Multivariate Data Analysis

Loukia Meligkotsidou,
National and Kapodistrian University of Athens

Department of Mathematics

Discriminant Analysis (DA)

The objective of **Discriminant Analysis** is the **description** and **classification** of multivariate observations coming from two or more **groups**.

Discriminant Analysis (DA)

The objective of **Discriminant Analysis** is the **description** and **classification** of multivariate observations coming from two or more **groups**.

Suppose that the population of interest is divided in q **subpopulations** (or groups), from each of which we have observed a sample of size n_k, $k = 1, \ldots, q$, respectively. Every sampled observation is **multivariate** (i.e we collect data on p variables).

Discriminant Analysis (DA)

The objective of **Discriminant Analysis** is the **description** and **classification** of multivariate observations coming from two or more **groups**.

Suppose that the population of interest is divided in q **subpopulations** (or groups), from each of which we have observed a sample of size n_k, $k = 1, \ldots, q$, respectively. Every sampled observation is **multivariate** (i.e we collect data on p variables).

Aims: (1) The description of the differences among groups.
(2) The classification of new observations.

Note: We know the groups that the observations of the original sample belong to and we want to classify (predict the groups) of a new observations.

Applications

- Credit Scoring
- Medicine
- Insurance Risk Management
- Classification of undecided voters in gallup polls

In each case, we are interested in constructing a **classification rule**, based on the original data, according to which new observations will be assigned to subpopulations/groups.

Sample Analysis

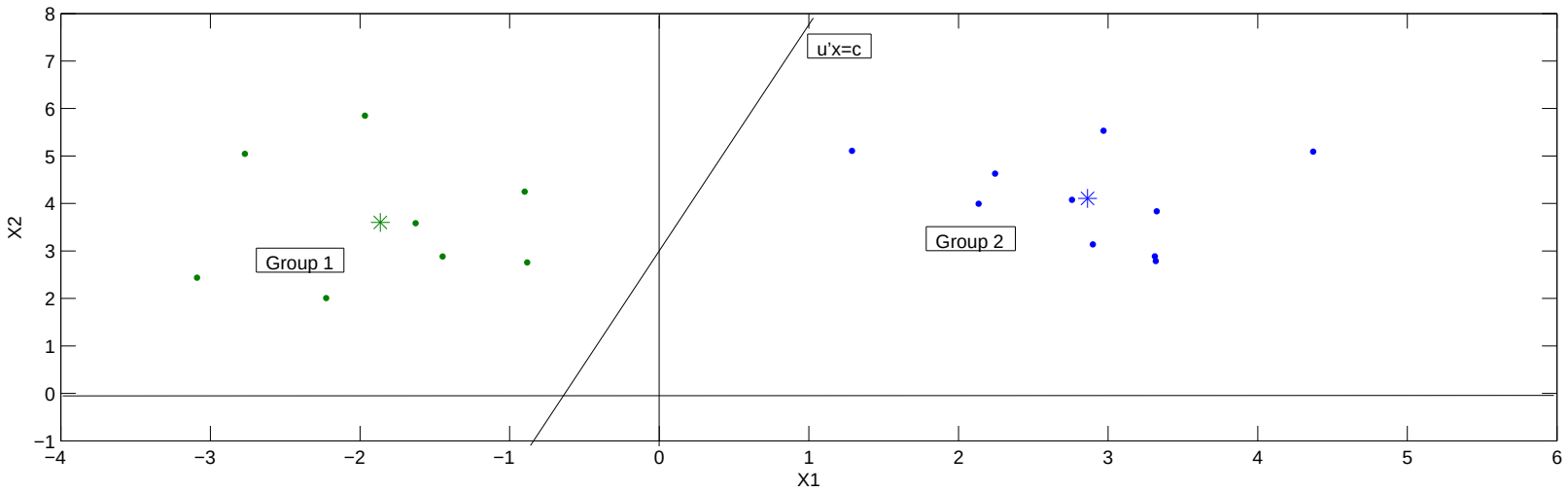
Consider data on p variables, $X_1, \ldots, X_p$, coming from q groups. From each group we have a p-variate sample of size n_k, $k = 1, \ldots, q$.

Hence, the **data matrix** from the kth group is

$$X^{(k)} = \begin{pmatrix} x_{11}^{(k)} & \ldots & x_{1p}^{(k)} \\ x_{21}^{(k)} & \ldots & x_{2p}^{(k)} \\ \vdots & & \vdots \\ x_{n_k 1}^{(k)} & \ldots & x_{n_k p}^{(k)} \end{pmatrix} = \begin{pmatrix} (\mathbf{x}_1^{(k)})' \\ (\mathbf{x}_2^{(k)})' \\ \vdots \\ (\mathbf{x}_{n_k}^{(k)})', \end{pmatrix}$$

while $\bar{\mathbf{x}}^{(k)} = (\bar{x}_1^{(k)}, \bar{x}_2^{(k)}, \ldots, \bar{x}_p^{(k)})'$, $\bar{x}_j^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij}^{(k)}$, $j = 1, \ldots, p$ is the **sample mean or centroid** of the kth group.

Example: $p = q = 2$



Classification Rule: Fisher's Discriminant Function

Consider the case $q = 2$. We require a classification rule to distinguish between two groups.

Fisher's Linear Discriminant Function: Define a new variable Y as a linear combination of the variables $X_1, \ldots, X_p$, i.e $Y = \mathbf{u}'\mathbf{X}$, $\mathbf{u} \in \mathbf{R^p}$, and let c be a constant, such that a new observation, $\mathbf{x}$, to be allocated to group 1 if $\mathbf{u}'\mathbf{x} \leq \mathbf{c}$ and to group 2 if $\mathbf{u}'\mathbf{x} > \mathbf{c}$.

In the example with $p = q = 2$, we require a straight line $\mathbf{u}'\mathbf{x} = \mathbf{c}$ which divides the plane in two half-planes, such that all of the observations of group 1 to lie on one half-plane and all the observations of group 2 to lie on the other.

