Chapter 12
Classification Problems

12.1. Introduction

We will use the same notations as in the previous chapters. Lower-case letters $x, y, \ldots$ will denote real scalar variables, whether mathematical or random. Capital letters $X, Y, \ldots$ will be used to denote real matrix-variate mathematical or random variables, whether square or rectangular matrices are involved. A tilde will be placed on top of letters such as $\tilde{x}, \tilde{y}, \tilde{X}, \tilde{Y}$ to denote variables in the complex domain. Constant matrices will for instance be denoted by A, B, C. A tilde will not be used on constant matrices unless the point is to be stressed that the matrix is in the complex domain. The determinant of a square matrix A will be denoted by $|A|$ or $\det(A)$ and, in the complex case, the absolute value or modulus of the determinant of A will be denoted as $|\det(A)|$. When matrices are square, their order will be taken as $p \times p$, unless specified otherwise. When A is a full rank matrix in the complex domain, then AA^* is Hermitian positive definite where an asterisk designates the complex conjugate transpose of a matrix. Additionally, $\mathrm{d}X$ will indicate the wedge product of all the distinct differentials of the elements of the matrix X. Thus, letting the $p \times q$ matrix $X = (x_{ij})$ where the x_{ij}'s are distinct real scalar variables, $\mathrm{d}X = \wedge_{i=1}^{p} \wedge_{j=1}^{q} \mathrm{d}x_{ij}$. For the complex matrix $\tilde{X} = X_1 + iX_2,\ i = \sqrt{(-1)}$, where X_1 and X_2 are real, $\mathrm{d}\tilde{X} = \mathrm{d}X_1 \wedge \mathrm{d}X_2$.

Historically, classification problems arose in anthropological studies. By taking a set of measurements on skeletal remains, anthropologists wanted to classify them as belonging to a certain racial group such as being of African or European origin. The measurements might have been of the following type: $x_1 =$ width of the skull, $x_2 =$ volume of the skull, $x_3 =$ length of the thigh bone, $x_4 =$ width of the pelvis, and so on. Let the measurements be represented by a $p \times 1$ vector X, with $X' = (x_1, \ldots, x_p)$ where a prime denotes the transpose. Nowadays, classification procedures are employed in all types of problems occurring in various contexts. For example, consider the situation of a battery of tests in an entrance examination to admit students into a professional program such as medical sciences, law studies, engineering science or management studies. Based on the $p \times 1$

A. M. Mathai et al., *Multivariate Statistical Analysis in the Real and Complex Domains*,
https://doi.org/10.1007/978-3-030-95864-0_12

vector of test scores, a statistician would like to classify an applicant as to whether or not he/she belongs to the group of applicants who will successfully complete a given program. This is a 2-group situation. If a third category is added such as those who are expected to complete the program with flying colors, this will become a 3-group situation. In general, one will have a k-group situation when an individual is classified into one of k classes.

Let us begin with the 2-group situation. The problem consists of classifying the $p \times 1$ vector X into one of two, groups, classes or categories. Let the categories be denoted by population π_1 and population π_2. This means X will either belong to π_1 or to π_2, no other options being considered. The $p \times 1$ vector X may be taken as a point in a p-space R_p or p-dimensional Euclidean space $\Re^p$. In a two-group situation when it is decided that the candidate either belongs to the population π_1 or the population π_2, two subspaces A_1 and A_2 within the p-space R_p are determined: $A_1 \subset R_p$ and $A_2 \subset R_p$, with $A_1 \cap A_2 = O$ (the empty set) or a decision rule can be symbolically written as $A = (A_1, A_2)$. If X falls in A_1, the candidate is classified into π_1 and if X falls in A_2, then the candidate is classified into π_2. In other words, $X \in A_1$ means the individual is classified into population π_1 and $X \in A_2$ means that the individual is classified into population π_2. The regions A_1 and A_2 or the rule $A = (A_1, A_2)$ are not known beforehand. These are to be determined by employing certain decision rules. Criteria for determining A_1 and A_2 will be subsequently put forward. Let us now consider the consequences. When a decision is made to classify X as coming from π_1, either the decision is correct or the decision is erroneous. If the population is actually π_1 and the decision rule classifies X into π_1, then the decision is correct. If X is classified into π_2 when in reality the population is π_1, then a mistake has been committed or a misclassification occurred. Misclassification will involve penalties, costs or losses. Let such a penalty, cost or loss of classifying an individual into group i when he/she actually belongs to group j, be denoted by $C(i|j)$. In a 2-group situation, i and j can only equal 1 or 2. That is, $C(1|2) > 0$ and $C(2|1) > 0$ are the costs of misclassifying, whereas $C(1|1) = 0$ and $C(2|2) = 0$ since there is no cost or penalty associated with correct decisions. The following table summarizes this discussion:

Table 12.1: Cost of misclassification $C(i|j)$

		Statistician's decision to classify into	
		π_1	π_2
Population	π_1	0	$C(2\|1)$
In reality	π_2	$C(1\|2)$	0

12.2. Probabilities of Classification

The vector random variable corresponding to the observation vector X may have its own probability/density function. The real scalar variables as well as the observations on them will be denoted by the lower-case letters $x_1, \ldots, x_p$. When dealing with the probability/density function of X, X is taken as vector random variable, whereas when looked upon as a point in the p-space, R_p, X is deemed to be an observation vector. The $p \times 1$ vector X may have a probability/density function $P(X)$. In a 2-group or two classes situation, $P(X)$ is either $P_1(X)$, the population density of π_1 or $P_2(X)$, the population density of π_2. For convenience, it will be assumed that X of the continuous type, the derivations in the discrete case being analogous. In the 2-group situation, $P(X)$ can only be $P_1(X)$ or $P_2(X)$. What is then the probability of achieving a correct classification under the rule $A = (A_1, A_2)$? If the sample point X falls in A_1, we classify the candidate as belonging to π_1, and if the true population is also π_1, then a correct decision is made. In that instance, the corresponding probability is

$$Pr\{1|1, A\} = \int_{A_1} P_1(X)\mathrm{d}X \tag{12.2.1}$$

where $\mathrm{d}X = \mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \ldots \wedge \mathrm{d}x_p$, $A = (A_1, A_2)$ denoting one decision rule or one given set of subspaces of the p-space R_p. The probability of misclassification in this case is

$$Pr\{2|1, A\} = \int_{A_2} P_1(X)\mathrm{d}X. \tag{12.2.2}$$

Similarly, the probabilities of correctly selecting and misclassifying $P_2(X)$ are respectively given by

$$Pr\{2|2, A\} = \int_{A_2} P_2(X)\mathrm{d}X \tag{12.2.3}$$

and

$$Pr\{1|2, A\} = \int_{A_1} P_2(X)\mathrm{d}X. \tag{12.2.4}$$

In a Bayesian setting, there is a prior probability q_1 of selecting the population π_1 and q_2 of selecting the population π_2, with $q_1 + q_2 = 1$. Then, what will be the probability of drawing an observation from π_1 and misclassifying it as belonging to π_2? It is $q_1 \times Pr\{2|1, A\} = q_1 \int_{A_2} P_1(X)\mathrm{d}X$ and, similarly, the probability of drawing an observation from π_2 and misclassifying it as coming from π_1 is $q_2 \times Pr\{1|2, A\} = q_2 \int_{A_1} P_2(X)$, with the respective costs of misclassifications being $C(2|1) = C(2|1, A)$ and $C(1|2) = C(1|2, A)$. What is

then the expected cost of misclassification? It is the sum of the costs multiplied by the corresponding probabilities. Thus,

$$\text{the expected cost } = q_1\, C(2|1)\, Pr\{2|1,\, A\} + q_2\, C(1|2)\, Pr\{1|2,\, A\}. \tag{12.2.5}$$

So, an advantageous criterion to rely on, when setting up A_1 and A_2 would consist in minimizing the expected cost as given in (12.2.5). A rule could be devised for determining A_1 and A_2 accordingly. In this regard, this actually corresponds to Bayes' rule. How can one interpret this expected cost? For example, in the case of admitting students to a particular program of study based on a vector X of test scores, it is the cost of admitting potentially incompetent students or students who would not have successfully completed the program of study and training them, plus the projected cost of losing good students who would have successfully completed the program of study.

