High Dimensional Data Analysis

Chapter 2: Discrimination and classification

Discrimination and classification

- Introduction
- Distance Discriminant
- Bayesian Discriminant
- Fisher Discriminant
- Stepwise Discriminant

Objectives of Discriminant Analysis

- Objective 1 (Prediction): Classification (or Allocation).
  In cases where the objects of study have been divided into several groups (also called classes or populations) using certain methods historically, to determine the group to which a new observed sample belongs.
- Objective 2 (Description): Separation.
  That is, using graphical methods (usually two-dimensional, sometimes three-dimensional or one-dimensional, generally achieved through dimensionality reduction) or algebraic methods to describe the differences between samples from each group, separating the groups to the maximum extent.

Introduction

Examples of discriminant classification:

- 1. Solvent and insolvent property liability insurance companies.
  Measured variables: Total assets, value of stocks and bonds, market value of stocks and bonds, loss expenses, surplus, signed premium amount.
- 2. Non-ulcer dyspepsia group (those with gastric dysfunction) and control group ("normal" individuals).
  Measured variables: Measures of anxiety, dependency, guilt, perfectionism.

Introduction (cont.)

- 3. Two kinds of weeds.
  Measured variables: Length of sepals and petals, depth of petal cleft, length of bract, pollen diameter.
- 4. Fast buyers and slow buyers of new products.
  Measured variables: Education, income, family size, number of past brand switches.
- 5. Good credit and bad credit risks.
  Measured variables: Income, age, number of credit cards, family size.
- The p-dimensional indicator values of all samples in each group form a p-variate population distribution for that group. We attempt to determine which group a new sample x comes from, primarily based on the population distributions or their distributional characteristics of each group.

$$\mathbf{x}' = (x_1, x_2, \ldots, x_p)$$

Distance Discriminant

Distance Discriminant

- I. Two-group Distance Discriminant
- II. Multi-group Distance Discriminant

I. Two-group Distance Discriminant

Assume the means of groups π_1 and π_2 are μ_1 and μ_2 respectively, and the covariance matrices are Σ_1 and Σ_2 ($\Sigma_1, \Sigma_2 > 0$), x is a new sample (p-dimensional), now to determine which group it comes from.

- 1. Discriminant when $\Sigma_1 = \Sigma_2 = \Sigma$
- 2. Discriminant when $\Sigma_1 \neq \Sigma_2$

1. Discriminant when $\Sigma_1 = \Sigma_2 = \Sigma$

Discriminant rule:

$$\begin{cases} \mathbf{x} \in \pi_1 & \text{if } d^2(\mathbf{x}, \pi_1) \leq d^2(\mathbf{x}, \pi_2) \\ \mathbf{x} \in \pi_2 & \text{if } d^2(\mathbf{x}, \pi_1) > d^2(\mathbf{x}, \pi_2) \end{cases}$$

where

$$d^2(\mathbf{x}, \pi_i) = (\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i), \ i = 1, 2$$

This can be simplified to:

$$W(\mathbf{x}) = \mathbf{a}'\mathbf{x} - \mathbf{a}'\bar{\mu} = \mathbf{a}'(\mathbf{x} - \bar{\mu})$$

where

$$\mathbf{a} = \Sigma^{-1}(\mu_1 - \mu_2), \ \bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$$

Rule:

$$\begin{cases} \mathbf{x} \in \pi_1 & \text{if } W(\mathbf{x}) \geq 0 \\ \mathbf{x} \in \pi_2 & \text{if } W(\mathbf{x}) < 0 \end{cases}$$

W(x) is the (linear) discriminant function for two-group distance discriminant, a is the discriminant coefficient vector.

Misclassification Probability

$$P(2|1) = P(W(\mathbf{x}) < 0|\mathbf{x} \in \pi_1)$$
$$P(1|2) = P(W(\mathbf{x}) \geq 0|\mathbf{x} \in \pi_2)$$

Assume $\pi_1 \sim N_p(\mu_1, \Sigma)$, $\pi_2 \sim N_p(\mu_2, \Sigma)$, then

$$P(2|1) = P(1|2) = \Phi\left(-\frac{\Delta}{2}\right)$$

where

$$\Delta^2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

is the Mahalanobis distance between the two groups.

- The more separated the two normal groups are, the smaller the two misclassification probabilities are, and the better the discriminant effect is at this time.
- When the two normal groups are very close, both misclassification probabilities will be large, and discriminant analysis will have no practical significance.

