Capítulo 13

DISCRIMINANT ANALYSIS

Ronald Aylmer Fisher (1890-1962)

British scientist, inventor of the techniques of discriminant analysis and maximum likelihood as well as design of scientific experiments. He worked at the Rothamsted Experimental Station in Hertfordshire, England and was Professor of Eugenics at the University of London. In addition to his numerous contributions to all branches of Statistics, which resulted in his being known as the father of the discipline in the 20th century, he was also an outstanding geneticist, agricultural researcher and biologist.

13.1 INTRODUCTION

The problem of discrimination or classification, which we will deal with in this chapter, can be approached in a variety of ways and is present in many areas of human activity: from medical diagnosis to credit scoring or recognizing art forgeries. The statistical approach

to the problem is as follows. We have a set of elements which may come from two or more populations. In each element a $p-$ dimensional random variable $\mathbf{x}$ is observed whose distribution is known in the considered populations. We want to classify a new element, with known values of the variables, in one of the populations. For example, the first application of discriminant analysis consists of classifying the remains of a skull found in an excavation as human, utilizing the distribution of physical measurements for human skulls and those of other anthropoids.

The problem of discrimination appears in many situations in which elements must be classified using incomplete information. For example, the automatic credit scoring systems of financial institutions today are based on using many measurable variables (income, seniority in place of work, wealth) in order to predict future behavior. In other cases the information for classification might be available, but it could result in the destruction of the element, as in the case of quality control for the tension resistance of certain components which should be classified as good or defective. Finally, there are cases where the information is quite costly to acquire. In engineering this problem is studied as *pattern recognition*, for the design of machines capable of automatic classification. This can include voice recognition, classification of bills or coins, on-screen character recognition or the classification of letters by postal code. Other examples of applications of discriminant analysis are: assigning a written text of unknown origin to one of several authors using word frequency, assigning a musical score or painting to an artist, recognizing a tax declaration as potentially fraudulent or not, a business as a bankruptcy risk or not, the teachings of a center as theoretical or applied, a patient as having cancer or not, a new manufacturing process as efficient or not.

The techniques we will study here are also known as *supervised classification*, in order to indicate that we know of a sample of well-classified elements which serve as a standard or model for the classification of subsequent observations.

There are various possible approaches to this problem. The first, which is presented here, is the classical discriminant analysis developed by Fisher, based on multivariate normality of the considered variables and which is optimal under that assumption. If all the variables are continuous it often happens that, although the original data are not normal, it is possible to transform the variables so that they are, and the methods in this chapter can be applied to the transformed variables. Nevertheless, when we have discrete and continuous variables to classify, the hypothesis of multivariate normality is rather unrealistic and in the next chapter we will present other approaches to the problem which work better in these cases.

13.2 CLASSIFICATION BETWEEN TWO POPULATIONS

13.2.1 The Problem

Let P_1 and P_2 be two populations where we have defined a random vector variable, $\mathbf{x}$, p-variate. We assume that $\mathbf{x}$ is absolutely continuous and that the density functions of both populations, f_1 and f_2, are known. We are going to study the problem of classifying a new element, $\mathbf{x}_0$, with known values of the p variables, in one of these two populations. If we

know the prior probabilities π_1, π_2, with $\pi_1 + \pi_2 = 1$, that the element comes from one of the two populations, its probability distribution will be a mixed distribution

$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})$$

and once $\mathbf{x}_0$ has been observed we can compute the posterior probabilities that the element has been generated by each of the two populations, $P(i/\mathbf{x}_0)$, with $i = 1, 2$. These probabilities are calculated using Bayes' Theorem

$$P(1|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|1)\pi_1}{\pi_1 P(\mathbf{x}_0|1) + \pi_2 P(\mathbf{x}_0|2)}$$

and since $P(\mathbf{x}_0|1) = f_1(\mathbf{x}_0)\mathbf{\Delta}\mathbf{x}_0$, we have:

$$P(1/\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)\pi_1}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}, \tag{13.1}$$

and for the second population

$$P(2|\mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)\pi_2}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}. \tag{13.2}$$

We classify $\mathbf{x}_0$ in the most probable posterior population. Since the denominators are equal, we classify $\mathbf{x}_0$ in P_2 if:

$$\pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

If the prior probabilities are equal, the condition for classifying in P_2 is simplified to:

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

which means that we classify $\mathbf{x}_0$ in the most probable population, or where its likelihood is highest.

Consideration of the consequences

In many classification problems the errors we might commit have consequences which can be quantified. For example, if an automatic teller mistakenly classifies a 10 Euro note as a 20 Euro note and gives the wrong change, then the cost of that error is 10 Euros. In other cases estimating the cost can be much more complicated: if we do not approve a loan which will be paid back, we can lose a client as well as future income which that client will generate, whereas if the loan is not repaid the cost is the amount outstanding on that loan. As a third example, if we classify a production process as in a state of control, the cost of being wrong is a defective production, whereas if we mistakenly stop a correctly functioning process the cost will be that of the stoppage and revision.

In general we assume that there are only two possible decisions in the problem: assign to P_1 or to P_2. One decision rule is a partition of the sampling space E_x (which is usually R^p) in two regions A_1 and $A_2 = E_x - A_1$, so that:

Figura 13.1: Illustration of a classification problem between two groups as a decision problem.

$$\text{if } \mathbf{x}_0 \in A_1 \Longrightarrow d_1 \text{ (classify in } P_1).$$
$$\text{if } \mathbf{x}_0 \in A_2 \Longrightarrow d_2 \text{ (classify in } P_2).$$

If the consequences of an error can be quantified, we can include them in the solution and formulate a Bayesian decision problem. We assume that:

1. The consequences associated with classification errors are, $c(2|1)$ and $c(1|2)$, where $c(i|j)$ is the cost of classifying a unit in P_i which belongs in P_j. These costs are assumed to be known;

2. The decision maker wants to maximize the utility function and this is equivalent to minimizing the expected cost.

With these two hypotheses the best decision is to minimize the expected costs, or functions of lost opportunity, using Wald's terminology. The results of each decision are presented schematically in Figure 13.1. If we classify the element into group 2 the possible consequences are:

(a) choose correctly, with probability $P(2|\mathbf{x}_0)$, in which case there is no penalization cost;

(b) make a mistake, with probability $P(1|\mathbf{x}_0)$, in which case we incur the associated cost $c(2|1)$.

The average cost, or expected value, of the decision "d_2: classify $\mathbf{x}_0$ in P_2" is:

$$E(d_2) = c(2|1)P(1|\mathbf{x}_0) + 0P(2|\mathbf{x}_0) = c(2|1)P(1|\mathbf{x}_0). \tag{13.3}$$

Analogously, the expected cost of the decision "d_1: classify in group 1" is:

$$E(d_1) = 0P(1|\mathbf{x}_0) + c(1|2)P(2|\mathbf{x}_0) = c(1|2)P(2|\mathbf{x}_0). \tag{13.4}$$

We will assign the element to group 2 if its expected cost is less. In other words, using (13.1) and (13.2), if:

$$\frac{f_2(\mathbf{x}_0)\pi_2}{c\,(2|1)} > \frac{f_1(\mathbf{x}_0)\pi_1}{c\,(1|2)}. \tag{13.5}$$

This condition indicates that, other terms being equal, we classify the item in population P_2 if

(a) its prior probability is higher;

(b) the likelihood that $\mathbf{x}_0$ comes from P_2 is higher;

(c) the cost of incorrectly classifying it in P_2 is lower.

Appendix 13.1 shows that this criterion is equivalent to minimizing the total probability of error in the classification.

13.2.2 Normal Populations: Linear discriminant function

We are going to apply the above analysis to the case in which f_1 and f_2 are normal distributions with different mean vectors but with identical covariance matrices. In order to establish a general rule, suppose that we wish to classify a generic element $\mathbf{x}$, which, if it belongs to the population $i = 1, 2$ its density function is:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}.$$

The optimal decision is, according to the above section, to classify the element in population P_2 if:

$$\frac{f_2(\mathbf{x})\pi_2}{c(2|1)} > \frac{f_1(\mathbf{x})\pi_1}{c(1|2)}. \tag{13.6}$$

Since both terms are always positive, taking the logarithms and replacing $f_i(\mathbf{x})$ with its expression, the above equation becomes:

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \log\frac{\pi_2}{c(2|1)} > -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \log\frac{\pi_1}{c\,(1|2)},$$

Letting D_i^2 be the Mahalanobis distance between the observed point, $\mathbf{x}$, and the mean of the population i; defined by:

$$D_i^2 = (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

we can write:

$$D_1^2 - \log\frac{\pi_1}{c\,(1|2)} > D_2^2 - \log\frac{\pi_2}{c(2|1)} \tag{13.7}$$

and assuming that the costs and prior probabilities are equal, $c(1/2) = c(2/1)$; $\pi_1 = \pi_2$, the above rule is simplified as:

$$\boxed{\text{Classify in 2 if } D_1^2 > D_2^2}$$

or rather, classify the observation in the population whose mean is closest, using the Mahalanobis distance as a measurement. We observe that if the variables $\mathbf{x}$ had $\mathbf{V} = \mathbf{I}\sigma^2$, the rule is equivalent to using the Euclidean distance. Figure 13.2 shows the equidistant curves with the Mahalanobis distance for two normal populations with centers in the origin and the point (5,10).

Figura 13.2: Equidistant curves with the Mahalanobis distance for classifying

13.2.3 Geometric Interpretation

The general rule presented above can be written so that the method of classification used can be interpreted geometrically. Equation (13.7) indicates that we need to calculate the Mahalanobis distance, correct it using the term corresponding to the prior probabilities and costs, and classify in the population where this modified distance is minimum. Since the distances always have the common term, $\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}$, which does not depend on the population, we can eliminate it from the comparison and calculate the indicator

$$-\boldsymbol{\mu}_i'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_i'\mathbf{V}^{-1}\boldsymbol{\mu}_i - \log\frac{\pi_i}{c(i|j)},$$

which will be a linear function in $\mathbf{x}$ and classify the individual in the population where this function is minimum. This rule divides the set of possible values of $\mathbf{x}$ into two regions whose divisions are given by:

$$-\boldsymbol{\mu}_1'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_1'\mathbf{V}^{-1}\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_2'\mathbf{V}^{-1}\boldsymbol{\mu}_2 - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1},$$

which, as a function of $\mathbf{x}$, is equivalent to:

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}\mathbf{x} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}\left(\frac{\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1}{2}\right) - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1}. \tag{13.8}$$

Letting:

$$\boxed{\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)} \tag{13.9}$$

the division can be written as:

$$\mathbf{w}'\mathbf{x} = \mathbf{w}'\frac{\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1}{2} - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \tag{13.10}$$

which is the equation of a hyperplane. In the particular case in which $c(1|2)\pi_2 = c(211)\pi_1$, we classify in P_2 if

$$\mathbf{w}'\mathbf{x} > \mathbf{w}'\left(\frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right). \tag{13.11}$$

or what is equivalent, if

$$\mathbf{w}'\mathbf{x} - \mathbf{w}'\boldsymbol{\mu}_1 > \mathbf{w}'\boldsymbol{\mu}_2 - \mathbf{w}'\mathbf{x} \tag{13.12}$$

This equation indicates that the procedure for classifying an element $\mathbf{x}_0$ can be summarized as follows:

(1) calculate the vector $\mathbf{w}$ with (13.9);

(2) construct the indicator variable:

$$z = \mathbf{w}'\mathbf{x} = w_1x_1 + ... + w_px_p$$

which transforms the multivariate variable $\mathbf{x}$ into the scalar variable z , which is a linear combination of the values of the multivariate variable with coefficients given by the vector $\mathbf{w}$;

(3) calculate the value of the indicator variable for the individual to be classified, $\mathbf{x}_0 = (x_{10}, ..., x_{p0})$, with $z_0 = \mathbf{w}'\mathbf{x}_0$ and the value of the indicator variable for the means of the populations, $m_i = \mathbf{w}'\boldsymbol{\mu}_i$. Classify in the population where the distance $|z_0 - m_i|$ is minimum.

