Chapter 6
Discriminant Analysis: Classification of Observations

Classification and discriminant analysis in multivariate analysis are useful for separating distinct sets of observations or objects and for allocating new observations or objects to previously defined groups.

Suppose an investigator makes several measurements on an individual and wants to classify the individual into one of several categories on the basis of these observations. It is not possible to directly identify and classify an individual into a category.

The difference between the techniques of cluster analysis and discriminant analysis is that the categories in discriminant analysis, in which the observation is to be classified, are known a priori, whereas the categories in which the observation is to be classified in cluster analysis are unknown a priori.

It is assumed that there is a finite number of categories or populations, and each category or population is characterized by a probability distribution of measurements. It is also assumed that the individual belongs to one of such known categories or populations. Thus, an individual is considered as a random observation from this population. The question to be answered is, given an individual with certain measurements, from which population or category did it arise?

For example, suppose that there are two areas or fields in which a mineral is found. The minerals in the two fields differ in their chemical properties, i.e., the minerals from both fields have the same chemical compounds, but their proportions differ. The chemical composition of the minerals from the two fields is supposed to be known. Now, a sample of mineral arrives, and it is to be found in which of the fields it belongs to. A possible solution is to determine the chemical properties of the new sample and compare them with those of the two known fields. Based on such a comparison, the new sample can be classified into one of the two known fields. The classification procedure is said to be good if the sample is correctly classified to the originating field, i.e., if the sample originating from, say, field 1 is classified to field 1 or the sample originating from, say,

field 2 is classified to field 2. On the other hand, the misclassification occurs when the sample is classified to the field from which it does not originate, i.e., if the sample arose from field 1 and is classified to field 2 and vice versa. Each field can be characterized by a probability distribution of measurements. So the observations on the chemical analysis of the sample can be considered as a random observation from one of the two mineral populations in the two fields.

Finally, the problem is that the measurements about the individual are given, and it is to be determined from which population they come. A good solution is to make a decision such that the chances of misclassification, or equivalently, the probability of misclassification, are minimum.

Classification as a Problem of Decision Functions

The classification problem can be considered a problem of "statistical decision functions". Suppose there are a number of hypotheses, and each hypothesis is that the distribution of the observation is a given one. One of the hypotheses must be accepted and the other hypothesis rejected. If there are only two populations, there are only two hypotheses. The test of a hypothesis is conducted to test one hypothesis about a specified distribution against another.

Since a population or a category is specified by a probability distribution, the categories are specified beforehand in the sense that the probability distributions of the measurements are completely known. Another possibility is that the form of each probability distribution is known, but the corresponding distribution parameters are unknown; hence, they must be estimated from a sample from that population.

Considerations for Classification

It is desired to have a classification procedure that is constructed such that the probability of misclassification is minimum or the bad effects of misclassification are, on average,

minimized.

We consider a simple case of two populations. The two populations are denoted as π_1 and π_2, and an observation is from either π_1 or π_2. The observation is a measurement denoted as $\underset{\sim}{x} = (x_1, x_2, \ldots, x_p)'$, which is a point in p-dimensional space, and the classification depends upon $\underset{\sim}{x}$. The aim is to set up a rule that if an individual is characterized by a set of measurement values $x_1, x_2, \ldots, x_p$, then classify it into π_1, i.e., the given observation is coming from the population π_1, and for all other values, it is classified into π_2.

As the observation $\underset{\sim}{x}$ is considered as a point in the p-dimensional space, the entire p-dimensional space is divided into two regions, say R_1 and R_2. Then the classification rule is as follows: If the observation falls in R_1, it is classified as coming from π_1; if it falls in R_2, it is classified as coming from π_2. The error of misclassifying the observation is inevitable in this procedure. There are two types of misclassifications that can occur: If the observation is actually coming from π_1, it is classified to π_2 or If the observation is coming from π_2, it is classified to π_1.

Both types of misclassification yield undesirable results, and it is necessary to know the relative undesirability of each misclassification. Such undesirable outcomes can be described by some associated costs.

Let the cost of misclassifying an individual from π_2 classified into π_1 be $C(1|2)$ and the cost of misclassifying an individual from π_1 classified into π_2 be $C(2|1)$. Both $C(1|2)$ and $C(2|1)$ are assumed to be positive and measured in some units. The statistician may not always know these costs exactly, but may have a rough idea about them. The misclassification may lead to a loss in a statistical sense.

The following notations are used for an observation $\underset{\sim}{x}$ and populations π_1 and π_2.

(i) $\pi_1|\pi_1$: $\underset{\sim}{x} \in \pi_1$ and actually $\underset{\sim}{x} \in \pi_1$.

So the observation is from π_1 and is correctly classified into π_1. This is the correct decision. So the associated cost of misclassification is $C(1|1) = 0$.

(ii) $\pi_2|\pi_2 : \underset{\sim}{x} \in \pi_2$ and actually $\underset{\sim}{x} \in \pi_2$.

The observation is from π_2 and is correctly classified into π_2. This is the correct decision. So the associated cost of misclassification is $C(2|2) = 0$.

(iii) $\pi_1|\pi_2 : \underset{\sim}{x} \in \pi_2$ and actually $\underset{\sim}{x} \in \pi_1$.

The observation is from π_2 but is incorrectly classified into π_1. This is an incorrect decision. The associated cost of misclassification is $C(1|2) > 0$.

(iv) $\pi_2|\pi_1 : \underset{\sim}{x} \in \pi_1$ and actually $\underset{\sim}{x} \in \pi_2$.

The observation comes from π_1 but is incorrectly classified as π_2. This is an incorrect decision. The associated cost of misclassification is $C(2|1) > 0$.

A good classification procedure is one that minimizes, in some sense or other, the cost of misclassification. The statistical loss due to classification is considered, and the "minimum cost" is defined for two cases - when the a priori probabilities of the two populations are

(i) known and

(ii) unknown.

We first consider the case in which the a priori probabilities of the two populations are known.

Case 1: When A Priori Probabilities are Known

Suppose the a priori probabilities that an observation comes from π_1 and π_2 are q_1 and q_2 respectively, i.e.,

$$P(\underset{\sim}{x} \in \pi_1) = q_1 \ ,$$

$$P(\underset{\sim}{x} \in \pi_2) = q_2 \ ,$$

$$q_1 + q_2 = 1 \ .$$

The probability properties of populations are specified by a distribution function which, for convenience, is assumed to have a density. Let $p_1(\underset{\sim}{x})$ and $p_2(\underset{\sim}{x})$ are the densities of π_1 and π_2, respectively.

Suppose R_1 and R_2 are the regions of classification of observation from π_1 and π_2, respectively.

The probability that an observation from π_1 is correctly classified into π_1 is denoted as

$$P(1|1,R) = \int_{R_1} p_1(\underset{\sim}{x}) d\underset{\sim}{x}$$

where $d\underset{\sim}{x} = dx_1 dx_2 \ldots dx_p$

The probability that an observation from π_1 is misclassified into π_2 is denoted as

$$P(2|1,R) = \int_{R_2} p_1(\underset{\sim}{x}) d\underset{\sim}{x}.$$

Similarly, the probability that an observation from π_2 is correctly classified into π_2 is denoted as

$$P(2|2,R) = \int_{R_2} p_2(\underset{\sim}{x}) d\underset{\sim}{x}.$$

The probability that an observation from π_2 is misclassified into π_1 is denoted as

$$P(1|2,R) = \int_{R_1} p_2(\underset{\sim}{x}) d\underset{\sim}{x}.$$

Since $P(\underset{\sim}{x} \in \pi_1) = q_1$, i.e., the probability of drawing an observation from π_1 is q_1, so the probability of drawing an observation from π_1 and correctly classifying it is

$$q_1 \int_{R_1} p_1(\underset{\sim}{x}) d\underset{\sim}{x} = q_1 P(1|1,R).$$

Similarly, the probability of drawing an observation from π_1 and misclassifying it is $q_1 P(2|1,R)$.

The probability of drawing an observation from π_2 and correctly classifying it is $q_2 P(2|2,R)$.

The probability of drawing an observation from π_2 and misclassifying it is $q_2 P(1|2,R)$.

The average or expected loss from costs of misclassification is given by

$$EL = C(2|1) \cdot P(2|1,R) \cdot q_1 + C(1|2) \cdot P(1|2,R) \cdot q_2.$$

It is an aim to minimize this average loss. In other words, divide the space into regions R_1 and R_2 such that the expected loss EL is as small as possible.

A procedure that minimizes the expected loss EL for a given q_1 and q_2 is called a Bayes procedure, and the average expected loss is called the Bayes risk.

We have assumed so far that the errors of misclassification are equally important to avoid hanging an innocent person as compared to letting a murderer free.

Here $C(1|2)$ and $C(2|1)$ are the penalties of misclassification. For example, if a potentially good candidate for admission to a medical programme is rejected, then the nation will suffer a shortage of good doctors. If a bad candidate is admitted to a medical programme, the candidate may not be able to complete it successfully, the college resources used on the candidate will be wasted, and the nation will suffer a shortage of medical doctors.

Case 2: When A Priori Probabilities are Unknown

When the priori probabilities are not known, the expected loss if the observation is from π_1 is

$$C(2|1)P(2|1,R) = r(1,R)$$

and the expected loss if the observation is from π_2 is

$$C(1|2)P(1|2,R) = r(2,R).$$

We do not know whether the observation comes from π_1 or π_2, nor the probabilities of these two instances.

A procedure is at least as good as R^* if

$$r(1,R) \leq r(1,R^*) \qquad \text{and} \qquad r(2,R) \leq r(2,R^*);$$

the procedure R is better than R^* if at least one of these inequalities is a strict inequality.