2. Discriminant Function for $\Sigma_1 \neq \Sigma_2$

- Discriminant rule:

$$\begin{cases} x \in \pi_1, & \text{if } d^2(x, \pi_1) \leq d^2(x, \pi_2) \\ x \in \pi_2, & \text{if } d^2(x, \pi_1) > d^2(x, \pi_2) \end{cases}$$

2. Discriminant Function for $\Sigma_1 \neq \Sigma_2$

- Discriminant rule:

$$\begin{cases} x \in \pi_1, & \text{if } d^2(x, \pi_1) \leq d^2(x, \pi_2) \\ x \in \pi_2, & \text{if } d^2(x, \pi_1) > d^2(x, \pi_2) \end{cases}$$

- Another form:

$$\begin{aligned} W(x) =& d^2(x, \pi_1) - d^2(x, \pi_2) \\ =& (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1) - (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2) \end{aligned}$$

The corresponding discriminant rule is

$$\begin{cases} x \in \pi_1, & \text{if } W(x) \leq 0 \\ x \in \pi_2, & \text{if } W(x) > 0 \end{cases}$$

2. Discriminant Function for $\Sigma_1 \neq \Sigma_2$

Assuming that the variances of the two groups π_1 and π_2 are different, the discriminant function can be simply taken as (Assume $\mu_1 < x < \mu_2$):

$$W(x) = d(x, \pi_1) - d(x, \pi_2) = \frac{x - \mu_1}{\sigma_1} - \frac{\mu_2 - x}{\sigma_2}$$
$$= \frac{\sigma_1 + \sigma_2}{\sigma_1 \sigma_2} \left(x - \frac{\sigma_2 \mu_1 + \sigma_1 \mu_2}{\sigma_1 + \sigma_2} \right) = \frac{\sigma_1 + \sigma_2}{\sigma_1 \sigma_2} (x - \mu^*)$$

where

$$\mu^* = \frac{\sigma_2 \mu_1 + \sigma_1 \mu_2}{\sigma_1 + \sigma_2}$$

μ^* is the weighted average of μ_1 and μ_2, the threshold point.

- If $x \leq u^*$, then $x \in \pi_1$.
- If $x > u^*$, then $x \in \pi_2$.

II. Multi-group Distance Discriminant

Suppose there are k m-dimensional populations: $\pi_1, \pi_2, \cdots, \pi_k$ $(k > 2)$. Their mean vectors and covariance matrices are $\mu^{(i)}$ and Σ_i $(i = 1, \cdots, k)$, respectively. For any given m-dimensional sample $X = (x_1, x_2, \cdots, x_m)^T$, we want to determine which population it comes from.
In the case of multiple populations, when classifying X according to the nearest distance criterion, we first calculate the Mahalanobis distances $d_i^2(X)$ $(i = 1, 2, \cdots, k)$ from the sample X to the k populations, then compare them and assign X to the population with the smallest distance. If for $i = j$, we have

$$d_j^2(X) = \min_{i=1,\cdots,k} \{d_i^2(X)\},$$

then $X \in \pi_j$.

Bayesian Discrimination

Bayesian Discrimination

- I. Maximum A Posteriori (MAP) Method
- II. Minimum Expected Misclassification Cost Method

An Example Where Distance Discrimination is Inappropriate

- π_1 (Graduate Student Group): $N_1 = 2000$, $\mu_1 = 500$
- π_2 (Undergraduate Student Group): $N_2 = 8000$, $\mu_2 = 400$
- There are 1,000 graduate students with x $\geq$ 500, and 2,000 undergraduate students with x $\geq$ 500.
- A student has $x = 500$. To which group should they be assigned?

An Example Where Distance Discrimination is Inappropriate

- π_1 (Graduate Student Group): $N_1 = 2000$, $\mu_1 = 500$
- π_2 (Undergraduate Student Group): $N_2 = 8000$, $\mu_2 = 400$
- There are 1,000 graduate students with x $\geq$ 500, and 2,000 undergraduate students with x $\geq$ 500.
- A student has $x = 500$. To which group should they be assigned?
- Using distance discrimination is clearly inappropriate.
- Consider using prior probabilities:

$$P_1 = \frac{2000}{10000} = 0.2, \quad P_2 = \frac{8000}{10000} = 0.8$$

I. Maximum A Posteriori (MAP) Method

- Let there be k groups $\pi_1, \pi_2, \cdots, \pi_k$.
- Group π_i has probability density $f_i(\mathbf{x})$.
- The prior probability that sample $\mathbf{x}$ comes from π_i is p_i, $i = 1, 2, \cdots, k$, with $\sum_{i=1}^{k} p_i = 1$.
- The posterior probability of belonging to π_i is:

$$P(\pi_i|\mathbf{x}) = \frac{p_i f_i(\mathbf{x})}{\sum_{j=1}^{k} p_j f_j(\mathbf{x})}, \quad i = 1, 2, \cdots, k$$

- The MAP discriminant rule is:

$$\mathbf{x} \in \pi_I \quad \text{if} \quad P(\pi_I|\mathbf{x}) = \max_{1 \le i \le k} P(\pi_i|\mathbf{x})$$

Example 1

- Groups: π_1, π_2, π_3. Discriminate sample $\mathbf{x}_0$.
- Given: $p_1 = 0.05, p_2 = 0.65, p_3 = 0.30$,
  $f_1(\mathbf{x}_0) = 0.10, f_2(\mathbf{x}_0) = 0.63, f_3(\mathbf{x}_0) = 2.4$.
- Compute posterior probabilities:

$$P(\pi_1|\mathbf{x}_0) = \frac{0.05 \times 0.10}{0.05 \times 0.10 + 0.65 \times 0.63 + 0.30 \times 2.4} = \frac{0.005}{1.1345} = 0.004$$

$$P(\pi_2|\mathbf{x}_0) = \frac{0.65 \times 0.63}{1.1345} = 0.361$$

$$P(\pi_3|\mathbf{x}_0) = \frac{0.30 \times 2.4}{1.1345} = 0.635$$

- Assign $\mathbf{x}_0$ to group π_3.