In terms of the scalar variable, z, since the average value of z in P_i is :

$$\mathbf{E}(z|P_i) = m_i = \mathbf{w}'\boldsymbol{\mu}_i, \quad i = 1, 2$$

The decision rule (13.12) is equivalent to classifying in P_2 if:

$$|z - m_1| > |z - m_2| \tag{13.13}$$

The variance of this indicator variable, z, is:

$$Var(z) = \mathbf{w}'Var(\mathbf{x})\mathbf{w} = \mathbf{w}'\mathbf{V}\mathbf{w} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = D^2. \tag{13.14}$$

Figura 13.3: Illustration of the optimum direction of projection for discriminating between two populations.

and the square of the scalar distance between the projected means is the Mahalanobis distance between the vectors of the original means:

$$(m_2 - m_1)^2 = (\mathbf{w}'(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1))^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = D^2. \tag{13.15}$$

The indicator variable z can be interpreted as a projection if we standardize the vector $\mathbf{w}$. Dividing the two members of (13.11) by the norm of $\mathbf{w}$ and letting $\mathbf{u}$ be the unit vector $\mathbf{w}/\left\|\mathbf{w}\right\|$, the classification rule becomes classify in P_2 if

$$\mathbf{u}'\mathbf{x} - \mathbf{u}'\boldsymbol{\mu}_1 > \mathbf{u}'\boldsymbol{\mu}_2 - \mathbf{u}'\mathbf{x}, \tag{13.16}$$

where, since $\mathbf{u}$ is a unit vector, $\mathbf{u}'\mathbf{x}$ is simply the projection of $\mathbf{x}$ in the direction of $\mathbf{u}$, and $\mathbf{u}'\boldsymbol{\mu}_1$ and $\mathbf{u}'\boldsymbol{\mu}_2$ are the projections of the population means in this direction.

In Figure 13.3 it can be seen that the hyperplane perpendicular to $\mathbf{u}$ at the midpoint $\mathbf{u}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ divides the sample space into two regions A_1 and A_2 which constitute the optimal partition desired. If $c(1|2)\pi_2 \neq c(2|1)\pi_1$ the interpretation is the same, but the hyperplane division moves parallel to itself, increasing or decreasing the region A_2.

The direction of the projection, $\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ has a clear geometric interpretation. We consider first the case in which the variables are uncorrelated and standardized so that $\mathbf{V} = \mathbf{I}$. Then, the optimal projection is that defined by $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Generally, the direction of the projection can be calculated in two steps: first, the variables are standardized multivariately, in order to transform them into uncorrelated variables with unit variance; second, the transformed data are projected over the direction which joins the means of the standardized variables.

The calculation of $\mathbf{w}'\mathbf{x}$ can be written as:

$$\mathbf{w}'\mathbf{x} = \left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1/2}\right](\mathbf{V}^{-1/2}\mathbf{x})$$

where $\mathbf{V}^{-1/2}$ exists if $\mathbf{V}$ is positive definite. This expression indicates that this operation is equivalent to: (1) standardizing the variables $\mathbf{x}$ to obtain new variables $\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{x}$ whose covariance matrix is the identity and mean vector is $\mathbf{V}^{-1/2}\boldsymbol{\mu}$; (2) project the new variables $\mathbf{y}$ over the direction $\boldsymbol{\mu}_2(\mathbf{y}) - \boldsymbol{\mu}_1(\mathbf{y}) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1/2}$.

Figure 13.4 illustrates some directions of projections. In (a) and (b) the direction of the lines which join the means coincides with some of the principal axes of the ellipsis and thus the direction $\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ coincides with $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, since this is an eigenvector of $\mathbf{V}$, and hence, also of $\mathbf{V}^{-1}$. In (c) the optimal direction is a compromise between $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ and the directions defined by the eigenvectors of $\mathbf{V}^{-1}$.

Figura 13.4: In cases (a) and (b) the optimum direction coincides with the line of the means and the axes of the ellipsis. In case (c) it is a compromise between the two.

13.2.4 Calculation of error probabilities

The usefulness of the classification rule depends on the expected errors. Since the distribution of the variable $z = \mathbf{w}'\mathbf{x}$ is normal, with mean $m_i = \mathbf{w}'\boldsymbol{\mu}_i$ and variance $D^2 = (m_2 - m_1)^2$, we can calculate the probabilities of erroneously classifying an observation in each of the two populations. Specifically, the probability of an erroneous decision when $\mathbf{x} \in P_1$ is:

$$P(2|1) = P\left\{ z \geq \frac{m_1 + m_2}{2} \mid z \text{ is } N(m_1; D) \right\}$$

and letting $y = (z - m_1)/D$ be a random variable $N(0,1)$, and Φ its distribution function:

$$P(2|1) = P\left\{ y \geq \frac{\frac{m_1+m_2}{2} - m_1}{D} \right\} = 1 - \Phi\left(\frac{D}{2}\right)$$

Analogously, the probability of an erroneous decision when $\mathbf{x} \in P_2$ is:

$$P(1|2) = P\left\{ z \leq \frac{m_1 + m_2}{2} \mid z \text{ is } N(m_2; D) \right\} =$$

$$= P\left\{y \leq \frac{\frac{m_1+m_2}{2} - m_2}{D}\right\} = \Phi\left(-\frac{D}{2}\right)$$

and both error probabilities are identical due to the symmetry of the normal distribution. We can conclude that the rule obtained makes the error probabilities equal and minimum (see Appendix 13.1) and makes it so the classification errors depend only on the Mahalanobis distance between the means.

13.2.5 A posteriori probabilities

The degree of confidence when classifying an observation depends on the probability of being right. The a posteriori probability that the observation belongs to the first population is calculated with:

$$\begin{aligned} P(1|\mathbf{x}) &= \frac{\pi_1 f_1(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})} = \\ &= \frac{\pi_1 \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{\left(\pi_1 \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\} + \pi_2 \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}\right.} \end{aligned}$$

and letting D_1^2 and D_2^2 be the Mahalanobis distances between the point and each of the two means, this expression can be written as:

$$P(1|\mathbf{x}) = \frac{1}{1 + \frac{\pi_2}{\pi_1}\exp\left\{-\frac{1}{2}(D_2^2 - D_1^2)\right\}}$$

and it depends only on the prior probabilities and on the distances between the point and the means of both populations. We observe that if $\pi_2/\pi_1 = 1$, the farther the point is from the first population, or rather the greater D_1^2 is with respect to D_2^2,the larger the denominator and the smaller the probability of its belonging to that population, $P(1|\mathbf{x})$, and vice versa.

Example: We wish to classify a portrait as being the work of two possible artists. To do this, two variables are measured: the depth of the tracing and the proportion of the canvas taken up by the portrait. The means of these variables for the first painter, A, are (2 and .8) and for the second, B, (2.3 and .7), the standard deviations of the variables are .5 and .1 and the correlation between these measurements is .5. The measurements of the work to be classified for these variables are (2.1 and .75). Calculate the probabilities of error.

The Mahalanobis distances will be, calculating the covariance as the product of the correlation for the standard deviations:

$$D_A^2 = (2.1-2, .75-.8)\begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1}\begin{pmatrix} 2.1-2 \\ 75-.8 \end{pmatrix} = 0,52$$

and for the second

$$D_B^2 = (2.1-2.3, .75-.7)\begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1}\begin{pmatrix} 2.1-2.3 \\ .75-.7 \end{pmatrix} = 0,8133$$

Thus, we assign the work to the first artist. The expected classification error with this rule depends on the Mahalanobis distance between the means which is

$$D^2 = (2. - 2.3, .8 - .7) \begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1} \begin{pmatrix} 2. - 2.3 \\ .8 - .7 \end{pmatrix} = 2,6133$$

and $D = 1.6166$. The probability of being wrong is

$$P(A/B) = 1 - \Phi(\frac{1.6166}{2}) = 1 - \Phi(.808) = 1 - 0,8106 = 0,1894.$$

In this way classification using these variables is not very accurate, since we have an 18.94% probability of error. Let us calculate the a posteriori probability that the painting belongs to painter A, assuming, a priori, that both painters are equally probable.

$$P(A/\mathbf{x}) = \frac{1}{1 + \exp(-0.5(0,8133 - 0,52)} = \frac{1}{1.86} = 0,5376$$

This probability indicates that by classifying the painting as belonging to painter A there is a great deal of uncertainty in the decision since the probability of it belonging to either artist are quite similar (0.5376 and 0.4624).

13.3 SEVERAL NORMAL POPULATIONS

13.3.1 General Approach

The generalization of these ideas for G populations is simple: the objective is now to divide the space E_x into G regions $A_1, \ldots, A_g, \ldots, A_G$ so that if $\mathbf{x}$ belongs to A_i the point is classified in population P_i. We assume that the cost of classification are constant and do not depend on the population in which it has been classified. Then, region A_g will be defined by those points with the maximum probability of being generated by P_g, or rather, where the product of the prior probability and the likelihood are maximum:

$$A_g = \{\mathbf{x} \in E_x | \pi_g f_g(\mathbf{x}) > \pi_i f_i(\mathbf{x}); \forall i \neq g\} \tag{13.17}$$

If the prior probabilities are equal, $\pi_i = G^{-1}, \forall i$, and the distributions $f_i(\mathbf{x})$ are normal with the same covariance matrix, the condition (13.17) is equivalent to calculating the Mahalanobis distance from the observed point to the center of each population and classifying it in the population which makes this distance minimum. Minimizing the Mahalanobis distances $\left(\mathbf{x} - \boldsymbol{\mu}_g\right)' \mathbf{V}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_g\right)$ is equivalent, after eliminating the term $\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}$ which appears in all of the equations, to minimizing the linear indicator

$$L_g(\mathbf{x}) = -2\boldsymbol{\mu}_g'\mathbf{V}^{-1}\mathbf{x} + \boldsymbol{\mu}_g'\mathbf{V}^{-1}\boldsymbol{\mu}_g. \tag{13.18}$$

and letting

$$\mathbf{w}_g = \mathbf{V}^{-1}\boldsymbol{\mu}_g$$

the rule is

$$\min_g(\mathbf{w}_g^{'}\boldsymbol{\mu}_g - 2\mathbf{w}_g^{'}\mathbf{x})$$

In order to interpret this rule, we observe that the division between the two populations, (ij), is defined by:

$$A_{ij}(\mathbf{x}) = L_i(\mathbf{x}) - L_j(\mathbf{x}) = 0 \tag{13.19}$$

substituting with (13.18) and re-ordering the terms we obtain:

$$A_{ij}(\mathbf{x}) = 2(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\mathbf{V}^{-1}\mathbf{x} + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\mathbf{V}^{-1}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) = 0$$

and letting

$$\mathbf{w}_{ij} = \mathbf{V}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = \mathbf{w}_i - \mathbf{w}_j$$

the division can be written as:

$$\mathbf{w}_{ij}'\mathbf{x} = \mathbf{w}_{ij}'\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j).$$

This equation allows the same projection interpretation as in the case of two populations. A direction $\mathbf{w}_{ij}$ is constructed and the means and the point $\mathbf{x}$ which we try to classify are projected over this direction. The region of indifference is when the projected point is equidistant from the projected means, and we assign the point to the population whose projected mean is closest.