If prior probabilities q_1 and q_2 are not involved, then the expected cost of misclassifying an observation from π_1 as coming from π_2 is

$$C(2|1)\, Pr\{2|1,\, A\} \equiv E_1(A), \tag{12.2.6}$$

and the expected cost of misclassifying an observation from π_2 as coming from π_1 is

$$C(1|2)\, Pr\{1|2,\, A\} \equiv E_2(A). \tag{12.2.7}$$

We would like to have $E_1(A)$ and $E_2(A)$ as small as possible. In this case, a procedure, rule or criterion $A = (A_1, A_2)$ corresponds to determining suitable subspaces A_1 and A_2 in the p-space R_p. If there is another procedure $A^{(j)} = (A_1^{(j)}, A_2^{(j)})$ such that $E_1(A) \le E_1(A^{(j)})$ and $E_2(A) \le E_2(A^{(j)})$, then procedure A is said to be as good as $A^{(j)}$, and if at least one of the inequalities above is a strict inequality, that is $<$, then A is preferable to $A^{(j)}$. If procedure A is preferable to all other available procedures $A^{(j)}$, $j = 1, 2, \ldots$, A is said to be *admissible*. We are seeking an admissible class $\{A\}$ of procedures.

12.3. Two Populations with Known Distributions

Let π_1 and π_2 be the two populations. Let $P_1(X)$ and $P_2(X)$ be the known p-variate probability/density functions associated with π_1 and π_2, respectively. That is, $P_1(X)$ and $P_2(X)$ are two p-variate probability/density functions which are fully known in the sense that all their parameters are known in addition to their functional forms. Consider the Bayesian situation where it is assumed that the prior probabilities q_1 and q_2 of selecting π_1 and π_2, respectively, are known. Suppose that a particular p-vector X is at hand. What

is the probability that this given X is an observation from π_1? This probability is $q_1\ P_1(X)$ if X is discrete or $q_1 P_1(X)\mathrm{d}X$ if X is continuous. What is the probability that the given vector X is an observation vector either from π_1 or from π_2? This probability is $q_1 P_1(X)+q_2 P_2(X)$ or $[q_1 P_1(X)+q_2 P_2(X)]\mathrm{d}X$. What is then the probability that the vector X at hand is from $P_1(X)$, given that it is an observation vector from π_1 or π_2? As this is a conditional statement, it is given by the following in the discrete or continuous case:

$$\frac{q_1 P_1(X)}{q_1 P_1(X)+q_2 P_2(X)} \quad \text{or} \quad \frac{q_1 P_1(X)\mathrm{d}X}{[q_1 P_1(X)+q_2 P_2(X)]\mathrm{d}X} = \frac{q_1 P_1(X)}{q_1 P_1(X)+q_2 P_2(X)} \tag{12.3.1}$$

where $\mathrm{d}X$, which is the wedge product of differentials and positive in this case, cancels out. If the conditional probability that a given X is an observation from π_1 is larger than or equal to the conditional probability that the given vector X is an observation from π_2 and if we assign X to π_1, then the chance of misclassification is reduced. Our main objective is to minimize the probability of misclassification and then come up with a decision rule. This statement is equivalent to the following: If

$$\frac{q_1 P_1(X)}{q_1 P_1(X)+q_1 P_2(X)} \geq \frac{q_2 P_2(X)}{q_1 P_1(X)+q_2 P_2(X)} \Rightarrow q_1 P_1(X) \geq q_2 P_2(X) \tag{12.3.2}$$

then we assign X to π_1, meaning that our subspace A_1 is specified by the following rule:

$$\begin{aligned} A_1 : q_1 P_1(X) \geq q_2 P_2(X) &\Rightarrow \frac{P_1(X)}{P_2(X)} \geq \frac{q_2}{q_1} \\ A_2 : q_1 P_1(X) < q_2 P_2(X) &\Rightarrow \frac{P_1(X)}{P_2(X)} < \frac{q_2}{q_1}. \end{aligned} \tag{12.3.3}$$