Classification Rule: Fisher's Discriminant Function

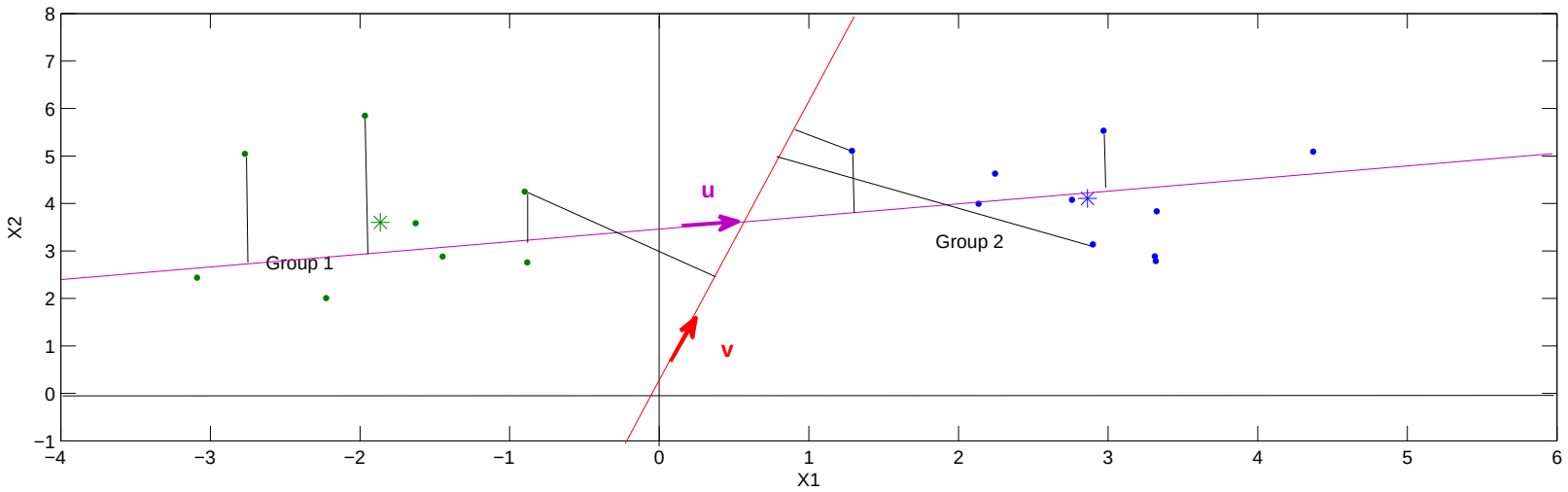
Note: Finding the best linear function which discriminates between the two populations is based on the study of the variability of the projections of the original observations on a line!

Consider the line in the direction of a unit vector, $\mathbf{u}$ or $\mathbf{v}$.

Let $\mathbf{y_1^{(1)}}, \ldots, \mathbf{y_{n1}^{(1)}}$ and $\mathbf{y_1^{(2)}}, \ldots, \mathbf{y_{n2}^{(2)}}$ be the projections of the original observations from groups 1 and 2, respectively, on the direction of $\mathbf{u}$.

Let $\mathbf{z_1^{(1)}}, \ldots, \mathbf{z_{n1}^{(1)}}$ and $\mathbf{z_1^{(2)}}, \ldots, \mathbf{z_{n2}^{(2)}}$ be the projections of the original observations from groups 1 and 2, respectively, on the direction of $\mathbf{v}$.

Example: $p = q = 2$



Classification Rule: Fisher's Discriminant Function

Note: A classification rule **discriminates well** between groups if **the variability of the projections within each group is small, while the distances between the projections of different groups are large**.

In our example, the line defined by **u** corresponds to a good discrimination between the groups, while that define by **v** doesn't. (**Note:** the projected observations on the direction of **v** are mixed!)

Classification Rule: Fisher's Discriminant Function

Note: A classification rule **discriminates well** between groups if **the variability of the projections within each group is small, while the distances between the projections of different groups are large**.

In our example, the line defined by **u** corresponds to a good discrimination between the groups, while that define by **v** doesn't. (**Note:** the projected observations on the direction of **v** are mixed!)

The discrimination according to the projections is equivalent to the classification according to Fisher's discriminant function!
Choosing the linear discriminant function $\mathbf{u}'\mathbf{x} = c$ is equivalent to finding the direction of **u** that best discriminates the projected observations of the groups.

Classification Rule: Fisher's Discriminant Function

The perpendicular in the middle of the line segment that connects the projections of the centroids of the two groups correpsonds to a linear discriminant function of the form $\mathbf{u}'\mathbf{x} = c$.

Classification Rule: Fisher's Discriminant Function

The perpendicular in the middle of the line segment that connects the projections of the centroids of the two groups correpsonds to a linear discriminant function of the form $\mathbf{u}'\mathbf{x} = c$.

Classification rule: Any new observation is allocated to the group with centroid that lies on the same half-plane with the observation with respect to the above described perpendicular.

Classification Rule: Fisher's Discriminant Function

The perpendicular in the middle of the line segment that connects the projections of the centroids of the two groups correpsonds to a linear discriminant function of the form $\mathbf{u}'\mathbf{x} = c$.

Classification rule: Any new observation is allocated to the group with centroid that lies on the same half-plane with the observation with respect to the above described perpendicular.

And now some maths!

