathbf{z})}\right) = \boldsymbol{\beta}'\mathbf{z},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)$.

Likelihood Function:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} p^{y_i}(\mathbf{z}_j)[1 - p(\mathbf{z}_j)]^{1-y_j} = \frac{\prod_{j=1}^{n} e^{y_j(\beta_0+\beta_1 z_{j1}+\cdots+\beta_r z_{jr})}}{\prod_{j=1}^{n}[1 + e^{\beta_0+\beta_1 z_{j1}+\cdots+\beta_r z_{jr}}]}.$$

Statistical Inferences:

- Maximum Likelihood Estimation.

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}).$$

- When the sample size is large,

$$\hat{\boldsymbol{\beta}} \rightsquigarrow N_{r+1}\left(\boldsymbol{\beta}, \left[\sum_{j=1}^{n} \hat{p}(\mathbf{z}_j)(1-\hat{p}(z_j))\mathbf{z}_j\mathbf{z}_j'\right]^{-1}\right).$$

- Confidence interval for β_k is

$$\hat{\beta}_k \pm z_{1-\alpha/2}\text{SE}(\hat{\beta}_k), \quad k = 0, 1, \ldots, r.$$

- Likelihood ratio test for $H_0 : \beta_k = 0$.
  Let $\hat{\boldsymbol{\beta}}$ denote the MLE of $L(\beta_0, \beta_1, \ldots, \beta_r)$, and $\tilde{\boldsymbol{\beta}}$ denote the MLE of $(\beta_0, \ldots, \beta_{k-1}, \beta_k, \ldots, \beta_r)$ for the reduced model $L(\beta_0, \ldots, \beta_{k-1}, \beta_k, \ldots, \beta_r)$.
  Then the deviance

$$-2[\log L(\tilde{\boldsymbol{\beta}}) - \log L(\hat{\boldsymbol{\beta}})] \sim \chi_1^2.$$

Classification via Logistic Regression

Classification vis Logistic Regression

Assign $\mathbf{z}$ to population 1 if the estimated odds ratio is greater than 1 or

$$\frac{\hat{p}(\mathbf{z})}{1-\hat{p}(\mathbf{z})} = \exp(\hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r) > 1$$

or

$$\hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r > 0.$$

Part II: Clustering

Outline

Searching the data for a structure of natural clusters is an important exploratory technique. Clusters can provide an exploratory means for assessing dimensionality, identifying outliers, and suggesting interesting hypotheses concerning relationships.

Clustering is distinct from the classification methods.

Classification assumes that the groups are known, and the objective is to assign new observations to one of these groups. In statistical machine learning term, classification is a “supervised learning method”.

Clustering is a more primitive technique in that no assumptions are made concerning the number of groups. In statistical machine learning term, classification is a “unsupervised learning method”.

Clustering is done on the basis of similarities or dissimilarities (distances), so meaningful clusterings depend on the definition of similarity.

Outline

When items (units/cases/observations) are clustered, similarity is usually indicated by some sort of distance; while when variables are grouped, similarity is usually measured by correlation coefficients.

Similarity or Dissimilarity of Two Items: Suppose two items are

$$\mathbf{x}' = [x_1, x_2, \ldots, x_p], \quad \mathbf{y}' = [y_1, y_2, \ldots, y_p].$$

Some commonly used distances:

- Euclidean Distance: $d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} (x_i - y_i)^2\right]^{1/2}$.
- Manhattan (City-Block) Distance: $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} |x_i - y_i|$.
- Canberra Distance (for nonnegative variables only)

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}.$$

- Czekanowski Distance (for nonnegative variables only)

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}.$$

Whenever possible, it is advisable to use "true" distances.

The distances defined previously are suitable for "continuous variables".

When items cannot be represented by continuous values, they are often compared on the basis of the presence or absence of certain characteristics, that is, the binary variable,

$$1 = \text{present}, \quad 0 = \text{absent}.$$

Let x_{ij} be the score (0 or 1) of the j-th binary variable on the i-th item. Then

$$\sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

provides a count of the number of mismatches.

Note: The above distance suffers from weighting the 1-1 and 0-0 matches equally.

To allow for different treatment of the 1-1 and 0-0 matches, several schemes of defining similarity coefficients have been proposed.

For two items i and k, suppose

		Item k 1	Item k 0	Totals
Item i	1	a	b	$a+b$
	0	c	d	$c+d$
Totals		$a+c$	$b+d$	$p=a+b+c+d$

a represents the frequency of 1–1 matches,

b is the frequency of 1–0 matches, and so forth.

SIMILARITY COEFFICIENTS FOR CLUSTERING ITEMS

Coefficient	Rationale
1. $\dfrac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2. $\dfrac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3. $\dfrac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\dfrac{a}{p}$	No 0–0 matches in numerator.
5. $\dfrac{a}{a+b+c}$	No 0–0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant.)
6. $\dfrac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7. $\dfrac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\dfrac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded.

