other criteria

- total probability of misclassification (TPM)

\[
TPM = P(\text{misclassifying a } \pi_1 \text{ observation or misclassifying a } \pi_2 \text{ observation})
\]

\[
= P(\text{observation comes from } \pi_1 \text{ and is misclassified}) + P(\text{observation comes from } \pi_2 \text{ and is misclassified})
\]

\[
= p_1 \int_{R_1} f_1(x) \, dx + p_2 \int_{R_2} f_2(x) \, dx \xrightarrow{\text{cf.}} \text{ECM in LNp8-5}
\]

⇒ equivalent to minimizing ECM when costs of misclassification are equal

- “posterior” probability approach

\[
P(\pi_1 | x_0) = \frac{P(\pi_1 \text{ occurs and we observe } x_0)}{P(\text{we observe } x_0)}
\]

\[
= \frac{P(\text{we observe } x_0 | \pi_1)P(\pi_1)}{P(\text{we observe } x_0 | \pi_1)P(\pi_1) + P(\text{we observe } x_0 | \pi_2)P(\pi_2)}
\]

\[
= \frac{p_1f_1(x_0)}{p_1f_1(x_0) + p_2f_2(x_0)}
\]

⇒ Classifying an observation \( x_0 \) as \( \pi_1 \) when \( P(\pi_1 | x_0) > P(\pi_2 | x_0) \)

⇒ equivalent to minimizing ECM when costs of misclassification are equal

---

• Classification with Two Multivariate Normal Populations

- now, further assume that \( f_1(x) \) and \( f_2(x) \) are multivariate normal densities,

- \( f_1(x) \) with mean vector \( \mu_1 \) and covariance matrix \( \Sigma_1 \)

- \( f_2(x) \) with mean vector \( \mu_2 \) and covariance matrix \( \Sigma_2 \)

- Classification of Normal Populations When \( \Sigma_1 = \Sigma_2 = \Sigma \)

- Suppose that the joint densities of \( X' = [X_1, X_2, \ldots, X_p] \) for populations \( \pi_1 \) and \( \pi_2 \) are given by

\[
f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu_i)'\Sigma^{-1}(x - \mu_i) \right]
\]

for \( i = 1, 2 \)

- minimum ECM regions become

\[ R_1: \quad \exp \left[ -\frac{1}{2} (x - \mu_1)'\Sigma^{-1}(x - \mu_1) + \frac{1}{2} (x - \mu_2)'\Sigma^{-1}(x - \mu_2) \right] \geq \frac{c_{12}}{c_{21}} \frac{p_2}{p_1} \]

\[ R_2: \quad \exp \left[ -\frac{1}{2} (x - \mu_1)'\Sigma^{-1}(x - \mu_1) + \frac{1}{2} (x - \mu_2)'\Sigma^{-1}(x - \mu_2) \right] < \frac{c_{12}}{c_{21}} \frac{p_2}{p_1} \]

**Result 11.2.** Let the populations \( \pi_1 \) and \( \pi_2 \) be described by multivariate normal densities of the form (11.10). Then the allocation rule that minimizes the ECM is as follows: Allocate \( x_0 \) to \( \pi_1 \) if

\[
(\mu_1 - \mu_2)'\Sigma^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln \left( \frac{c_{12}}{c_{21}} \frac{p_2}{p_1} \right)
\]

Allocate \( x_0 \) to \( \pi_2 \) otherwise.

**proof.**

\[
\ln(c_{12}) = -\frac{1}{2} \left[ x_0'\Sigma^{-1}x_0 - x_0'\Sigma^{-1}x_0 + \Sigma^{-1}x_0'\mu_1 + \Sigma^{-1}x_0'\mu_2 + \frac{1}{2} \left( x_0'\Sigma^{-1}x_0 - x_0'\Sigma^{-1}x_0 + \Sigma^{-1}x_0'\mu_1 + \Sigma^{-1}x_0'\mu_2 \right) \right]
\]
in practical situation, $\mu_1, \mu_2$, and $\Sigma$ are usually unknown $\Rightarrow$ replacing the population parameters by their counterparts (Q: how?)

Suppose, then, that we have $n_1$ observations of the multivariate random variable $X' = [X_1, X_2, \ldots, X_p]$ from $\pi_1$ and $n_2$ measurements of this quantity from $\pi_2$, with $n_1 + n_2 - 2 \geq p$. Then the respective data matrices are

$$X_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n_1} \end{bmatrix}, \quad X_2 = \begin{bmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n_2} \end{bmatrix}$$

dataset = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}

the sample mean vectors and covariance matrices are

$$\overline{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j}, \quad S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - \overline{x}_1)(x_{1j} - \overline{x}_1)'$$

$$\overline{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}, \quad S_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \overline{x}_2)(x_{2j} - \overline{x}_2)'$$

Since it is assumed that the parent populations have the same covariance matrix $\Sigma$, the sample covariance matrices $S_1$ and $S_2$ are combined (pooled) to derive a single, unbiased estimate of $\Sigma$

$$S_{\text{pooled}} = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} S_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} S_2$$

The Estimated Minimum ECM Rule for Two Normal Populations
Allocate $x_0$ to $\pi_1$ if

$$(\overline{x}_1 - \overline{x}_2)'S_{\text{pooled}}^{-1}x_0 - \frac{1}{2} (\overline{x}_1 - \overline{x}_2)'S_{\text{pooled}}^{-1}(\overline{x}_1 + \overline{x}_2) \geq \ln \left( \frac{c(1,2)}{c(2,1)} \right) \left( \frac{p_2}{p_1} \right)$$

Allocate $x_0$ to $\pi_2$ otherwise.

