Kapitel 7
Discriminant Analysis

In this section we will address the problem of classifying an individual in one of two (or more) known populations based on measurements of some characteristics of the individual.

We first consider the problem of discriminating between two groups (classes).

7.1 Discrimination between two populations

7.1.1 Bayes and minimax solutions

We consider the **populations** π_1 and π_2 and wish to conclude whether a given individual is a member of group one or group two. We perform measurements of p different characteristics of the individual and hereby get the result

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}.$$

If the individual comes from π_1 the frequency function of $\mathbf{X}$ is $\mathrm{f}_1(\mathbf{x})$ and if it comes from π_2 it is $\mathrm{f}_2(\mathbf{x})$.

Let us furthermore assume that we have given a **loss function** L:

		Choose:	
		π_1	π_2
State	π_1	0	$\mathrm{L}(1,2)$
	π_2	$\mathrm{L}(2,1)$	0

We will assume that there is no loss if we take the correct decision.

In certain situations one also knows approximately what the **prior probability** is to have an individual from each of the groups i.e. we haven given a prior distribution g:

$$\mathrm{g}(\pi_1) = p_1, \quad \mathrm{g}(\pi_2) = p_2.$$

We now seek a **decision function** d: $R^p \to \{\pi_1, \pi_2\}$. d: is defined by

$$\mathrm{d}(\mathbf{x}) = \mathrm{d}_{R_1}(\mathbf{x}) = \begin{cases} \pi_1 & \text{hvis } \mathbf{x} \in R_1 \\ \pi_2 & \text{hvis } \mathbf{x} \in R_2 = \complement R_1. \end{cases}$$

We divide R^p in two regions R_1 and R_2. If our observation lies in R_1 we will choose π_1 and if our observation lies in R_2 we will chose π_2.

If we have a **prior distribution** we define the posterior distribution k by

$$\mathrm{k}(\pi_i|\mathbf{x}) = \frac{\mathrm{f}_i(\mathbf{x})\mathrm{g}(\pi_i)}{p_1\mathrm{f}_1(\mathbf{x}) + p_2\mathrm{f}_2(\mathbf{x})} = \frac{p_i\mathrm{f}_i(\mathbf{x})}{p_1\mathrm{f}_1(\mathbf{x}) + p_2\mathrm{f}_2(\mathbf{x})},$$

cf. p. 6.6 in Vol. 1.

The expected loss in this distribution is

$$\begin{aligned} \mathrm{E}_{\mathbf{x}}(\mathrm{L}(\pi_i, \mathrm{d}_{R_1}(\mathbf{x}))) &= \mathrm{L}(\pi_1, \mathrm{d}_{R_1}(\mathbf{x}))\mathrm{k}(\pi_1|\mathbf{x}) + \mathrm{L}(\pi_2, \mathrm{d}_{R_1}(\mathbf{x}))\mathrm{k}(\pi_2|\mathbf{x}) \\ &= \begin{cases} \mathrm{L}(\pi_2, \pi_1)\mathrm{k}(\pi_2|\mathbf{x}), & \mathbf{x} \in R_1 \\ \mathrm{L}(\pi_1, \pi_2)\mathrm{k}(\pi_1|\mathbf{x}), & \mathbf{x} \in R_2 \end{cases}. \end{aligned}$$

The Bayes solution is defined by that we have to minimise this quantity for any $\mathbf{x}$ (p. 6.9 in Vol. 1), i.e. we must define R_1 by

$$\begin{aligned} \mathbf{x} \in R_1 \quad &\Leftrightarrow \quad \mathrm{L}(2,1)\mathrm{k}(\pi_2|\mathbf{x}) \leq \mathrm{L}(1,2)\mathrm{k}(\pi_1|\mathbf{x}) \\ &\Leftrightarrow \quad \frac{\mathrm{L}(1,2)\mathrm{f}_1(\mathbf{x})p_1}{\mathrm{L}(2,1)\mathrm{f}_2(\mathbf{x})p_2} \geq 1 \\ &\Leftrightarrow \quad \frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq \frac{\mathrm{L}(2,1)}{\mathrm{L}(1,2)}\frac{p_2}{p_1}. \end{aligned}$$

We collect these considerations in

SÆTNING 7.1. The Bayes solution to classification problem is given by the region

$$R_1 = \{\mathbf{x} | \frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq \frac{\mathrm{L}(2,1)}{\mathrm{L}(1,2)} \frac{p_2}{p_1}\}.$$

▲

BEMÆRKNING 7.1. This result is exactly the same as given in theorem 5, chapter 6 in Vol. 1. ▼

If we do not have a prior distribution we can determine a minimax strategy i.e. determine an R_1 so that the maximal risk is minimised. The risk is (cf. p. 6.3, Vol 1)

$$\begin{aligned} R(\pi_1, \mathrm{d}_{R_1}) &= \mathrm{E}_{\pi_1} \mathrm{L}(\pi_1, \mathrm{d}_{R_1}(\mathbf{X})) = \mathrm{L}(1,2) P\{\mathbf{X} \in R_2 | \pi_1\}. \\ R(\pi_2, \mathrm{d}_{R_1}) &= \mathrm{E}_{\pi_2} \mathrm{L}(\pi_2, \mathrm{d}_{R_1}(\mathbf{X})) = \mathrm{L}(2,1) P\{\mathbf{X} \in R_1 | \pi_2\}. \end{aligned}$$

One can now show (see e.g. the proof for theorem 4, chapter 6 in Vol. 1)

SÆTNING 7.2. The minimax solution for the classification problem is given by the region

$$R_1 = \{\mathbf{x} | \frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c\},$$

where c is determined by

$$\mathrm{L}(1,2) P\{\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} < c | \pi_1\} = \mathrm{L}(2,1) P\{\frac{\mathrm{f}_1(\mathbf{X})}{\mathrm{f}_2(\mathbf{X})} \geq c | \pi_2\}.$$

▲

BEMÆRKNING 7.2. The relation for the determination for c can be written

$$\begin{aligned} & \mathrm{L}(1,2) \cdot \text{ (the probability for misclassification if } \pi_1 \text{ is true)} \\ = \; & \mathrm{L}(2,1) \cdot \text{ (the probability for misclassification if } \pi_2 \text{ is true)} \end{aligned}$$

Since one is an increasing and the other a decreasing function of c it is obvious that we will minimise the maximal risk when we have equality. If we do not have any idea about the size of the losses we can let them both equal one. The minimax solution gives us the region which minimises the maximal probability from this classification. ▼

We will now consider the important special case where f_1 and f_2 are normal distributions.