Classification Rule: Fisher's Discriminant Function

We define the global centroid of the $n = \sum_{k=1}^{q} n_k$ observations as

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{k=1}^{q} n_k \bar{\mathbf{x}}^{(k)}, \quad \text{where} \quad \bar{\mathbf{x}}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_i^{(k)}.$$

Within each group, we define the (group) sample covariance matrix

$$S_k = \frac{1}{n_k - 1} \mathbf{X}_k^{*\prime} \mathbf{X}_k^*, \quad k = 1, \ldots, q,$$

where $\mathbf{X}_k^*$ is the matrix of the centered observations from the kth group, that is

$$\mathbf{X}_k^* = \begin{pmatrix} x_{11}^{(k)} - \bar{x}_1^{(k)} & \ldots & x_{1p}^{(k)} - \bar{x}_p^{(k)} \\ \vdots & & \vdots \\ x_{n_1 1}^{(k)} - \bar{x}_1^{(k)} & \ldots & x_{n_1 p}^{(k)} - \bar{x}_p^{(k)} \end{pmatrix}.$$

The Pooled Covariance Matrix

The pooled (weighted) covariance matrix of the total sample is defined as

$$S_p = \frac{\sum_{k=1}^{q}(n_k - 1)S_k}{n - q}.$$

In the denominator of this relationship we have $n - q$ because for obtaining the estimator S_p we have used q estimated parameters, $\bar{\mathbf{x}}^{(1)}, \ldots, \bar{\mathbf{x}}^{(q)}$ (we have $n - q$ degrees of freedom).

The pooled covariance matrix, S_p, is different from the global sample covariance matrix, $S = \frac{1}{n-1}\mathbf{X}^{*\prime}\mathbf{X}^{*}$, where $\mathbf{X}^{*}$ is the matrix of centered (with respect to the global centroid) observations.

Total Variation Decomposition

Let $\mathbf{x}_i^{(k)} = (x_{i1}^{(k)}, \ldots, x_{ip}^{(k)})'$, $k = 1, \ldots, q$, $i = 1, \ldots, n_k$, be the data vector corresponding to the ith observation from the kth group. Then,

$$\mathbf{X}_k^* = \begin{pmatrix} x_{11}^{(k)} - \bar{x}_1^{(k)} & \ldots & x_{1p}^{(k)} - \bar{x}_p^{(k)} \\ \vdots & & \vdots \\ x_{n_1 1}^{(k)} - \bar{x}_1^{(k)} & \ldots & x_{n_1 p}^{(k)} - \bar{x}_p^{(k)} \end{pmatrix}, k = 1, \ldots, q,$$

Total Variation Decomposition

and

$$\mathbf{X}^* = \begin{pmatrix} x_{11}^{(1)} - \bar{x}_1 & \dots & x_{1p}^{(1)} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n_1 1}^{(1)} - \bar{x}_1 & \dots & x_{n_1 p}^{(1)} - \bar{x}_p \\ x_{11}^{(2)} - \bar{x}_1 & \dots & x_{1p}^{(2)} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n_2 1}^{(2)} - \bar{x}_1 & \dots & x_{n_2 p}^{(2)} - \bar{x}_p \\ \vdots & & \vdots \\ x_{11}^{(q)} - \bar{x}_1 & \dots & x_{1p}^{(q)} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n_q 1}^{(q)} - \bar{x}_1 & \dots & x_{n_q p}^{(q)} - \bar{x}_p \end{pmatrix}.$$

Total Variation Decomposition

The respective matrices of sums of squares and cross-products are given by

$$\mathbf{X}_k^{*\prime}\mathbf{X}_k^* = \sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})', \quad k = 1, \ldots, q,$$

and

$$\mathbf{X}^{*\prime}\mathbf{X}^* = \sum_{k=1}^{q}\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}})'$$

Total Variation Decomposition

Proposition: $\mathbf{X}^{*\prime}\mathbf{X}^{*} = \sum_{k=1}^{q} \mathbf{X}_k^{*\prime}\mathbf{X}_k^{*} + \sum_{k=1}^{q} n_k(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})$

Proof. We have

$$
\begin{aligned}
\mathbf{X}^{*\prime}\mathbf{X}^{*} &= \sum_{k=1}^{q}\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}})' \\
&= \sum_{k=1}^{q}\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)} + \bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)} + \bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})' \\
&= \sum_{k=1}^{q}\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})' \\
&\quad + \sum_{k=1}^{q}\sum_{i=1}^{n_k}(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}) \\
&\quad + 2\sum_{k=1}^{q}\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})
\end{aligned}
$$

Total Variation Decomposition

That is

$$\mathbf{X}^{*\prime}\mathbf{X}^{*} = \sum_{k=1}^{q} \mathbf{X}_k^{*\prime}\mathbf{X}_k^{*} + \sum_{k=1}^{q} n_k(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})$$

or

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

since $\sum_{k=1}^{q}\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}) =$
$\sum_{k=1}^{q}\left\{\left[\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})\right](\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})\right\} =$
$\sum_{k=1}^{q}\left\{\left[\sum_{i=1}^{n_k}\mathbf{x}_i^{(k)} - n_k\bar{\mathbf{x}}^{(k)}\right](\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})\right\} = 0.$

Total Variation Decomposition

We have shown that

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

where

$\mathbf{T} = \mathbf{X}^{*\prime}\mathbf{X}^{*}$ is the total variation matrix, i.e the matrix of the sums of squared distances of the observations from the global centroid $\bar{\mathbf{x}}$,

$\mathbf{W} = \sum_{k=1}^{q} \mathbf{X}_k^{*\prime}\mathbf{X}_k^{*}$ is the within groups variation matrix, i.e the matrix of the sums of squared distances of the observations from the group centroids $\bar{\mathbf{x}}^{(k)}$, $k = 1, \ldots, q$,

$\mathbf{B} = \sum_{k=1}^{q} n_k(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})$ is the between group variation matrix, i.e the matrix of the sums of squared distances of the group centroids from the global centroid.

Projections

We work with the centered (with respect to the global centroid) data $\mathbf{X}^*$.

