Discrimination amongst several populations

We want to determine if an observation vector

$$\boldsymbol{X} = x_1, \dots , x_p$$

comes from one of the g populations:

$$\pi_1 : f_1(x_1, \dots , x_p)$$

$$\dots$$

$$\pi_g : f_g(x_1, \dots , x_p)$$

Usually, the densities $f_1, \dots , f_g$ will be assumed to be multivariate normal.

For this purpose we need to partition p-dimensional space into g regions $R_1, R_2, \ldots, R_g$

We will make the decision $D_i = \{\boldsymbol{X} \text{ came from } \pi_i\}$

if $\boldsymbol{X}$ belongs to R_i

Misclassification probabilities

$$P(k|i) = P[\text{ classify the case in } \pi_k \text{ when case is from } \pi_i]$$
$$= P(X \in R_k|\pi_i) = \int_{R_k} f_i(\boldsymbol{x})d\boldsymbol{x}$$

Cost of Misclassification

$c_{k|i}$ = Cost classifying the case in π_k when case is from π_i

Prior probabilities of inclusion

$P(i) = P[$ classify the case is from π_i initially$]$

Expected Cost of Misclassification of a case from population *i*

We assume that we know the case came from π_i

$$ECM(i) = c_{1|i}P\left[1|i\right] + \ldots + c_{i-1|i}P\left[i-1|i\right] + c_{i+1|i}P\left[i+1|i\right] + \\ + \ldots + c_{k|i}P[g|i]$$

$$= \sum_{j \neq i} c_{j|i} P\left[j|i\right]$$

Total Expected Cost of Misclassification

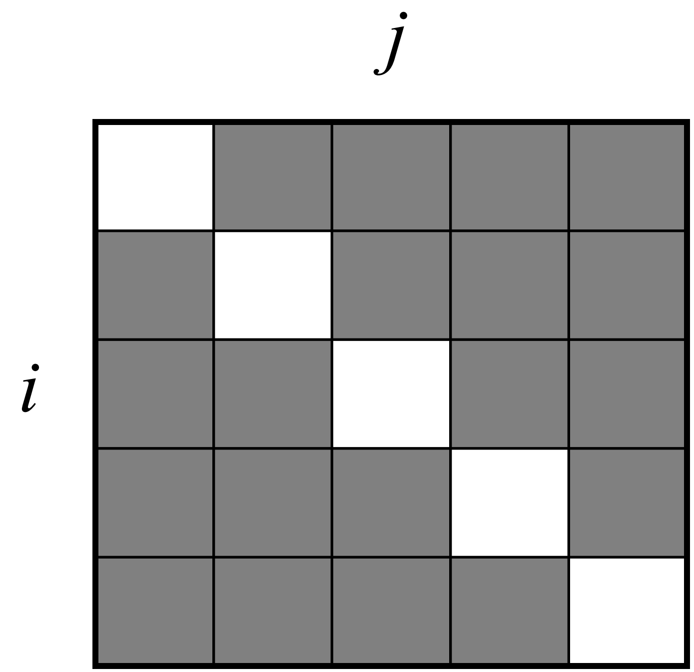
$$\text{ECM} = P[1]\text{ECM}(1) + P[2]\text{ECM}(2) + \ldots + P[g]\text{ECM}(g)$$

$$= \sum_i P[i] ECM(i)$$

$$= \sum_i P[i] \sum_{j \neq i} c_{j|i} P[j|i]$$

$$= \sum_i P[i] \sum_{j \neq i} c_{j|i} \int_{C_j} f_i(\vec{x}) d\vec{x}$$

$$= \sum_j \int_{C_j} \sum_{i \neq j} P[i] f_i(\vec{x}) c_{j|i} d\vec{x}$$



Optimal Classification Rule

The optimal classification rule will find the regions C_j that will minimize:

$$ECM = \sum_j \int_{C_j} \sum_{i \neq j} P[i] f_i(\vec{x}) c_{j|i} d\vec{x}$$

$$= c \sum_j \int_{C_j} \sum_{i \neq j} P[i] f_i(\vec{x}) d\vec{x} \quad \text{if } c_{j|i} = c$$

$$= c \sum_j \int_{C_j} \left[\sum_{i=1}^{k} P[i] f_i(\vec{x}) - P[j] f_j(\vec{x}) \right] d\vec{x}$$