We are going to prove that if we have G populations we only need to find

$$r = \min(G-1, p)$$

directions of projection. In the first place we observe that, although we can construct $\binom{G}{2} = G(G-1)/2$ vectors $\mathbf{w}_{ij}$ from G means, once we have $G-1$ vectors the rest will be determined by these. We can determine the $G-1$ vectors $\mathbf{w}_{i,i+1}$, for $i = 1, ..., G-1$, and obtain any other from these $G-1$ directions. For example:

$$\mathbf{w}_{i,i+2} = \mathbf{V}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+2}) = \mathbf{V}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+1}) - \mathbf{V}^{-1}(\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_{i+2}) = \mathbf{w}_{i,i+1} - \mathbf{w}_{i+1,i+2}.$$

In conclusion, if $p > G-1$, the maximum number of vectors $\mathbf{w}$ we can have is $G-1$, since all the rest are deduced from them. When $p \leq G-1$, since these vectors belong to R^p, the maximum number of linearly independent vectors is p.

It is important to point out that, logically, the decision rule obtained fulfills the transitive property. For example, if $G = 3$, and we find that for a point $(\mathbf{x})$

$$\begin{aligned} D_1^2(\mathbf{x}) &> D_2^2(\mathbf{x}) \\ D_2^2(\mathbf{x}) &> D_3^2(\mathbf{x}) \end{aligned}$$

then we are forced to conclude that $D_1^2(\mathbf{x}) > D_3^2(\mathbf{x})$ and this is the result that we obtain if we calculate these distances, thus the analysis is coherent. Moreover, if $p = 2$, each of the

three equations $A_{ij}(\mathbf{x}) = 0$ will be a straight line and the three will be intersected in the same point. Note that any straight line which passes through the cut off point of the straight lines $A_{12}(\mathbf{x}) = 0$ and $A_{23}(\mathbf{x}) = 0$ has the expression

$$a_1 A_{12}(\mathbf{x}) + a_2 A_{23}(\mathbf{x}) = 0$$

since $\mathbf{x}_0^*$ is the cut off point as $A_{12}(\mathbf{x}^*) = 0$, by belonging to the first straight line, and $A_{23}(\mathbf{x}^*) = 0$, by belonging to the second will belong to the linear combination. Since, according to (13.19), $A_{13}(\mathbf{x}) = L_1(\mathbf{x}) - L_3(\mathbf{x}) = L_1(\mathbf{x}) - L_2(\mathbf{x}) + L_2(\mathbf{x}) - L_3(\mathbf{x})$, we have:

$$A_{13}(\mathbf{x}) = A_{12}(\mathbf{x}) + A_{23}(\mathbf{x})$$

and the straight line $A_{13}(\mathbf{x})$ always has to pass through the intersecting point of the other two.

13.3.2 Operative procedure

In order to illustrate the operative procedure, we assume we have five populations with $p > 4$, so that there will be four independent classification rules and the rest will be deduced from them. There are two ways of carrying out the analysis. The first is to calculate the Mahalanobis distances for the G populations (or what is equivalent, the projections (13.18)) and classify the element in the closest one. The second is to make an analysis comparing the populations two by two. We assume that we have obtained the following results from the comparisons in twos: $i > j$ indicates that the population i is preferable to j, or rather that the point is closer to the mean of population i than to that of j:

$$\begin{aligned} 1 &> 2 \\ 2 &> 3 \\ 4 &> 3 \\ 5 &> 4 \end{aligned}$$

Populations 2, 3 and 4 are rejected (since $1 > 2 > 3$ and $5 > 4$). The unresolved doubt is that of populations 1 and 5. Constructing (from the previous rules) the rule for discriminating between the latter two populations, we assume that

$$5 > 1$$

and we classify in population 5.

When $p < G - 1$ the maximum number of linearly independent projections that we can construct is p, and this will be the maximum number of variables to be defined. For example, we assume that $p = 2$ and $G = 5$. We can define the direction of any projection, for example

$$\mathbf{w}_{12} = \mathbf{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and project all of the means $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_5)$ and the point $\mathbf{x}$ over this direction. Then, we classify the point in the population whose projected mean is closest. However, it is possible

that the projected means of several populations coincide in this direction. If this occurs with $\boldsymbol{\mu}_4$ and $\boldsymbol{\mu}_5$ for example, we can solve the problem by projecting over the direction defined by another pair of populations.

Example: A machine that takes coins carries out three measurements of each coin to determine its value: weight (x_1), thickness (x_2) and the thickness of the grooves on its side (x_3). The instruments used to measure these variables are not too precise and it has been proved in extensive experimenting using three types of coins, M_1, M_2, M_3 , that the measurements are distributed normally with means for each type of coin given by:

$$\begin{array}{rccc} \boldsymbol{\mu}_1 = & 20 & 8 & 8 \\ \boldsymbol{\mu}_2 = & 19.5 & 7.8 & 10 \\ \boldsymbol{\mu}_3 = & 20.5 & 8.3 & 5 \end{array}$$

and covariance matrix

$$V = \begin{bmatrix} 4 & .8 & -5 \\ .8 & .25 & -.9 \\ -5 & -.9 & 9 \end{bmatrix}$$

Indicate how a coin with measurements (22, 8.5 ,7) would be classified, and analyze the classification rule. Calculate the probabilities of error.

Apparently the coin to be classified is closest to M_3 in the first two coordinates, but closer to M_1 for x_3, the thickness of the grooves. The indicator variable for classifying between M_1 and M_3 is

$$z = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} = 1.77x_1 - 3.31x_2 + .98x_3$$

the mean of this variable for the first coin, M_1,is $1.77 \times 20 - 3.31 \times 8 + .98 \times 8$ =16.71 and for the third, M_3, $1.77 \times 20.5 - 3.31 \times 8.3 + .98 \times 5$ =13.65. The cut-off point is the mean, 15.17. Since the mean for the coin to be classified is

$$z = 1.77 \times 22 - 3.31 \times 8.5 + .98 \times 7 = 17.61$$

we classify it as M_1. This analysis is equivalent to calculating the Mahalanobis distances to each population which are $D_1^2 = 1.84$, $D_2^2 = 2.01$ and $D_3^2 = 6.69$. Therefore, we classify first in M_1, then in M_2 and finally as M_3. The rule for classifying between the first and second is

$$z = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\mathbf{V}^{-1}\mathbf{x} = -.93x_1 + 1.74x_2 - .56x_3$$

from these two rules we quickly deduce the rule for classifying between the second and third, since

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\mathbf{V}^{-1}\mathbf{x}$$

Now let's analyze the classification rules we have obtained. We are going to express the initial rules for classifying between M_1 and M_3 for the standardized variables, thus avoiding the problem of units. Letting $\widetilde{x}_i$ be the variables divided by their standard deviations $\widetilde{x}_1 = x_1/2; \widetilde{x}_2 = x_2/.5$, and $\widetilde{x}_3 = x_3/3$, the rule in standardized variables is

$$z = 3.54\widetilde{x}_1 - 1.65\widetilde{x}_2 + 2.94\widetilde{x}_3$$

which indicates that the variables with more weight in classification decisions are the first and the third, which are the ones with larger coefficients. We observe that with standardized variables the covariance matrix is that of the correlation

$$R = \begin{bmatrix} 1 & .8 & -.83 \\ .8 & 1 & -.6 \\ -.83 & -.6 & 1 \end{bmatrix}$$

The origin of these correlations between the measurement errors is that if the coin gets dirty and its weight increases slightly, this also increases its thickness and makes it more difficult to determine the thickness of the grooves. This is why there are positive correlations between weight and thickness, increased weight increases the thickness, but negative correlations with thickness of the grooves. Although the coin we wish to classify has greater weight and thickness, which would indicate that it belongs in class 3, the thickness of the grooves should then be measured as low, since there are negative correlations between both measurements, yet nevertheless it has a relatively high measurement in the coin. The three measurements are consistent with a dirty coin of type 1, and for this reason it is easily classified into this group.

We are going to calculate the posterior probability that the observation belongs in class M_1. Assuming that the prior probabilities are equal the probability will be

$$P(1/x_0) = \frac{\exp(-D_1^2/2)}{\exp(-D_1^2/2) + \exp(-D_2^2/2) + \exp(-D_3^2/2)}$$

and substituting in the Mahalanobis distances

$$P(1/x_0) = \frac{\exp(-1.84/2)}{\exp(-1.84/2) + \exp(-2.01/2) + \exp(-6.69/2)} = .50$$

and analogously $P(2/x_0) = .46$, and $P(3/x_0) = .04$.

We can calculate the probabilities of error of classifying a coin in any other group. For example, the probability of classifying an M_3 coin with this rule as type M_1 is

$$P(z > 15.17/N(13.64, \sqrt{3.07})) = P(y > \frac{15.17 - 13.64}{1.75}) = P(y > .87) = .192$$

and we see that the probability is quite high. If we want to reduce it we have to increase the Mahalanobis distance between the means of the groups, which means "increasing"the matrix $\mathbf{V}^{-1}$ or "reducing"the matrix $\mathbf{V}$. For example, if we reduce the error by half in the measurement of the grooves by introducing more accurate measuring devices, but the correlations with the other measurements are kept the same, we turn to the covariance matrix

$$V_2 = \begin{bmatrix} 4 & .8 & -2.5 \\ .8 & .25 & -.45 \\ -1 & -.2 & 2.25 \end{bmatrix}$$

the classification rule between the first and third is now

$$z = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} = 3.44x_1 - 4.57x_2 + 4.24x_3$$

and the Mahalanobis distance between the populations 1 and 3 (coins M_1 and M_3) has gone from 3.01 to 12.38, which implies that the probability of error between the two populations has decreased to $1 - \Phi(\sqrt{12.38}/2) = 1 - \Phi(1.76) = .04$ and we see that the probability of error has decreased considerably. We can calculate accuracy in this way in the measurements that we would need to come up with determined error probabilities.