7.1.2 Discrimination between two normal populations

If f_1 and f_2 are normal with the same variance-covariance matrix we have

SÆTNING 7.3. Let $\pi_1 \simeq \mathrm{N}(\mu_1, \mathbf{\Sigma})$ and $\pi_2 \simeq \mathrm{N}(\mu_2, \mathbf{\Sigma})$. Then we have

$$\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c \Leftrightarrow \mathbf{x}'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2'\mathbf{\Sigma}^{-1}\mu_2 \geq \log c.$$

▲

BEVIS 7.1. We introduce the inner product $(\cdot|\cdot)$ and the norm $\| \quad \|$ by

$$(\mathbf{x}|\mathbf{y}) = \mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{y}$$

and

$$\|x\|^2 = (\mathbf{x}|\mathbf{x}).$$

We then have

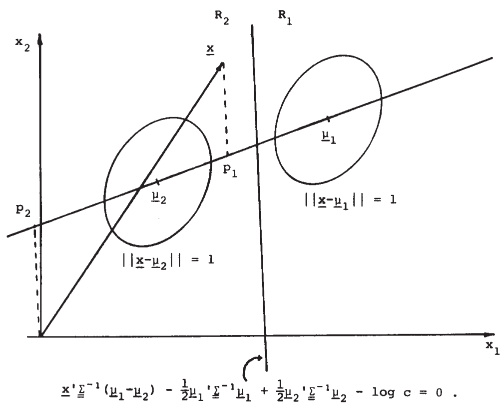
$$\mathrm{f}_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^p \sqrt{\det \mathbf{\Sigma}}} \exp(-\frac{1}{2}\|\mathbf{x} - \mu_i\|^2).$$

From this we readily get

$$\begin{aligned} & \frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c \Leftrightarrow \log \frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq \log c \\ \Leftrightarrow \quad & -\|\mathbf{x} - \mu_1\|^2 + \|\mathbf{x} - \mu_2\|^2 \geq 2 \log c \\ \Leftrightarrow \quad & -(\mathbf{x} - \mu_1 | \mathbf{x} - \mu_1) + (\mathbf{x} - \mu_2 | \mathbf{x} - \mu_2) \geq 2 \log c \\ \Leftrightarrow \quad & 2(\mathbf{x}|\mu_1) - 2(\mathbf{x}|\mu_2) - (\mu_1|\mu_1) + (\mu_2|\mu_2) \geq 2 \log c \\ \Leftrightarrow \quad & 2(\mathbf{x}|\mu_1 - \mu_2) - (\mu_1|\mu_1) + (\mu_2|\mu_2) \geq 2 \log c. \end{aligned}$$

By using the link between $(|)$ and $\mathbf{\Sigma}^{-1}$ we have that the theorem readily follows.

■



BEMÆRKNING 7.3. The expression $\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c$ is seen to define a subset of R^p which is delimited by a hyper-plane (for $p = 2$ a straight line and for $p = 3$ a plane)

The vector $\vec{p_1 p_2}$ is the orthogonal projection (NB! The orthogonal projection with respect to $\mathbf{\Sigma}^{-1}$) of $\mathbf{x}$ on the line which connects μ_1 and μ_2. (It can be shown that the slope of the projection lines etc. are equal to the slope of the ellipse- (ellipsoid-) tangents in the at the points where they intersect the line (μ_1, μ_2)). Since the length of a projection of a vector is equal to the inner product between the vector and a unity vector on the line we see that we have classified the observation as coming from π_1 iff the projection of $\mathbf{x}$ is large enough (computed with sign). Otherwise we will classify the observation as coming from π_2.

The function

$$\mathbf{x}'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2'\mathbf{\Sigma}^{-1}\mu_2 - \log c$$

is called the discriminator or the discriminant function.

We then have that the discriminator is the linear projection which - after the addition of suitable constants - minimises the expected loss (the Bayes situation) or the probability of misclassification (the minimax situation). ▼

In order to make the reader more confident with the content - we will now give a slightly different interpretation of a discriminator. If we let

$$\delta = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2),$$

we have the following

SÆTNING 7.4.
$$\varphi(\mathbf{d}) = \frac{[\mathrm{E}_1(\mathbf{X}'\mathbf{d}) - \mathrm{E}_2(\mathbf{X}'\mathbf{d})]^2}{\mathrm{V}(\mathbf{X}'\mathbf{d})} = \frac{[(\mu_1 - \mu_2)'\mathbf{d}]^2}{\mathbf{d}'\mathbf{\Sigma}\mathbf{d}}.$$
▲

BEVIS 7.2. The proof is not very interesting but fairly simple. Since we readily have that $\varphi(k \cdot d) = k \cdot \varphi(\mathbf{d})$ we can determine extremes for φ by determining extremes for the numerator under the following constraint

$$\mathbf{d}'\mathbf{\Sigma}\mathbf{d} = 1.$$

We introduce a Lagrange multiplier λ and seek the maximum of

$$\psi(\mathbf{d}) = [(\mu_1 - \mu_2)'d]^2 - \lambda(\mathbf{d}'\mathbf{\Sigma}\mathbf{d} - 1).$$

Now we have that

$$\frac{\partial \psi}{\partial \mathbf{d}} = 2(\mu_1 - \mu_2)(\mu_1 - \mu_2)'\mathbf{d} - 2\lambda\mathbf{\Sigma}\mathbf{d}.$$

If we let this equal 0, we have

$$(\mu_1 - \mu_2)(\mu_1 - \mu_2)'\mathbf{d} = \lambda\mathbf{\Sigma}\mathbf{d},$$

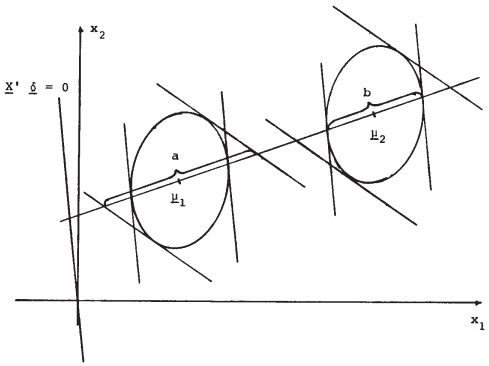
i.e.

$$\mathbf{d} = \frac{(\mu_1 - \mu_2)'\mathbf{d}}{\lambda}\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) = k \cdot \delta,$$

where k is a scalar. ■

BEMÆRKNING 7.4. The content of the theorem is that the linear function determined by δ

$$\mathbf{X}'\delta = \delta_1 X_1 + \cdots + \delta_p X_p,$$



is the projection that “moves” π_1 furthest possible away from π_2 or - in the language of analysis of variance - the projection which maximises the variance between populations divided by the total variance.

The geometric content of the theorem is indicated in the above figure where

b: is the projection of the ellipse on the line μ_1, μ_2 in the direction determined by $\mathbf{x}'\delta = 0$

a: is the projection of the ellipse on the line μ_1, μ_2 on another direction.