ECM will be minimized if C_j is chosen where the term that is omitted:

$$P[j] f_j(\vec{x})$$

is the largest

Optimal Regions when misclassification costs are equal

Allocate $\boldsymbol{X}$ to π_k if

$$P[k]f_k(\boldsymbol{X}) > P[i]f_i(\boldsymbol{X}) \text{ for all } i \neq k$$

Or, equivalently if

$$\ln[P[k]f_k(\boldsymbol{X})] > \ln[P[i]f_i(\boldsymbol{X})] \text{ for all } i \neq k$$

Optimal Regions when misclassification costs are equal and distributions are p-variate Normal with common covariance matrix Σ

$$R_k = \{\boldsymbol{X}: P[k]f_k(\boldsymbol{X}) > P[i]f_i(\boldsymbol{X}) \text{ for all } i \neq k\}$$

$$= \{\boldsymbol{X}: \ln[P[k]f_k(\boldsymbol{X})] > \ln[P[i]f_i(\boldsymbol{X})] \text{ for all } i \neq k\}$$

In the case of normality

$$f_i(\vec{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\Sigma^{-1}(\vec{x}-\vec{\mu}_i)}$$

$$\ln P[i] f_i(\vec{x})$$

$$= \ln P[i] - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \tfrac{1}{2}(\vec{x}-\vec{\mu}_i)'\Sigma^{-1}(\vec{x}-\vec{\mu}_i)$$

and $\ln P[j] f_j(\vec{x}) > \ln P[i] f_i(\vec{x})$ if:

$$\ln P[j] - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \tfrac{1}{2}\left(\vec{x}-\vec{\mu}_j\right)'\Sigma^{-1}\left(\vec{x}-\vec{\mu}_j\right)$$

$$> \ln P[i] - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \tfrac{1}{2}\left(\vec{x}-\vec{\mu}_i\right)'\Sigma^{-1}\left(\vec{x}-\vec{\mu}_i\right)$$

that is

$$\vec{\mu}_j'\Sigma^{-1}\vec{x} - \tfrac{1}{2}\vec{\mu}_j'\Sigma^{-1}\vec{\mu}_j + \ln P[j] > \vec{\mu}_i'\Sigma^{-1}\vec{x} - \tfrac{1}{2}\vec{\mu}_i'\Sigma^{-1}\vec{\mu}_i + \ln P[i]$$