13.4 UNKNOWN POPULATIONS: GENERAL CASE

13.4.1 Estimated classification rule

We are going to study how to apply the above theory when, instead of working with two populations, we have samples. We will go directly to the case of G possible populations. As a particular case, the classical discrimination is for $G = 2$. The general matrix of data $\mathbf{X}$, $n \times p$ (n individuals and p variables), can now be thought of as divided into G matrices corresponding to the subpopulations. We will let x_{ijg} be the elements of these submatrices where i represents the individual, j the variable, and g the group or submatrix. We let n_g be the number of elements in group g and the total number of observations is:

$$n = \sum_{g=1}^{G} n_g$$

We are going to let $\mathbf{x}'_{ig}$ be the row vector $(1 \times p)$ which contains the p values of the variables for the individual i in group g, that is, $\mathbf{x}'_{ig} = (x_{i1g}, ..., x_{ipg})$. The vector of means within each class or subpopulation will be:

$$\overline{\mathbf{x}}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x}_{ig} \tag{13.20}$$

and is a column vector of dimension p which contains the p means for the observations of the class g. The covariance matrix for the elements of class g is:

$$\widehat{\mathbf{S}}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)' \tag{13.21}$$

where we have divided by $n_g - 1$ to obtain central estimates of the variances and covariances. If we assume that the G subpopulations have the same covariance matrix, their best central estimation with all the data will be a linear combination of the central estimates of each population weighted proportionally to their precision. Therefore:

$$\widehat{\mathbf{S}}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \widehat{\mathbf{S}}_g$$

and we let $\mathbf{W}$ be the *matrix of the sum of squares within the classes* which are given by:

$$\mathbf{W} = (n - G)\widehat{\mathbf{S}}_w \tag{13.22}$$

In order to obtain the discriminant functions we will use $\overline{\mathbf{x}}_g$ as an estimation of μ_g, and $\widehat{\mathbf{S}}_w$ as an estimation of $\mathbf{V}$. Specifically, assuming that the prior probabilities and the classification costs are equal, we classify an element in the group which leads to a minimum Mahalanobis distance value between point $\mathbf{x}$ and the mean of the group. In other words, letting $\widehat{\mathbf{w}}_g = \widehat{\mathbf{S}}_w^{-1}\overline{\mathbf{x}}_g$ we classify a new element $\mathbf{x}_0$ in that population g where

$$\min_g (\mathbf{x}_0 - \overline{\mathbf{x}}_g)' \widehat{\mathbf{S}}_w^{-1} (\mathbf{x}_0 - \overline{\mathbf{x}}_g) = \min_g \widehat{\mathbf{w}}_g' (\overline{\mathbf{x}}_g - \mathbf{x}_0)$$

which is equivalent to constructing the scalar indicator variables

$$z_{g,g+1} = \widehat{\mathbf{w}}_{g,g+1}' \mathbf{x}_0 \quad g = 1, ..., G$$

where

$$\widehat{\mathbf{w}}_{g,g+1} = \widehat{\mathbf{S}}_w^{-1} (\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_{g+1}) = \widehat{\mathbf{w}}_g - \widehat{\mathbf{w}}_{g+1}$$

and classify in g as opposed to $g+1$ if

$$|z_{g,g+1} - \widehat{m}_g| < |z_{g,g+1} - \widehat{m}_{g+1}|$$

where $\widehat{m}_g = \widehat{\mathbf{w}}_{g,g+1}' \overline{\mathbf{x}}_g$.

It is advisable before constructing the classification rule to carry out a test to see whether the two groups really are different, or rather, that not all of the means $\boldsymbol{\mu}_g$ are equal. This test can be carried out following what was shown in 10.7. In Appendix 13.2 there is a proof that in the case of two groups the function of the linear discrimination $\widehat{w} = \widehat{\mathbf{S}}_\omega^{-1} (\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)$ can be obtained by regression, defining a dummy variable which takes values zero or one according to whether the element belongs to one population or the other.

13.4.2 Calculation of Error Probabilities

The calculation of error probabilities can be done by replacing the unknown parameters with the estimated ones and applying the formulas in section 13.2, but this method is not advisable since it greatly underestimates the error probabilities by not taking into account the uncertainty of the estimation of the parameters. A better procedure, which in addition, does not depend on the hypothesis of normality, is to apply the discriminant function to the n observations and classify them. In the case of two groups, we obtain the table:

		Classified	
		P_1	P_2
Reality	P_1	n_{11}	n_{12}
	P_2	n_{21}	n_{22}

where n_{ij} is the amount of data which, coming from the population i is classified in j. The apparent error in the rule is:

$$\text{Error} = \frac{n_{12} + n_{21}}{n_{11} + n_{22}} = \frac{\text{Total incorrectly classified}}{\text{Total correctly classified}}.$$

This method tends to underestimate the probabilities of error since the same data are used to estimate the parameters and to evaluate the resulting procedure. A better procedure is to classify each element with a rule that was not constructed using it. To do this, we can construct n discriminant functions with the n samples of size $n-1$ which are the result of eliminating each element of the population one by one and later classifying each piece of information with the rule constructed without it. This method is known as *cross validation* and leads to a better estimate of the classification error. If the number of observations is high, the computational cost of the cross validation is high as well; a quicker solution is to subdivide the sample into k equal groups and carry out the cross validation by, instead of eliminating an observation, eliminating one of the groups.

Example: We are going to use the MEDIFIS data to classify people by their gender knowing the body measurements of the variables (medifis.dat file). Since the data for the whole population of men and women are unknown, we are going to work with sample data. In the sample there are 15 women (variable sex=0) and 12 men (sex=1).

In example 10.2 we proved that the means of the populations of body measurements for men and women are different. The discriminant functions $\widehat{\mathbf{w}}_g = \widehat{\mathbf{S}}_w^{-1}\overline{\mathbf{x}}_g$ are shown in the following table.

	ht	wt	ftl	$arml$	bkw	crd	$kn-al$
men	−1.30	−4.4	20.0	10.0	−2.1	24.4	−4.4
women	−1.0	−4.4	17.7	9.5	−2.5	25.1	−4.7
difference	−.3	0	2.3	.5	.4	−.7	.3

The difference between these two functions provides us with a linear discriminant function. We observe that the variable with greatest weight in the discrimination is foot length. To interpret this result, the following table shows the standardized differences between the means of each variable in both populations. For example, the standardized difference between heights is $(177.58-161.73)/6.4 = 2.477$

	ht	wt	ftl	$arml$	bkw	crd	$kn-al$
diff. means	15.8	18.65	4.83	7.72	5.67	1.36	4.56
std. dev.	6.4	8.8	1.5	3.1	2.9	1.7	2.2
stand. diff.	2.47	2.11	3.18	2.48	1.97	.78	2.07

,

The variable that most separates both populations is foot length. Since foot length is also highly correlated with height and arm length, knowing foot length means that these other variables are not as informative, which explains their low weight in the discriminant function.

If we apply the discriminant function to classify sample data our success rate is 100%. All the observations are well classified. Applying the cross validation we get

		Classified	
		M	H
Reality	M	13	2
	H	2	10

which means a proportion of correct decisions of 23/27=0.852. The incorrectly classified observations are 2, 7, 9, and 18. We see that the cross validation method gives us a more realistic idea of the efficiency of the classification procedure.

13.5 CANONICAL DISCRIMINANT VARIABLES

13.5.1 The case of two groups

The linear discriminant function for two groups was obtained for the first time by Fisher using an intuitive reasoning which we will briefly outline. The criterion proposed by Fisher is to find a scalar variable:

$$z = \boldsymbol{\alpha}'\mathbf{x} \tag{13.23}$$

so that it maximizes the distance between the projected means as related to the resulting variability in the projection. Intuitively, the scalar z allows the greatest possible separation between the two groups.

The mean of variable z in group 1, which is the projection of the vector of means over the direction of $\boldsymbol{\alpha}$, is $\widehat{m}_1 = \boldsymbol{\alpha}'\overline{\mathbf{x}}_1$, and the mean of group 2 is $\widehat{m}_2 = \boldsymbol{\alpha}'\overline{\mathbf{x}}_2$. The variance of variable z will be the same in both groups, $\boldsymbol{\alpha}'\mathbf{V}\boldsymbol{\alpha}$, and we will estimate it with $s_z^2 = \boldsymbol{\alpha}' S_w \boldsymbol{\alpha}$. We wish to select $\boldsymbol{\alpha}$ in such a way that the separation between the means m_1 and m_2 is maximum. An adimensional measurement of this separation is:

$$\phi = \left(\frac{\widehat{m}_2 - \widehat{m}_1}{s_z}\right)^2,$$

and this expression is equivalent to:

$$\phi = \frac{(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1))^2}{\boldsymbol{\alpha}' S_w \boldsymbol{\alpha}}. \tag{13.24}$$

In this equation $\boldsymbol{\alpha}$ represents a direction, since ϕ is invariant to multiplications of $\boldsymbol{\alpha}$ by a constant: if $\boldsymbol{\beta} = p\boldsymbol{\alpha}$, $\phi(\boldsymbol{\beta}) = \phi(\boldsymbol{\alpha})$. In order to find the direction $\boldsymbol{\alpha}$ which maximizes ϕ, taking the derivative in (13.24) and setting the result to zero:

$$\frac{d\phi}{d\boldsymbol{\alpha}} = \mathbf{0} = \frac{2\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)'\boldsymbol{\alpha}' S_w \boldsymbol{\alpha} - 2S_w\boldsymbol{\alpha}\,(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1))^2}{(\boldsymbol{\alpha}' S_w \boldsymbol{\alpha})^2}$$

which we write:

$$(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)\boldsymbol{\alpha}'\mathbf{S_w}\boldsymbol{\alpha} = \mathbf{S_w}\boldsymbol{\alpha}\,(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1))$$

or also

$$(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = S_w\boldsymbol{\alpha}\frac{(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1))}{\boldsymbol{\alpha}' S_w \boldsymbol{\alpha}}$$

which results in

$$\boldsymbol{\alpha} = \lambda {S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)$$

where $\lambda = (\boldsymbol{\alpha}'\mathbf{S}_w\boldsymbol{\alpha})/\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)$. Since, given $\boldsymbol{\alpha}$, λ is a constant and the function for optimizing is invariant to constants, we can take $\boldsymbol{\alpha}$, normalizing so that $\lambda = 1$, which gives us:

$$\boldsymbol{\alpha} = \mathbf{S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) \tag{13.25}$$

which is the direction $\mathbf{w}$ of projection that we found in the previous section. Furthermore:

$$\boldsymbol{\alpha}'\mathbf{S}_w\boldsymbol{\alpha} = (\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)'\mathbf{S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = D^2(\overline{\mathbf{x}}_2, \overline{\mathbf{x}}_1) = (\widehat{m}_2 - \widehat{m}_1)^2$$

and the variance of the resulting variable of the projection is the Mahalanobis distance between the means. Also:

$$\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = (\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)'\mathbf{S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = D^2(\overline{\mathbf{x}}_2, \overline{\mathbf{x}}_1)$$

and comparing with (13.24) we see that ϕ is the Mahalanobis distance between the means. This procedure leads to a search for the direction of projection which maximizes the Mahalanobis distance between centers of both populations. We observe that if $\mathbf{S}_w = \mathbf{I}$ the Mahalanobis distance is reduced to that of the Euclidean and the direction of projection is parallel to the vector which joins both means. Finally, we see that this rule was obtained without imposing any hypothesis on the distribution of the variable $\mathbf{x}$ in the populations.