Note that if $q_1 P_1(X) = q_2 P_2 X)$, then X can be assigned to either π_1 or π_2; however, we have assigned it to π_1 for convenience. Observe that, it is assumed that $q_1 P_1(X) + q_2 P_2(X) \neq 0$, $q_1 > 0$, $q_2 > 0$ and $q_1 + q_2 = 1$ in (12.3.2). The conditional statement made in (12.3.2), which can also be written as

$$\frac{q_i P_i(X)}{q_1 P_1(X)+q_2 P_2(X)} = \frac{\eta_i P_i(X)}{\eta_1 P_1(X)+\eta_2 P_2(X)},\ \eta_i > 0,\ \eta_1+\eta_2 = \eta > 0,\ \frac{\eta_i}{\eta} = q_i,\ i = 1, 2,$$

holds for some weight functions $\eta_i,\ i = 1, 2$.

If the observation is from $\pi_1 : P_1(X)$, then the expected cost of misclassification is $q_1 P_1(X)C(2|1) + q_2 P_2(X)C(2|2) = q_1 P_1(X)C(2|1)$ since $C(i|i) = 0,\ i = 1, 2$. Similarly, the expected cost of misclassifying of the observation X from $\pi_2 : P_2(X)$ is $q_2 P_2(X)C(1|2)$. If $P_1(X)$ is our preferred distribution, then we would like the associated expected cost of misclassification to be the lesser one, that is,

$$\begin{aligned} q_1 P_1(X)C(2|1) < q_2 P_2(X)C(1|2) \text{ in } A_2 \Rightarrow \\ \frac{P_1(X)}{P_2(X)} < \frac{q_2 C(1|2)}{q_1 C(2|1)} \text{ in } A_2 \text{ or} \\ \frac{P_1(X)}{P_2(X)} \geq \frac{q_2 C(1|2)}{q_1 C(2|1)} \text{ in } A_1, \end{aligned} \tag{12.3.4}$$

which is the same rule as in (12.3.3) where q_1 is replaced by $q_1 C(2|1)$ and q_2, by $q_2 C(1|2)$.

12.3.1. Best procedure

It can be established that the procedure $A = (A_1, A_2)$ in (12.3.3) is the best one for minimizing the probability of misclassification. To this end, consider any other procedure $A^{(j)} = (A_1^{(j)}, A_2^{(j)}),\ j = 1, 2, \ldots$. The probability of misclassification under the procedure $A^{(j)}$ is the following:

$$\begin{aligned} & q_1 \int_{A_2^{(j)}} P_1(X)\mathrm{d}X + q_2 \int_{A_1^{(j)}} P_2(X)\mathrm{d}X \\ & = \int_{A_2^{(j)}} [q_1 P_1(X) - q_2 P_2(X)]\mathrm{d}X + q_2 \int_{A_1^{(j)} \cup A_2^{(j)}} P_2(X)\mathrm{d}X. \end{aligned} \tag{12.3.5}$$