Case of Normal Groups

- Let $\pi_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $\boldsymbol{\Sigma}_i > 0$, $i = 1, 2, \cdots, k$.
- Density for π_i:

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_i|^{-1/2} \exp\left[-0.5 d^2(\mathbf{x}, \pi_i)\right]$$

  where $d^2(\mathbf{x}, \pi_i) = (\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$ is the squared Mahalanobis distance.
- Case 1: $p_1 = p_2 = \cdots = p_k = 1/k$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$

$$P(\pi_i|\mathbf{x}) = \frac{\exp\left[-\frac{1}{2} d^2(\mathbf{x}, \pi_i)\right]}{\sum_{j=1}^{k} \exp\left[-\frac{1}{2} d^2(\mathbf{x}, \pi_j)\right]}$$

Cases for Normal Groups

- Case 2: $p_1 = p_2 = \cdots = p_k = 1/k$, $\mathbf{\Sigma}_i$ not all equal

$$P(\pi_i|\mathbf{x}) = \frac{\exp\left\{-\frac{1}{2}\left[d^2(\mathbf{x},\pi_i) + \ln|\mathbf{\Sigma}_i|\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\mathbf{x},\pi_j) + \ln|\mathbf{\Sigma}_j|\right]\right\}}$$

- Case 3: p_i not all equal, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_k = \mathbf{\Sigma}$

$$P(\pi_i|\mathbf{x}) = \frac{\exp\left\{-\frac{1}{2}\left[d^2(\mathbf{x},\pi_i) - 2\ln p_i\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\mathbf{x},\pi_j) - 2\ln p_j\right]\right\}}$$

- Case 4: p_i not all equal, $\mathbf{\Sigma}_i$ not all equal

$$P(\pi_i|\mathbf{x}) = \frac{\exp\left\{-\frac{1}{2}\left[d^2(\mathbf{x},\pi_i) + \ln|\mathbf{\Sigma}_i| - 2\ln p_i\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\mathbf{x},\pi_j) + \ln|\mathbf{\Sigma}_j| - 2\ln p_j\right]\right\}}$$

Unified Expression for Posterior Probability

The posterior probability of the above cases can be uniformly expressed as

$$P(\pi_i|\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}D^2(\mathbf{x},\pi_i)\right]}{\sum_{j=1}^{k}\exp\left[-\frac{1}{2}D^2(\mathbf{x},\pi_j)\right]}, \quad i = 1, 2, \cdots, k$$

where

$$D^2(\mathbf{x},\pi_i) = d^2(\mathbf{x},\pi_i) + g_i + h_i$$

$$g_i = \begin{cases} \ln|\mathbf{\Sigma}_i|, & \text{if } \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \cdots, \mathbf{\Sigma}_k \text{ not all equal} \\ 0, & \text{if } \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_k = \mathbf{\Sigma} \end{cases}$$

$$h_i = \begin{cases} -2\ln p_i, & \text{if } p_1, p_2, \cdots, p_k \text{ not all equal} \\ 0, & \text{if } p_1 = p_2 = \cdots = p_k = \frac{1}{k} \end{cases}$$

$D^2(\mathbf{x},\pi_i)$ is the generalized squared distance from $\mathbf{x}$ to π_i. Under normality, the rule is equivalent to:

$$\mathbf{x} \in \pi_I \quad \text{if} \quad D^2(\mathbf{x},\pi_I) = \min_{1\le i\le k} D^2(\mathbf{x},\pi_i)$$

II. Minimum Expected Misclassification Cost Method

- Motivation Example:
  π_1: Qualified medicine, π_2: Unqualified medicine.
- For new sample $\mathbf{x}$: $P(\pi_1|\mathbf{x}) = 0.6$, $P(\pi_2|\mathbf{x}) = 0.4$.
- The cost (loss) of different misclassifications is usually not the same. Discriminating solely based on posterior probability is often inappropriate.

Classification Table and Misclassification Costs

- Let the pdf of groups π_1 and π_2 be $f_1(x)$ and $f_2(x)$, respectively.
- The prior probabilities of groups π_1 and π_2 be p_1 and p_2, respectively, with $p_1 + p_2 = 1$.
- Furthermore, let the cost of classifying x from π_i as π_l be $c(l|i)$, where l,i = 1,2.