13.5.2 Several Groups

Fisher's approach can be generalized to find canonical variables with maximum discriminant power to classify new elements among G populations. The objective is, instead of working with the p original variables $\mathbf{x}$, to define a vector $\mathbf{z} = (z_1, ..., z_r)'$ of r canonical variables, where $r = min(G-1, p)$, which is obtained as a linear combination of the originals, $z_i = \mathbf{u}_i'\mathbf{x}$, and which allows the classification problem to be solved in the following way:

(1) We project the means of the variables in the groups, $\overline{\mathbf{x}}_g$, over the space determined by the r canonical variables. Let $\overline{\mathbf{z}}_1, ..., \overline{\mathbf{z}}_g$ be the variables $r \times 1$ whose coordinates are these projections.

(2) We project the point $\mathbf{x}_0$ to be classified and let $\mathbf{z}_0$ be its projection over the space defined by the canonical variables.

(3) We classify the point in that population whose mean is closest. The distances are measured using the Euclidean distance in the space of the canonical variables z. This means that we classify in population i if:

$$(\mathbf{z}_0 - \overline{\mathbf{z}}_i)'(\mathbf{z}_0 - \overline{\mathbf{z}}_i) = \min_g (\mathbf{z}_0 - \overline{\mathbf{z}}_g)'(\mathbf{z}_0 - \overline{\mathbf{z}}_g)$$

With several groups the relative separation between the means is measured by the ratio between the variability between groups, or the explained variability by groups, and the variability within the groups, or the unexplained or residual variability. This is the usual criterion for comparing several means in analysis of variance and leads to Fisher's F-test. In order to obtain the canonical discriminant variables we begin by looking for a vector $\mathbf{u}_1'$, of norm one, so that the groups of points projected on to it have a maximum relative separation. The projection of the mean of the observations from group g in this direction will be the scalar variable:

$$\overline{z}_g = \mathbf{u}_1'\overline{\mathbf{x}}_g$$

and the projection of the mean for all of the data will be:

$$z_T = \mathbf{u}_1'\overline{\mathbf{x}}_T$$

where $\overline{\mathbf{x}}_T$ is the vector $p \times 1$ which contains the means of the p variables for the n observations of the sample uniting all of the groups. The total variability among the means of the projected

groups, is given by $\sum_{g=1}^{G} n_g(\overline{z}_g - \overline{z}_T)^2$. Comparing this quantity with the variability between the groups, given by $\sum\sum(z_{ig} - \overline{z}_g)^2$, the relative separation between the means will be given by the statistic:

$$\phi = \frac{\sum n_g(\overline{z}_g - \overline{z}_T)^2}{\sum\sum(z_{ig} - \overline{z}_g)^2}.$$

and if all the data come from the same population and no distinct groups exist, this variable is distributed as an F with $G-1$ and $n-G+1$ degrees of freedom. We are going to express this criterion according to the original data. The sum of squares *within* the groups, or unexplained variability (VNE), for the projected points is:

$$VNE = \sum_{j=1}^{n_g}\sum_{g=1}^{G}(z_{jg} - \overline{z}_g)^2 = \sum_{j=1}^{n_g}\sum_{g=1}^{G}\mathbf{u}'(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)'\mathbf{u} = \mathbf{u}'\mathbf{W}\mathbf{u}$$

where $\mathbf{W}$ is given by

$$\mathbf{W} = \sum_{j=1}^{n_g}\sum_{g=1}^{G}(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)'$$

which coincides with (13.22). This matrix has dimensions $p \times p$ and, in general, will have rank p, assuming $n - G \geq p$. We estimate the variability of the data with respect to their group means, which is the same, as hypothesized, in all of them.

The sum of squares *between* groups, or the explained variability (VE), for the projected points is:

$$\begin{aligned} VE &= \sum_{g=1}^{G} n_g(\overline{z}_g - \overline{z}_T)^2 = \\ &= \sum n_g\mathbf{u}'(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)'\mathbf{u} = \\ &= \mathbf{u}'\mathbf{B}\mathbf{u} \end{aligned} \tag{13.26}$$

where $\mathbf{B}$ is the matrix of the sum of squares between groups, which can be written as:

$$\mathbf{B} = \sum_{g=1}^{G} n_g\mathbf{a}_g\mathbf{a}_g'$$

and $\mathbf{a}_g = \overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T$. The $p \times p$ matrix $\mathbf{B}$ is square and symmetric, and is obtained as the sum of G matrices of rank one formed by the vectors $\mathbf{a}_g$, which are not independent since they are linked by the equation $\sum_{g=1}^{G} n_g\mathbf{a}_g = \mathbf{0}$. This implies that the rank of $\mathbf{B}$ will be $G-1$.

To summarize, the matrix $\mathbf{W}$ measures the differences within the groups and $\mathbf{B}$ the differences between groups. The quantity to be maximized can also be written as:

$$\phi = \frac{\mathbf{u}_1'\mathbf{B}\mathbf{u}_1}{\mathbf{u}_1'\mathbf{W}\mathbf{u}_1}, \tag{13.27}$$

taking the derivative and setting to zero in the usual way:

$$\frac{d\phi}{du_1} = 0 = \frac{2\mathbf{B}\mathbf{u}_1(\mathbf{u}_1'\mathbf{W}\mathbf{u}_1) - 2(\mathbf{u}_1'\mathbf{B}\mathbf{u}_1)\mathbf{W}\mathbf{u}_1}{(\mathbf{u}_1'\mathbf{W}\mathbf{u}_1)^2} = 0$$

thus:

$$\mathbf{B}\mathbf{u}_1 = \mathbf{W}\mathbf{u}_1 \left(\frac{\mathbf{u}_1' \mathbf{B}\mathbf{u}_1}{\mathbf{u}_1' \mathbf{W}\mathbf{u}_1} \right)$$

that is, using

$$\mathbf{B}\mathbf{u}_1 = \phi \mathbf{W}\mathbf{u}_1$$

and assuming $\mathbf{W}$ is non-singular:

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{u}_1 = \phi \mathbf{u}_1$$

which implies that $\mathbf{u}_1$ has to be an eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ and thus, ϕ is the associated eigenvalue. Since we want to maximize ϕ, which is the value of the F-test in a scalar test on projected means, $\mathbf{u}$ will be the eigenvector associated with the largest eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$.

We can consider obtaining a second axis which maximizes the partition ϕ, but with the condition that the new canonical variable $z_2 = \mathbf{u}_2'\mathbf{x}$ is uncorrelated with the first, $z_1 = \mathbf{u}_1'\mathbf{x}$. It can be shown that the same thing happens if we take the second eigenvector (linked to the second eigenvalue) of the matrix $\mathbf{W}^{-1}\mathbf{B}$. In general, $\alpha_1, \ldots, \alpha_r$ are the non-null eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are the eigenvectors linked to the non-null eigenvalues. The scalar variables $z_j = \mathbf{u}_j'\mathbf{x}$ provide maximum separation just as the F-test for testing whether there are differences between the G projected groups. Furthermore, these scalar variables z_j are uncorrelated both within the groups as well as in the sample as a whole. In order to prove this, let $\mathbf{z}_j$ be the vector $n \times 1$ resulting from projecting the sampling points in the direction $\mathbf{u}_j'$, or rather, $\mathbf{z}_j = \mathbf{X}\mathbf{u}_j$. The mean of this variable is $\overline{z_j} = \mathbf{1}'\mathbf{z}_j/n = \mathbf{1}'\mathbf{X}\mathbf{u}_j/n = \overline{\mathbf{x}}_T'\mathbf{u}_j$ and the covariance between the two scalar variables, z_j and z_h will be given by

$$cov(z_j, z_h) = \frac{1}{n}\sum_{i=1}^{n}(z_{ji} - \overline{z}_j)(z_{hi} - \overline{z}_h) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}_j'(\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)'\mathbf{u}_h$$

and letting $\mathbf{T}$ be the matrix of the total sum of squares, the covariances between the canonical variables are $\mathbf{u}_j'\mathbf{T}\mathbf{u}_h$. If we decompose these variables into groups in such a way that each variable $\mathbf{z}_j$ produces G variables z_{jg} where g indicates the group, it can be proven analogously that the covariances between z_{jg} and z_{hg}, added up for all the groups are given by $\mathbf{u}_j'\mathbf{W}\mathbf{u}_h$. We are going to show that, for two different eigenvectors, $h \neq j$:

$$\mathbf{u}_h'\mathbf{W}\mathbf{u}_j = \mathbf{u}_h'\mathbf{T}\mathbf{u}_j = 0,$$

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

In order to prove this property, let us assume that $\alpha_h > \alpha_j$. The eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ verify that

$$(\mathbf{W}^{-1}\mathbf{B})\mathbf{u}_h = \alpha_h \mathbf{u}_h$$

or rather

$$\mathbf{B}\mathbf{u}_h = \alpha_h \mathbf{W}\mathbf{u}_h. \tag{13.28}$$

Therefore, for another, different eigenvector $\mathbf{u}_j$, where $\alpha_h \neq \alpha_j$, we have:

$$\mathbf{B}\mathbf{u}_j = \alpha_j \mathbf{W}\mathbf{u}_j \tag{13.29}$$

multiplying (13.28) by $\mathbf{u}'_j$ and (13.29) by $\mathbf{u}'_h$:

$$\begin{aligned} \mathbf{u}'_j\mathbf{B}\mathbf{u}_h &= \alpha_h \mathbf{u}'_j\mathbf{W}\mathbf{u}_h \\ \mathbf{u}'_h\mathbf{B}\mathbf{u}_j &= \alpha_j \mathbf{u}'_h\mathbf{W}\mathbf{u}_j \end{aligned}$$

Since the first members are equal, the second ones must be so as well, and as $\alpha_h \neq \alpha_j$, the only possibility is $\mathbf{u}'_j\mathbf{W}\mathbf{u}_h = 0 = \mathbf{u}'_j\mathbf{B}\mathbf{u}_h = \mathbf{u}'_j\mathbf{T}\mathbf{u}_h$.