If $A_1^{(j)} \cup A_2^{(j)} = R_p$, then $\int_{A_1^{(j)} \cup A_2^{(j)}} P_2(X)\mathrm{d}X = 1$; it is otherwise a given positive constant. However, $q_1 P_1(X) - q_2 P_2(X)$ can be negative, zero or positive, whereas the left-hand side of (12.3.5) is a positive probability. Accordingly, the left-hand side is minimum if

$$q_1 P_1(X) - q_2 P_2(X) < 0 \Rightarrow \frac{P_1(X)}{P_2(X)} < \frac{q_2}{q_1}, \tag{i}$$

which actually is the rejection region A_2 of the procedure $A = (A_1, A_2)$. Hence, the procedure $A = (A_1, A_2)$ minimizes the probabilities of misclassification; in other words, it is the best procedure. If cost functions are also involved, then *(i)* becomes the following:

$$\frac{P_1(X)}{P_2(X)} < \frac{C(1|2)\, q_2}{C(2|1)\, q_1}. \tag{ii}$$

The region where $q_1 P_1(X) - q_2 P_2(X) = 0$ or $q_1 C(2|1) P_1(X) - q_2 C(1|2) P_2(X) = 0$ need not be empty and the probability over this set need not be zero. If

$$Pr\left\{\frac{P_1(X)}{P_2(X)} = \frac{q_2\, C(1|2)}{q_1\, C(2|1)}\Big|\pi_i\right\} = 0,\ i = 1, 2, \tag{12.3.6}$$

it can also be shown that the above Bayes procedure $A = (A_1, A_2)$ is unique. This is stated as a theorem:

Theorem 12.3.1. *Let q_1 be the prior probability of drawing an observation X from the population π_1 with probability/density function $P_1(X)$ and let q_2 be the prior probability of selecting an observation X from the population π_2 with probability/density function $P_2(X)$. Let the cost or loss associated with misclassifying an observation from π_1 as coming from π_2 be $C(2|1)$ and the cost of misclassifying an observation from π_2 as originating from π_1 be $C(1|2)$. Letting*

$$Pr\left\{\frac{P_1(X)}{P_2(X)} = \frac{C(1|2)\, q_2}{C(2|1)\, q_1}\Big|\pi_i\right\} = 0,\ i = 1, 2,$$

the classification rule given by $A = (A_1, A_2)$ of (12.3.4) is unique and best in the sense that it minimizes the probabilities of misclassification.

Example 12.3.1. Let π_1 and π_2 be two univariate exponential populations whose parameters are θ_1 and θ_2 with $\theta_1 \neq \theta_2$. Let the prior probability of drawing an observation from π_1 be $q_1 = \frac{1}{2}$ and that of selecting an observation from π_2 be $q_2 = \frac{1}{2}$. Let the costs or loss associated with misclassifications be $C(2|1) = C(1|2)$. Compute the regions and probabilities of misclassification if (1): a single observation x is drawn; (2): iid observations $x_1, \ldots, x_n$ are drawn.

Solution 12.3.1.(1). In this case, one observation is drawn and the populations are

$$P_i(x) = \frac{1}{\theta_i}\mathrm{e}^{-\frac{x}{\theta_i}},\ x \geq 0,\ \theta_i > 0,\ i = 1, 2.$$

Consider the following inequality on the support of the density:

$$\frac{P_1(x)}{P_2(x)} \geq \frac{C(1|2)\, q_2}{C(2|1)\, q_1} = 1,$$

or equivalently,

$$\frac{\theta_2}{\theta_1}\mathrm{e}^{-x(\frac{1}{\theta_1}-\frac{1}{\theta_2})} \geq 1 \Rightarrow \mathrm{e}^{-x(\frac{1}{\theta_1}-\frac{1}{\theta_2})} \geq \frac{\theta_1}{\theta_2}.$$

On taking logarithms, we have

$$-x\Big(\frac{1}{\theta_1}-\frac{1}{\theta_2}\Big)\geq \ln\frac{\theta_1}{\theta_2}\Rightarrow x\Big(\frac{1}{\theta_2}-\frac{1}{\theta_1}\Big)\geq \ln\frac{\theta_1}{\theta_2}$$
$$\Rightarrow x\geq \frac{\theta_1\theta_2}{\theta_1-\theta_2}\ln\frac{\theta_1}{\theta_2}\ \text{ for } \theta_1>\theta_2.$$