Usually, there is no procedure that is better than all others, or at least as good as all of them.

A procedure R is called admissible if there is no procedure better than R. Under certain conditions, the class of admissible procedures is the same as the class of Bayes procedures (obtained when the a priori probabilities are known).

The minimax principle leads to a unique procedure. A procedure is minimax if the maximum expected loss $r(i, R)$ is a minimum. From a conservative point of view, this may be considered an optimum procedure.

Procedure of Classification into One of Two Populations with Known Probability Distributions and A Priori Probabilities are Known:

The aim is now to choose the regions R_1 and R_2 such that the expected loss

$$C(2|1) \cdot P(2|1, R) \cdot q_1 + C(1|2) \cdot P(1|2, R) \cdot q_2$$

is minimized.

Since the a priori probabilities are assumed to be known, the joint probabilities of the population and the observed set of variables can be defined. The probability that an observation comes from π_1 and that each variate is less than the corresponding component in $\underset{\sim}{y}$ is

$$\int_{-\infty}^{y_p} \int_{-\infty}^{y_{p-1}} \dots \int_{-\infty}^{y_1} q_1 p_1(\underset{\sim}{x}) dx_1 dx_2 \dots dx_p.$$

The conditional probability that an observation came from a given population, given the observed values of the variates, is defined as follows.

The conditional probability that an observation is coming from π_1, given an observation $\underset{\sim}{x}$, is

$$\frac{q_1 p_1(\underset{\sim}{x})}{q_1 p_1(\underset{\sim}{x}) + q_2 p_2(\underset{\sim}{x})}.$$

Recall the following definition of conditional probability for two sets A and B

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

which is used to find the conditional probability, and A^c is the complement of set A.

Suppose for a moment that the costs of misclassifications $C(1|2) = C(2|1) = 1$. Then the expected loss is

$$q_1 \int_{R_2} p_1(\underset{\sim}{x}) d\underset{\sim}{x} + q_2 \int_{R_1} p_2(\underset{\sim}{x}) d\underset{\sim}{x}.$$

This is also the probability of a misclassification. Hence, we wish to minimize the probability of misclassification.

Similarly, the conditional probability that an observation is coming from π_2, given an observation $\underset{\sim}{x}$, is

$$\frac{q_2 p_2(\underset{\sim}{x})}{q_1 p_1(\underset{\sim}{x}) + q_2 p_2(\underset{\sim}{x})}.$$

For a given observed point $\underset{\sim}{x}$, we minimize the probability of misclassification by assigning the population that has the higher conditional probability. If

$$\frac{q_1 p_1(\underset{\sim}{x})}{q_1 p_1(\underset{\sim}{x}) + q_2 p_2(\underset{\sim}{x})} \geq \frac{q_2 p_2(\underset{\sim}{x})}{q_1 p_1(\underset{\sim}{x}) + q_2 p_2(\underset{\sim}{x})},$$

we choose population π_1. Otherwise, choose population π_2. Since the probability of misclassification is minimized at each point, we minimize it over the whole space. Thus, the rule is as follows:

When q_1 and q_2 are known, then an individual is classified into π_1, if

$$R_1 : q_1 p_1(\underset{\sim}{x}) \geq q_2 p_2(\underset{\sim}{x})$$

and the individual is classified into π_2, if

$$R_2 : q_1 p_1(\underset{\sim}{x}) < q_2 p_2(\underset{\sim}{x}).$$

If $q_1 p_1(\underset{\sim}{x}) = q_2 p_2(\underset{\sim}{x})$, then the individual could be classified as either π_1 or π_2. We have arbitrarily classified into R_1. If $q_1 p_1(\underset{\sim}{x}) + q_2 p_2(\underset{\sim}{x}) = 0$ for a given $\underset{\sim}{x}$, the individual may be classified into π_1 or π_2. Note that R_2 is the complement of R_1.

This is the Bayes procedure against a priori q_1 and q_2.

Next we prove that the Bayes procedure given by $R_1 : q_1 p_1(\underset{\sim}{x}) \geq q_2 p_2(\underset{\sim}{x})$ and $R_2 : q_1 p_1(\underset{\sim}{x}) < q_2 p_2(\underset{\sim}{x})$ for assigning an individual into π_1 and π_2, respectively is the best procedure.

Consider any other procedure $R^* = (R_1^*, R_2^*)$. Assume that the individual must belong to either of the two populations. The average expected loss or the probability of misclassification is

$$\begin{aligned}
& q_1 \cdot \int_{R_2^*} p_1(\underset{\sim}{x}) d\underset{\sim}{x} + q_2 \cdot \int_{R_1^*} p_2(\underset{\sim}{x}) d\underset{\sim}{x} \\
&= q_1 \cdot \int_{R_2^*} p_1(\underset{\sim}{x}) d\underset{\sim}{x} + q_2 \cdot \left[\int_{R_1^*} p_2(\underset{\sim}{x}) d\underset{\sim}{x} + \int_{R_2^*} p_2(\underset{\sim}{x}) d\underset{\sim}{x} - \int_{R_2^*} p_2(\underset{\sim}{x}) d\underset{\sim}{x} \right] \\
&= \int_{R_2^*} \left(q_1 p_1(\underset{\sim}{x}) - q_2 p_2(\underset{\sim}{x}) \right) d\underset{\sim}{x} + q_2 \cdot \int_{R} p_2(\underset{\sim}{x}) d\underset{\sim}{x}
\end{aligned}$$

where $R = R_1^* \cup R_2^*$.

Note that the second term $q_2 \int_R p_2(\underset{\sim}{x}) d\underset{\sim}{x}$ is a given number and a constant, as we are integrating over the whole space. So minimize only the first term

$$\int_{R_2^*} (q_1 p_1(\underset{\sim}{x}) - q_2 p_2(\underset{\sim}{x})) d\underset{\sim}{x}.$$

If we define

$$R_2^* : q_1 p_1(\underset{\sim}{x}) - q_2 p_2(\underset{\sim}{x}) < 0$$

then the term $\int_{R_2^*} (q_1 p_1(\underset{\sim}{x}) - q_2 p_2(\underset{\sim}{x})) d\underset{\sim}{x}$ can be positive, negative or equal to zero. So $\int_{R_2^*} (q_1 p_1(\underset{\sim}{x}) - q_2 p_2(\underset{\sim}{x})) d\underset{\sim}{x}$ will be minimized if R_2^* includes the point $\underset{\sim}{x}$ such that

$$q_1 p_1(\underset{\sim}{x}) - q_2 p_2(\underset{\sim}{x}) < 0$$

and excludes the points for which

$$q_1 p_1(\underset{\sim}{x}) - q_2 p_2(\underset{\sim}{x}) > 0.$$

Hence, the average expected loss is minimum when

$$R_2^* : q_1 p_1(\underset{\sim}{x}) < q_2 p_2(\underset{\sim}{x})$$

or

$$R_2^* : \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} < \frac{q_2}{q_1}.$$

Obviously, then

$$R_1^* : q_1 p_1(\underset{\sim}{x}) \geq q_2 p_2(\underset{\sim}{x})$$

or

$$R_1^* : \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} \geq \frac{q_2}{q_1},$$

which is the same as the regions R_1 and R_2 obtained in the Bayes procedure.

If

$$P\left(\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} = \frac{q_2}{q_1}\middle|\pi_i\right) = 0, i = 1, 2,$$

then the Bayes procedure is unique except for sets of probability zero.

Now the problem of classification can be stated mathematically as follows:

Given nonnegative constants q_1 and q_2 and nonnegative functions $p_1(\underset{\sim}{x})$ and $p_2(\underset{\sim}{x})$, choose regions R_1 and R_2 so as to minimize

$$q_1 \int_{R_2} p_1(\underset{\sim}{x}) d\underset{\sim}{x} + q_2 \int_{R_1} p_2(\underset{\sim}{x}) d\underset{\sim}{x}.$$

The solution is

$$R_1 : q_1 p_1(\underset{\sim}{x}) \geq q_2 p_2(\underset{\sim}{x})$$

$$R_2 : q_1 p_1(\underset{\sim}{x}) < q_2 p_2(\underset{\sim}{x}).$$

So far, we have assumed that the costs of misclassification are equal. If the costs $C(1|2)$ and $C(2|1)$ are unequal, then the expected loss function is given by

$$C(2|1) \cdot q_1 \cdot \int_{R_2} p_1(\underset{\sim}{x}) d\underset{\sim}{x} + C(1|2) \cdot q_2 \cdot \int_{R_1} p_2(\underset{\sim}{x}) d\underset{\sim}{x}$$

which is minimized and the R_1 and R_2 are chosen as follows

$$R_1 : [C(2|1) \cdot q_1]\, p_1(\underset{\sim}{x}) \geq [C(1|2) \cdot q_2]\, p_2(\underset{\sim}{x})$$

$$R_2 : [C(2|1) \cdot q_1]\, p_1(\underset{\sim}{x}) < [C(1|2) \cdot q_2]\, p_2(\underset{\sim}{x}),$$

since $[C(2|1)q_1]$ and $[C(1|2)q_2]$ are nonnegative constants.

In another way,

$$R_1 : \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} \geq \frac{C(1|2)q_2}{C(2|1)q_1}$$
$$R_2 : \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} < \frac{C(1|2)q_2}{C(2|1)q_1}$$

and the Bayes procedure will be unique, except for sets of probability zero, if

$$P\left[\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} = \frac{C(1|2)q_2}{C(2|1)q_1}\right] = 0.$$

Here we note that $\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})}$ is the likelihood ratio for a sample of size one.