We can see that the eigenvectors of the matrix $\mathbf{W}^{-1}\mathbf{B}$ are *not,* in general, orthogonal because despite the fact that the matrices $\mathbf{W}^{-1}$ and $\mathbf{B}$ are symmetric, their product is not necessarily so. Furthermore, the rank of this matrix $\mathbf{W}^{-1}\mathbf{B}$, will be $r = \min(p, G-1)$, (remember that the rank of the product of two matrices is lesser than or equal to that of the originals) and this is the maximum number of discriminant factors we can obtain.

The matrix $\mathbf{W}^{-1}\mathbf{B}$ has been called by Rao the matrix of generalized Mahalanobis distance, since its trace is the sum of the Mahalanobis distances between the mean of each group and the total mean. We have

$$tr(\mathbf{W}^{-1}\mathbf{B}) = \mathbf{tr}\sum(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)'(\mathbf{W}/n_g)^{-1}(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)$$

13.5.3 Canonical discriminant variables

This procedure provides $r = min(p, G-1)$ canonical discriminant variables which are given by

$$\mathbf{z} = \mathbf{U}'_r\mathbf{x} \tag{13.30}$$

where $\mathbf{U}_r$ is a $p \times r$ matrix whose columns contain the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ and $\mathbf{x}$ a $p \times 1$ vector. The $r \times 1$ vector, $\mathbf{z}$, contains the values of the canonical variables for the item $\mathbf{x}$, which are the coordinates of the point in the space defined by the canonical variables.

The canonical variables obtained in this way solve the classification problem. In order to classify a new individual $\mathbf{x}_0$ it is only necessary to calculate its coordinates $\mathbf{z}_0$ with (13.30) and assign it to the group whose transformed mean is closest with the Euclidean distance.

A significant problem is that of finding out how many dimensions we need for the discrimination, since it is possible that most of the separating capacity of the populations is achieved with the first two canonical variables. To study this problem we assume that instead of taking the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ with unit norm $\mathbf{u}_i$, we standardize them using $\mathbf{v}_i = \mathbf{u}_i / \left|\mathbf{u}'_i\mathbf{W}\mathbf{u}_i\right|^{1/2}$ in such a way that these vectors $\mathbf{v}_i$ are still eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ but will now verify $\mathbf{v}'_i\mathbf{W}\mathbf{v}_i = 1$. Then, the variability explained by the canonical variable $\mathbf{v}_i$ is, because of (13.26),

$$VE(\mathbf{v}_i) = \mathbf{v}'_i\mathbf{B}\mathbf{v}_i$$

but $\mathbf{v}_i$ being an eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ verifies

$$\mathbf{B}\mathbf{v}_i = \alpha_i\mathbf{W}\mathbf{v}_i$$

and multiplying by $\mathbf{v}_i'$ and taking into account that by construction $\mathbf{v}_i'\mathbf{W}\mathbf{v}_i = 1$:

$$VE(\mathbf{v}_i) = \mathbf{v}_i'\mathbf{B}\mathbf{v}_i = \alpha_i,$$

which indicates that the variability explained by the canonical variable $\mathbf{v}_i$ is equal to its associated eigenvalue. Therefore, the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ standardized so that $\mathbf{v}_i'\mathbf{W}\mathbf{v}_i = 1$ tell us the explained variability which each canonical variable contributes to the discrimination problem. When p and G are large it often happens that greater discriminatory capacity is achieved with a few canonical variables.

The results of classification which are obtained with canonical variables are identical to those obtained with the Mahalanobis distance (see Hernández and Velilla, 2000, for a complete study of this problem). It is easy to prove if $G = 2$, in the case of two populations, or when the means are colinear. In both cases the matrix $\mathbf{B}$ has rank one and the eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ linked to the non-null eigenvalue automatically provides Fisher's linear discriminant function. In order to prove it we have only to note that if $G = 2$ the matrix $\mathbf{B}$ is:

$$\mathbf{B} = \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'$$

and the eigenvector associated with the non-null eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ is $\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$, which we obtained previously. If the means of the G populations are in a straight line, then,

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T) = p_1(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_T) = ... = k_j(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T) = c(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

and the matrix $\mathbf{B}$ can be written

$$\mathbf{B} = \sum_{g=1}^{G}(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)' = k^*(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'$$

and its eigenvector associated with the non-null eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ is proportional to $\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

Example: We are going to study geographic discrimination among the countries of the world in the MUNDODES data bank. The 91 countries included have been classified a priori as being from Eastern Europe (9 countries, code 1), Central and South America (12 countries, code 2), Western Europe plus Canada and the US (18 countries, code 3), Asia (25 countries, code 4) and Africa (27 countries, code 5). The GNP has been expressed in Napierian logarithms, in accordance with the descriptive results we obtained in Chapter 3.

The output of the SPSS program is shown for the multiple discrimination that provides the results of the discriminant analysis using canonical variables.

The averages of the five groups in each variable are:

C	BR	MR	IMR	LEM	LEW	GDP
1	15.15	10.52	18.14	67.35	74.94	7.48
2	29.17	9.41	51.32	62.70	68.53	7.25
3	13.01	9.58	8.04	71.25	77.97	9.73
4	30.31	8.07	56.49	63.08	65.86	7.46
5	44.52	14.62	99.79	50.63	54.14	6.19

Total 29.46 10.73 55.28 61.38 66.03 7.51
and the total column indicates the means for the set of data.
The standard deviations in the groups are:

G	BR	MR	IMR	LEM	LEW	GDP
1	3.97	2.16	6.97	2.22	1.50	.48
2	7.38	5.51	31.69	4.92	5.31	.66
3	1.85	1.37	1.734	2.51	2.18	.44
4	10.01	3.77	46.02	7.92	9.73	1.69
5	5.68	4.79	30.58	7.09	7.03	1.04
Total	13.69	4.68	46.30	9.72	11.13	1.64

and the matrix W with 86 degrees of freedom is

	BR	MR	IMR	LEM	LEW	GDP
BR	46.90					
MR	11.89	15.63				
IMR	139.41	87.49	1007.64			
LEM	-27.42	-18.76	-169.71	37.55		
LEW	-31.88	-21.08	-194.29	40.52	46.20	
GDP	-3.54	-2.18	-22.42	4.60	5.43	1.25

which we can express with the correlation matrix:

	BR	MR	IMR	LEM	LEW	GDP
BR	1.00000					
MR	.43930	1.00000				
IMR	.64128	.69719	1.00000			
LEM	-.65345	-.77451	-.87247	1.00000		
LEW	-.68487	-.78452	-.90052	.97278	1.00000	
GDP	-.46341	-.49350	-.63275	.67245	.71588	1.00000

The linear classification functions for each group are:

G	1	2	3	4	5
BR	3.4340	3.7363	3.3751	3.6194	3.9314
MR	9.7586	9.1856	9.6773	8.6879	8.9848
IMR	1.7345	1.7511	1.7387	1.7107	1.6772
LEM	-.1319	.28153	.7638	1.7363	.59934
LEW	16.962	16.3425	15.780	14.347	15.342
GDP	-9.422	-8.2661	-5.999	-6.703	-7.053
(Constant)	-690.658	-683.135	-690.227	-642.071	-647.1495

and the eigenvalues of $W^{-1}B$ and the proportion of explained variability are:

Fcn	Val.	pr.	Var. %	
1*	3.9309	69.33	69.33	
2*	1.1706	20.65	89.97	
3*	.4885	8.62	98.59	
4*	.0802	1.41	100.00	

We observe that the first discriminant function or canonical variable, defined by the first eigenvector of the matrix $\mathbf{W}^{-1}\mathbf{B}$ explains 69% of the variability and that the first two together explain 89.97%.

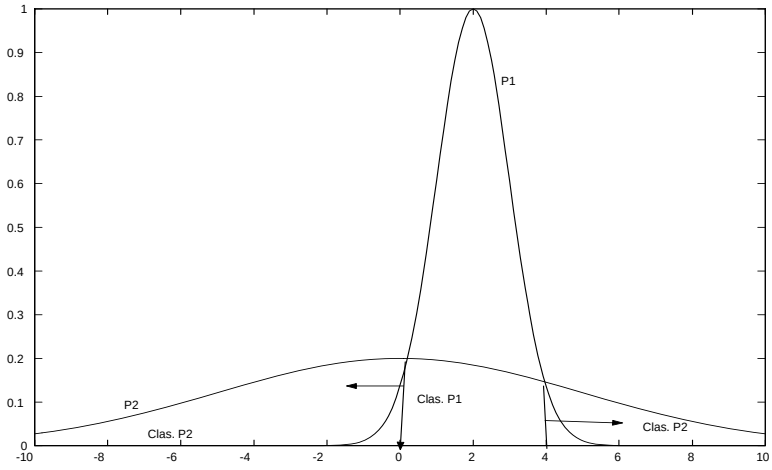

Figura 13.5: Graph of the projection of points over the first two canonical variables.

The coefficients of the canonical variables indicate that the most important variables are globally, life expectancy for women and birth rate.

In the graph we can see the projections of the countries over the first two canonical variables. The results of the classification with the 4 canonical variables are summarized in the following table, where r represents the true classification and p the predictions of the model.

	$p1$	$p2$	$p3$	$p4$	$p5$
$r1$	8	1			
$r2$	1	9	1		1
$r3$			18		
$r4$		2	2	19	2
$r5$		1		4	22

We observe that the European countries are well classified and that it is among the Asian countries where more variability appears.

In this case, the apparent errors obtained by classifying with the Mahalanobis distance (without cross-validation) and using the canonical variables are the same. The attached output from MINITAB includes the basic information. First, the results of classifying using discriminant functions are presented:

	True Group				
Group	1	2	3	4	5
1	8	1	0	0	0
2	1	9	0	2	1
3	0	1	18	2	0
4	0	0	0	19	4
5	0	1	0	2	22
N Total	9	12	18	25	27
N Correct	8	9	18	19	22
Propor.	0.89	0.75	1.0	0.76	0.82

N = 91 N Correct = 76 Proportion Correct = 0.835

and next the result of applying cross-validation:

Placed in True group

Group	1	2	3	4	5
1	8	1	1	0	0
2	1	8	0	4	1
3	0	1	17	2	0
4	0	1	0	16	5
5	0	1	0	3	21
N. Total	9	12	18	25	27
N Correct	8	8	17	16	21
Propor.	0.89	0.67	0.94	0.64	0.78

N = 91 N Correct = 70 Propor. Correct = 0.769

discriminant linear functions

	1	2	3	4	5
Con.	-689.05	-681.53	-688.62	-640.46	-645.54
C2	3.43	3.74	3.38	3.62	3.93
C3	9.76	9.19	9.68	8.69	8.98
C4	1.73	1.75	1.74	1.71	1.68
C5	-0.13	0.28	0.76	1.74	0.60
C6	16.96	16.34	15.78	14.35	15.34
C9	-9.42	-8.27	-6.00	-6.70	-7.05

The following table summarizes the results with cross-validation.