Cost matrix:

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

Visualizing a Classification Rule

Let Ω denote the sample space, i.e., all possible values of x, and

- $R_1 \subset \Omega$ is the subset for which we classify x as π_1
- $R_2 = \Omega - R_1$ is the subset for which we classify x as π_2

$$\Omega = R_1 \cup R_2$$

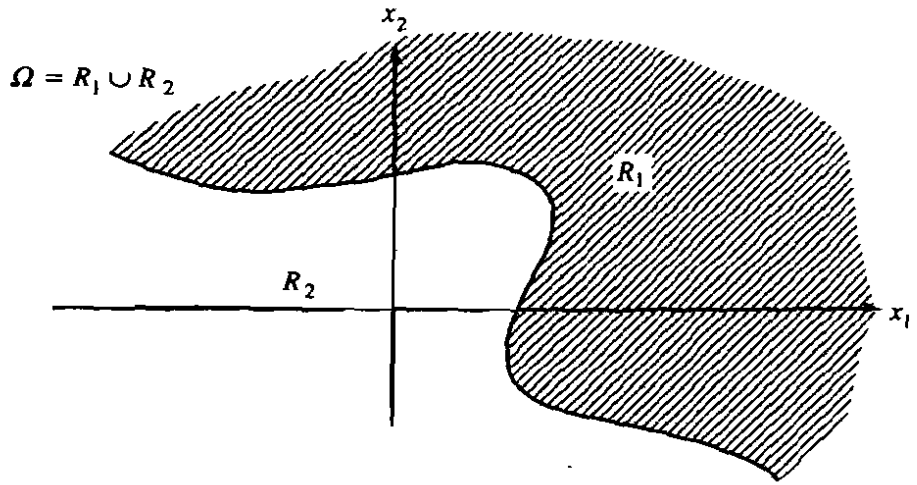


Figure: Classification regions for two populations.

Visualization is for $p = 2$ variables.

Probability of Misclassification

The conditional probability $P(2|1)$ of classifying an object as π_2 when it belongs to π_1 is:

$$P(2|1) = P(X \in R_2|\pi_1) = \int_{R_2} f_1(x)dx$$

The conditional probability $P(1|2)$ of classifying an object as π_1 when it belongs to π_2 is:

$$P(1|2) = P(X \in R_1|\pi_2) = \int_{R_1} f_2(x)dx$$

Visualizing the Probability of Misclassification

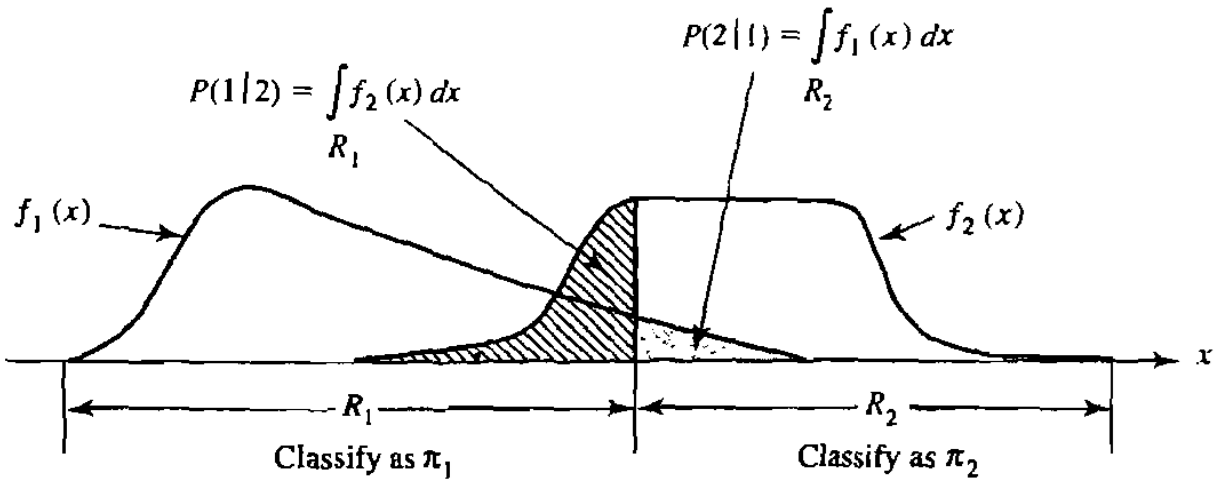


Figure: Misclassification probabilities for hypothetical classification regions when $p = 1$.

$$P(1|2) = \int_{R_1} f_2(x)\,dx$$

Visualization is for $p = 1$ variable.

Classification Rule (Region) Minimizing ECM

The expected cost of misclassification (ECM) is:

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

The R_1 and R_2 that minimize the ECM are defined via:

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

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

- If $c(1|2) = c(2|1)$, then classifying via posterior probabilities.
- If $\frac{p_1}{p_2} = \frac{c(1|2)}{c(2|1)}$, then:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1 \quad R_2 : \frac{f_1(x)}{f_2(x)} < 1$$

Example 2

- Let the probability density functions of groups π_1 and π_2 be $f_1(x)$ and $f_2(x)$, respectively.
- Given: $c(1|2) = 12$ units, $c(2|1) = 4$ units.
- Prior probabilities: $p_1 = 0.6$, $p_2 = 0.4$.
- The minimum ECM discriminant rule is:

$$\begin{cases} x \in \pi_1, & \text{if } \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{c(1|2) \cdot p_2}{c(2|1) \cdot p_1} = \dfrac{12 \times 0.4}{4 \times 0.6} = 2 \\ x \in \pi_2, & \text{if } \dfrac{f_1(x)}{f_2(x)} < \dfrac{12 \times 0.4}{4 \times 0.6} = 2 \end{cases}$$