Consider a line passing through the origin, defined by a vector $\mathbf{u}$. Let $\mathbf{y}_i^{(k)}$ be the projection of the ith centered observation from the kth group, $\mathbf{x}_i^{*(k)}$, $i = 1, \ldots, n_k$, $k = 1, \ldots, q$. We have

$$\mathbf{y}_i^{(k)} = (\mathbf{u}'\mathbf{x_i}^{*(k)} \frac{\mathbf{u}}{||\mathbf{u}||}).$$

Therefore,

$$||\mathbf{y}_i^{(k)}||^2 = ((\mathbf{u}'\mathbf{x_i}^{*(k)})^2 = \mathbf{u}'\mathbf{x_i}^{*(k)}\mathbf{x_i}^{*(k)'}\mathbf{u}.$$

Then, the total variability of the projected observations is given by

$$Dis_T(\mathbf{u}) = \sum_{\mathbf{k=1}}^{\mathbf{q}} \sum_{\mathbf{i=1}}^{\mathbf{n_k}} ||\mathbf{y_i}^{(\mathbf{k})}||^2 = \mathbf{u}' \left[\sum_{\mathbf{k=1}}^{\mathbf{q}} \sum_{\mathbf{i=1}}^{\mathbf{n_k}} \mathbf{x_i}^{*(\mathbf{k})}\mathbf{x_i}^{*(\mathbf{k})'} \right] \mathbf{u} = \mathbf{u}'\mathbf{X}^{*'}\mathbf{X}^*\mathbf{u}$$

Projections' Variation Decomposition

We have $Dis_T(\mathbf{u}) = \mathbf{u}'\mathbf{X}^{*\prime}\mathbf{X}^*\mathbf{u}$ and from the above proposition

$$Dis_T(\mathbf{u}) = \mathbf{u}'\mathbf{X}^{*\prime}\mathbf{X}^*\mathbf{u} = \mathbf{u}'\mathbf{T}\mathbf{u} = \mathbf{u}'\mathbf{W}\mathbf{u}+\mathbf{u}'\mathbf{B}\mathbf{u} = Dis_W(\mathbf{u})+Dis_B(\mathbf{u})$$

where

$Dis_W(\mathbf{u}) = \mathbf{u}'\mathbf{W}\mathbf{u} = \mathbf{u}'\mathbf{X}_k^{*\prime}\mathbf{X}_k^*\mathbf{u}$ is the variability of the projections within groups

$Dis_B(\mathbf{u}) = \mathbf{u}'\mathbf{B}\mathbf{u} = \mathbf{u}'\left[\sum_{k=1}^{q} n_k(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})\right]\mathbf{u}$ is the variability of the projections between groups

Fisher's Discriminant Function

Fisher's discrimination rule is based on the criterion that the groups are better distinguished if the variability between groups is large compared to the variability within groups. The rule results from the maximization of the ratio

$$Q(\mathbf{u}) = \frac{Dis_B(\mathbf{u})}{Dis_W(\mathbf{u})} = \frac{\mathbf{u'Bu}}{\mathbf{u'Wu}}.$$

Hence, we look for the direction of $\mathbf{u}$ so as $Q(\mathbf{u})$ to be maximized. If $\mathbf{u}$ is the vector that maximizes $Q(\mathbf{u})$, the function $\mathbf{u'x}^*$ is called (first) canonical discriminant function.

Fisher's Discriminant Function

Proposition: The ratio $Q(\mathbf{u}) = \frac{Dis_B(\mathbf{u})}{Dis_W(\mathbf{u})} = \frac{\mathbf{u'Bu}}{\mathbf{u'Wu}}$ is maximized for $\mathbf{u} = \mathbf{u}_1$, where $\mathbf{u}_1$ is the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$.

Proof. The matrices $\mathbf{B}, \mathbf{W}$ are positive definite, therefore $\mathbf{u'Bu} \geq 0$ and $\mathbf{u'Wu} \geq 0$, $\forall \mathbf{u} \in \mathcal{R}^p$. The quantity $Q(\mathbf{u})$ does not change if we multiply the vector $\mathbf{u}$ by some non-zero constant, α say, i.e $Q(\alpha\mathbf{u}) = \frac{\alpha^2\mathbf{u'Bu}}{\alpha^2\mathbf{u'Wu}} = Q(\mathbf{u})$, $\forall \alpha \neq 0$ and $\forall \mathbf{u} \in \mathcal{R}^p$.

Therefore, without loss of generality, we set $\mathbf{u'Wu} = 1$ and we require

$$\max_{\mathbf{u} \in \mathcal{R}^p} \left\{ Q(\mathbf{u}) : \mathbf{u'Wu} = 1 \right\} = \max_{\mathbf{u} \in \mathcal{R}^p} \left\{ \mathbf{u'Bu} : \mathbf{u'Wu} = 1 \right\}.$$

Note: This is an optimization under restriction problem!

Fisher's Discriminant Function

We define the Lagrangean: $L(\mathbf{u}, \lambda) = \mathbf{u}'\mathbf{B}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{W}\mathbf{u} - 1)$.

A necessary condition in order for a vector $\mathbf{u}$ to be solution of the above optimization problem is the existence of some $\lambda \in \mathcal{R}$ such that

$$\frac{\partial L(\mathbf{u}, \lambda)}{\partial \mathbf{u}} = \mathbf{0} \quad \text{and} \quad \frac{\partial L(\mathbf{u}, \lambda)}{\partial \lambda} = 0$$

We have

$$\frac{\partial L(\mathbf{u}, \lambda)}{\partial \mathbf{u}} = 2\mathbf{B}\mathbf{u} - 2\lambda\mathbf{W}\mathbf{u} = \mathbf{0} \Rightarrow \mathbf{B}\mathbf{u} = \lambda\mathbf{W}\mathbf{u} \qquad (1)$$

$$\frac{\partial L(\mathbf{u}, \lambda)}{\partial \lambda} = \mathbf{u}'\mathbf{W}\mathbf{u} - 1 = 0 \Rightarrow \mathbf{u}'\mathbf{W}\mathbf{u} = 1 \quad \text{(the restriction)}$$

If the matrix $\mathbf{W}$ is invertible, then $(\mathbf{W}^{-1}\mathbf{B})\mathbf{u} = \lambda\mathbf{u}$, therefore λ is an eigenvalue and $\mathbf{u}$ is an eigenvector of the matrix $\mathbf{W}^{-1}\mathbf{B}$.