	$p1$	$p2$	$p3$	$p4$	$p5$
$r1$	8	1			
$r2$	1	8	1	1	1
$r3$	1		17		
$r4$		4	2	16	3
$r5$		1		5	21

Finally, the matrix of distances between the means of the groups with the Mahalanobis distance is

	$EO(1)$	$AL(2)$	$E(3)$	$AS(4)$	$AF(5)$
$EO(1)$		7.2	7.8	20.3	25.2
$AL(2)$			10.9	6.5	7.6
$E(3)$				15.4	30.0
$AS(4)$					1.9
$AF(5)$					

It is observed that the greatest distance appears between group E (which includes Western European countries plus Canada and the US) and Africa. The second is between Western Europe (EO) and Africa. The smallest distance is between Asia and Africa.

13.6 QUADRATIC DISCRIMINATION. DISCRIMINATION OF NON-NORMAL POPULATIONS

If by admitting the normality of the observations the hypothesis of equality of variances were not admissible, the procedure for solving the problem is to classify the observation in the group with maximum posterior probabilities. This is equivalent to classifying the observation $\mathbf{x}_0$ in the group where the function is minimized:

$$\min_{j\in(1,...,G)}\left[\frac{1}{2}\log|\mathbf{V}_j|+\frac{1}{2}(\mathbf{x}_0-\boldsymbol{\mu}_j)'\mathbf{V}_j^{-1}(\mathbf{x}_0-\boldsymbol{\mu}_j)-\ln(C_j\pi_j)\right]$$

When $\mathbf{V}_j$ and $\boldsymbol{\mu}_j$ are unknown we estimate by $\mathbf{S}_j$ and $\overline{\mathbf{x}}_j$ in the usual way. Now the term $\mathbf{x}_0{}'\mathbf{V}_j^{-1}\mathbf{x}_0$ cannot be annulled, since it depends on the group, and the determinant functions are not linear and will have a second degree term. Assuming that the classification costs are equal in all of the groups, we will classify the new observations using the rule:

$$\min_{j\in(1,...,G)}\left[\frac{1}{2}\log|\widehat{\mathbf{V}}_j|+\frac{1}{2}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_j)'\widehat{\mathbf{V}}_j^{-1}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_j)-\ln\pi_j\right]$$

In the particular case of two populations and assuming the same prior probabilities, we will classify a new observation in population 2 if

$$\log|\widehat{\mathbf{V}}_1|+(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_1)'\widehat{\mathbf{V}}_1^{-1}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_1)>\log|\widehat{\mathbf{V}}_2|+(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_2)'\widehat{\mathbf{V}}_2^{-1}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_2)$$

which is equivalent to

$$\mathbf{x}_0'(\widehat{\mathbf{V}}_1^{-1}-\widehat{\mathbf{V}}_2^{-1})\mathbf{x}_0-2\mathbf{x}_0'(\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1-\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2)>c \tag{13.31}$$

where $c=\log(|\widehat{\mathbf{V}}_2|/|\widehat{\mathbf{V}}_1|)+\widehat{\boldsymbol{\mu}}_2{}'\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2-\widehat{\boldsymbol{\mu}}_1{}'\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1$. Letting

$$\widehat{\mathbf{V}}_d^{-1}=(\widehat{\mathbf{V}}_1^{-1}-\widehat{\mathbf{V}}_2^{-1})$$

and

$$\widehat{\boldsymbol{\mu}}_d=\widehat{\mathbf{V}}_d(\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1-\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2)$$

and defining the new variables

$$\mathbf{z}_0=\widehat{\mathbf{V}}_d^{-1/2}\mathbf{x}_0$$

and letting $\mathbf{z}_0=(z_{01},...,z_{0p)'}$ and defining the vector $\mathbf{m}=(m_1,...,m_p)'=\widehat{\mathbf{V}}_d^{1/2}(\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1-\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2)$, equation (13.31) can be written

$$\sum_{i=1}^{p}z_{0i}^2-2\sum_{i=1}^{p}z_{0i}m_i>c$$

This is a second degree equation in the new variables z_{0i}. The resulting regions using these second degree functions are typically disjointed and sometimes difficult to interpret in several dimensions. For example, Figure (13.6) shows a unidimensional example of the type of regions which are obtained with quadratic discrimination.

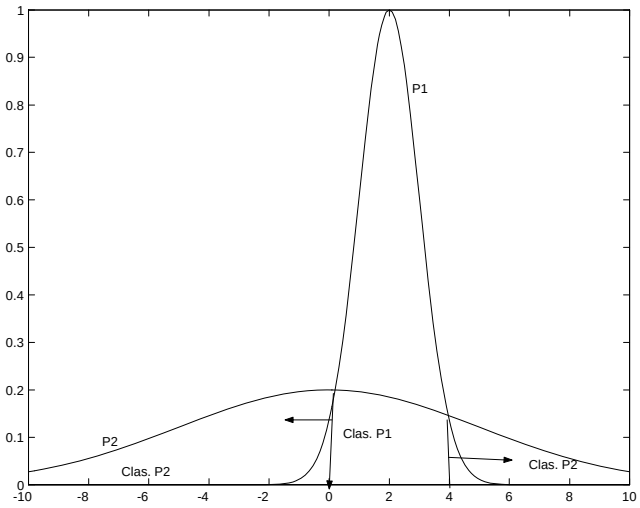

Figura 13.6: Example of quadratic discrimination. The classification zone of P1 is the central zone and that of P2 the tails.

The number of parameters to estimate in the quadratic case is much greater than in the linear. In the linear case we have to estimate $Gp+p(p+1)/2$ and in the quadratic case, $G(p+p(p+1)/2)$. For example with 10 variables and 4 groups we go from estimating 95 parameters in the linear case to 260 in the quadratic. This large number of parameters makes it so, except for cases with very large samples, the quadratic discrimination is relatively unstable and, although the covariance matrices are very different, we frequently obtain better results using the linear function than the quadratic. An additional problem with the quadratic discriminant function is that it is very sensitive to deviations from normality in the data. The available evidence indicates that linear classification is more robust in these cases. We recommend always calculating the classification errors with both rules using cross-validation and in case of very small differences, go with the linear.

A problem of quadratic discrimination also appears in the analysis of determined non-normal populations (See Lachenbruch (1975)). In the general case of arbitrary populations we have two alternatives: (a) apply the general theory presented in 13.2 and obtain the discriminant function which can be complicated, (b) apply the theory of normal populations, take the Mahalanobis distance as a measure of distance and classify $\mathbf{x}$ in population P_j so that for $\mathbf{D}^2$:

$$\mathbf{D}^2 = (\mathbf{x} - \overline{\mathbf{x}}_j)' \widehat{\mathbf{V}}_j^{-1} (\mathbf{x} - \overline{\mathbf{x}}_j)$$

is minimum.

For discrete populations these approximations are not good. Alternative methods have been proposed based on multinomial distribution or the χ^2 distance whose efficiency is yet to be determined.

If we apply the quadratic discrimination to the body measurements data we obtain the table of classification errors by cross-validation (without applying cross-validation it is 100% accurate as in the linear case)

		Classified	
		M	H
True	M	11	4
	H	5	7

which assumes an accuracy percentage of 67%, less than in the linear case. There is no evidence that quadratic discrimination provides any advantages in this case.

13.7 BAYESIAN DISCRIMINATION

We saw in section 13.2 that the Bayesian approach allows us a general solution to the classification problem when the parameters are known. When the parameters must be estimated from the data, the Bayesian approach also provides a direct solution which takes into account the uncertainty in the estimation of the parameters, unlike the classical approach which ignores this uncertainty. The solution is valid whether or not the covariance matrices are equal. The procedure for classifying an observation, $\mathbf{x}_0$, given the training sample, $\mathbf{X}$, is to assign it to the most probable population. To do this, we obtain the maximum of the posterior probabilities that the observation to be classified, $\mathbf{x}_0$, comes from each of the populations given the sample $\mathbf{X}$. These probabilities are calculated by

$$P(i/\mathbf{x}_0, \mathbf{X}) = \frac{f_i(\mathbf{x}_0|\mathbf{X})\boldsymbol{\pi}_i}{\sum_{g=1}^{G} f_g(\mathbf{x}|\mathbf{X})\boldsymbol{\pi}_j}$$

where the densities $f_g(\mathbf{x}_0|\mathbf{X})$, called posterior predictives, or simply predictives, are proportional to the probabilities that the observation $\mathbf{x}_0$ is generated by population g. These densities are generated from the likelihood averaging over the possible values of the parameters in each population with its posterior distribution:

$$f_g(\mathbf{x_0}|\mathbf{X}) = \int f(\mathbf{x_0}|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g|\mathbf{X})\mathbf{d}\boldsymbol{\theta}_g \tag{13.32}$$

where $\boldsymbol{\theta}_g$ are the parameters of population g.

We are going to study how to obtain these probabilities. First, the posterior distribution of the parameters is calculated in the usual way using

$$p(\boldsymbol{\theta}_g|\mathbf{X}) = kf(\mathbf{X}|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g).$$

As we saw in section 9.2.2 the likelihood for the population g with n_g sample elements and sample mean $\overline{\mathbf{x}}_g$ and variance S_g is

$$f(\mathbf{X}|\boldsymbol{\theta}_g) = k\left|\mathbf{V}_g^{-1}\right|^{n_g/2} \exp(-\frac{n_g}{2} tr(\mathbf{V}_g^{-1}\left\{S_g + (\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)(\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)'\right\})$$

and with the reference prior

$$p(\boldsymbol{\mu}_g, \mathbf{V}_g^{-1}) = k\left|\mathbf{V}_g^{-1}\right|^{-(p+1)/2}$$

we obtain the posterior

$$p(\boldsymbol{\mu}_g, \mathbf{V}_g^{-1}/\mathbf{X}) = k \left|\mathbf{V}_g^{-1}\right|^{(n_g-p-1)/2} \exp(-\frac{n_g}{2} tr(\mathbf{V}_g^{-1} \left\{S_g + (\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)(\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)'\right\}$$

The predictive distribution is obtained with (13.32), where now $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_g, \mathbf{V}_g^{-1})$. Integrating with respect to these parameters it can be obtained (see Press, 1989, for details of the integration), that the predictive distribution is multivariate t

$$p(\mathbf{x}_0/\mathbf{X}, \mathbf{g}) = \left[\frac{n\pi(n_g+1)}{n_g}\right]^{-p/2} \frac{\Gamma\left(\frac{n_g}{2}\right)}{\Gamma\left(\frac{n_g-p}{2}\right)} |\mathbf{S}_g|^{-1/2} \left[1 + \frac{1}{n_g+1}(\mathbf{x}_0 - \overline{\mathbf{x}}_g)'\mathbf{S}_g^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_g)\right]^{-n_g/2}$$

With this distribution we can calculate the posterior probabilities for each population. Alternatively, to decide between population i and j we can calculate the ratio of posterior probabilities, given by:

$$\frac{P(i|\mathbf{x})}{P(j|\mathbf{x})} = c_{ij}\frac{\pi_i}{\pi_j} \cdot \frac{|\mathbf{S}_j|^{1/2}}{|\mathbf{S}_i|^{1/2}} \frac{\left(1 + \frac{1}{n_j+1}(\mathbf{x}_0 - \overline{\mathbf{x}}_j)'\mathbf{S}_j^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_j)\right)^{n_j/2}}{\left(1 + \frac{1}{n_i+1}(\mathbf{x}_0 - \overline{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_i)\right)^{n_i/2}}$$

where π_i are the prior probabilities, $\mathbf{S}_j$ the estimated covariance matrices, and

$$c_{ij} = \left[\frac{n_i(n_j+1)}{n_j(n_i+1)}\right]^{p/2} \frac{\Gamma\left(\frac{n_i}{2}\right)\Gamma\left(\frac{n_j-p}{2}\right)}{\Gamma\left(\frac{n_j}{2}\right)\Gamma\left(\frac{n_i-p}{2}\right)}$$

If the sample sizes are approximately equal, $n_i \simeq n_j$, then $c_{ij} \simeq 1$. The optimal classifier is quadratic. If we suppose that the covariance matrices of the groups are equal we again obtain the discriminant linear function (see Aitchinson and Dunsmore, 1975).