$\star$ If

$$\begin{bmatrix} c(1,2) \\ c(2,1) \end{bmatrix} \begin{bmatrix} p_2 \\ p_1 \end{bmatrix} = 1$$

then $\ln(1) = 0$, and the estimated minimum ECM rule for two normal populations amounts to comparing the scalar variable

$$\hat{y} = (\overline{x}_1 - \overline{x}_2)'S_{\text{pooled}}^{-1}x = \hat{a}'x$$

evaluated at $x_0$, with the number

$$\hat{m} = \frac{1}{2} (\overline{x}_1 - \overline{x}_2)'S_{\text{pooled}}^{-1}(\overline{x}_1 + \overline{x}_2) = \frac{1}{2} (\overline{y}_1 + \overline{y}_2)$$

where

$$\overline{y}_1 = (\overline{x}_1 - \overline{x}_2)'S_{\text{pooled}}^{-1}\overline{x}_1 = \hat{a}'\overline{x}_1$$

$$\overline{y}_2 = (\overline{x}_1 - \overline{x}_2)'S_{\text{pooled}}^{-1}\overline{x}_2 = \hat{a}'\overline{x}_2$$

$\star$ the estimated minimum ECM rule is to

1. creating two univariate populations for the $y$ values by taking an appropriate linear combination of the observations from two populations
2. assign a new observation $x_0$ to $\pi_1$ or $\pi_2$ depending upon whether $\hat{y}_0 = \hat{a}'x_0$ falls to the right or left of the midpoint $\hat{m}$ between the two univariate means $\overline{y}_1$ and $\overline{y}_2$

$\Rightarrow$ it is called linear discriminant analysis
Fisher’s approach to classification with two populations

Fisher’s idea was to transform the multivariate variables $X_1, \ldots, X_p$ to a univariate variable $Y$, which is a linear function of the $X$ variables, i.e.,

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p = a^T X, \quad (a_1, \ldots, a_p)^T$$

such that the $Y$ observations derived from the two populations were separated as much as possible.

A fixed linear combination of the $x$’s takes the values $y_{11}, y_{12}, \ldots, y_{1n_1}$ for the observations from the first population and the values $y_{21}, y_{22}, \ldots, y_{2n_2}$ for the observations from the second population. The separation of these two sets of univariate $y$’s is assessed in terms of the difference between $\bar{y}_1$ and $\bar{y}_2$, expressed in standard deviation units. That is,

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{\sqrt{s_y^2}}$$

where $s_y^2 = \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$

is the pooled estimate of the variance.

Result 11.3. The linear combination $\hat{y} = \hat{a}^T x = (\bar{x}_1 - \bar{x}_2)^T S^{-1}_{\text{pooled}} x$ maximizes the ratio

$$\frac{\text{between sample means of } y}{\text{within group variation}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{a}^T \bar{x}_1 - \hat{a}^T \bar{x}_2)^2}{\hat{a}^T S^{-1}_{\text{pooled}} \hat{a} = \hat{a}^T S^{-1}_{\text{pooled}} \hat{a}}$$

over all possible coefficient vectors $\hat{a}$ where $d = (\bar{x}_1 - \bar{x}_2)$. The maximum of the ratio is $D^2 = (\bar{x}_1 - \bar{x}_2)^T S^{-1}_{\text{pooled}} (\bar{x}_1 - \bar{x}_2)$.

Proof. By maximization Lemma in LNp.2-17,

$$\hat{a} = c \cdot S^{-1}_{\text{pooled}} d,$$

maximum value $= \hat{a}^T S^{-1}_{\text{pooled}} \hat{a}$.

Allocation rule based on Fisher’s discriminant function

Allocate $x_0$ to $\pi_1$ if $\hat{y}_0 = (\bar{x}_1 - \bar{x}_2)^T S^{-1}_{\text{pooled}} x_0$ is defined by

$$\text{1st LDF minimize } E(\text{ECM})$$

Fisher’s LDF maximize “separation”

Note: do not need normality assumption.

Allocate $x_0$ to $\pi_2$ if $\hat{y}_0 < \hat{m}$

Note. Fisher’s linear discriminant function was developed under the assumption that the two populations, whatever their form, have a common covariance matrix.
Is classification a good idea for your data?