Fisher's Discriminant Function

From (1)

$$\mathbf{Bu} = \lambda \mathbf{Wu} \Rightarrow \mathbf{u}'\mathbf{Bu} = \lambda \mathbf{u}'\mathbf{Wu} \Rightarrow Q(\mathbf{u}) = \frac{\mathbf{u}'\mathbf{Bu}}{\mathbf{u}'\mathbf{Wu}} = \frac{\lambda \mathbf{u}'\mathbf{Wu}}{\mathbf{u}'\mathbf{Wu}} = \lambda$$

We have seen thatλ is an eigenvalue and $\mathbf{u}$ is an eigenvector of the matrix $\mathbf{W}^{-1}\mathbf{B}$. Therefore, the quantity $Q(\mathbf{u})$ is maximized for $\lambda = \lambda_1$, the maximum eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$.

Notes

1. One discriminant function is sufficient for discriminating between two populations (it corresponds to one straight line). In general, if there are q different populations or groups, the function $\mathbf{u}_j'\mathbf{x}$, where $\mathbf{u}_j$ is the eigenvector corresponding to the jth ordered (in descending order) eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$, is called jth canonical discriminant function.

2. It can be shown that the matrix $\mathbf{W}^{-1}\mathbf{B}$ has at most r positive eigenvalues (the remaining ones being equal to zero), where $r = \min(q-1, p)$. Therefore, there are at most r resulting discriminant functions. (For $q = 2$, $r = 1$).

Classification of New Observations

Let $\lambda_1, \ldots, \lambda_r > 0$ be the (ordered) positive eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and $\mathbf{u}_1, \ldots, \mathbf{u}_r$ be the respective eigenvectors. Hence, we obtain r discriminant functions $\mathbf{u}_1'\mathbf{x}^*, \ldots, \mathbf{u}_r\mathbf{x}^*$.

Consider a new (centered) observation $\mathbf{x}_0^*$ for which we do not now which group it belongs to. Moreover, consider the projections of $\mathbf{x}_0^*$ on the straight lines defined by $\mathbf{u}_1, \ldots, \mathbf{u}_r$, i.e

$$\mathbf{y}_{0j} = \mathbf{u}_j'\mathbf{x}_0^* \frac{\mathbf{u}_j}{||\mathbf{u}_j||}, \quad j = 1, \ldots, r.$$

The observation $\mathbf{x}_0^*$ will be allocated to the group for which the projections $\mathbf{y}_{0j}$ are closer to the respective projections of the group centroid than those of the projections of all other groups' centroids.

Classification of New Observations

We define $d_k(\mathbf{x}_0^*)$ as the sum of squared distances of the projections of $\mathbf{x}_0^*$ on the directions of $\mathbf{u}_1, \ldots, \mathbf{u}_r$ from the respective projections of the (centered with respect to the global mean) centroids $\bar{\mathbf{x}}^{*(k)}$ of the kth group. We have

$$
\begin{aligned}
d_k(\mathbf{x}_0^*) &= \sum_{j=1}^{r} (\mathbf{y}_j - \mathbf{u}_j' \mathbf{x}^{*(k)} \frac{\mathbf{u}_j}{||\mathbf{u}_j||})^2 \\
&= \sum_{j=1}^{r} (\mathbf{u}_j' \mathbf{x}_0^* \frac{\mathbf{u}_j}{||\mathbf{u}_j||} - \mathbf{u}_j' \mathbf{x}^{*(k)} \frac{\mathbf{u}_j}{||\mathbf{u}_j||})^2 \\
&= \sum_{j=1}^{r} (\mathbf{u}_j' (\mathbf{x}_0^* - \mathbf{x}^{*(k)}))^2 \\
&= \sum_{j=1}^{r} (\mathbf{u}_j' (\mathbf{x}_0 - \mathbf{x}^{(k)}))^2, \quad k = 1, \ldots, q.
\end{aligned}
$$

Classification of New Observations

The new observation $\mathbf{x}_0^*$ is allocated to the group for which the quantity $d_k(\mathbf{x}_0^*)$ is maximized, i.e

$$r(\mathbf{x}_0^*) = i, \quad \text{if} \quad d_i(\mathbf{x}_0^*) = \min_k \{ d_k(\mathbf{x}_0^*), \ k = 1, \ldots, q \}.$$

For $q = 2$ (discrimination of two populations), we have

$$r(\mathbf{x}_0^*) = 1, \quad \text{if} \quad | \mathbf{u}_1'(\mathbf{x}_0 - \mathbf{x}^{(1)}) | \ < \ | \mathbf{u}_1'(\mathbf{x}_0 - \mathbf{x}^{(2)}) |$$

Example: $q = 2$ groups, $p = 3$ variables

Consider the Within and Between group variability matrices:

$$W = \begin{bmatrix} 0.6 & -0.4 & -0.2 \\ -0.4 & 0.4 & 0.2 \\ -0.2 & 0.2 & 1.2 \end{bmatrix}, \quad B = \begin{bmatrix} 0.16 & 0.08 & 0.04 \\ 0.08 & 0.10 & 0.02 \\ 0.04 & 0.02 & 0.01 \end{bmatrix}$$

Recall: We want the variability within groups to be small and the variability between groups to be large!

$$W^{-1}B = \begin{bmatrix} 0.597 & 0.285 & 0.165 \\ 0.738 & 0.352 & 0.203 \\ 0.004 & 0.002 & 0.001 \end{bmatrix}$$

The largest eigenvalue of the matrix $W^{-1}B$ is $\lambda_1 = 0.9507$ and the respective eigenvector is $\mathbf{u}_1 = (0.629, 0.777, 0.004)'$.

The third variable has large variability within groups and small variability between groups, therefore it is not good for discriminating between these two populations.
(**Note:** its coefficient in the discriminant function is small!)

Classification Rule: Maximum Likelihood

Normality Assumption: We assume that the observations of each group come from a different (multivariate) normal distribution.