When q_1 and q_2 are Unknown

When q_1 and q_2 are unknown, we look for any other procedure, since the Bayes procedure is based on an a priori distribution. So we use the minimax rule.

When q_1 and q_2 are unknown, then the minimax rule can be obtained by equating the two losses due to misclassification. Thus

$$C(1|2)\int_{R_1} p_2(\underset{\sim}{x})d\underset{\sim}{x} = C(2|1)\int_{R_2} p_1(\underset{\sim}{x})d\underset{\sim}{x}.$$

It costs of classification are equal, i.e., $C(1|2) = C(2|1)$, then

$$\int_{R_1} p_2(\underset{\sim}{x})d\underset{\sim}{x} = \int_{R_2} p_1(\underset{\sim}{x})d\underset{\sim}{x}.$$

Another approach to minimise the average expected loss is due to the Neyman-Pearson theory. A linear combination of α and β, say $a\alpha + b\beta$ is minimized such that $a + b = 1$. This will lead to Bayes' decision rule against a priori distribution (a, b). So there is no loss of generality in setting $C(1|2) = C(2|1)$. Also, α and β are the errors of the first and second type in the context of testing a hypothesis.

Classification into One of Two Known Multivariate Normal Populations

Now we use the Bayes procedure in the case of two multivariate normal populations with equal covariance matrices, say $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$. Then for $i = 1, 2$, the i^{th} density is

$$p_i(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_i)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_i)\right].$$

The ratio of densities is

$$\begin{aligned}
\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} &= \frac{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)\right]}{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)\right]} \\
&= \exp\left[-\frac{1}{2}\left\{(\underset{\sim}{x} - \underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_1) - (\underset{\sim}{x} - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)\right\}\right] \\
&= \exp\left[-\frac{1}{2}\left\{-2\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_1 + 2\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_2 + \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_1\right\}\right] \\
&= \exp\left[\underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\right].
\end{aligned}$$

The region of classification into π_1 is R_1 and R_1 is the set of $\underset{\sim}{x}$'s for which

$$\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} \geq k$$

for suitably chosen k. Using the monotonicity property of the logarithmic function, we consider

$$\log\left(\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})}\right) \geq \log k$$

or

$$\left[\underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\right] \geq \log k.$$

Here $\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ is a scalar as $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ and Σ are known quantities. Thus $\underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ is a linear function of $x_1, x_2, \ldots, x_p$. It is called as discriminant function. It is a linear function of the component of the observation vector.

Thus, the best regions of classification are given by

$$R_1 : \underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \geq \log k$$
$$R_2 : \underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) < \log k.$$

If a priori probabilities q_1 and q_2 are known, then

$$k = \frac{q_2 C(1|2)}{q_1 C(2|1)}.$$

Note that the same k was obtained using the Bayes rule as well.

If the two populations are equally likely and the costs are equal, then $k = 1$ and $\log k = 0$. Then the regions of classification are

$$R_1 : \underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \geq \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$
$$R_2 : \underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) < \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2).$$

If we do not have a priori probabilities, we may choose $\log k = c$, say, on the basis of making the expected losses due to misclassification equal.

Let $\underset{\sim}{X}$ be a random observation. There are now two cases.

Case (i): If $\underset{\sim}{X} \in \pi_1$, then $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}_1, \Sigma)$.

Case (ii): If $\underset{\sim}{X} \in \pi_2$, then $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$.

Let

$$U = \underset{\sim}{X}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2).$$

Now we find the distribution of U under two cases - when $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}_1, \Sigma)$ and when $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$.

Case (i): When $\underset{\sim}{X} \in N_p(\underset{\sim}{\mu}_1, \Sigma)$

Since U is a linear function of normally distributed random variables, so U is also normally distributed. The mean of U under $N_p(\underset{\sim}{\mu}_1, \Sigma)$ is

$$\begin{aligned} E_1(U) &= \underset{\sim}{\mu}_1'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= \left[\underset{\sim}{\mu}_1 - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)\right]' \Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= \frac{\Delta^2}{2} \end{aligned}$$

where Δ^2 is the Mahalanobis distance between $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$.

The variance of U under $N_p(\underset{\sim}{\mu}_1, \Sigma)$ is

$$\begin{aligned} V_1(U) &= E[(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}_1)(\underset{\sim}{X} - \underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)] \\ &= (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= \Delta^2. \end{aligned}$$

Thus if $\underset{\sim}{X} \in \pi_1$, then $U \sim N(\frac{\Delta^2}{2}, \Delta^2)$.

Case (ii): When $\underset{\sim}{X} \in N_p(\underset{\sim}{\mu}_2, \Sigma)$

Since U is a linear function of normally distributed random variables, so U is also normally distributed.

The mean of U under $N_p(\underset{\sim}{\mu}_2, \Sigma)$ is

$$\begin{aligned} E_2(U) &= \underset{\sim}{\mu}_2'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= -\frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= -\frac{\Delta^2}{2}. \end{aligned}$$

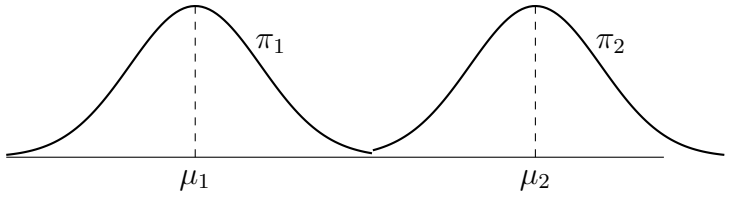
The variance of U under $N_p(\underset{\sim}{\mu}_2, \Sigma)$ is

$$\begin{aligned} V_2(U) &= (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}E(\underset{\sim}{X} - \underset{\sim}{\mu}_2)(\underset{\sim}{X} - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \\ &= \Delta^2. \end{aligned}$$

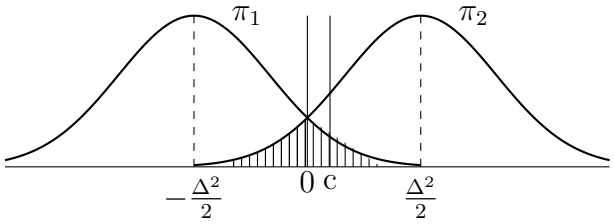
Thus if $\underset{\sim}{X} \in \pi_2$, then $U \sim N(-\frac{\Delta^2}{2}, \Delta^2)$.

Probability of Classification

Two well classified normal populations π_1 and π_2 with no misclassification with means μ_1 and μ_2 with same variances looks as follows:



The same populations with misclassification look like follows, and we want to find the probabilities of misclassification.



If any observation lies in the shaded region, then the observation can belong to π_1 or π_2 with certain probabilities of misclassification.

If $\underset{\sim}{X} \in \pi_1$, then the probability of misclassification is

$$\begin{aligned} P(2|1) &= \int_{R_2} N\left(\frac{\Delta^2}{2}, \Delta^2\right) \\ &= \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\Delta^2}} \exp\left[-\frac{1}{2}\left(\frac{z-\frac{\Delta^2}{2}}{\Delta^2}\right)^2\right] dz \\ &= \int_{-\infty}^{\frac{c-\frac{\Delta^2}{2}}{\Delta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy. \end{aligned}$$

If $\underset{\sim}{X} \in \pi_2$, then the probability of misclassification is

$$\begin{aligned} P(1|2) &= \int_{R_1} N\left(-\frac{\Delta^2}{2}, \Delta^2\right) \\ &= \int_{c}^{\infty} \frac{1}{\sqrt{2\pi\Delta^2}} \exp\left[-\frac{1}{2}\left(\frac{z+\frac{\Delta^2}{2}}{\Delta^2}\right)^2\right] dz \\ &= \int_{\frac{c+\frac{\Delta^2}{2}}{\Delta}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy. \end{aligned}$$

The probabilities $P(1|2)$ and $P(2|1)$ can be calculated with the help of a normal probability table. Here, c is known because we are considering the Bayes' solution in which a priori probabilities and costs are known. Also, given the probability, we can find the value of c.

When prior probabilities are unknown, we use the minimax rule. If the costs are known, then the expected loss is

$$C(1|2) \cdot P(1|2) + C(2|1) \cdot P(2|1).$$

For the minimax solution, we choose c such that

$$C(1|2) \int_{\frac{c+\frac{\Delta^2}{2}}{\Delta}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy = C(2|1) \int_{-\infty}^{\frac{c-\frac{\Delta^2}{2}}{\Delta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy.$$

After finding c, we can easily find R_1 and R_2.

Fisher's Approach to find the Linear Discriminant Function

The linear discriminant function $\underset{\sim}{x}'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ was also derived by Fisher with a different approach.

Suppose there are two populations $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$. These populations, if they are well classified, will look like those shown in Figure 1.

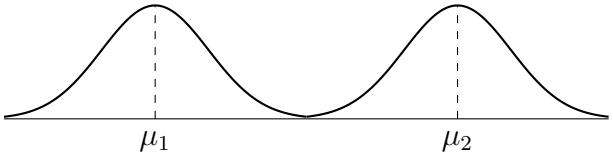


Figure 1: No misclassification

When misclassification is possible, the same populations will appear as in Figure 2.