- For two populations, the maximum relative separation that can be obtained by considering linear combinations of the multivariate observations is equal to the distance $D^2$

- Note. $D^2$ can be used to test whether the population means differ significantly (Hotelling’s $T^2$ test)
  $\Rightarrow$ A test for differences in mean vectors can be viewed as a test for the “significance” of the separation that can be achieved

- Note. Significant separation does not necessarily imply good classification. By contrast, if the separation is not significant, the search for a useful classification rule will probably prove fruitless

- Classification of Normal Populations When $\Sigma_1 \neq \Sigma_2$

  \[
  f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)/2}
  \]

  **Result 11.4.** Let the populations $\pi_1$ and $\pi_2$ be described by multivariate normal densities with mean vectors and covariance matrices $\mu_1, \Sigma_1$ and $\mu_2, \Sigma_2$, respectively. The allocation rule that minimizes the expected cost of misclassification is given by

  $\begin{align*}
  R_1: & -\frac{1}{2} x' (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1}) x - k \geq \ln \left[ \frac{c(1|2)}{c(2|1)} \right] \\
  R_2: & -\frac{1}{2} x' (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1}) x - k < \ln \left[ \frac{c(1|2)}{c(2|1)} \right]
  \end{align*}$

  where

  $k = \frac{1}{2} \ln \left( \frac{\Sigma_1}{\Sigma_2} \right) + \frac{1}{2} (\mu_1' \Sigma_1^{-1} \mu_1 - \mu_2' \Sigma_2^{-1} \mu_2)$

- It is called *quadratic* classification because of the quadratic term

  \[
  \text{QP: } -\frac{1}{2} x' (\Sigma_1^{-1} - \Sigma_2^{-1}) x \rightarrow \mu_1' \Sigma_1^{-1} \mu_1 - \mu_2' \Sigma_2^{-1} \mu_2
  \]

  **Allocate** $x_0$ to $\pi_1$ if

  $-\frac{1}{2} x_0' (S_1^{-1} - S_2^{-1}) x_0 + (\mu_1' S_1^{-1} - \mu_2' S_2^{-1}) x_0 - k(\mu_1, \mu_2) \geq \ln \left[ \frac{c(1|2)}{c(2|1)} \right]$

  **Allocate** $x_0$ to $\pi_2$ otherwise.

- Evaluating classification functions

  - One important way of judging the performance of any classification procedure is to calculate its “error rate,” or misclassification probabilities

  - Total probability of misclassification (population)

    $TPM = p_1 \int_{R_2} f_1(x) \, dx + p_2 \int_{R_1} f_2(x) \, dx$

    - Actual error rate (sample)

      $AER = p_1 \int_{R_2} f_1(x) \, dx + p_2 \int_{R_1} f_2(x) \, dx$

    - Apparent error rate (do not dependent on population densities)

      $APER = \text{proportion of items in the training set that are misclassified}$

made by S.-W. Cheng (NTHU, Taiwan)
Q: does $n_1 : n_2$ reflect $P_1 : P_2$?

Actual membership

\[
\begin{array}{c|c|c|c}
\pi_1 & \pi_2 \\
\hline
n_1 & n_2 & \pi_1 \\
\hline
n_1C & n_2M & n_1 \\
\hline
n_1M & n_2C & n_2 \\
\hline
\end{array}
\]

APER = \[ \frac{n_1M + n_2M}{n_1 + n_2} = \frac{n_2}{n_1 + n_2} \cdot \frac{n_1M}{n_1} + \frac{n_2M}{n_2} \cdot \frac{n_1}{n_1 + n_2} \]

- it is easy to calculate and can be calculated for any classification procedure
- it tends to underestimate the AER because the data used to build the classification function are also used to evaluate it
- one procedure is to split the total sample into a training sample and a validation sample, but it required large sample and the information in the validation sample is not used to construct the classification function

> cross-validation method (leave-one-out method)

1. Start with the $\pi_1$ group of observations. Omit one observation from this group, and develop a classification function based on the remaining $n_1 - 1$, $n_2$ observations.
2. Classify the “holdout” observation, using the function constructed in Step 1.
3. Repeat Steps 1 and 2 until all of the $\pi_1$ observations are classified. Let $n_1^{(H)}$ be the number of holdout (H) observations misclassified in this group.

4. Repeat Steps 1 through 3 for the $\pi_2$ observations. Let $n_2^{(H)}$ be the number of holdout observations misclassified in this group.

\[
\hat{P}(2 | 1) = \frac{n_1^{(H)}}{n_1}, \quad \hat{P}(1 | 2) = \frac{n_2^{(H)}}{n_2}
\]

\[
\hat{E}(\text{AER}) = \frac{n_1^{(H)} + n_2^{(H)}}{n_1 + n_2}
\]

\[
\frac{\hat{P}_1}{\hat{P}_2} = \frac{n_1^{(H)}}{n_1 + n_2} \cdot \frac{n_2^{(H)}}{n_1 + n_2}
\]

- Reading: textbook, 11.1, 11.2, 11.3, 11.4