It is seen that the projection determined by δ on the line which connects μ_1 and μ_2 is the one which “moves” the projection of the contour ellipsoids of the two populations distribution furthest possible away from each other. ▼

We now give a theorem which is very useful in the determination of misclassification probabilities.

SÆTNING 7.5. We consider the criterion in theorem 7.3

$$Z = \mathbf{X}'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2'\mathbf{\Sigma}^{-1}\mu_2.$$

On this we have

$$Z \in \begin{cases} \mathrm{N}(+\frac{1}{2}\|\mu_1 - \mu_2\|^2, \|\mu_1 - \mu_2\|^2), & \text{hvis } \pi_1 \text{ sand} \\ \mathrm{N}(-\frac{1}{2}\|\mu_1 - \mu_2\|^2, \|\mu_1 - \mu_2\|^2), & \text{hvis } \pi_2 \text{ sand} \end{cases}.$$

▲

BEVIS 7.3. The proof is straight forward. Let us e.g. consider the case π_1 true. We then have that $\mathrm{E}(\mathbf{X}) = \mu_1$ and then

$$\begin{aligned} \mathrm{E}(Z) &= \mu_1'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2'\mathbf{\Sigma}^{-1}\mu_2 \\ &= \frac{1}{2}(\mu_1 - \mu_2)'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \\ &= \frac{1}{2}\|\mu_1 - \mu_2\|^2. \end{aligned}$$

$$\begin{aligned} \mathrm{V}(Z) &= (\mu_1 - \mu_2)'\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \\ &= (\mu_1 - \mu_2)'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \\ &= \|\mu_1 - \mu_2\|^2. \end{aligned}$$

The result regarding π_2 is shown analogously. ■

We will now consider some examples.

EKSEMPEL 7.1. We consider the case where

$$\begin{aligned} \pi_1 &\leftrightarrow \mathrm{N}(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}) \\ \pi_2 &\leftrightarrow \mathrm{N}(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}), \end{aligned}$$

and we want to determine a “best” discriminator function. Since we know nothing about the prior probabilities and the like, we will use the function which corresponds to the constant c in theorem 7.3 being 1. Since

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix},$$

we get the following function

$$(x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix} - \frac{1}{2}(2\cdot 16 + 1\cdot 4 - 2\cdot 8) + \frac{1}{2}(2\cdot 1 + 1\cdot 1 - 2\cdot 1) = 0$$

or

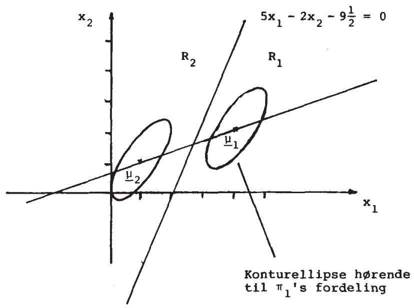
$$5x_1 - 2x_2 - 9\frac{1}{2} = 0.$$

If we enter an arbitrary point, e.g. $\begin{pmatrix} 5 \\ 6 \end{pmatrix}$ we get

$$5\cdot 5 - 2\cdot 6 - 9\frac{1}{2} = 3\frac{1}{2} > 0.$$

This point is therfore classified as coming from π_1.

We have indicated the situation in the following figure



♦

If we have a loss function, the procedure is a bit different which is seen from

EKSEMPEL 7.2. Let us assume that we have losses assigned for the different decisions:

		Choose: π_1	π_2
Nature:	π_1	0	2
	π_2	1	0

Since we have no prior probabilities we will determine the minimax solution. We will need

$$\|\mu_1 - \mu_2\|^2 = 2\cdot 9 + 1\cdot 1 - 2\cdot 3\cdot 1 = 13.$$

From theorem 7.2 follows that we must determine c so

$$\begin{aligned}
& 2\cdot P\left\{\frac{\mathrm{f}_1(\mathbf{X})}{\mathrm{f}_2(\mathbf{X})} < c|\pi_1\right\} = P\left\{\frac{\mathrm{f}_1(\mathbf{X})}{\mathrm{f}_2(\mathbf{X})} \geq c|\pi_2\right\} \\
\Leftrightarrow\quad & 2\cdot P\{Z < \log c|\pi_1\} = P\{Z \geq \log c|\pi_2\} \\
\Leftrightarrow\quad & 2\cdot P\{\mathrm{N}(\tfrac{1}{2}13, 13) < \log c\} = P\{\mathrm{N}(-\tfrac{1}{2}13, 13) \geq \log c\} \\
\Leftrightarrow\quad & 2\cdot P\left\{\mathrm{N}(0,1) < \frac{\log c - 6.5}{\sqrt{13}}\right\} = P\left\{\mathrm{N}(0,1) \geq \frac{\log c + 6.5}{\sqrt{13}}\right\}.
\end{aligned}$$

By trying with different values of c we see that

$$c \simeq 0.5617.$$

Using this value the misclassification probabilities are

If π_1 is true: $\quad P\{\mathrm{N}(0,1) < \frac{\log 0.5617 - 6.5}{\sqrt{13}}\} \simeq 0.025.$

If π_2 is true: $\quad P\{\mathrm{N}(0,1) < \frac{\log 0.5617 + 6.5}{\sqrt{13}}\} \simeq 0.050.$

The discriminating line is now determined by

$$5x_1 - 2x_2 - 9\frac{1}{2} = \log 0.5617,$$

or

$$5x_1 - 2x_2 - 8.92 = 0.$$

This line intersects the line connecting μ_1 and μ_2 in $((2.36, 1.46)$ i.e. it is moved towards μ_2 compared to the mid-point $(2.5, 1.5)$. It is also obvious that the line is moved parallelly in this direction since we see from the loss matrix that it is more serious to be wrong if μ_1 is true than if μ_1 is true. We must therefore expand R_1 i.e. move the limiting line towards μ_2. ♦

We must stipulate that it is of importance that the variance-covariance matrices for the two populations are equal. If this is not the case we will get a completely different result which will be seen from the following example.

EKSEMPEL 7.3. Let us assume that the variance-covariance matrix for population 2 is changed to an identity matrix i.e.

$$\begin{array}{lll} \pi_1 & \leftrightarrow & \mathrm{N}\left(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right) \\ \pi_2 & \leftrightarrow & \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \end{array}$$

Again we want to classify an observation $\mathbf{X}$ which comes from one of the above mentioned distributions. Since the variance covariance matrices are not equal we cannot use the result in theorem 7.3but have to start from the beginning with theorem 7.2.