We define a random variable G representing the group which a random observations derives from. Then, the distribution of the p-variate random variable $\mathbf{X} = (X_1, \ldots, X_p)'$ depends on the value of G. That is $\mathbf{X} \mid G = k \sim N_p(\mu_k, \Sigma_k)$, $k = 1, \ldots, q$, i.e

$$f_k(\mathbf{x}) = f(\mathbf{x} \mid G = k) = \frac{\mid \Sigma_k \mid^{-1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k)\right\}$$

Classification Rule: Maximum Likelihood

If the values of the parameters μ_k, Σ_k are known, then, as a rule for discriminating among the groups can be used the **maximum likelihood classification rule**.

According to this rule, an observation **x** is allocated to the population which gives to the observed **x** the largest value of the likelihood. That is

$$r(\mathbf{x}) = i, \quad \text{if} \quad f_i(\mathbf{x}) = \max_k \{f_k(\mathbf{x}), k = 1, \ldots, q\}.$$

Example 1. The Maximum Likelihood Classification Rule for two univariate normal populations with common variance

Let $X \mid G = 1 \sim N(\mu_1, \sigma^2)$ and $X \mid G = 2 \sim N(\mu_1, \sigma^2)$, with $\mu_1 < \mu_2$, i.e

$$f_k(x) = f(x \mid G = k) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu_k)^2\right\}, \quad k = 1, 2.$$

The likelihood $f_1(x)$ is larger than $f_2(x)$ and, therefore, the rule allocates the observation x to the first population, if

$$\frac{f_1(x)}{f_2(x)} = \exp\left\{-\frac{1}{2\sigma^2}\left[(x - \mu_1)^2 - (x - \mu_2)^2\right]\right\} > 1,$$

or equivalently if

$$-(x^2 - 2\mu_1 x + \mu_1^2 - x^2\mu_2 x - \mu_2^2) > 0 \Rightarrow 2(\mu_2 - \mu_1)x < \mu_2^2 - \mu_1^2$$

Example 1. The Maximum Likelihood Classification Rule for two univariate normal populations with common variance

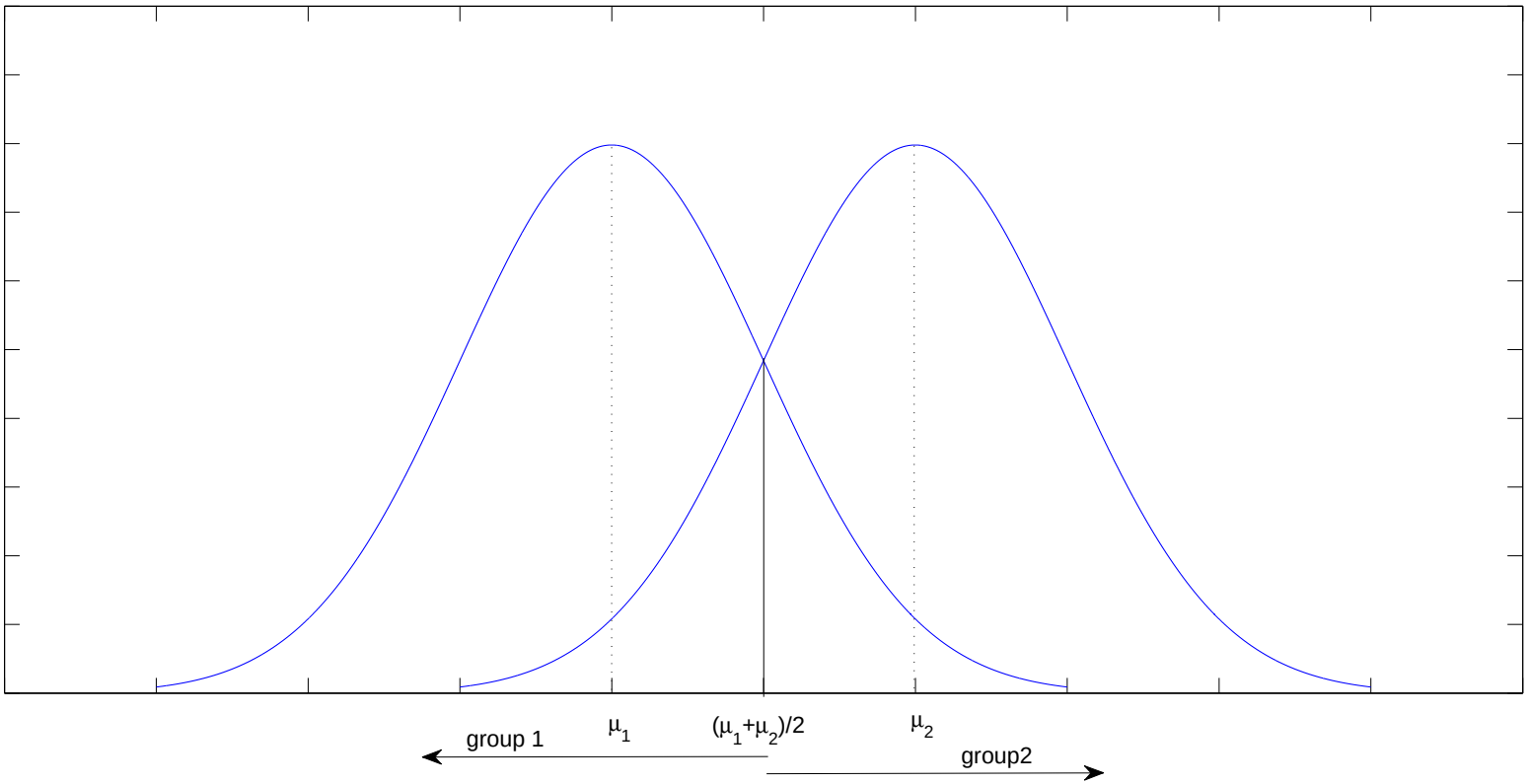
Since $\mu_2 > \mu_1$ the inequality is satisfied for

$$x < \frac{\mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)} \Rightarrow x < \frac{\mu_2 + \mu_1}{2}$$