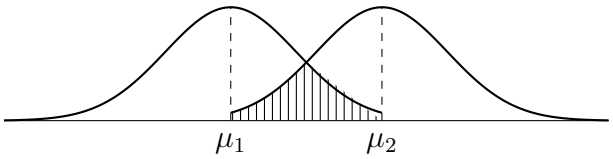


Figure 2: The misclassification area is indicated by the shaded region

In such a case, the misclassification is possible if an individual lies in the shaded region. So the question arises: what criterion should we use to classify the observations into either population? It is to be observed that in such a situation, the populations are differing with respect to the centre of location only, i.e., the means μ_1 and μ_2. So, a criterion where the distance between μ_1 and μ_2 is maximum, i.e., $(\mu_1 - \mu_2)$ is maximum, can be used to decide the misclassification. Since σ^2 also has an important role as a scale parameter, considering $(\frac{\mu_1-\mu_2}{\sigma})$ will make the distance independent of σ. Thus the choice for the criterion becomes now "$(\frac{\mu_1-\mu_2}{\sigma})$ is maximum". Fisher used this idea in a multivariate normal population.

It is intuitively clear that an observation $\underset{\sim}{x}$ will be classified to $\pi_1 \equiv N(\mu, \sigma^2)$ if $\underset{\sim}{x}$ is nearer to μ_1 rather than μ_2. So the risk of misclassification is smaller if $(\frac{\mu_1-\mu_2}{\sigma})^2$, which is unit-free, is large because the two normal density curves are farther apart and the overlap is smaller. The reverse is true otherwise. Also note that maximizing $(\frac{\mu_1-\mu_2}{\sigma})$ is the same as to maximize $(\frac{\mu_1-\mu_2}{\sigma})^2$.

So, suppose two multivariate normal populations are given as $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$ which differ in mean vectors $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$ and having the same covariance matrix Σ. We

want a linear function of $\underset{\sim}{x}$, viz., $\underset{\sim}{a}'\underset{\sim}{x}$ where $\underset{\sim}{a}$ is unknown and the choice of $\underset{\sim}{a}$ has to be made such that

$$\frac{[E_1(\underset{\sim}{a}'\underset{\sim}{x}) - E_2(\underset{\sim}{a}'\underset{\sim}{x})]^2}{\underset{\sim}{a}'\Sigma\underset{\sim}{a}} \tag{1}$$

is maximum. Fisher used this idea in finding such an $\underset{\sim}{a}$. Now, similar to (1) , we can write

$$\frac{(\underset{\sim}{a}'\underset{\sim}{\mu}_1 - \underset{\sim}{a}'\underset{\sim}{\mu}_2)^2}{\underset{\sim}{a}'\Sigma\underset{\sim}{a}} = \frac{\underset{\sim}{a}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\underset{\sim}{a}}{\underset{\sim}{a}'\Sigma\underset{\sim}{a}} \tag{2}$$

So we have to maximize (2) with an aim to obtain $\underset{\sim}{a}$, using the following lemma:

Lemma 1: Let B be a $p \times p$ positive definite matrix and $\underset{\sim}{d}$ is a $p \times 1$ given vector. Then for any arbitrary nonzero vector $\underset{\sim}{x}$ of order $p \times 1$,

$$\max_{\underset{\sim}{x} \neq 0} \frac{(\underset{\sim}{x}'\underset{\sim}{d})^2}{\underset{\sim}{x}'B\underset{\sim}{x}} = \underset{\sim}{d}B^{-1}\underset{\sim}{d}$$

with the maximum attained when $\underset{\sim}{x} = cB^{-1}\underset{\sim}{d}$ for any constant $c \neq 0$.

Using Lemma 1, we get

$$\max_{\underset{\sim}{a} \neq 0} \frac{[\underset{\sim}{a}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)]^2}{\underset{\sim}{a}'\Sigma\underset{\sim}{a}} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) = \Delta^2$$

with the maximum attained with $\underset{\sim}{a} = c\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ for any constant $c \neq 0$. Thus the linear discriminant function $\underset{\sim}{a}'\underset{\sim}{x}$ is now the same as earlier $\underset{\sim}{a}'\underset{\sim}{x} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}$. Any linear function which is proportional to $\underset{\sim}{a}'\underset{\sim}{x}$ can be used as a discriminant function.

Note that the Mahalanobis distance Δ^2 is given by

$$\Delta^2 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2).$$

The same result can also be alternatively proved by maximizing the function using calculus

$$\underset{\sim}{a}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\underset{\sim}{a}' - \lambda(\underset{\sim}{a}'\Sigma\underset{\sim}{a} - 1)$$

where without loss of generality, we can take $\underset{\sim}{a}'\Sigma\underset{\sim}{a} = 1$. This gives $\underset{\sim}{a}'\underset{\sim}{x} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}$ which is again the same result as obtained earlier.

Performance of Classification Rules

The performance of a discriminant function depends, in principle, on how well it correctly classifies observations into the correct populations.

The performance of a sample classification rule is evaluated by calculating the Actual Error Rate (AER) defined as

$$AER = q_1 \int_{R_2} f_1(\underset{\sim}{x}) d\underset{\sim}{x} + q_2 \int_{R_1} f_2(\underset{\sim}{x}) d\underset{\sim}{x}$$

where R_1 and R_2 represent the classification regions determined by samples of sizes N_1 and N_2. For example,

$$R_1 : (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} \underset{\sim}{x} - \frac{1}{2}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} (\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2) \geq \log_e \left(\frac{C(1|2)}{C(2|1)} \cdot \frac{q_2}{q_1} \right)$$
$$R_2 : (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} \underset{\sim}{x} - \frac{1}{2}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} (\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2) < \log_e \left(\frac{C(1|2)}{C(2|1)} \cdot \frac{q_2}{q_1} \right).$$

AER indicates how the sample classification function will perform in future samples, but it depends on the unknown density functions $f_1(\underset{\sim}{x})$ and $f_2(\underset{\sim}{x})$.

However, an estimate of a quantity related to AER, called Apparent Error Rate ($APER$), is defined as the fraction of observations in the training sample that are misclassified by the sample classification function. The $APER$ does not depend on the form of the parent populations and can be calculated for any classification procedure. The $APER$ can be calculated from the **confusion matrix**, which shows the actual versus predicted group membership.

The confusion matrix for N_1 observations from population π_1 and N_2 observations from population π_2 has the following form:

		Predicted membership		
		π_1	π_2	
Actual membership	π_1	n_{1c}	$n_{1m} = n_1 - n_{1c}$	n_1
	π_2	$n_{2m} = n_2 - n_{2c}$	n_{2c}	n_2

Here

n_{1c} : Number of π_1 items correctly classified as π_1 items.

n_{1m} : Number of π_1 items misclassified as π_2 items.

n_{2c} : Number of π_2 items correctly classified as π_2 items.

n_{2m} : Number of π_2 items misclassified as π_1 items.

The $APER$ is defined as

$$APER = \frac{n_{1m} + n_{2m}}{n_1 + n_2}$$

which is the proportion of items in the training set that are misclassified.

Violation of Assumptions

Now, we consider situations in which the assumptions underlying the earlier discriminant function are violated.

When the Populations do not have the Same Covariance Matrices

Instead of having two multivariate normal populations with equal covariance matrices, suppose there are multivariate normal populations $N_p(\underset{\sim}{\mu}_1, \Sigma_1)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma_2)$ having different covariance matrices Σ_1 and Σ_2. In such a case, we will not have a linear discriminant function but a quadratic function in $\underset{\sim}{x}$ is obtained. This can be seen from the following

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = \exp\left[(\underset{\sim}{x} - \underset{\sim}{\mu}_2)'\Sigma_2^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_2) - (\underset{\sim}{x} - \underset{\sim}{\mu}_1)'\Sigma_1^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)\right]$$

which is a quadratic function in $\underset{\sim}{x}$.

Hence, the earlier defined R_1 and R_2 will change accordingly. Anderson and Bahadur (1962) considered a linear discriminant function $R_1 : \underset{\sim}{b}'\underset{\sim}{x} < c$ and $R_2 : \underset{\sim}{b}'\underset{\sim}{x} > c$, but c is difficult to determine.

2. When the Parameters $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2$ and Σ are Unknown

Consider two multivariate normal populations $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$. If $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2$ and Σ are unknown, then they are replaced by their respective maximum likelihood estimates,

i.e., $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2$ and Σ are replaced by $\bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2$ and S, respectively and the discriminant function is obtained as follows:

Let the two samples of sizes N_1 and N_2 are obtained from $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$ respectively and these samples are called as training samples.

Let $\underset{\sim}{x}_1^{(1)}, \underset{\sim}{x}_2^{(1)}, \ldots, \underset{\sim}{x}_{N_1}^{(1)} \sim N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $\underset{\sim}{x}_1^{(2)}, \underset{\sim}{x}_2^{(2)}, \ldots, \underset{\sim}{x}_{N_2}^{(2)} \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$. Then the sample means based on N_1 and N_2 observations are obtained as $\bar{\underset{\sim}{x}}^{(1)}$ and $\bar{\underset{\sim}{x}}^{(2)}$ which are the maximum likelihood estimators of $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$, respectively. The maximum likelihood estimator of Σ is based on pooled covariance matrix as

$$S = \frac{1}{N_1 + N_2 - 2}\left[\sum_{\alpha=1}^{N_1}(\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)' + \sum_{\alpha=1}^{N_2}(\underset{\sim}{x}_\alpha^{(2)} - \bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}_\alpha^{(2)} - \bar{\underset{\sim}{x}}_2)'\right].$$

Our aim is to classify the observation $\underset{\sim}{x}$ into $\pi_1 \equiv N_p(\underset{\sim}{\mu}_1, \Sigma)$ or $\pi_2 \equiv N_p(\underset{\sim}{\mu}_2, \Sigma)$.