For $c > 0$ we have

$$\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c \quad \Leftrightarrow$$

$$-(\mathbf{x}-\mu_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\mu_1) + (\mathbf{x}-\mu_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\mu_2) \geq 2\log c.$$

Since

$$\begin{aligned} (\mathbf{x}-\mu_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\mu_1) &= 2(x_1-4)^2 - (x_2-2)^2 - 2(x_1-4)(x_2-2) \\ &= 2x_1^2 + x_2^2 - 2x_1x_2 - 12x_1 + 4x_2 + 20, \end{aligned}$$

and

$$\begin{aligned} (\mathbf{x}-\mu_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\mu_2) &= (x_1-1)^2 + (x_2-1)^2 \\ &= x_1^2 + x_2^2 - 2x_1 - 2x_2 + 2, \end{aligned}$$

then

$$\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c \Leftrightarrow -x^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 \geq 2\log c.$$

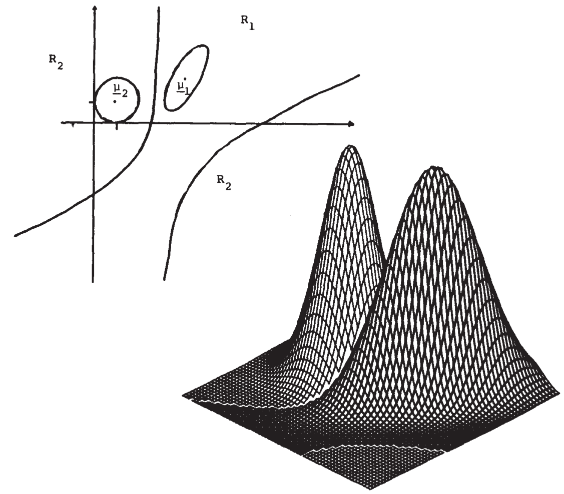
If we choose $c = 1$, we note that the curve which separates R_1 and R_2 is the hyperbola

$$\{\mathbf{x} | -x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 = 0\}.$$

It has centre in $(3, -2)$ and asymptotes

$$x_1 - 3 = 0,$$

$$x_1 - 2x_2 - 7 = 0.$$



These curves are shown in the above figure together with the contour ellipses for the two normal distributions. Note e.g. that a point such as $(9, 0)$ is in R_2 and therefore will be classified as coming from the distribution with centre in $(1, 1)$. Furthermore the frequency functions are shown.

♦

We will not consider the problem of misclassification probabilities in cases as the above mentioned where we have quadratic discriminators.

7.1.3 Discrimination with unknown parameters

If one does not know the two distributions f_1 and f_2 one must estimate them based on some observations and then construct discriminators from the estimated distributions the same way we did for the exact distributions.

Let us consider the normal case

$$\begin{aligned}\pi_1 &\leftrightarrow \mathrm{N}(\mu_1, \mathbf{\Sigma})\\ \pi_2 &\leftrightarrow \mathrm{N}(\mu_2, \mathbf{\Sigma}),\end{aligned}$$

where the parameters are unknown. If we have observations $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ which we know come from π_1 and observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ which we know come from π_2 we can estimate the parameters as follows

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n_1}\sum_i \mathbf{X}_i = \bar{\mathbf{X}}\\ \hat{\mu}_2 &= \frac{1}{n_2}\sum_i \mathbf{Y}_i = \bar{\mathbf{Y}}\\ \hat{\mathbf{\Sigma}} &= \frac{1}{n_1+n_2-2}\left(\sum_i(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})' + \sum_i(\mathbf{Y}_i-\bar{\mathbf{Y}})(\mathbf{Y}_i-\bar{\mathbf{Y}})'\right)\end{aligned}$$

In complete analogy to theorem p. 206 we have the discriminator

$$\mathbf{x}'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1-\hat{\mu}_2) - \frac{1}{2}\hat{\mu}_1'\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1 + \frac{1}{2}\hat{\mu}_2'\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2$$

The exact distribution of this quantity if we substitute $\mathbf{x}$ with a stochastic variable $\mathbf{X} \in \mathrm{N}(\mu_i, \mathbf{\Sigma})$ is fairly complicated but for large sample sizes it is asymptotically equal to the distribution of Z in theorem 7.5 so for reasonable sample sizes we can use the theory we have derived.

The estimated norm between the expected values is

$$\|\hat{\mu}_1-\hat{\mu}_2\|^2 \simeq \mathrm{D}^2 = (\hat{\mu}_1-\hat{\mu}_2)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1-\hat{\mu}_2) = \|\hat{\mu}_1-\hat{\mu}_2\|^2_{\hat{\mathbf{\Sigma}}^{-1}}.$$

This is called **Mahalanobis**' distance. It should here be noted that a number of authors use the expression Mahalabobis' distance also on the quantity $\|\mu_1-\mu_2\|^2$. This is after the Indian statistician P.C. Mahalanobis who developed discriminant analysis at the same time as the English statistician R.A. Fisher in the 30'es.

By means of D^2 we can test if $\mu_1 = \mu_2$ since

$$Z = \frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \cdot \frac{n_1 n_2}{n_1+n_2}\,\mathrm{D}^2$$

is $\mathrm{F}(p, n_1+n_2-p-1)$-distributed if $\mu_1 = \mu_2$. If $\mu_1 \neq \mu_2$ then Z has a larger mean value so the critical region become large values of Z. This test is of course equivalent to Hotelling's T^2-test in section 6.1.2.

We give an example (data come from K.R. Nair: A biometric study of the desert locust, Bull. Int. Stat. Inst. 1951).

EKSEMPEL 7.4. In an investigation of dessert locusts one measured the following biometric characteristics they were

x_1: length of hind femur
x_2: maximum width of the head in the genal region
x_3: length of pronotum at the scull

The two species which were examined are gregaria and an intermediate phase between gregaria and solotaria.

The following mean values were found.

	Mean values	
	Gregaria $n_1 = 20$	Intermediate phase $n_2 = 72$
x_1	25.80	28.35
x_2	7.81	7.41
x_3	10.77	10.75

The estimated variance-covariance matrix is

	x_1	x_2	x_3
x_1	4.7350	0.5622	1.4685
x_2	0.5622	0.1413	0.2174
x_3	1.4685	0.2174	0.5702

One is now interested in determining a discrimination function for classification of future locusts by means of measurements of x_1, x_2, x_3.

First it would, however, be reasonable to investigate if the three measurements from the two populations are different at all i.e. we must investigate if it can be assumed that $\mu_1 = \mu_2$. We have

$$\mathrm{D}^2 = (\hat{\mu}_1-\hat{\mu}_2)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1-\hat{\mu}_2) = 9.7421.$$

This value is inserted in the test statistic p. 216 and we get

$$Z = \frac{20+72-3-1}{3(20+72-2)} \cdot \frac{20 \cdot 72}{20+72} \cdot 9.7421 = 49.70.$$

Since

$$\mathrm{F}(3, 88)_{0.999} \simeq 6,$$

we will reject the hypothesis of the two mean values being equal. It is therefore sensible to try constructing a discriminator.