- For a new sample x_0, we compute:

$$\frac{f_1(x_0)}{f_2(x_0)} = \frac{0.36}{0.24} = 1.5 < 2$$

- According to the discriminant rule:

$$\text{Since } 1.5 < 2, \quad x_0 \text{ is assigned to group } \pi_2.$$

Case of Multiple Groups

- Let $f_i(\mathbf{x})$ be the probability density function for group π_i, $i = 1, 2, \cdots, k$.
- Define:
  - p_i: Prior probability of group π_i, $i = 1, 2, \cdots, k$
  - $c(l|i)$: Cost of misclassifying a sample from π_i into π_l, $l, i = 1, 2, \cdots, k$
  - For correct classification: $c(i|i) = 0$, $i = 1, 2, \cdots, k$
  - R_l: Region where samples are classified into π_l, $l = 1, 2, \cdots, k$
  - For $l, i = 1, 2, \cdots, k$, the conditional probability of classifying a sample from π_i into π_l is:

$$P(l|i) = P(\mathbf{x} \in R_l | \mathbf{x} \in \pi_i) = \int_{R_l} f_i(\mathbf{x}) d\mathbf{x}$$

Expected Cost of Misclassification (ECM)

$$
\begin{aligned}
ECM = E[c(l|i)] &= \sum_{i=1}^{k}\sum_{l=1}^{k} c(l|i)P(x \in \pi_i, x \in R_l) \\
&= \sum_{i=1}^{k}\sum_{l=1}^{k} c(l|i)P(x \in R_l|x \in \pi_i)P(x \in \pi_i) \\
&= \sum_{i=1}^{k}\sum_{l=1}^{k} c(l|i)P(l|i)p_i = \sum_{i=1}^{k} p_i \sum_{l=1, l\neq i}^{k} c(l|i)P(l|i)
\end{aligned}
$$

Minimizing the ECM

The discriminant rule that minimizes ECM is:

$$x \in \pi_l \quad \text{if} \quad \sum_{j=1, j\neq l}^{k} p_j f_j(x) c(l|j) = \min_{1\leq i\leq k} \sum_{j=1, j\neq i}^{k} p_j f_j(x) c(i|j)$$

If all misclassification costs are equal (i.e., $c(l|i) = 1$ for $l \neq i$), then:

$$ECM = \sum_{i=1}^{k} p_i \sum_{\substack{l=1 \\ l\neq i}}^{k} P(l|i) = 1 - \sum_{i=1}^{k} p_i P(i|i)$$

This is the **total probability of misclassification**.

Minimum Total Misclassification Probability

The rule becomes:

$$x \in \pi_l \quad \text{if} \quad \sum_{j=1, j\neq l}^{k} p_j f_j(x) = \min_{1\leq i\leq k} \sum_{j=1, j\neq l}^{k} p_j f_j(x)$$

Subtracting $\sum_{j=1}^{k} p_j f_j(x)$ from both sides gives:

$$x \in \pi_l \quad \text{if} \quad p_l f_l(x) = \max_{1\leq i\leq k} p_i f_i(x)$$

The maximum a posteriori probability method can be regarded as the minimum expected misjudgment cost method when all misjudgment costs are the same.

Example 3

True Group	Classified Group		
	π_1	π_2	π_3
π_1	$c(1\|1) = 0$	$c(2\|1) = 10$	$c(3\|1) = 200$
π_2	$c(1\|2) = 20$	$c(2\|2) = 0$	$c(3\|2) = 100$
π_3	$c(1\|3) = 60$	$c(2\|3) = 50$	$c(3\|3) = 0$

Parameter	Value
p_1	0.05
p_2	0.65
p_3	0.30
$f_1(\mathbf{x}_0)$	0.10
$f_2(\mathbf{x}_0)$	0.63
$f_3(\mathbf{x}_0)$	2.4

$$l = 1: \quad p_2 f_2(x_0) c(1|2) + p_3 f_3(x_0) c(1|3)$$
$$= 0.65 \times 0.63 \times 20 + 0.30 \times 2.4 \times 60 = 51.39$$
$$l = 2: \quad p_1 f_1(x_0) c(2|1) + p_3 f_3(x_0) c(2|3)$$
$$= 0.05 \times 0.10 \times 10 + 0.30 \times 2.4 \times 50 = 36.05$$
$$l = 3: \quad p_1 f_1(x_0) c(3|1) + p_2 f_2(x_0) c(3|2)$$
$$= 0.05 \times 0.10 \times 200 + 0.65 \times 0.63 \times 100 = 41.95$$

Since the minimum ECM value (36.05) occurs for $l = 2$, assign x_0 to population π_2.

Fisher's Discriminant

Fisher's Discriminant (Canonical Discriminant)

- Basic idea: **Projection** (or Dimension Reduction).
- Use a few linear combinations of the p-dimensional vector $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ (called Fisher discriminant functions or canonical variables):

$$y_1 = \mathbf{a}_1'\mathbf{x}, \quad y_2 = \mathbf{a}_2'\mathbf{x}, \quad \cdots, \quad y_r = \mathbf{a}_r'\mathbf{x}$$

  (Typically $r \ll p$) to replace the original p variables $x_1, x_2, \cdots, x_p$.
- Purpose: Reduce dimensionality and classify samples based on $y_1, y_2, \cdots, y_r$.
- Successful reduction facilitates more effective grouping/separation and allows visualization (e.g., plotting the first two or three discriminant functions).