Substitute $\bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2$ and S in place of $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2$ and Σ, respectively in the earlier obtained linear discriminant function U and we obtain

$$V(\underset{\sim}{x}) = \underset{\sim}{x}' S^{-1}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) - \frac{1}{2}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)' S^{-1}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2).$$

This V statistic is called as **Anderson statistic**.

Now the regions R_1 and R_2 are defined accordingly as in U :

$$R_1 : V > c$$

$$R_2 : V < c$$

where c is a constant. The V statistic can be used as a classification criterion, as we did with U earlier.

Note that V is no longer a linear discriminant function. Both the terms in V are functions of random variables. Since $\bar{\underset{\sim}{x}}_1$ and $\bar{\underset{\sim}{x}}_2$ are normally distributed whereas S has a Wishart distribution, so the exact distribution of V is difficult to find but approximate results for the distribution of V are available in literature.

When the population is known, we can argue that the classification criterion is best, in the sense that it minimises expected loss when a priori probabilities are known, and

generates the class of admissible procedures when a priori probabilities are not known. We cannot justify V in the same way. However, intuitively, V should perform well.

Classification of Whole Sample

Suppose we have a sample $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_N$ from either π_1 or π_2 and we wish to classify the sample as a whole, then we proceed as follows:

Define S as

$$(N_1 + N_2 + N - 3)S = \sum_{\alpha=1}^{N_1} (\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)' + \sum_{\alpha=1}^{N_2} (\underset{\sim}{x}_\alpha^{(2)} - \bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}_\alpha^{(2)} - \bar{\underset{\sim}{x}}_2)' + \sum_{\alpha=1}^{N} (\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})'$$

where $\bar{\underset{\sim}{x}} = \frac{1}{N} \sum_{\alpha=1}^{N} \underset{\sim}{x}_\alpha$. Then the criterion is

$$\left[\bar{\underset{\sim}{x}} - \frac{1}{2}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)\right]' S^{-1}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2).$$

The larger N is, the smaller the probabilities of misclassification.

On the Distribution of V Statistic

Now we discuss the distribution of V statistic given as

$$V = \underset{\sim}{X}' S^{-1}(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2) - \frac{1}{2}(\bar{\underset{\sim}{X}}_1 + \bar{\underset{\sim}{X}}_2)' S^{-1}(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2)$$

for random $\underset{\sim}{X}$, $\bar{\underset{\sim}{X}}_1$ and $\bar{\underset{\sim}{X}}_2$. The distribution of V is very complicated and depends on the sample sizes and unknown Δ^2. We attempt, as follows, to gain some information about the distribution of V. Let

$$\underset{\sim}{Y}_1 = C_1 \left[\underset{\sim}{X} - \frac{(N_1 \bar{\underset{\sim}{X}}_1 + N_2 \bar{\underset{\sim}{X}}_2)}{N_1 + N_2}\right]$$

$$\underset{\sim}{Y}_2 = C_2 \left[\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2\right]$$

where $C_1 = \sqrt{\frac{N_1+N_2}{N_1+N_2+1}}$ and $C_2 = \sqrt{\frac{N_1N_2}{N_1+N_2}}$, $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_2$ are independently and normally distributed with the following mean vectors and covariance matrices:

$$\begin{aligned}
E(\underset{\sim}{Y}_1) &= C_1\left[E(\underset{\sim}{X}) - \frac{N_1\underset{\sim}{\mu}_1 + N_2\underset{\sim}{\mu}_2}{N_1+N_2}\right] \\
&= \begin{cases} C_1\left(\frac{N_2}{N_1+N_2}\right)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \text{ if } \underset{\sim}{X} \in \pi_1 \\ -C_1\left(\frac{N_1}{N_1+N_2}\right)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \text{ if } \underset{\sim}{x} \in \pi_2 \end{cases} \\
E(\underset{\sim}{Y}_2) &= C_2(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \text{ irrespective of either } \underset{\sim}{X} \in \pi_1 \text{ or } \underset{\sim}{X} \in \pi_2.
\end{aligned}$$

where $E(\bar{\underset{\sim}{X}}_1) = \underset{\sim}{\mu}_1$, $E(\bar{\underset{\sim}{X}}_2) = \underset{\sim}{\mu}_2$.

$$\begin{aligned}
Var(\underset{\sim}{Y}_1) &= C_1^2\left[Var(\underset{\sim}{X}) + \frac{1}{(N_1+N_2)^2}\{N_1^2 Var(\bar{\underset{\sim}{X}}_1) + N_2^2 Var(\bar{\underset{\sim}{X}}_2)\}\right] \\
&= C_1^2\left[\Sigma + \frac{1}{(N_1+N_2)^2}\left(\frac{N_1^2}{N_1} + \frac{N_2^2}{N_2}\right)\Sigma\right] \\
&= C_1^2\left(\frac{N_1+N_2+1}{N_1+N_2}\right)\Sigma \\
&= \Sigma \text{ irrespective of either } \underset{\sim}{X} \in \pi_1 \text{ or } \underset{\sim}{X} \in \pi_2. \\
Var(\underset{\sim}{Y}_2) &= C_2^2\left[Var(\bar{\underset{\sim}{X}}_1) + Var(\bar{\underset{\sim}{X}}_2)\right] \\
&= C_2^2\left(\frac{\Sigma}{N_1} + \frac{\Sigma}{N_2}\right) \\
&= \Sigma \text{ irrespective of either } \underset{\sim}{X} \in \pi_1 \text{ or } \underset{\sim}{X} \in \pi_2.
\end{aligned}$$

Further, also

$$Cov(\underset{\sim}{Y}_1, \underset{\sim}{Y}_2) = 0.$$

Thus $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_2$ are independently distributed. Both $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_2$ follow multivariate normal distribution with mean vectors given by $E(\underset{\sim}{Y}_1)$ and $E(\underset{\sim}{Y}_2)$ respectively and the same covariance matrix Σ.

Let

$$\begin{aligned}
\underset{\sim}{Y} &= (\underset{\sim}{Y}_1, \underset{\sim}{Y}_2)' \\
M &= \underset{\sim}{Y}' S^{-1} \underset{\sim}{Y} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}.
\end{aligned}$$

Then V can be expressed as

$$V = \sqrt{\frac{N_1 + N_2 + 1}{N_1 N_2}} \cdot m_{12} + \frac{N_1 - N_2}{2N_1 N_2} \cdot m_{22}.$$

The density of M has been studied by Sitgreaves (1952) and Anderson (1951). The density of V has been studied by Wald (1944).

If $N_1 = N_2$, then the distribution of V for $\underset{\sim}{X}$ from π_1 is the same as that of -V for $\underset{\sim}{X}$ from π_2. Then, if $V \geq 0$ is the region of classification as π_1, then the probability of misclassifying $\underset{\sim}{X}$ when it is from π_1 is equal to the probability of misclassifying it when it is from π_2.

The Asymptotic Distribution of V

In the case of large samples from $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$, we can apply limiting distribution theory. We denote $\hat{\theta}_n \xrightarrow{Pr} \theta$, i.e., $\hat{\theta}_n$ converges to θ in probability, denoted as $\text{plim}_{n\to\infty}\hat{\theta}_n = \theta$.

Since $\bar{\underset{\sim}{X}}_1$ and $\bar{\underset{\sim}{X}}_2$ are the means of N_1 and N_2 observations, respectively from $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$, respectively, so $\bar{\underset{\sim}{X}}_1 \xrightarrow{Pr} \underset{\sim}{\mu}_1$ and $\bar{\underset{\sim}{X}}_2 \xrightarrow{Pr} \underset{\sim}{\mu}_2$ or $\text{plim}_{N_1\to\infty}\bar{\underset{\sim}{X}}_1 = \underset{\sim}{\mu}_1$, $\text{plim}_{N_2\to\infty}\bar{\underset{\sim}{X}}_2 = \underset{\sim}{\mu}_2$. So plim $S = \Sigma$ as $N_1 \to \infty$ or $N_2 \to \infty$ or both $N_1, N_2 \to \infty$.

Also plim $S^{-1} = \Sigma^{-1}$. Thus

$\text{plim}_{N_1,N_2\to\infty} S^{-1}(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2) = \Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$,

$\text{plim}_{N_1,N_2\to\infty}(\bar{\underset{\sim}{X}}_1 + \bar{\underset{\sim}{X}}_2)' S^{-1}(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2) = (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$.

Thus $V \xrightarrow{Pr} U$.

For sufficiently large samples from π_1 and π_2, we can use this criterion as if we knew the population exactly and make only a small error.

So the limiting distribution V as $N_1 \to \infty$ and $N_2 \to \infty$ is

- $N(\frac{\Delta^2}{2}, \Delta^2)$ if $\underset{\sim}{X}$ is distributed according to $N_p(\underset{\sim}{\mu}_1, \Sigma)$
- $N(-\frac{\Delta^2}{2}, \Delta^2)$ if $\underset{\sim}{X}$ is distributed according to $N_p(\underset{\sim}{\mu}_2, \Sigma)$.

Testing of Hypothesis and Likelihood Ratio Test

The classification problem can also be viewed as a hypothesis-testing problem. The main aim of classification is to assign an observation to one of the two populations, which can be expressed in terms of the null and alternative hypotheses.