Therefore, according to the maximum likelihood classification rule,

$$r(x) = 1, \quad \text{if} \quad x < \frac{\mu_2 + \mu_1}{2}.$$

Example 1



Example 2. The Maximum Likelihood Classification Rule for two univariate normal populations with different variances

Let $X \mid G = 1 \sim N(\mu_1, \sigma_1^2)$ and $X \mid G = 2 \sim N(\mu_1, \sigma_2^2)$, with $\mu_1 < \mu_2$ and $s_1^2 < \sigma_2^2$, i.e

$$f_k(x) = f(x \mid G = k) = (2\pi\sigma_k^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right\}, \quad k = 1, 2.$$

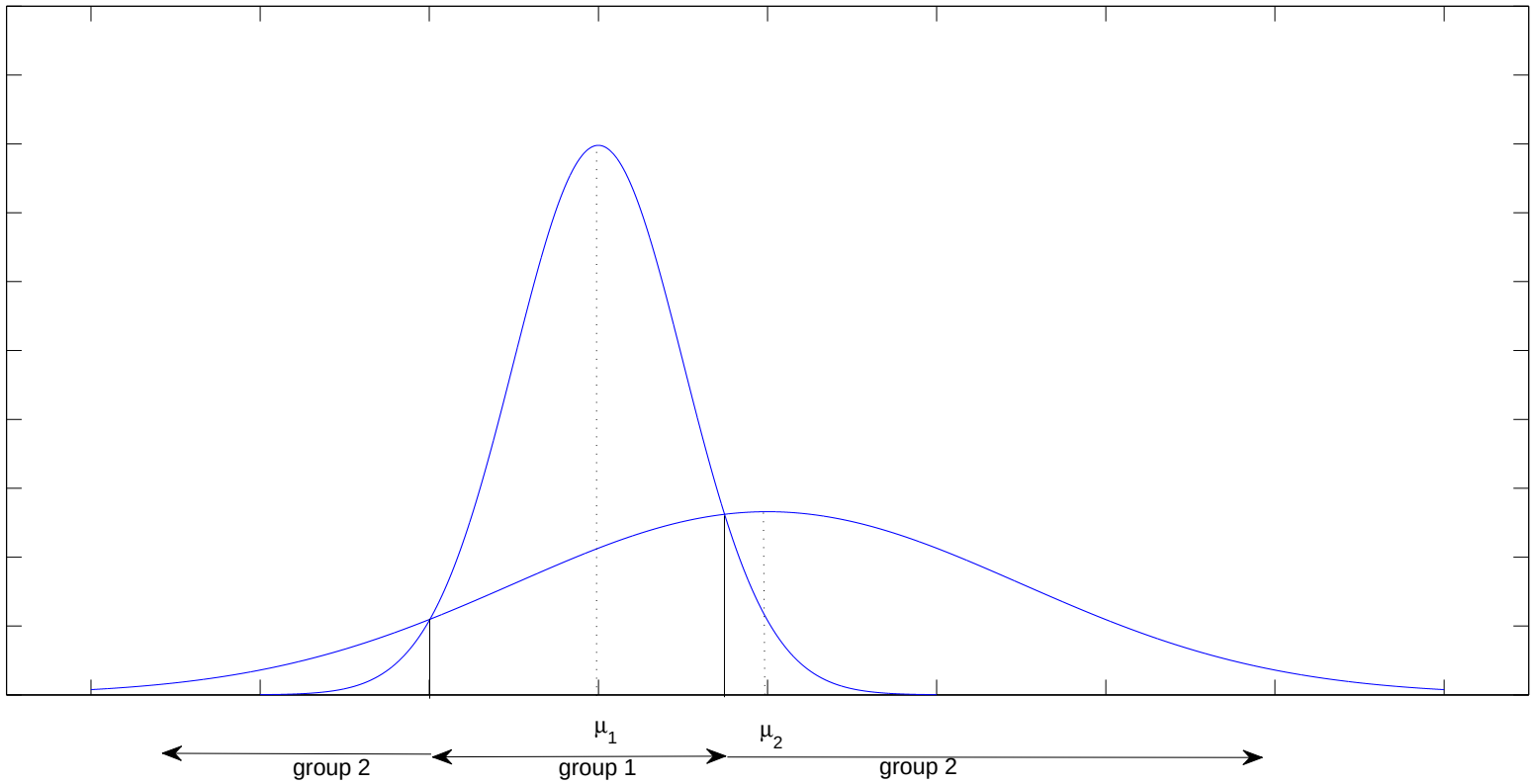
The likelihood $f_1(x)$ is larger than $f_2(x)$ and, therefore, the rule allocates the observation x to the first population, if

$$\frac{f_1(x)}{f_2(x)} = \frac{\sigma_2}{\sigma_1} \exp\left\{-\frac{1}{2}\left[\left(\frac{x - \mu_1}{\sigma_1}\right)^2 - \left(\frac{x - \mu_2}{\sigma_2}\right)^2\right]\right\} > 1,$$

or equivalently if

$$x^2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) - 2x\left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) + \left(\frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2}\right) < 2\log\frac{\sigma_1}{\sigma_2}$$

Example 2



Example 3. The Maximum Likelihood Classification Rule for q multivariate normal populations with common covariance matrices

Let $\mathbf{X} \mid G = k \sim N_p(\mu_k, \Sigma)$, $k = 1, \ldots, q$, i.e

$$f_k(\mathbf{x}) = f(\mathbf{x} \mid G = k) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma^{-1}(\mathbf{x} - \mu_k)\right\}$$

A random observation $\mathbf{x}$ is allocated to the population for which the **Mahalanobis' distance** is maximized. The later is given by

$$\Delta_k^2 = (\mathbf{x} - \mu_k)'\Sigma^{-1}(\mathbf{x} - \mu_k), \quad k = 1, \ldots, q.$$