Visualizing Fisher's Linear Discriminant Function

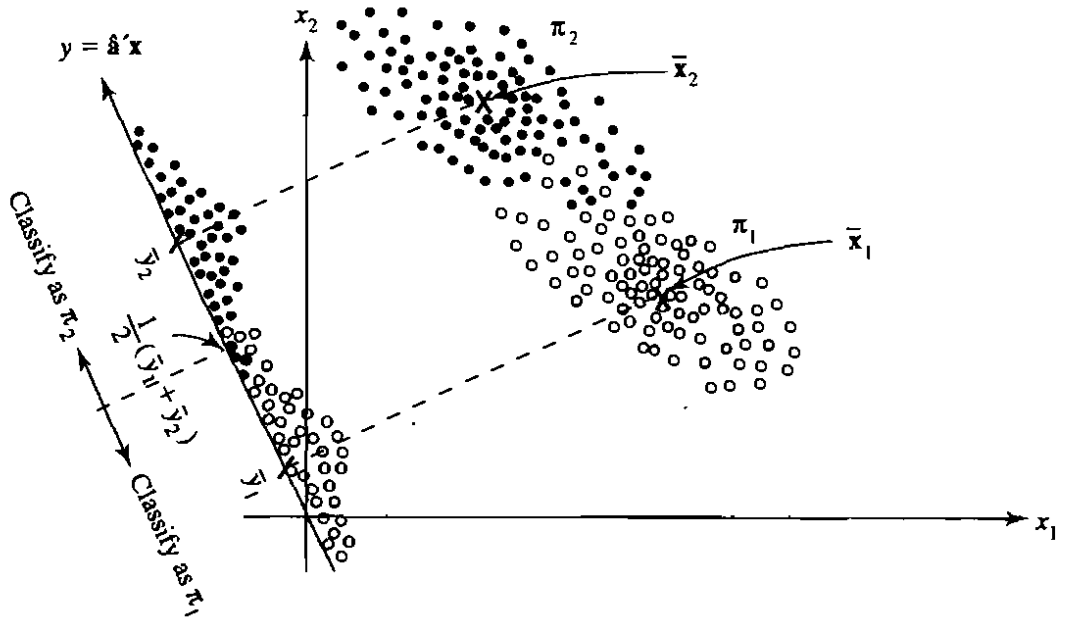


Figure: A pictorial representation of Fisher's procedure for two populations with p = 2. (Johnson & Wichern)

Projection and Notation

- Let $\mathbf{x}_{ij}$ be the p-dimensional observation j from group π_i, $j = 1, \cdots, n_i$, $i = 1, \cdots, k$. Project them onto a constant vector $\mathbf{a}$, yielding projected points:

$$y_{ij} = \mathbf{a}'\mathbf{x}_{ij}, \quad j = 1, \cdots, n_i, \; i = 1, \cdots, k$$

- Group and overall means:

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \mathbf{a}'\bar{\mathbf{x}}_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i = \mathbf{a}'\bar{\mathbf{x}}$$

  where

$$n = \sum_{i=1}^{k} n_i, \quad \bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}, \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{\mathbf{x}}_i$$

- **Assumption:** $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$.

Between-Group and Within-Group Variation

- The between-group sum of squares and within-group sum of squares of y_{ij} are:

$$\text{SSTR} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 = \mathbf{a}'\mathbf{H}\mathbf{a}, \quad \text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \mathbf{a}'\mathbf{E}\mathbf{a}$$

where

$$\mathbf{H} = \sum_{i=1}^{k} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$$

$$\mathbf{E} = \sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$$

- Measure of separation:

$$\Delta(\mathbf{a}) = \frac{\text{SSTR}}{\text{SSE}} = \frac{\mathbf{a}'\mathbf{H}\mathbf{a}}{\mathbf{a}'\mathbf{E}\mathbf{a}}$$

Maximizing Separation

- Find $\mathbf{a}$ that maximizes $\Delta(\mathbf{a})$ under constraint $\mathbf{a}'\mathbf{S}_p\mathbf{a} = 1$, where

$$\mathbf{S}_p = \frac{1}{n-k}\mathbf{E}$$

  is the pooled unbiased estimate of $\mathbf{\Sigma}$.
- Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s > 0$ be the nonzero eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$, with $s = \text{rank}(\mathbf{H}) \leq \min(k-1, p)$.
- Corresponding eigenvectors: $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_s$ (standardized: $\mathbf{t}_i'\mathbf{S}_p\mathbf{t}_i = 1$).
- Maximum $\Delta(\mathbf{a})$ is achieved when $\mathbf{a} = \mathbf{t}_1$. Then $y_1 = \mathbf{t}_1'\mathbf{x}$ is the **first Fisher discriminant function**.
- If one function is insufficient, construct $y_2 = \mathbf{a}_2'\mathbf{x}$ such that

$$\text{Cov}(y_1, y_2) = \mathbf{t}_1'\mathbf{\Sigma}\mathbf{a}_2 = 0$$

Second and Subsequent Discriminant Functions

- Using $\mathbf{S}_p \approx \mathbf{\Sigma}$, the constraint becomes:

$$\mathbf{t}_1' \mathbf{S}_p \mathbf{a}_2 = 0 \quad (\text{or } \mathbf{t}_1' \mathbf{E} \mathbf{a}_2 = 0)$$