Suppose we have two samples of sizes N_1 and N_2 as follows:

$$\underset{\sim}{x}_1^{(1)}, \underset{\sim}{x}_2^{(1)}, \ldots, \underset{\sim}{x}_{N_1}^{(1)} \sim N_p(\underset{\sim}{\mu}_1, \Sigma)$$
$$\underset{\sim}{x}_1^{(2)}, \underset{\sim}{x}_2^{(2)}, \ldots, \underset{\sim}{x}_{N_2}^{(2)} \sim N_p(\underset{\sim}{\mu}_2, \Sigma).$$

An individual observation $\underset{\sim}{x}$ is coming from either $N_p(\underset{\sim}{\mu}_1, \Sigma)$ or $N_p(\underset{\sim}{\mu}_2, \Sigma)$. This can be framed as a problem of testing of hypothesis as follows:

$$H_0 : \underset{\sim}{x}, \underset{\sim}{x}_1^{(1)}, \underset{\sim}{x}_2^{(1)}, \ldots, \underset{\sim}{x}_{N_1}^{(1)} \sim N_p(\underset{\sim}{\mu}_1, \Sigma) \quad \text{and} \quad \underset{\sim}{x}_1^{(2)}, \underset{\sim}{x}_2^{(2)}, \ldots, \underset{\sim}{x}_{N_2}^{(2)} \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$$
$$H_1 : \underset{\sim}{x}_1^{(1)}, \underset{\sim}{x}_2^{(1)}, \ldots, \underset{\sim}{x}_{N_1}^{(1)} \sim N_p(\underset{\sim}{\mu}_1, \Sigma) \quad \text{and} \quad \underset{\sim}{x}, \underset{\sim}{x}_1^{(2)}, \underset{\sim}{x}_2^{(2)}, \ldots, \underset{\sim}{x}_{N_2}^{(2)} \sim N_p(\underset{\sim}{\mu}_2, \Sigma).$$

We know that $\bar{\underset{\sim}{x}}_1$ and $\bar{\underset{\sim}{x}}_2$ are the maximum likelihood estimators of $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$, respectively.

The maximum likelihood estimators under H_0 are as follows:

$$\hat{\underset{\sim}{\mu}}_1 = \frac{\underset{\sim}{x} + \sum_{\alpha=1}^{N_1} \underset{\sim}{x}_\alpha^{(1)}}{N_1 + 1} = \frac{\underset{\sim}{x} + N_1 \bar{\underset{\sim}{x}}_1}{N_1 + 1}$$
$$\hat{\underset{\sim}{\mu}}_2 = \frac{\sum_{\alpha=1}^{N_2} \underset{\sim}{x}_\alpha^{(2)}}{N_2} = \bar{\underset{\sim}{x}}_2$$
$$\hat{\Sigma}_1 = \frac{1}{N_1 + N_2 + 1}\Big[\sum_{\alpha=1}^{N_1} (\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)' + \sum_{\alpha=1}^{N_2} (\underset{\sim}{x}_\alpha^{(2)} - \hat{\underset{\sim}{\mu}}_2)(\underset{\sim}{x}_\alpha^{(2)} - \hat{\underset{\sim}{\mu}}_2)' + (\underset{\sim}{x} - \hat{\underset{\sim}{\mu}}_1)(\underset{\sim}{x} - \hat{\underset{\sim}{\mu}}_1)'\Big].$$

The maximum likelihood estimators under H_1 are as follows:

$$\hat{\underset{\sim}{\mu}}_1 = \bar{\underset{\sim}{x}}_1$$
$$\hat{\underset{\sim}{\mu}}_2 = \frac{\underset{\sim}{x} + \sum_{\alpha=1}^{N_2} \underset{\sim}{x}_\alpha^{(2)}}{N_2 + 1} = \frac{\underset{\sim}{x} + N_2 \bar{\underset{\sim}{x}}_2}{N_2 + 1}$$
$$\hat{\Sigma}_2 = \frac{1}{N_1 + N_2 + 1}\Big[\sum_{\alpha=1}^{N_1} (\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)' + \sum_{\alpha=1}^{N_2} (\underset{\sim}{x}_\alpha^{(2)} - \hat{\underset{\sim}{\mu}}_2)(\underset{\sim}{x}_\alpha^{(2)} - \hat{\underset{\sim}{\mu}}_2)' + (\underset{\sim}{x} - \hat{\underset{\sim}{\mu}}_2)(\underset{\sim}{x} - \hat{\underset{\sim}{\mu}}_2)'\Big].$$

First we simplify the terms of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ as follows:

Let

$$A_{11} = \sum_{\alpha=1}^{N_1} (\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)'$$

$$A_{22} = \sum_{\alpha=1}^{N_2} (\underset{\sim}{x}_\alpha^{(2)} - \bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}_\alpha^{(2)} - \bar{\underset{\sim}{x}}_2)'$$

$$A = A_{11} + A_{22}.$$

Consider

$$\begin{aligned}
&\sum_{\alpha=1}^{N_1} (\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)' \\
&\quad = \sum_{\alpha=1}^{N_1} \left[(\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1) + (\bar{\underset{\sim}{x}}_1 - \hat{\underset{\sim}{\mu}}_1)\right] \left[(\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1) + (\bar{\underset{\sim}{x}}_1 - \hat{\underset{\sim}{\mu}}_1)\right]' \\
&\quad = \sum_{\alpha=1}^{N_1} (\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \bar{\underset{\sim}{x}}_1)' + N_1(\bar{\underset{\sim}{x}}_1 - \hat{\underset{\sim}{\mu}}_1)(\bar{\underset{\sim}{x}}_1 - \hat{\underset{\sim}{\mu}}_1)' \\
&\quad = A_{11} + N_1 \left[\left(\bar{\underset{\sim}{x}}_1 - \frac{N_1\bar{\underset{\sim}{x}}_1 + \underset{\sim}{x}}{N_1+1}\right)\left(\bar{\underset{\sim}{x}}_1 - \frac{N_1\bar{\underset{\sim}{x}}_1 + \underset{\sim}{x}}{N_1+1}\right)'\right] \\
&\quad = A_{11} + \frac{N_1}{(N_1+1)^2}(N_1\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_1 - N_1\bar{\underset{\sim}{x}}_1 - \underset{\sim}{x})(N_1\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_1 - N_1\bar{\underset{\sim}{x}}_1 - \underset{\sim}{x})' \\
&\quad = A_{11} + \frac{N_1}{(N_1+1)^2}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)'.
\end{aligned}$$

So

$$\begin{aligned}
&\sum_{\alpha=1}^{N_1} (\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)(\underset{\sim}{x}_\alpha^{(1)} - \hat{\underset{\sim}{\mu}}_1)' + (\underset{\sim}{x} - \hat{\underset{\sim}{\mu}}_1)(\underset{\sim}{x} - \hat{\underset{\sim}{\mu}}_1)' \\
&\quad = A_{11} + \frac{N_1}{(N_1+1)^2}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)' + \left(\underset{\sim}{x} - \frac{N_1\bar{\underset{\sim}{x}}_1 + \underset{\sim}{x}}{N_1+1}\right)\left(\underset{\sim}{x} - \frac{N_1\bar{\underset{\sim}{x}}_1 + \underset{\sim}{x}}{N_1+1}\right)' \\
&\quad = A_{11} + \frac{N_1}{(N_1+1)^2}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)' + N_1^2(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)' \\
&\quad = A_{11} + \frac{N_1(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)'}{N_1+1}.
\end{aligned}$$

Thus, under H_0

$$\begin{aligned}\hat{\Sigma}_1 &= \frac{1}{N_1+N_2+1}\Big[\sum_{\alpha=1}^{N_1}(\underset{\sim}{x}_\alpha^{(1)}-\underset{\sim}{\hat{\mu}}_1)(\underset{\sim}{x}_\alpha^{(1)}-\underset{\sim}{\hat{\mu}}_1)'+(\underset{\sim}{x}-\underset{\sim}{\hat{\mu}}_1)(\underset{\sim}{x}-\underset{\sim}{\hat{\mu}}_1)' \\ &\quad +\sum_{\alpha=1}^{N_2}(\underset{\sim}{x}_\alpha^{(2)}-\underset{\sim}{\hat{\mu}}_2)(\underset{\sim}{x}_\alpha^{(2)}-\underset{\sim}{\hat{\mu}}_2)'\Big] \\ &= \frac{1}{N_1+N_2+1}\left[A_{11}+\left(\frac{N_1}{N_1+1}\right)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)'+A_{22}\right] \\ &= \frac{1}{N_1+N_2+1}\left[A+\left(\frac{N_1}{N_1+1}\right)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)'\right]\end{aligned}$$

where $A = A_{11} + A_{22}$.

Similarly, under H_1,

$$\hat{\Sigma}_2 = \frac{1}{N_1+N_2+1}\left[A+\left(\frac{N_2}{N_2+1}\right)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)'\right]$$

The likelihood ratio test criterion is

$$\lambda = \frac{\underset{H_0}{\text{Sup}}\ L}{\underset{H_1}{\text{Sup}}\ L} = \frac{|\hat{\Sigma}_2|^{\frac{N_1+N_2+1}{2}}}{|\hat{\Sigma}_1|^{\frac{N_1+N_2+1}{2}}}$$

$$\text{or} \qquad \lambda^{\frac{2}{N_1+N_2+1}} = \frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|}.$$

Now we first simplify

$$\begin{aligned}\frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} &= \frac{|A+\frac{N_2}{N_2+1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)'|}{|A+\frac{N_1}{N_1+1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)'|} \\ &= \frac{1+\frac{N_2}{N_2+1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)'A^{-1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)}{1+\frac{N_1}{N_1+1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)'A^{-1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)}\end{aligned}$$

which is obtained by using the following result:

Result 2: For a vector $\underset{\sim}{y}$,

$$\begin{vmatrix} C & \underset{\sim}{y} \\ \underset{\sim}{y}' & 1 \end{vmatrix} = |C - \underset{\sim}{y}\underset{\sim}{y}'| = \begin{vmatrix} 1 & \underset{\sim}{y}' \\ \underset{\sim}{y} & C \end{vmatrix} = |C|(1 - \underset{\sim}{y}'C^{-1}\underset{\sim}{y})$$

and then

$$\frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} = \frac{n+\frac{N_2}{N_2+1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)'S^{-1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_2)}{n+\frac{N_1}{N_1+1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)'S^{-1}(\underset{\sim}{x}-\underset{\sim}{\bar{x}}_1)}$$

where $n = N_1 + N_2 - 2$ and $S = \frac{A}{n}$. The region of classification into π_1 consists of those points for which

$$\lambda > K_n \text{ (where } K_n \text{ is given)}$$

or

$$R_1 : n + \frac{N_2}{N_2+1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2)'S^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2) > K_n \left[n + \frac{N_1}{N_1+1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)'S^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1) \right].$$

If $K_n = 1$, then the rule is the maximum likelihood rule. Let

$$Z = \frac{1}{2}\Big[\frac{N_2}{N_2+1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2)'S^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2) - \frac{N_1}{N_1+1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)'S^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1) \Big].$$

Then the maximum likelihood rule becomes:

$$\text{Classify to } \pi_1 \text{ if } Z > 0$$

$$\text{Classify to } \pi_2 \text{ if } Z < 0$$

as $\log K_n = 0$.