We have

$$\mathbf{x}'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) = -2.7458x_1 + 6.6217x_2 + 4.5820x_3$$

and

$$\frac{1}{2}(\hat{\mu}_1'\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1 - \hat{\mu}_2'\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2) = 25.3506.$$

Since there is no information on prior probabilities we will use $c = 1$, i.e. : $\log c = 0$, and we will therefore use the function

$$\mathrm{d}(\mathbf{x}) = -2.7458x_1 + 6.6217x_2 + 4.582x_3 - 25.3506$$

in classifying the two possible species of locust.

If we for instance have caught a specimen with the measured characteristics

$$\mathbf{x} = \begin{pmatrix} 27.06 \\ 8.03 \\ 11.36 \end{pmatrix}$$

we get $\mathrm{d}(\mathbf{x}) = 5.5715 > 0$ meaning we will classify the individual as being a gregaria.

♦

7.1.4 Test for best discrimination function

We remind that the best discrimination

$$\hat{\delta} = \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2),$$

can be found by maximising the function

$$\hat{\varphi}(\mathbf{d}) = \frac{[(\hat{\mu}_1 - \hat{\mu}_2)'\mathbf{d}]^2}{\mathbf{d}'\hat{\mathbf{\Sigma}}\mathbf{d}}.$$

The maximum value is

$$\hat{\varphi}(\hat{\delta}) = \frac{[(\hat{\mu}_1 - \hat{\mu}_2)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)]^2}{(\hat{\mu}_1 - \hat{\mu}_2)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)} = \mathrm{D}^2,$$

i.e. Mahalanobis' D^2 is the maximum value of $\hat{\varphi}(\mathbf{d})$. For an arbitrary (fixed) $\mathbf{d}$ we now let

$$\mathrm{D}_1^2 = \hat{\varphi}(\mathbf{d}) = \frac{[(\hat{\mu}_1 - \hat{\mu}_2)'\mathbf{d}]^2}{\mathbf{d}'\hat{\mathbf{\Sigma}}\mathbf{d}}.$$

We can then test the hypothesis that the linear projection determined by $\mathbf{d}$ is the best discriminator by means of the test statistic

$$Z = \frac{n_1 + n_2 - p - 1}{p - 1} \cdot \frac{n_1 n_2(\mathrm{D}^2 - \mathrm{D}_1^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 \mathrm{D}_1^2},$$

which is $\mathrm{F}(p-1, n_1+n_2-p-1)$-distributed under the hypothesis. Large values of Z are critical.

We will not come into the reason why the distribution for the 0-hypothesis looks the way it does but just note that Z gives a measure of how much the "distance" between the two populations is reduced by using $\mathbf{d}$ instead of $\hat{\delta}$. If this reduction is too big i.e. if Z is large we will not be able to assume that $\mathbf{d}$ gives essentially as good a discrimination between the two populations as $\hat{\delta}$.

EKSEMPEL 7.5. In the following table we give averages of 50 measurements of different characteristics of two different types of Iris, Iris versicolor and Iris setosa. (The data come from Fisher's investigations in 1936.)

	Versicolor	Setosa	Differens
Bægerblads længde	5.936	5.006	0.930
Bægerblads bredde	2.770	3.428	−0.658
Kronblads længde	4.260	1.462	2.789
Kronblads bredde	1.326	0.246	1.080

The estimated variance covariance matrix (based on 98 degrees of freedom) is

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} 0.19534 & 0.09220 & 0.099626 & 0.03306 \\ & 0.12108 & 0.04718 & 0.02525 \\ & & 0.12549 & 0.039586 \\ & & & 0.02511 \end{bmatrix}$$

From this it readily follows that

$$\hat{\delta} = \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) = \begin{bmatrix} -3.0692 \\ -18.0006 \\ 21.7641 \\ 30.7549 \end{bmatrix}.$$

Mahalanobis' distance between the mean values is

$$\mathrm{D}^2 = [0.930, -0.658, 2.789, 1.080] \begin{bmatrix} -3.0692 \\ -18.0006 \\ 21.7641 \\ 30.7549 \end{bmatrix} = 103.2119.$$

We first test if we can assume that $\mu_1 = \mu_2$. The test statistic is

$$\frac{50+50-4-1}{4(50+50-2)} \frac{50 \cdot 50}{50+50} \cdot 103.2119 = 625.3256$$
$$> \mathrm{F}(4,95)_{0.9995} \simeq 5.5.$$

It will not be reasonable to assume $\mu_1 = \mu_2$.

By looking at the differences between the components in μ_1 and μ_2 we note that the number for versicolor is largest except for x_2 (the sepal's width). Since we are looking for a linear projection which takes a large value for $\mu_1 - \mu_2$ we could try with the projection

$$\mathbf{x}'\mathbf{d}_0 = x_1 - x_2 + x_3 + x_4,$$

where $\mathbf{d}_0$ here is the vector $\begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$.

We will now test if it can be assumed that the best discriminator has the form

$$\delta = \text{konstant} \cdot \begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix} = \text{konstant} \cdot \mathbf{d}_0.$$

We determine the value of φ corresponding to $\mathbf{d}_0$:

$$\frac{[(\hat{\mu}_1 - \hat{\mu}_2)'\mathbf{d}_0]^2}{\mathbf{d}_0'\hat{\mathbf{\Sigma}}\mathbf{d}_0} = 61.9479.$$

The test statistic becomes

$$\frac{50+50-4-1}{4-1} \cdot \frac{50 \cdot 50(103.2119 - 61.9479)}{(50+50)(50+50-2) + 50 \cdot 50 \cdot 61.9479}$$
$$= 1984 > \mathrm{F}(3,95)_{0.9995} \simeq 6.5.$$

We must therefore reject the hypothesis and note that we cannot assume that the best discriminator is of the form $x_1 - x_2 + x_3 + x_4$. ♦

7.1.5 Test for further information

Given one has measurements of a number of variables for some individuals with the goal of determining a discriminant function. One often has the question if it really is necessary with all the measurements, or if one can do with fewer variables in order to separate the populations from each other. One could e.g. think it might be sufficient to measure the length of sepal and petal in order to discriminate between Iris versicolor and Iris setosa.

We will reformulate these thoughts a bit more precisely. In the discrimination we measure the variables $X_1, \ldots, X_p$. We now will perform a test in order to investigate if it is possible that the last q variables are unnecessary for the discrimination.