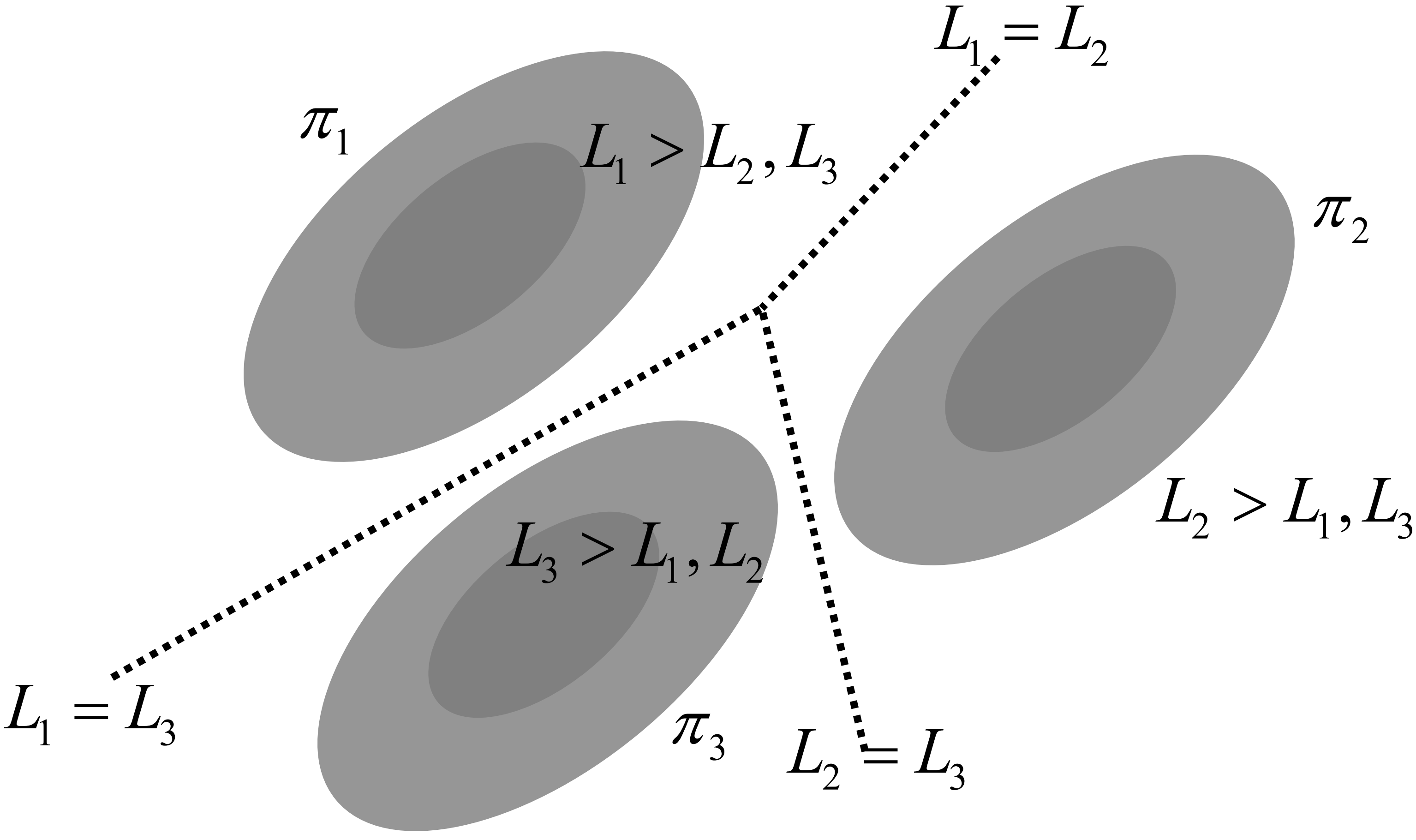
or $\qquad \vec{a}_j'\vec{x} + b_j > \vec{a}_i'\vec{x} + b_i$

where $\vec{a}_i = \Sigma^{-1}\vec{\mu}_i$ and $b_i = \ln P[i] - \tfrac{1}{2}\vec{\mu}_i'\Sigma^{-1}\vec{\mu}_i$

Summarizing

We will classify the observation vector in population π_j if: $L_j = \vec{a}_j'\vec{x} + b_j = \max_i L_i = \max_i \left(\vec{a}_i'\vec{x} + b_i\right)$

where $\vec{a}_i = \Sigma^{-1}\vec{\mu}_i$ and $b_i = \ln P\left[i\right] - \frac{1}{2}\vec{\mu}_i'\Sigma^{-1}\vec{\mu}_i$



Classification with Normal Populations with Different Covariance Matrices

Let

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}, i = 1, \dots, g$$

Assume further that $c(i|i) = 0$ and $c(k|i) = 0$ for $k \neq i$

Rule: Allocate $\boldsymbol{x}$ to π_k if:

$$\begin{aligned}
&\ln P[k] f_k(\boldsymbol{x}) \\
&= \ln P[k] - \left(\frac{p}{2}\right)\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) \\
&= \max_i \ln P[i] f_i(\boldsymbol{x})
\end{aligned}$$