- Maximize $\Delta(\mathbf{a}_2)$ under this constraint. The solution is $\mathbf{a}_2 = \mathbf{t}_2$, yielding the **second discriminant function** $y_2 = \mathbf{t}_2' \mathbf{x}$.
- Generally, the i-th discriminant function $y_i = \mathbf{a}_i' \mathbf{x}$ is constructed to be uncorrelated with all previous ones:

$$\mathbf{t}_j' \mathbf{S}_p \mathbf{a}_i = 0, \quad j = 1, 2, \cdots, i-1$$

The solution is $\mathbf{a}_i = \mathbf{t}_i$, and $\Delta(\mathbf{a}_i)$ reaches its maximum λ_i.

Number of Functions and Contribution

- For $k = 2$: only one discriminant function. For $k = 3$: at most two.
- $\Delta(\mathbf{t}_i) = \lambda_i$ indicates the contribution of the i-th discriminant function y_i.
- Contribution ratio of y_i:

$$\frac{\lambda_i}{\sum_{j=1}^{s} \lambda_j}$$

- Cumulative contribution of the first r functions:

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{s} \lambda_i}$$

- In practice, if the cumulative contribution of the first r functions is high enough (e.g., 75%–95%), use them for discrimination.

Classification Rule

- Since discriminant functions have unit variance and are uncorrelated, Mahalanobis distance reduces to Euclidean distance.
- Assign a new observation $\mathbf{x}$ to the group with the closest mean in the space of $(y_1, \ldots, y_r)$:

$$\mathbf{x} \in \pi_i \quad \text{if} \quad \sum_{j=1}^{r} (y_j - \bar{y}_{lj})^2 = \min_{1 \le i \le k} \sum_{j=1}^{r} (y_j - \bar{y}_{ij})^2$$

  where $\bar{y}_{ij} = \mathbf{t}_j' \bar{\mathbf{x}}_i$, and $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$.
- Equivalently:

$$\mathbf{x} \in \pi_i \quad \text{if} \quad \sum_{j=1}^{r} [\mathbf{t}_j'(\mathbf{x} - \bar{\mathbf{x}}_l)]^2 = \min_{1 \le i \le k} \sum_{j=1}^{r} [\mathbf{t}_j'(\mathbf{x} - \bar{\mathbf{x}}_i)]^2$$

Stepwise Discriminant Analysis

Stepwise Discriminant Analysis

- Stepwise discriminant analysis is a method for automatically selecting a subset of variables in discriminant analysis.
- It is not necessarily optimal but is often effective and is one of the most widely used methods for variable selection in discriminant analysis.
- The basic idea and steps are similar to stepwise regression in regression analysis.

Stepwise Discriminant Analysis

- I. Additional Information Test
- II. Variable Selection Methods

Additional Information Test

- Let $x = (x_1', x_2')'$, where $x_1 = (x_1, x_2, \cdots, x_r)$ are the original discriminant variables, and $x_2 = (x_{r+1}, x_{r+2}, \cdots, x_p)$ are newly introduced variables.
- We want to know whether the additional information provided by x_2 (beyond that contained in x_1) significantly improves the ability to distinguish between groups.
- If not, introducing x_2 is considered not worthwhile.
- Suppose there are k groups $\pi_1, \pi_2, \cdots, \pi_k$, each with a p-variate normal distribution and the same covariance matrix. Independent samples are drawn from each group; n is the total sample size.
- Test the hypothesis:

$$H_0 : \text{The groups have equal } E(x_2|x_1)$$

Test Statistic

Partition the within-group sum of squares and cross-products matrix E, and the between-group matrix H:

$$E = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \quad H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

The test statistic is:

$$\Lambda(x_2|x_1) = \frac{\Lambda(x_1, x_2)}{\Lambda(x_1)}$$

where

$$\Lambda(x_1, x_2) = \frac{|E|}{|E + H|}, \quad \Lambda(x_1) = \frac{|E_{11}|}{|E_{11} + H_{11}|}$$

Distribution and Partial Statistics

Under H_0, $\Lambda(x_2|x_1) \sim \Lambda(p-r, k-1, n-k-r)$.
When $p-r=1$ (i.e., $r=p-1$):

$$\Lambda = \Lambda(x_p|x_1, \cdots, x_{p-1}) = \frac{\Lambda(x_1, \cdots, x_p)}{\Lambda(x_1, \cdots, x_{p-1})} \sim \Lambda(1, k-1, n-k-p+1)$$

The partial F statistic:

$$F = \frac{1-\Lambda}{\Lambda} \cdot \frac{n-k-p+1}{k-1} \sim F(k-1, n-k-p+1)$$

Reject H_0 if $F \geq F_\alpha(k-1, n-k-p+1)$.