13.8 Additional Reading

The classical discriminant analysis presented here can be found in all multivariate analysis textbooks. Presentations of a similar level to that in this book can be found in Cuadras (1991), Flury (1997), Johnson and Wichern (1998), Mardia et al (1979), Rechner (1998) and Seber (1984). A very detailed basic textbook and one with many extensions and references is McLachlan (1992). Lachenbruch, (1975) contains many historic references. More applied approaches, centered in the analysis of examples and computer output, are presented in Huberty (1994), Hair el al (1999) and Tabachnick and Fidell (1996). Hernandez and Velilla (2001) study dimensionality reduction. A Bayesian approach to the problem of classification can be found in Press (1989).

Exercises

13.1 Suppose that we wish to discriminate between two normal populations with mean vectors (0,0) and (1,1) , variances (2,4) and linear correlation coefficient r=.8. Build the discriminant linear function and interpret it.

13.2 Discuss how the probabilities of error in the above problem vary as a function of the correlation coefficient. Does the correlation help the discrimination?

13.3 The prior probabilities in problem 13.1 are 0.7 for the first population and 0.3 for the second. Calculate the discriminant linear function in this case.

13.4 We wish to discriminate between three normal populations with mean vectors (0,0), (1,1) and (0,1) with variances (2,4) and linear correlation coefficient r =.5. Calculate and plot the discriminant functions and find their cut-off point.

13.5 If the costs of being mistaken in the above problem are not the same, such that the cost of classifying something in the third population when it comes from the first is twice the cost of the others, calculate the discriminant functions.

13.6 Justify that the eigenvalues of $W^{-1}B$ are positive, proving that this matrix has the same eigenvalues as the matrix $W^{-1/2}BW^{-1/2}$.

13.7 Justify that the same discriminant canonical variables are obtained using the matrices W and B, as with the associated variance matrices corrected by degrees of freedom.

13.8 Prove that it is the same to obtain the largest eigenvector $W^{-1}B$ and the smallest of $T^{-1}W$.

13.9 Prove that the first principal component when there are two groups is given by $\mathbf{v} = c(\mathbf{W} - \boldsymbol{\lambda}\mathbf{I})^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ (suggestion: If $\mathbf{T} = \mathbf{W} + \mathbf{B}$, the first component is the largest eigenvector (associated with the largest eigenvalue) of $\mathbf{T}$ and verifies $\mathbf{T}\ \mathbf{v} = \mathbf{W}\mathbf{v} + \mathbf{B}\mathbf{v} = \lambda\mathbf{v}$. Since $\mathbf{B} = \mathbf{k}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'$, we have $\mathbf{W}\mathbf{v} + c(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = \lambda\mathbf{v}$).

13.10 Prove that if $(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ is an eigenvector of $\mathbf{W}^{-1}$ the discriminant direction is the natural axis of distance between the means and coincides with the first principal component.

13.11 Prove that the Mahalanobis distance is invariant to linear transformations showing that if $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, with $\mathbf{A}$ squared and non-singular, it is verified that $D^2(y_i, y_j) = D^2(x_i, x_j)$. (Suggestion: use $\mathbf{V}_y = \mathbf{A}\mathbf{V}_\mathbf{x}\mathbf{A}'$ and $\mathbf{V}_\mathbf{y}^{-\mathbf{1}} = (\mathbf{A}')^{-1}\mathbf{V}_\mathbf{x}^{-\mathbf{1}}\mathbf{A}'$)

APPENDIX 13.1: THE CRITERION OF MINIMIZING THE PROBABILITY OF ERROR

The criterion of minimizing the probability of error can be written as minimizing P_T, where:

$$P_T(error) = P(1|\mathbf{x} \in 2) + P(2|\mathbf{x} \in 1)$$

with $P(i|\mathbf{x} \in j)$ being the probability of classifying and observation coming from j in population i. This probability is given by the area enclosed by the distribution j in the classification zone of i, that is:

$$P(i|\mathbf{x} \in j) = \int_{A_i} f_j(\mathbf{x})d\mathbf{x}$$

therefore:

$$P_T = \int_{A_1} f_2(\mathbf{x})d\mathbf{x} + \int_{A_2} f_1(\mathbf{x})d\mathbf{x}$$

and since A_1 and A_2 are complementary:

$$\int_{A_1} f_2(\mathbf{x})d\mathbf{x} = \mathbf{1} - \int_{A_2} f_2(\mathbf{x})d\mathbf{x}$$

which leads to:

$$P_T = 1 - \int_{A_2} (f_2(\mathbf{x}) - f_1(\mathbf{x}))d\mathbf{x}$$

and to minimize the probability of error we must maximize the integral. This is done by defining A_2 as a set of points where the integrand is positive, that is:

$$A_2 = \{\mathbf{x}|f_2(\mathbf{x}) > f_1(\mathbf{x})\}$$

and we again obtain the criterion established earlier.

APPENDIX 13.2: DISCRIMINATION AND REGRESSION

An interesting result is that the construction of a discriminant function in the case of two populations can be approached as a problem of regression.

Let us consider the n observations as data in a linear model and we define a response variable and it takes the value $+k_1$, when $\mathbf{x} \in P_1$ and $-k_2$, when $\mathbf{x} \in P_2$. We can assign any values to k_1 and k_2 although, as we will see, the computations are simplified if we make these constants equal to the number of elements of the sample in each class. The model will be:

$$y_i = \beta_1(x_{1i} - \overline{x}_1) + \beta_2(x_{2i} - \overline{x}_2) + ... + \beta_p(x_{pi} - \overline{x}_p) + u_i \qquad i = 1, 2 \tag{13.33}$$

where we have expressed the x in deviations. The least squares estimator is:

$$\widehat{\beta} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Y} \tag{13.34}$$

where $\widetilde{\mathbf{X}}$ is the matrix of the data in deviations.

Let $\overline{\mathbf{x}}_1$ be the vector of means in the first group, $\overline{\mathbf{x}}_2$ in the second, and $\overline{\mathbf{x}}_T$ that which corresponds to all the observations. We suppose that in the sample there are n_1 data points from the first group and n_2 from the second. Then,

$$\overline{\mathbf{x}}_T = \frac{n_1\overline{\mathbf{x}}_1 + n_2\overline{\mathbf{x}}_2}{n_1 + n_2}. \tag{13.35}$$

Substituting (13.35) in the first term of (13.34):

$$\begin{aligned}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} &= \sum_{i=1}^{n_1+n_2} (\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)' = \\ &= \sum_{i=1}^{n_1}(\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)' + \sum_{i=1+n_1}^{n_1+n_2} (\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)'.\end{aligned}$$

Since

$$\begin{aligned}\sum_{i=1}^{n_1}(\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)' &= \sum_{i=1}^{n_1}(\mathbf{x}_i - \overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)' = \\ &= \sum_{i=1}^{n_1}(\mathbf{x}_i - \overline{\mathbf{x}}_1)(\mathbf{x}_i - \overline{\mathbf{x}}_1)' + n_1(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)'\end{aligned}$$

because the crossed terms are cancelled out by the fact that $\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}}_1) = 0$. Proceeding analogously for the other group, we can write:

$$\begin{aligned}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} &= \sum_{i=1}^{n_1}(\mathbf{x}_i - \overline{\mathbf{x}}_1)(\mathbf{x}_i - \overline{\mathbf{x}}_1)' + \sum_{i=n_1+1}^{n_1+n_2}(\mathbf{x}_i - \overline{\mathbf{x}}_2)(\mathbf{x}_i - \overline{\mathbf{x}}_2)' \\ &+ n_1(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)' + n_2(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_T)'\end{aligned} \tag{13.36}$$

The first two terms lead to the matrix $\mathbf{W}$ of sum of squares within groups which, as we have seen estimates $\mathbf{V}$, using:

$$\widehat{\mathbf{V}} = \mathbf{S} = \frac{1}{n_1 + n_2 - 2}\mathbf{W}. \tag{13.37}$$

The second two terms are the sums of squares between groups. Replacing $\overline{\mathbf{x}}_T$ with (13.35):

$$\overline{\mathbf{x}}_1 - \frac{n_1\overline{\mathbf{x}}_1 + n_2\overline{\mathbf{x}}_2}{n_1 + n_2} = \frac{1}{n_1 + n_2}n_2(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2), \tag{13.38}$$

which results in:

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T)' = \left(\frac{n_2}{n_1 + n_2}\right)^2(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \tag{13.39}$$

$$(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_T)' = \left(\frac{n_1}{n_1 + n_2}\right)^2(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)' \tag{13.40}$$

Substituting (13.39) and (13.40) in (13.36) gives us:

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = (n_1 + n_2 - 2)\mathbf{S} + \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'$$

which implies

$$\left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1} = (n_1 + n_2 - 2)^{-1}\mathbf{S}^{-1} + a\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{S}^{-1} \tag{13.41}$$

where a is a constant. On the other hand:

$$\widetilde{\mathbf{X}}'\mathbf{Y} = \sum_{i=1}^{n_1+n_2} y_i(\mathbf{x}_i - \overline{\mathbf{x}}_T) = k_1\sum_{i=1}^{n_1}(\mathbf{x}_i - \overline{\mathbf{x}}_T) - \ k_2\sum_{i=1}^{n_2}(\mathbf{x}_i - \overline{\mathbf{x}}_T)$$

replacing $\overline{\mathbf{x}}_T$ with its expression (13.35), yields, by (13.38):

$$\widetilde{\mathbf{X}}'\mathbf{Y} = \frac{k_1 n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) + \frac{k_2 n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)k_T \tag{13.42}$$

with $k_T = n_1 n_2(k_1 + k_2)/(n_1 + n_2)$. Substituting (13.41) and (13.42) in formula (13.34) we obtain:

$$\widehat{\beta} = k.\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

which is the expression of the classical discriminant function.