Example 3. Two Multivariate Normal Populations

For $q = 2$, $r(\mathbf{x}) = 1$ if $f_1(\mathbf{x}) > f_2(\mathbf{x})$, or equivalently if

$$
\begin{aligned}
& (\mathbf{x} - \mu_1)'\Sigma^{-1}(\mathbf{x} - \mu_1) < (\mathbf{x} - \mu_2)'\Sigma^{-1}(\mathbf{x} - \mu_2) \\
\Rightarrow \quad & [(\mathbf{x} - \mu_2) - (\mathbf{x} - \mu_1)]' \Sigma^{-1} [(\mathbf{x} - \mu_2) + (\mathbf{x} - \mu_1)] > 0 \\
\Rightarrow \quad & (\mu_1 - \mu_2)'\Sigma^{-1}(2\mathbf{x} - \mu_1 - \mu_2) > 0 \\
\Rightarrow \quad & 2(\mu_1 - \mu_2)'\Sigma^{-1}(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}) > 0 \\
\Rightarrow \quad & \mathbf{u}'(\mathbf{x} - \mu) > 0, \quad \text{where} \quad \mathbf{u} = \Sigma^{-1}(\mu_1 - \mu_2), \quad \text{and} \quad \mu = \frac{1}{2}(\mu_1 + \mu_2)
\end{aligned}
$$

Note: This is a linear discriminant function!

In the general case of $q = 2$ populations, the maximum likelihood discrimination rule is defined by the discriminant function $h(\mathbf{x}) = \log f_1(\mathbf{x}) - \log f_2(\mathbf{x})$. Then,

$$r(\mathbf{x}) = 1, \quad \text{if} \quad h(\mathbf{x}) > 0.$$

The Bayesian Classification Rule

We have defined the random variable G, representing the group which a random observation derives from. Hence, G is a discrete random variable with probability function

$$\pi_k = P(G = k), \quad k = 1, \ldots, q.$$

If we can assume specific values for the prior classification probabilities, π_k, then we can apply Bayes' theorem to obtain the posterior classification probabilities. For a given observation, $\mathbf{x}$, the posterior probability that it comes from group k is

$$\tilde{\pi}_k(\mathbf{x}) = P(G = k \mid \mathbf{X} = \mathbf{x}) = \frac{P(G = k) f(\mathbf{x} \mid G = k)}{f(\mathbf{x})}, \quad k = 1, \ldots, q,$$

where $f(\mathbf{x}) = \sum_{j=1}^{q} P(G = j) f(\mathbf{x} \mid G = j)$. Therefore,

$$\tilde{\pi}_k(\mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^{q} \pi_j f_j(\mathbf{x})}, \quad k = 1, \ldots, q.$$

The Bayesian Classification Rule

A random observation $\mathbf{x}$ is allocated to the population for which the posterior classification probability is maximized, i.e

$$r(\mathbf{x}) = i, \quad \text{if} \quad \tilde{\pi}_i(\mathbf{x}) = \max_k \{\tilde{\pi}_k(\mathbf{x})\}.$$

Equivalently,

$$r(\mathbf{x}) = i, \quad \text{if} \quad d_i(\mathbf{x}) = \max_k \{d_k(\mathbf{x})\}, \quad \text{where} \quad d_k(\mathbf{x}) = \log\left(\pi_k f_k(\mathbf{x})\right).$$

Note:

$$d_k(\mathbf{x}) = \log \pi_k - \tfrac{p}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma_k| - \tfrac{1}{2}(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k).$$

The Bayesian Classification Rule: $q = 2$

For $q = 2$ (discriminating between two populations), the Bayesian classification rule is defined by the function

$$\tilde{h}(\mathbf{x}) = \log\left(\pi_1 f_1(\mathbf{x})\right) - \log\left(\pi_2 f_2(\mathbf{x})\right),$$

Then,

$$r(\mathbf{x}) = 1, \quad \text{if} \quad \tilde{h}(\mathbf{x}) > 0.$$

Equivalently,

$$\begin{aligned} r(\mathbf{x}) = 1, \quad & \text{if} \quad \log \pi_1 + \log f_1(\mathbf{x}) - \log \pi_2 - \log f_2(\mathbf{x}) > 0 \Rightarrow \\ r(\mathbf{x}) = 1, \quad & \text{if} \quad h(\mathbf{x}) > \log \pi_2 - \log \pi_1, \\ & \quad \text{where} \quad h(\mathbf{x}) = \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) \end{aligned}$$

Specification of Prior Probabilities

- If there exist some former estimates of the classification probabilities they can be used as priors.
- They can be subjectively specified if we have some prior knowledge, information or belief.
- They can be estimated, for each group, from the proportion of the observations in the sample belonging to that group, i.e $\pi_k = \frac{n_k}{n_1+...+n_q}$.
- If there is no prior information about the classification probabilities, then the groups can be assumed to be a-priori equally probable, i.e $\pi_1 = \ldots = \pi_q = \frac{1}{q}$. In this case, the Bayesian classification rule coincides with the maximum likelihood classification rule.

Normality Assumption: Unknown Parameters

We assume that $\mathbf{X} \mid G = k \sim N_p(\mu_k, \Sigma_k)$, $k = 1, \ldots, q$, where the parameters μ_k, Σ_k are unknown. Then, they have to be estimated from the observed data. Specifically, the parameters μ_k, Σ_k of the kth population are estimated using the n_k observations from that population (note that $\sum_{k=1}^{q} n_k = n$).

We use the sample mean of the kth group, $\bar{\mathbf{x}}^{(k)}$, as an estimate of μ_k and the repective sample covariance matrix, S_k, as an estimate of Σ_k. Then, the observation $\mathbf{x}$ is allocated to the ith group if

$$\hat{d}_i(\mathbf{x}) = \max_k \left\{ \hat{d}_k(\mathbf{x}) \right\}, \quad \text{where}$$

$$\hat{d}_k(\mathbf{x}) = \log \pi_k - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |S_k| - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}}^{(k)})' S_k^{-1} (\mathbf{x} - \bar{\mathbf{x}}^{(k)}).$$

Note: For $\pi_k = 1/q$, $k = 1, \ldots, q$, the Bayesian classification rule reduces to the maximum likelihood rule.