We still assume that there are n_1 observations from π_1 and n_2 observations from population π_2. We let

$$\begin{bmatrix} X_1 \\ \vdots \\ X_{p-q} \end{bmatrix} = \mathbf{X}_1 \quad \text{og} \quad \begin{bmatrix} X_{p-q+1} \\ \vdots \\ X_p \end{bmatrix} = \mathbf{X}_2,$$

and we perform the same partitioning of mean vectors and variance-covariance matrix

$$\begin{aligned} \mu_i &= \begin{bmatrix} \mu_i^{(1)} \\ \mu_i^{(2)} \end{bmatrix} \\ \mathbf{\Sigma} &= \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}. \end{aligned}$$

We now compute Mahalanobis' distance between the populations, first using full information i.e. all p variables and then using the reduced information i.e. the first $p-q$ variables. We then have

$$D_p^2 = (\hat{\mu}_1 - \hat{\mu}_2)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

and

$$D_{p-q}^2 = (\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)})' \hat{\boldsymbol{\Sigma}}_{11}^{-1} (\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)}).$$

A test for the hypothesis that the last q variables do not contribute to a better discrimination is based on

$$Z = \frac{n_1 + n_2 - p - 1}{q} \frac{n_1 n_2 (D_p^2 - D_{p-q}^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_{p-q}^2}.$$

It can be shown that $Z \in \mathrm{F}(q, n_1 + n_2 - p - 1)$ if H_0 is true. We will omit the proof, but just state that Z "measures" the relative larger distance there is between populations when going from $p-q$ variables to p variables. It is therefore also intuitively reasonable that we reject the hypothesis that it is sufficient with $p - q$ variables if Z is large.

We now give an illustrative

EKSEMPEL 7.6. We will investigate if it is sufficient to measure the length of sepal and petal in order to discriminate the types of Iris given in example 7.5.

We now perform an ordinary discriminant analysis on the data given but we do not consider the width measurements. The resulting Mahalanobis' distance is

$$D_2^2 = 76.7082,$$

so the test statistic for the hypothesis given is

$$\begin{aligned} &\frac{50 + 50 - 4 - 1}{2} \frac{50 \cdot 50(103.2119 - 76.7082)}{(50 + 50 - 2)(50 \cdot 50 \cdot 76.7082)} \\ &= 15.6132 > \mathrm{F}(2, 95)_{0.9995} \simeq 8.25. \end{aligned}$$

We must therefore assume that there is extra information in the width measurements which can help us in discriminating setosa from versicolor. ♦

7.2 Discrimination between several populations

7.2.1 The Bayes solution

The main idea in the generalisation in this section is that one compares the populations pairwise as in the previous section and then finally chooses the most probable population.

We consider the populations

$$\pi_1, \ldots, \pi_k$$

and on the basis of measurements of p characteristics (or variables) of a given individual we wish to classify it as coming from one of the populations $\pi_1, \ldots, \pi_k$.

The result of the observations is

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}.$$

If the individual comes from π_i then the frequency function for $\mathbf{X}$ is $\mathrm{f}_i(\mathbf{x})$.

We assume that a loss function L is given as shown in the following table.

		Vælger			
		π_1	π_2	$\cdots$	π_k
	π_1	0	$\mathrm{L}(1,2)$	$\cdots$	$\mathrm{L}(1,k)$
	π_2	$\mathrm{L}(2,1)$	0	$\cdots$	$\mathrm{L}(2,k)$
Tilstand :	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	π_k	$\mathrm{L}(k,1)$	$\mathrm{L}(k,2)$	$\cdots$	0

Finally we can assume that we have a prior distribution

$$\mathrm{g}(\pi_i) = p_i, \qquad i = 1, \ldots, k.$$

For an individual with the observation $\mathbf{x}$ we define the discriminant value or discriminant score for the i'th population as

$$S_i^*(\mathbf{x}) = S_i^* = -[p_1 \mathrm{f}_1(\mathbf{x}) \mathrm{L}(1,i) + \cdots + p_k \mathrm{f}_k(\mathbf{x}) \mathrm{L}(k,i)]$$

(note that $\mathrm{L}(i,i) = 0$ so that the sum has no term $p_i \mathrm{f}_i(\mathbf{x})$). Since the prior probability for π_ν is

$$\begin{aligned} \mathrm{k}(\pi_\nu | \mathbf{x}) &= \frac{p_\nu \mathrm{f}_\nu \mathbf{x})}{p_1 \mathrm{f}_1(\mathbf{x}) + \cdots + p_k \mathrm{f}_k(\mathbf{x})} \\ &= \frac{p_\nu \mathrm{f}_\nu(\mathbf{x})}{\mathrm{h}(\mathbf{x})}, \end{aligned}$$

we note that by choosing the i'th population S_i^* is a constant $(-\mathrm{h}(\mathbf{x}))$ times the expected loss with respect to the posterior distribution of π. Since the proportionality factor $-\mathrm{h}(\mathbf{x})$ is negative we note that the Bayes' solution to the decision problem is to choose the population which has the largest discriminant value (discriminant score) i.e. choose π_ν if

$$S_\nu^* \geq S_i^*, \qquad \forall i.$$

If all $\mathrm{L}(i,j)$ $(i \neq j)$ are equal we can simplify the expression for the discriminant score. We prefer π_i for π_j if

$$S_i^* > S_j^*,$$

i.e. if

$$\begin{aligned} & -(\sum_\nu p_\nu \mathrm{f}_\nu(\mathbf{x}) - p_i \mathrm{f}_i(\mathbf{x})) > -(\sum_\nu p_\nu \mathrm{f}_\nu(\mathbf{x}) - p_j \mathrm{f}_j(\mathbf{x})) \\ \Leftrightarrow \quad & p_i \mathrm{f}_i(\mathbf{x}) > p_j \mathrm{f}_j(\mathbf{x}). \end{aligned}$$

In this case we can therefore choose the discriminant score

$$S_i' = p_i \mathrm{f}_i(\mathbf{x}).$$

In this case the **Bayes' rule** is that we choose the population which has the largest posterior distribution i.e. choose group i, if $S_i' > S_j', \forall j \neq i$. This rule is not only used where the losses are equal but also where it has not been possible to determine such losses. If the p_is are unknown and it is not possible to estimate them one usually uses the discriminant score

$$S_i'' = \mathrm{f}_i(\mathbf{x}),$$

i.e. choose the population where the observed probability is the largest.

The minimax solutions are determined by choosing the strategy which makes all the misclassification probabilities equally large. (Still assuming that all losses are equal.) We will, however, not be going into more detail about this here.

7.2.2 The Bayes' solution in the case with several normal distributions

We will now consider the case where

$$\pi_i \quad \leftrightarrow \quad \mathrm{N}(\mu_i, \mathbf{\Sigma}_i),$$

i.e.

$$\mathrm{f}_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \mathbf{\Sigma}_i}} \exp(-\frac{1}{2}(\mathbf{x}-\mu_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)),$$

for $i = 1, \ldots, k$.