Roughly speaking, assign $\underset{\sim}{x}$ to π_1 or π_2 according to whether the distance to $\bar{\underset{\sim}{x}}_1$ is less or greater than the distance to $\bar{\underset{\sim}{x}}_2$.

The difference between V and Z is

$$V - Z = \frac{1}{2}\Big[\frac{1}{N_2+1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2)'S^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_2) - \frac{1}{N_1+1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1)'S^{-1}(\underset{\sim}{x} - \bar{\underset{\sim}{x}}_1) \Big]$$

for which

$$\text{plim}(V - Z) = 0$$

as $N_1, N_2 \to \infty$. The probability of misclassification with V is equivalent asymptotically to that with Z for large samples.

Classification into One of Several Populations when A Priori Probabilities are Known

Now we consider the case when there are more than two populations and an observation is to be classified into one of them when a priori probabilities are known.

Let $\pi_1, \pi_2, \ldots, \pi_m$ be m populations with probability density functions $p_1(\underset{\sim}{x}), p_2(\underset{\sim}{x}), \ldots, p_m(\underset{\sim}{x})$, respectively. Then the observation space is divided into m mutually exclusive and exhaustive regions $R_1, R_2, \ldots, R_m$.

We say that the observation comes from π_i if the observation falls in R_i.

Let

- $C(j|i)$: Cost of misclassification when an observation coming from π_j is misclassified into π_i.
- $P(j|i, R) = \int_{R_j} p_i(\underset{\sim}{x}) d\underset{\sim}{x}$,
- $q_1, q_2, \ldots, q_m$ are the a priori probabilities of the populations $\pi_1, \pi_2, \ldots, \pi_m$,

respectively.

The expected loss is given as

$$\sum_{i=1}^{m} q_i \left[\sum_{\substack{j=1 \\ i \neq j}}^{m} C(j|i) P(j|i, R) \right]$$

and $R_1, R_2, \ldots, R_m$ are chosen such that this expected loss is minimum. Since the a priori probabilities are known, the conditional probabilities of an observation coming from a population given the values of $\underset{\sim}{x}$ can be defined.

The probability of an observation coming from π_i given $\underset{\sim}{x}$ is

$$P[\text{Observation coming from } \pi_i | \underset{\sim}{x}] = \frac{q_i p_i(\underset{\sim}{x})}{\sum_{k=1}^{m} q_k p_k(\underset{\sim}{x})}.$$

If the observation is classified as from π_j, the expected loss is

$$EL = \sum_{\substack{i=1 \\ i \neq j}}^{m} \left[\frac{q_i p_i(\underset{\sim}{x})}{\sum_{k=1}^{m} q_k p_k(\underset{\sim}{x})} C(j|i) \right].$$

We minimize the expected loss at this point if we choose j so as to minimize the expected loss EL, i.e., we consider

$$\sum_{\substack{i=1 \\ i \neq j}}^{m} q_i p_i(\underset{\sim}{x}) C(j|i)$$

for all j and select that j which gives the minimum.

If two different indices give the minimum, it is irrelevant which index is selected. Follow this procedure for each $\underset{\sim}{x}$ and define $R_1, R_2, \ldots, R_m$. The classification rule is to classify an observation as coming from π_j if it falls into the region R_j.

Classification into One of Several Populations when A Priori Probabilities are Unknown

Now we consider the case when there are more than two populations and an observation is to be classified into one of them when a priori probabilities are unknown.

The unconditional expected loss for a classification procedure cannot be defined in such a case.

The conditional expected loss if the observation is coming from π_i is

$$\sum_{\substack{j=1 \\ j \neq i}}^{m} P(j|i, R) = r(i, R).$$

A procedure R is at least as good as another procedure R^* if

$$r(i, R) \leq r(i, R^*),\ i = 1, 2, \ldots, m.$$

The procedure R is **better** if at least one inequality is strict. R is **admissible** if there is no procedure R^* that is better than R. A class of procedures is **complete** if for every procedure R outside the class, there is a procedure in the class that is better.

Bayes procedure is admissible.

Classification Into one of Several Multivariate Normal Populations

Suppose there are m multivariate normal populations $\pi_i : N_p(\underset{\sim}{\mu}_i, \Sigma)$, $i = 1, 2, \ldots, m$ and an observation $\underset{\sim}{x}$ is to be classified into one of these populations.

Recall the details of classifying an observation into one of the two multivariate normal populations $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$. Based on that classification rule, we use

$$\begin{aligned} u_{jk} &= \log \frac{p_j(\underset{\sim}{x})}{p_k(\underset{\sim}{x})} \\ &= \left[\underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu}_j + \underset{\sim}{\mu}_k)\right]' \Sigma^{-1}(\underset{\sim}{\mu}_j - \underset{\sim}{\mu}_k) \end{aligned}$$

for classifying an observation $\underset{\sim}{x}$ between $N_p(\underset{\sim}{\mu}_i, \Sigma)$ and $N_p(\underset{\sim}{\mu}_j, \Sigma)$. If a priori probabilities are known, R_j is defined by those $\underset{\sim}{x}$ satisfying

$$R_j : u_{jk}(\underset{\sim}{x}) > \log \frac{q_k}{q_j}, k = 1, 2, \ldots, m, k \neq j.$$

Such regions $R_1, R_2, \ldots, R_m$ minimizes the expected cost and

$$u_{jk}(\underset{\sim}{x}) = -u_{kj}(\underset{\sim}{x}).$$

Based on earlier developments about the U statistic, we can find the following:

$$\begin{aligned} U_{ji} &= \left[\underset{\sim}{X} - \frac{1}{2}(\underset{\sim}{\mu}_i + \underset{\sim}{\mu}_j)\right]' \Sigma^{-1}(\underset{\sim}{\mu}_j - \underset{\sim}{\mu}_i) \\ U_{ji} &= -U_{ij} \\ U_{ji} &\sim N\left(\frac{\Delta_{ji}^2}{2}, \Delta_{ji}^2\right) \\ \Delta_{ji}^2 &= (\underset{\sim}{\mu}_j - \underset{\sim}{\mu}_i)'\Sigma^{-1}(\underset{\sim}{\mu}_j - \underset{\sim}{\mu}_i) \\ Cov(U_{ji}, U_{jk}) &= \Delta_{jk,ji} = (\underset{\sim}{\mu}_j - \underset{\sim}{\mu}_k)'\Sigma^{-1}(\underset{\sim}{\mu}_j - \underset{\sim}{\mu}_i). \end{aligned}$$

To determine the constant c_j, consider the integral

$$P(j|j, R) = \int_{c_j - c_m}^{\infty} \ldots \int_{c_j - c_1}^{\infty} f_j du_{j1} \ldots du_{j,j-1} du_{j,j+1} \ldots du_m$$

where f_j is the density of $U_{ji}, i = 1, 2, \ldots, m, i \neq j$.

When the parameters are unknown, then they are estimated from the training samples $\underset{\sim}{x}_1^{(i)}, \underset{\sim}{x}_2^{(i)}, \ldots, \underset{\sim}{x}_{N_i}^{(i)}$ drawn from $N_p(\underset{\sim}{\mu}_i, \Sigma), i = 1, 2, \ldots, m$ as

$$\begin{aligned} \hat{\underset{\sim}{\mu}}_i &= \bar{\underset{\sim}{x}}_i = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} \underset{\sim}{x}_\alpha^{(i)} \\ \hat{\Sigma} &= S \end{aligned}$$

where $(\sum_{i=1}^{m} N_i - m)S = \sum_{i=1}^{m} \sum_{\alpha=1}^{N_i} (\underset{\sim}{x}_\alpha^{(i)} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_\alpha^{(i)} - \bar{\underset{\sim}{x}}_i)'$.

An analogue of $U_{ij}(\underset{\sim}{X})$ when the parameters are unknown and estimated from the training sample is

$$W_{ij}(\underset{\sim}{X}) = \left[\underset{\sim}{X} - \frac{1}{2}(\bar{\underset{\sim}{x}}_i + \bar{\underset{\sim}{x}}_j)\right]' S^{-1}(\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}}_j).$$

As the variables in W_{ij} are random, so the distribution of W_{ij} differs from the distribution of U_{ij}. However, as $N_i \to \infty$, the joint distributions approach those of U_{ij}. So for sufficiently large samples, W_{ij} can be used.