Variable Selection Methods

- **Forward selection**: Starts with no variables; at each step, the variable that most significantly improves discrimination is added. Process stops when no remaining variable meets the entry criterion.
- **Backward elimination**: Starts with all variables; at each step, the variable that contributes least is removed. Process stops when all remaining variables are significant.
- **Stepwise method**: A combination of forward and backward steps, allowing variables to be added or removed. Most popular in practice.

Stepwise Procedure

For each variable x_i, compute the univariate ANOVA F-statistic $F(x_i)$.
Suppose:

$$F(x_1) = \max_i F(x_i)$$

- If $F(x_1) < F_\alpha(k-1, n-k)$, no variable is selected.
- If $F(x_1) \geq F_\alpha(k-1, n-k)$, select x_1 and proceed.

For each variable not yet selected, compute the partial F-statistic $F(x_i|x_1)$.
Suppose:

$$F(x_2|x_1) = \max_{2 \leq i \leq p} F(x_i|x_1)$$

- If $F(x_2|x_1) < F_\alpha(k-1, n-k-1)$, stop.
- If $F(x_2|x_1) \geq F_\alpha(k-1, n-k-1)$, select x_2 and proceed.

General Step and Verification

Suppose r variables $x_1, \ldots, x_r$ are selected.
Let:

$$F(x_{r+1}|x_1, \ldots, x_r) = \max_{r+1 \le i \le p} F(x_i|x_1, \ldots, x_r)$$

- If $F < F_\alpha(k-1, n-k-r)$, stop.
- If $F \ge F_\alpha(k-1, n-k-r)$, select x_{r+1}.

After adding a variable, verify significance of previously selected variables. Remove any that are no longer significant. Then return to selection step.

- The process ends when no new variable can be added and no existing variable can be removed.
- To avoid infinite loops, use a slightly lower removal threshold than entry threshold.
- Stepwise discriminant analysis is essentially stepwise MANOVA. Discriminant functions are computed only after variable selection is complete.

Example 4

For each variable, compute univariate ANOVA F and p-value.

Variable	x_1	x_2	x_3	x_4
F	119.26	49.16	1180.16	960.01
p-value	<0.0001	<0.0001	<0.0001	<0.0001

$$\Lambda(x_i|x_3) = \frac{\Lambda(x_3, x_i)}{\Lambda(x_3)}, \quad F(x_i|x_3) = \frac{1 - \Lambda(x_i|x_3)}{\Lambda(x_i|x_3)} \cdot \frac{n-k-1}{k-1}$$

x_3 is selected first.

Variable	x_1	x_2	x_4
Partial F	34.32	43.04	24.77
p-value	<0.0001	<0.0001	<0.0001

Table: Partial F and p-values after x_3 is selected

x_2 is selected next.

Example 4

Verify x_3 remains significant after adding x_2:

$$F(x_3|x_2) = 1112.95, \quad p < 0.0001 \quad \Rightarrow \quad \text{Retain } x_3.$$

$$\Lambda(x_i|x_2, x_3) = \frac{\Lambda(x_2, x_3, x_i)}{\Lambda(x_2, x_3)}, \quad F(x_i|x_2, x_3) = \frac{1 - \Lambda(x_i|x_2, x_3)}{\Lambda(x_i|x_2, x_3)} \cdot \frac{n - k - 2}{k - 1}$$

From the table below, x_4 is selected.

Variable	x_1	x_4
Partial F	12.27	34.57
p-value	<0.0001	<0.0001

Table: Partial F and p-values after x_2, x_3 are selected

Example 4

Verify x_2 and x_3 after adding x_4:

Variable	x_2	x_3
Partial F	54.58	38.72
p-value	< 0.0001	< 0.0001

Both remain significant.
Then compute for x_1:

$$\Lambda(x_1|x_2, x_3, x_4) = \frac{\Lambda(x_1, x_2, x_3, x_4)}{\Lambda(x_2, x_3, x_4)}$$

$$F(x_i|x_2, x_3, x_4) = \frac{1 - \Lambda(x_i|x_2, x_3, x_4)}{\Lambda(x_1|x_2, x_3, x_4)} \cdot \frac{n - k - 3}{k - 1}$$

$$\Rightarrow F(x_1|x_2, x_3, x_4) = 4.72, \quad p = 0.0103 \quad \Rightarrow \quad \text{Select } x_1.$$

Example 4

Verify whether x_2, x_3, and x_4 are still significant after x_1 is selected. Calculate the partial F statistics $F(x_2|x_1, x_3, x_4)$, $F(x_3|x_1, x_2, x_4)$, and $F(x_4|x_1, x_2, x_3)$.

Variable	x_2	x_3	x_4
Partial F	21.94	35.59	24.90
p-value	<0.0001	<0.0001	<0.0001

The calculation results show that none of the selected variables have been eliminated.

Example 4

Summary of Variable Selection

Step	1	2	3	4
Variable	x_3	x_2	x_4	x_1
F	1180.16	43.04	34.57	4.72
p-value	$<$0.0001	$<$0.0001	$<$0.0001	0.0103

Table: Summary of variable selection process