Since we get the same decision rule by choosing monotone transformations of our discriminant scores we will take the logarithm of the f_is and disregard the common factor $(2\pi)^{-\frac{p}{2}}$. This gives (assuming that the losses are equal)

$$S_i' = -\frac{1}{2}\log(\det \mathbf{\Sigma}_i) - \frac{1}{2}(\mathbf{x}-\mu_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mu_i) + \log p_i.$$

This function is quadratic in $\mathbf{x}$ and is called a quadratic discriminant function. If all the $\mathbf{\Sigma}_i$ are equal then the terms

$$-\frac{1}{2}\log\det\mathbf{\Sigma} - \frac{1}{2}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}.$$

are common for all S_is and can therefore be omitted. We then get

$$S_i = \mathbf{x}'\mathbf{\Sigma}^{-1}\mu_i - \frac{1}{2}\mu_i'\mathbf{\Sigma}^{-1}\mu_i + \log p_i.$$

This is seen to be a linear (affine) function in $\mathbf{x}$. If there are only two groups we note that we choose group 1 if

$$\begin{aligned} & S_1' > S_2' \Leftrightarrow S_1 - S_2 > 0 \\ \Leftrightarrow \quad & \mathbf{x}'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2'\mathbf{\Sigma}^{-1}\mu_2 \geq \log\frac{p_2}{p_1}, \end{aligned}$$

i.e. the same result as p. 206.

The posterior probability for the νth group becomes

$$\mathrm{k}(\pi_\nu|\mathbf{x}) = \frac{\exp(S_\nu)}{\sum_{i=1}^k \exp(S_i)}$$

It is of course possible to describe the decision rules by dividing R^p into sets $R_1, \ldots, R_k$ so that we choose π_i exactly if $\mathbf{x} \in R_i$. Among other things this will be seen from the following

EKSEMPEL 7.7. We consider populations π_1, π_2 and π_3 given by normal distributions with expected values

$$\mu_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{og} \quad \mu_3 = \begin{pmatrix} 2 \\ 6 \end{pmatrix},$$

and the common variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

cf. the example p. 210. Assuming that all p_i are equal so that we can disregard them in the discriminant scores - we then have

$$\begin{aligned} S'_{11} &= (x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \end{pmatrix} - \frac{1}{2}(4,2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \end{pmatrix} \\ &= 6x_1 - 2x_2 - 10 \end{aligned}$$

$$\begin{aligned} S'_{12} &= (x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2}(1,1)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= x_1 - \frac{1}{2} \end{aligned}$$

$$\begin{aligned} S'_{13} &= (x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 6 \end{pmatrix} - \frac{1}{2}(2,6)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 6 \end{pmatrix} \\ &= -2x_1 + 4x_2 - 10. \end{aligned}$$

We now choose to prefer π_1 for π_2 if

$$\begin{aligned} u_{12}(\mathbf{x}) &= 6x_1 - 2x_2 - 10 - (x_1 - \frac{1}{2}) \\ &= 5x_1 - 2x_2 - 9\frac{1}{2} \\ &> 0. \end{aligned}$$

We choose to prefer π_1 for π_3 if

$$\begin{aligned} u_{13}(\mathbf{x}) &= 6x_1 - 2x_2 - 10 - (-2x_1 + 4x_2 - 10) \\ &= 8x_1 - 6x_2 \\ &> 0, \end{aligned}$$

and finally we will choose to prefer π_2 for π_3 if

$$\begin{aligned} u_{23}(\mathbf{x}) &= x_1 - \frac{1}{2} - (-2x_1 + 4x_2 - 10) \\ &= 3x_1 - 4x_2 + 9\frac{1}{2} \\ &> 0. \end{aligned}$$

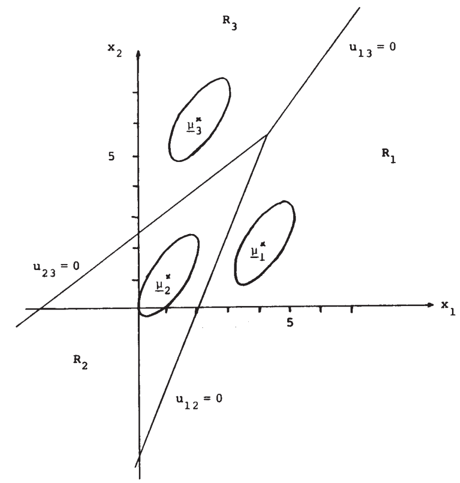
It is now evident that we will choose π_1 if both $u_{12}(\mathbf{x}) > 0$ and $u_{13}(\mathbf{x}) > 0$ and analogously with the others.

We can therefore define the regions

$$\begin{aligned} R_1 &= \{\mathbf{x} | u_{12}(\mathbf{x}) > 0 \quad \wedge \quad u_{13}(\mathbf{x}) > 0\} \\ R_2 &= \{\mathbf{x} | u_{12}(\mathbf{x}) < 0 \quad \wedge \quad u_{23}(\mathbf{x}) > 0\} \\ R_3 &= \{\mathbf{x} | u_{13}(\mathbf{x}) < 0 \quad \wedge \quad u_{23}(\mathbf{x}) < 0\}, \end{aligned}$$

and we have that we will choose π_i exactly if $\mathbf{x} \in R_i$.

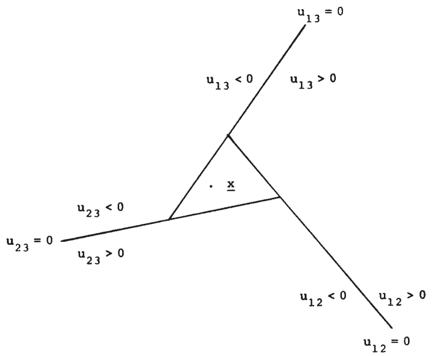
We have sketched the situation in the following figure.



One can easily prove that the lines will intersect in a point. It is, however, also possible to make a reasoning for this. Let us assume that the situation is as in figure 7.1.

We now note that

$$u_{ij} > 0 \Leftrightarrow S'_{1i} > S'_{1j} \Leftrightarrow \mathrm{f}_i > \mathrm{f}_j.$$



Figur 7.1:

For the point $\mathbf{x}$ we have

$$\left.\begin{array}{lll} u_{23}(\mathbf{x})<0 & \text{d.v.s.} & \mathrm{f}_2(\mathbf{x})<\mathrm{f}_3(\mathbf{x}) \\ u_{13}(\mathbf{x})>0 & \text{d.v.s.} & \mathrm{f}_1(\mathbf{x})>\mathrm{f}_3(\mathbf{x}) \end{array}\right\} \Rightarrow \mathrm{f}_1(\mathbf{x})>\mathrm{f}_2(\mathbf{x})$$
$$u_{12}(\mathbf{x})<0 \quad \text{d.v.s.} \quad \mathrm{f}_1(\mathbf{x})<\mathrm{f}_2(\mathbf{x})$$

i.e. we have now established a contradiction i.e. the three lines determined by u_{12}, u_{13} and u_{23} must intersect each other in a point. ♦

If the parameters are unknown and instead are estimated they are inserted in the estimating expressions in the above mentioned relations cf. the procedure in section 7.1.3.