Example (Calculating the values of a similarity coefficient)

Suppose five individuals possess the following characteristics:

	Height	Weight	Eye color	Hair color	Handedness	Gender
Individual 1	68 in	140 lb	green	blond	right	female
Individual 2	73 in	185 lb	brown	brown	right	male
Individual 3	67 in	165 lb	blue	blond	right	male
Individual 4	64 in	120 lb	brown	brown	right	female
Individual 5	76 in	210 lb	brown	brown	left	male

Define six binary variables $X_1, X_2, X_3, X_4, X_5, X_6$ as

$$X_1 = \begin{cases} 1 & \text{height} \geq 72 \text{ in.} \\ 0 & \text{height} < 72 \text{ in.} \end{cases} \qquad X_4 = \begin{cases} 1 & \text{blond hair} \\ 0 & \text{not blond hair} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{weight} \geq 150 \text{ lb} \\ 0 & \text{weight} < 150 \text{ lb} \end{cases} \qquad X_5 = \begin{cases} 1 & \text{right handed} \\ 0 & \text{left handed} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{brown eyes} \\ 0 & \text{otherwise} \end{cases} \qquad X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

Based on the similarity coefficient 1, the similarity coefficients for these 5 individuals can be summarized in the following matrix:

$$
\text{Individual}\quad
\begin{array}{c}
 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5
\end{array}
\begin{array}{c}
\text{Individual} \\
\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{bmatrix}
1 & & & & \\
\frac{1}{6} & 1 & & & \\
\frac{4}{6} & \frac{3}{6} & 1 & & \\
\frac{4}{6} & \frac{3}{6} & \frac{2}{6} & 1 & \\
0 & \frac{5}{6} & \frac{2}{6} & \frac{2}{6} & 1
\end{bmatrix}
\end{array}
$$

Similarity and Association Measures for Variables

In some applications, it is the variables, not the items, that must be grouped.

For continuous variables, the sample correlation coefficient is often used as the similarity measure.

For binary variables with n items, the n items are categorized, with the usual 0 and 1 coding, the contingency table becomes

		Variable k 1	Variable k 0	Totals
Variable i	1	a	b	$a+b$
	0	c	d	$c+d$
	Totals	$a+c$	$b+d$	$n=a+b+c+d$

The usual product moment correlation coefficient is given by

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Outline

There are two hierarchical clustering methods.

- Agglomerative hierarchical methods

  Starts with the individual objects, then the most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster.

- Divisive hierarchical methods

  An initial single group of objects is divided into two subgroups, these subgroups are then further divided into dissimilar subgroups; the process continues until there are as many subgroups as objects.

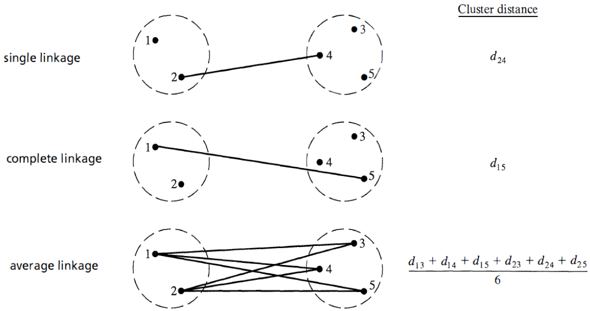
Steps for Agglomerative Hierarchical Clustering Algorithm (AHCA):

Suppose we have N objects.

1. Start with N clusters, each contains a single entity;
2. Compute the $N \times N$ symmetric matrix of distances (or similarities) $\mathbf{D} = \{d_{ij}\}$;
3. Search the distance matrix for the nearest pair of clusters. Let the distance between the "most similar" clusters U and V be d_{UV};
4. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by
   - a. deleting the rows and columns corresponding to clusters U and V;
   - b. adding a row and column giving the distances between cluster (UV) and the remaining cluster;
5. Repeat Steps 3-4 a total of $N-1$ times. Record the identity of clusters that are merged and the levels (distances or similarities) at which the merges take place.

- Single Linkage (minimum distance or nearest neighbor)
- Complete Linkage (maximum distance or farthest neighbor)
- Average Linkage (average distance)

Graphical illustration of linkages



Example (Clustering using single linkage): We use the following hypothetical distances between five objects to illustrate the single linkage algorithm.