Classification Into One of Two Known Multivariate Populations with Unequal Covariance Matrices

Let there be two multivariate normal populations with unequal mean vectors and unequal covariance matrices as

$$\pi_1 : N_p(\underset{\sim}{\mu}_1, \Sigma_1), \pi_2 : N_p(\underset{\sim}{\mu}_2, \Sigma_2),$$

where $\underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$ and $\Sigma_1 \neq \Sigma_2$. When the parameters are known, the likelihood ratio for classifying an observation $\underset{\sim}{x}$ into one of the populations is

$$\begin{aligned}\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} &= \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} \cdot \frac{\exp\left[-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)'\Sigma_1^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)\right]}{\exp\left[-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)'\Sigma_2^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)\right]} \\ &= |\Sigma_2|^{1/2}|\Sigma_1|^{-1/2} \exp\left[\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)'\Sigma_2^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_2) - \frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)'\Sigma_1^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)\right].\end{aligned}$$

Here $\log \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})}$ is a quadratic function in $\underset{\sim}{x}$. So use a transformation of $\underset{\sim}{x}$ such that the covariance matrix is an identity matrix and matrix of quadratic form is diagonal; then $\log\left(\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})}\right)$ is distributed as linear combination of noncentral χ^2 distribution and a constant in addition.

When the parameters are unknown, then under the same null and alternative hy-

potheses about the observation $\underset{\sim}{x}$, as earlier, use the likelihood ratio test criterion as

$$\begin{aligned}
&\frac{|\hat{\Sigma}_1(2)|^{\frac{N_1}{2}}|\hat{\Sigma}_2(2)|^{\frac{N_2+1}{2}}}{|\hat{\Sigma}_1(1)|^{\frac{N_1+1}{2}}|\hat{\Sigma}_2(1)|^{\frac{N_2}{2}}} \\
&= \frac{[1+(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_2)'A_2^{-1}(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_2)]^{\frac{N_2+1}{2}}}{[1+(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_1)'A_1^{-1}(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_1)]^{\frac{N_1+1}{2}}} \cdot \frac{|A_2|^{1/2}}{|A_1|^{1/2}} \cdot \frac{(N_1+1)^{\frac{(N_1+1)p}{2}}N_2^{\frac{N_2p}{2}}}{(N_2+1)^{\frac{(N_2+1)p}{2}}N_1^{\frac{N_1p}{2}}}
\end{aligned}$$

where

$$\begin{aligned}
\hat{\Sigma}_1(1) &= \frac{1}{N_1+1}\left[A_1 + \frac{N_1}{N_1+1}(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_1)'\right] \\
\hat{\Sigma}_2(1) &= \frac{A_2}{N_2} \\
\hat{\Sigma}_1(2) &= \frac{A_1}{N_1} \\
\hat{\Sigma}_2(2) &= \frac{1}{N_2+1}\left[A_2 + \frac{N_2}{N_2+1}(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_2)'\right] \\
A_i &= \sum_{\alpha=1}^{N_i}(\underset{\sim}{x}_\alpha^{(i)}-\bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_\alpha^{(i)}-\bar{\underset{\sim}{x}}_i)', i=1,2.
\end{aligned}$$

Tests Associated with Discriminant Function

1. Goodness of Fit of a Hypothetical Discriminant Function

Suppose we want to test whether the discriminant function is good enough. So the null hypothesis can be framed as

H_0 : A given function $\underset{\sim}{\alpha}'\underset{\sim}{x}$ is good enough for discriminating between π_1 and π_2.

Transform $\underset{\sim}{X}$ to Z_1 as

$$Z_1 = \underset{\sim}{\alpha}'\underset{\sim}{X}$$

and $Z_2, Z_3, \ldots, Z_p$ are $(p-1)$ other suitable linear functions of $\underset{\sim}{X}$.

First, we prove that a discriminant function is invariant under a linear transformation as follows:

Let $\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_p)' = A\underset{\sim}{X}$, then $Var(\underset{\sim}{Z}) = A\Sigma A'$ and difference in the means

of $\underset{\sim}{Z}$ in π_1 and π_2 is $A(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$. The discriminant function based on $\underset{\sim}{Z}$ is therefore

$$[A(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)]'(A\Sigma A')^{-1}\underset{\sim}{Z} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X}$$

as A is non-singular and chosen such that the first column is $\underset{\sim}{\alpha}$ and rest are arbitrary.

If H_0 is true, then $(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X}$ must be proportional to $\underset{\sim}{\alpha}'\underset{\sim}{X}$, i.e., Z_1. So

$$[A(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)]'(A\Sigma A')^{-1}\underset{\sim}{Z} = \gamma Z_1$$

where γ is some constant. The coefficients $Z_2, Z_3, \ldots, Z_p$ in $[A(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)]'(A\Sigma A')^{-1}\underset{\sim}{Z}$ are thus all zero.

So H_0 is equivalent to the hypothesis that the variables $Z_2, Z_3, \ldots, Z_p$ do not occur in the discriminant function based on $\underset{\sim}{Z}$.

Using the hypothesis related to $\Delta_p^2 = \Delta_k^2$, putting $k = 1$, the hypothesis reduces to in this case as $H_0 : \Delta_p^2 = \Delta_1^2$ which can be tested by Rao's U-statistic and under H_0,

$$\frac{N_1 + N_2 - p - 1}{p - 1} \cdot \frac{N_1 N_2 (D_p^2 - D_1^2)}{(N_1 + N_2)(N_1 + N_2 - 2) + N_1 N_2 D_1^2} \sim F_{p-1, N_1+N_2-p-1}$$

where D_p^2 is the distance (D^2) based on $Z_1, Z_2, \ldots, Z_p$ and D_1^2 is distance (D^2) based on Z_1 alone. But D^2 is invariant for a non-singular linear transformation like $\underset{\sim}{Z} = A\underset{\sim}{X}$ as $(A\underset{\sim}{d})'(ASA')^{-1}(A\underset{\sim}{d}) = \underset{\sim}{d}'S^{-1}\underset{\sim}{d}$. So D_p^2 is distance (D^2) based on $\underset{\sim}{X}$, viz., $(N_1 + N_2)\underset{\sim}{d}'S^{-1}\underset{\sim}{d}$ where $\underset{\sim}{d} = \bar{\underset{\sim}{X}} - \bar{\underset{\sim}{y}}$. However, as $\underset{\sim}{Z}_1 = \underset{\sim}{\alpha}'\underset{\sim}{X}$,

$$\frac{D_1^2}{N_1 + N_2} = (\underset{\sim}{\alpha}'\underset{\sim}{d})(\underset{\sim}{\alpha}'S\underset{\sim}{\alpha})^{-1}(\underset{\sim}{\alpha}'\underset{\sim}{d}) = \frac{(\underset{\sim}{\alpha}'\underset{\sim}{d})^2}{\underset{\sim}{\alpha}'S\underset{\sim}{\alpha}}.$$

If $H_0 : \Delta_p^2 = \Delta_1^2$ is accepted, then we conclude that the first observation, i.e., Z_1 is sufficient to discriminate between the two populations. Thus, we test the adequacy of the discriminant function.

2. Testing Discriminating Ability of Some Variables

To test the discriminating ability of some variables in a discriminant function, we consider the following three types of null hypotheses, which can be tested by Rao's U-statistic.

(I) $H_{01} : a_{k+1} = a_{k+2} = \ldots = a_p = 0$

where $\underset{\sim}{a} = \Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ is coefficient vector of discriminating function $\underset{\sim}{a}'\underset{\sim}{X}$ and $\underset{\sim}{a} = (\underset{\sim}{a}_1 \ \underset{\sim}{a}_2)'$, $\underset{\sim}{X} = (\underset{\sim}{X}_1 \ \underset{\sim}{X}_2)'$ and $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$ where $\underset{\sim}{X}_1$ and $\underset{\sim}{a}_1$ are of order $(k \times 1)$; and $\underset{\sim}{X}_2$ and $\underset{\sim}{a}_2$ are of order $(p-k) \times 1$ and orders of S_{11}, S_{12} and S_{22} are suitably determined. So H_{01} can be expressed as

$$H_{01} : \underset{\sim}{a}_2 = \underset{\sim}{0}$$

.

(II) $H_{02} : \Delta_p^2 = \Delta_q^2$

(III) $H_{03} : \underset{\sim}{\delta}_2 = B\underset{\sim}{\delta}_1$

where $\underset{\sim}{\delta} = (\underset{\sim}{\delta}_1 \ \underset{\sim}{\delta}_2)'$, $\underset{\sim}{\delta} = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ and $\underset{\sim}{\delta}_1$ and $\underset{\sim}{\delta}_2$ are of orders $(k \times 1)$ and $(p-k) \times 1$. So H_{03} can be expressed as

$$H_{03} : E(\underset{\sim}{X}_2 | \underset{\sim}{X}_1)$$

is the same in both π_1 and π_2.

All three null hypotheses (I), (II) and (III) can be tested by Rao's U statistic as follows:

$$\frac{N_1 + N_2 - p - 1}{p - 1} \cdot \frac{N_1 N_2 (D_p^2 - D_1^2)}{(N_1 + N_2)(N_1 + N_2 - 2) + N_1 N_2 D_1^2} \sim F_{p-1, N_1+N_2-p-1}$$

which is obtained using

$$\frac{N_1 + N_2 - p - 1}{p - 1} \cdot \left(\frac{1}{U} - 1 \right).$$

When H_0 is accepted, it means that the variables $x_{k+1}, x_{k+2}, \ldots, x_p$ do not have additional discriminating ability, once $x_1, x_2, \ldots, x_k$ have already been considered.