7.2.3 Alternative discrimination procedure for the case of several populations.

In the previous section we have given one form of the generalisation of discriminant analysis from 2 to several populations. We will now describe another procedure which instead generalises theorem 7.4.

We still consider k groups with $n_1, \ldots, n_k$ observations in each. The group averages are called $\bar{\mathbf{X}}_1, \ldots, \bar{\mathbf{X}}_k$. We define an “among groups” matrix

$$\mathbf{A} = \sum_{i=1}^{k} n_i(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})',$$

a ”within groups” matrix

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

and a “total” matrix

$$\mathbf{T} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \bar{\mathbf{X}})(\mathbf{X}_{ij} - \bar{\mathbf{X}})'.$$

A fundamental equation is

$$\mathbf{T} = \mathbf{A} + \mathbf{W}.$$

We can now go ahead with the discrimination. We seek a best discriminator function where best means that the function should maximise the ratio between variation among groups and variation within groups. I.e. we seek a function $\mathbf{y} = \mathbf{d}'\mathbf{x}$ so

$$\varphi(\mathbf{d}) = \frac{\mathbf{d}'\mathbf{A}\,\mathbf{d}}{\mathbf{d}'\mathbf{W}\,\mathbf{d}} \qquad (\mathbf{d} \text{ is chosen so } \mathbf{d}'\mathbf{d} = 1)$$

is maximised. We note from theorem **??** that the maximum value is the largest eigenvalue λ_1 and the corresponding eigenvector $\mathbf{d}_1$ to

$$\det(\mathbf{A} - \lambda\mathbf{W}) = 0$$

or

$$\det(\mathbf{W}^{-1}\mathbf{A} - \lambda\mathbf{I}) = 0.$$

We then seek a new discriminant function $\mathbf{d}_2$ so

$$\varphi(\mathbf{d}_2) = \frac{\mathbf{d}_2\mathbf{A}\,\mathbf{d}_2}{\mathbf{d}_2\mathbf{W}\,\mathbf{d}_2}$$

is maximised under the constraint that

$$\mathbf{d}_2'\mathbf{d}_1 = 0 \quad \text{eller} \quad \mathbf{d}_1 \perp \mathbf{d}_2 \quad \text{og} \quad \mathbf{d}_2'\mathbf{d}_2 = 1.$$

This corresponds to the second largest eigenvalue for $\mathbf{W}^{-1}\mathbf{A}$ and the corresponding eigenvector.

In this way one can continue until one gets an eigenvalue for $\mathbf{W}^{-1}\mathbf{A}$ which is 0 (or until $\mathbf{W}^{-1}\mathbf{A}$ is exhausted).

A plot of the projections of the single observations (normed with the total mean) onto the $\mathbf{d}_1, \mathbf{d}_2$ plane will be useful as a means of visualisation. This plan separates the points best in the sense described above.

The coordinates of the projections are

$$[\mathbf{d}_1'(\mathbf{x}_{ij} - \bar{\mathbf{x}}), \quad \mathbf{d}_2'(\mathbf{x}_{ij} - \bar{\mathbf{x}})].$$

Another useful plot is one of the vectors

$$\begin{pmatrix} d_{11} \\ d_{21} \end{pmatrix}, \ldots, \begin{pmatrix} d_{1p} \\ d_{2p} \end{pmatrix}.$$

These show with which weight the value of the single variable contributes to the plot on the $(\mathbf{d}_1, \mathbf{d}_2)$-plane.

E.g. in the programme BMD07M - STEPWISE DISCRIMINANT ANALYSIS - the plane $(\mathbf{d}_1, \mathbf{d}_2)$ is denoted the first two canonical variables.

In this programme variables can - as the name indicates - be included or removed from the analysis in a way which is completely similar to a stepwise regression analysis (The version which is called STEPWISE REGRESSION). Apart from controlling the inclusion and removal of variables by means of F-tests there are a number of intuitive criteria which are very well described in the BMD manual p. 243.

It should also be mentioned here that Wilk's Λ for the test of the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k \quad \text{against} \quad H_1 : \exists i, j : \mu_i \neq \mu_j,$$

is

$$\Lambda = \frac{\det \mathbf{W}}{\det \mathbf{T}} = \prod_{j=1}^{p} \frac{1}{1 + \lambda_j}.$$

The distribution of this quantity can be approximated by a χ^2 or F-distribution. The last possibility is probably the numerically best approximation. These are given in the BMD manual p. 242. Cf. with section 6.1.3.

EKSEMPEL 7.8. In the following table we give mean values and standard deviations for content of different elements for 208 washed soil samples collected in Jameson Land. The variable Sum gives the sum of Y and La contents.

Variable	Mean Value	Standard deviation
B	73	141
Ti	40563	22279
V	678	491
Cr	1135	1216
Mn	2562	2081
Fe	225817	122302
Co	62	26
Ni	116	54
Cu	69	56
Ga	21	10
Zr	14752	14771
Mo	29	20
Sn	56	99
Pb	351	786
Sum	–	–

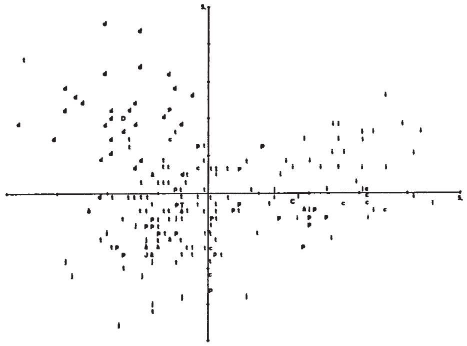
A distributional analysis showed that the data were best approximated by LN-distributions. Therefore all numbers were transformed and were standardised in order to obtain a mean of 0 and a variance of 1. The problem is to how great an extent the content of the elements characterises the difference geologic periods. The number of measurements from the different periods are given below.

Period	Number
Jura	17
Trias	80
Perm	30
Carbon	9
Devon	31
Tertiære intrusives	35
Caledonsk crystallic	4
Eleonora Bay Formation	2

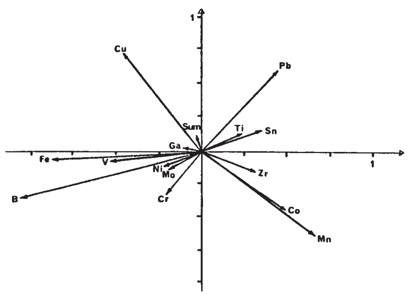
In order to examine this some discriminant analyses were performed. We will not pursue this further here. We will simply illustrate the use of the previously mentioned plot, see figure 7.2.

In the above figure the coefficient for the ordinary variables on the two canonical variables are given.

By comparing the two figures one can e.g. see that Cu is fairly specific for Devon, and



Figur 7.2:



Figur 7.3:

the figures give quite a good impression of how the distribution of elements is for the different periods. ♦