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

- $$\begin{array}{c|c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array} \Longrightarrow \min_{i,k}(d_{ik}) = d_{53} = 2 \Longrightarrow (35)$$

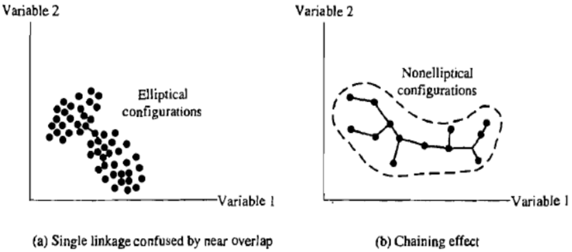
- $$\begin{aligned} d_{(35)1} &= \min\{d_{31}, d_{51}\} = 3 \\ d_{(35)2} &= \min\{d_{32}, d_{52}\} = 7 \\ d_{(35)4} &= \min\{d_{34}, d_{54}\} = 8 \end{aligned} \Longrightarrow \begin{array}{c|c} & \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{array}{c} (35) \\ 1 \\ 2 \\ 4 \end{array} & \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array} \Longrightarrow (135)$$

- $$\begin{aligned} d_{(135)2} &= \min\{d_{12}, d_{32}, d_{52}\} = 7 \\ d_{(135)4} &= \min\{d_{14}, d_{34}, d_{54}\} = 6 \end{aligned} \Longrightarrow \begin{array}{c|c} & \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{array}{c} (135) \\ 2 \\ 4 \end{array} & \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array} \Longrightarrow (24)$$

- $$d_{(135)(24)} = \min\{7, 6\} = 6 \Longrightarrow \begin{array}{c|c} & \begin{array}{cc} (135) & (24) \end{array} \\ \begin{array}{c} (135) \\ (24) \end{array} & \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array} \Longrightarrow (12345)$$

Remarks on Single Linkage

Refer to the following figure



Single linkage clusters.

- Since single linkage joins clusters by the shortest link between them, the technique cannot discern poorly separated clusters;
- Single linkage is one of the few clustering methods that can delineate non-ellipsoidal clusters. The tendency of single linkage to pick out long stringlike clusters is known as chaining.

Complete Linkage: Complete linkage proceeds in much the same manner as single linkage clusterings, with one important exception: at each stage, the distance (similarity) between clusters is determined by the distance (similarity) between the two elements, one from each cluster, that are most distant.

Average Linkage: The procedure is similar to that of the single linkage and complete linkage, except that the distance between two clusters are defined as

$$d_{UV} = \frac{\sum_i \sum_k d_{ik}}{N_U N_V},$$

where d_{ik} is the distance between object i in the cluster U and object k in the cluster W, and N_U and N_V are the number of items in clusters U and V.

Ward's method is based on minimizing the "loss of information" from joining two groups.

Suppose we have N items.

For a given cluster k, let ESS_k be the sum of squared deviations of every item in the cluster from the cluster mean (centroid).

Key Steps in Ward's Procedure

- Initially, each cluster consists of a single item. So

$$\mathrm{ESS}_k = 0, k = 1, 2, \ldots, N,$$

  thus the sum of the cluster ESS_k is $\mathrm{ESS} = 0$.
- At each step in the analysis, the union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS are joined.

Ward's method is based on the notion that the clusters of multivariate observations are expected to be roughly elliptically shaped.

Nonhierachical clustering methods are designed to group items, rather than variables, into K clusters.

K-means Method

Step 1: Partition the items into initial clusters.

We can start with a partition of all items into K preliminary groups, or specify K initial centroids (seed points).

Step 2: Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. Then recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.

Step 3: Repeat Step 2 until no more reassignment take place.

Mixture Model

Suppose a population has K clusters.

The probability density function of the k-th cluster is $f_k(\mathbf{x})$.

For any observation $\mathbf{x}$ from this population, with probability p_k it comes from the k-th cluster, $k = 1, 2, \ldots, K$.

Therefore, the density function of $\mathbf{x}$ is

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}),$$

where $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$.

Normal Mixture Model

The most common mixture model is the mixture of multivariate normal distributions where $f_k(\mathbf{x}) = N_p(\boldsymbol{\mu}_k, \Sigma_k)$.

The likelihood function L based a sample of size n, $\{\mathbf{x}_j\}_{j=1}^n$ is

$$\prod_{j=1}^{n}\left[\sum_{k=1}^{K}\frac{p_k}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}_j-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_k)\right)\right].$$

One can estimate the unknown parameters using maximum likelihood estimation procedure.

After all the parameters are estimated, the j-th observation will be assigned to the k-th cluster for which the conditional probability of membership

$$p(k|\mathbf{x}_j) = \frac{\hat{p}_k f_k(\mathbf{x}_j)}{\sum_{i=1}^{K}\hat{p}_i f_i(\mathbf{x}_j)}$$

is the largest.

Note:

- Computing the MLE is very complicated. Additional covariance structure are often considered to reduce the calculation complexity.
- How to decide the number of clusters K?
    - Minimizing the AIC

$$AIC = -2\log(L_{\max}) + 2\left[\frac{K(p+1)(p+2)}{2} - 1\right]$$

    - Minimizing the BIC

$$BIC = -2\log(L_{\max}) + 2\log(n)\left[\frac{K(p+1)(p+2)}{2} - 1\right]$$