Letting $\theta_1>\theta_2$, the steps in the case $\theta_1<\theta_2$ being parallel, we have

$$x\geq k,\ \ k=\frac{\theta_1\theta_2}{\theta_1-\theta_2}\ln\frac{\theta_1}{\theta_2}.$$

Accordingly,

$$A_1: x\geq k \text{ and } A_2: x<k.$$

The probabilities of misclassification are:

$$P(2|1)=\int_{A_2}P_1(x)\mathrm{d}x=\int_{x=0}^{k}\frac{1}{\theta_1}\mathrm{e}^{-\frac{x}{\theta_1}}\mathrm{d}x=1-\mathrm{e}^{-\frac{k}{\theta_1}}$$
$$P(1|2)=\int_{x=k}^{\infty}\frac{1}{\theta_2}\mathrm{e}^{-\frac{x}{\theta_2}}\mathrm{d}x=\mathrm{e}^{-\frac{k}{\theta_2}}.$$

Solution 12.3.1.(2). In this case, $X'=(x_1,\ldots,x_n)$ and

$$P_i(X)=\prod_{j=1}^{n}\frac{1}{\theta_i}\mathrm{e}^{-\frac{x_j}{\theta_i}}=\frac{1}{\theta_i^n}\mathrm{e}^{-\frac{u}{\theta_i}},\ i=1,2,$$

where $u=\sum_{j=1}^{n}x_j$ is gamma distributed with the parameters (n,θ_i), $i=1,2$. The density of u is then given by

$$g_i(u)=\frac{1}{\theta_i^n\Gamma(n)}u^{n-1}\mathrm{e}^{-\frac{u}{\theta_i}},\ i=1,2.$$

Proceeding as above, for $\theta_1>\theta_2$, $A_1: u\geq k_1$ and $A_2: u<k_1$, $k_1=\frac{\theta_1\theta_2}{\theta_1-\theta_2}\ln[\frac{\theta_1}{\theta_2}]^n=nk$ where k is as given in Solution 12.3.1(1). Consequently, the probabilities of misclassification are as follows:

$$P(2|1)=\int_{u=0}^{k_1}\frac{u^{n-1}}{\theta_1^n\Gamma(n)}\mathrm{e}^{-\frac{u}{\theta_1}}\mathrm{d}u=\int_{0}^{\frac{k_1}{\theta_1}}\frac{u^{n-1}}{\Gamma(n)}\mathrm{e}^{-u}\mathrm{d}u$$
$$P(1|2)=\int_{k_1}^{\infty}\frac{u^{n-1}}{\theta_2^n\Gamma(n)}\mathrm{e}^{-\frac{u}{\theta_2}}\mathrm{d}u=\int_{\frac{k_1}{\theta_2}}^{\infty}\frac{u^{n-1}}{\Gamma(n)}\mathrm{e}^{-u}\mathrm{d}u$$

where the integrals can be expressed in terms of incomplete gamma functions or determined by using integration by parts.

Example 12.3.2. Assume that no prior probabilities or costs are involved. Suppose that in a certain clinic, the waiting time before a customer is attended to, depends upon the manager on duty. If manager M_1 is on duty, the expected waiting time is 10 minutes, and if manager M_2 is on duty, the expected waiting time is 5 minutes. Assume that the waiting times are exponentially distributed with expected waiting time equal to $\theta_i,\ i = 1, 2$. On a particular day (1): a customer had to wait 6 minutes before she was attended to, (2): three customers had to wait 6, 6 and 8 minutes, respectively. Who between M_1 and M_2 was likely to be on duty on that day?

Solution 12.3.2.(1). In this case, $\theta_1 = 10,\ \theta_2 = 5$ and the populations are exponential with parameters θ_1 and θ_2, respectively. Thus, $k = \frac{\theta_1\theta_2}{\theta_1-\theta_2}\ln\frac{\theta_1}{\theta_2} = \frac{(10)(5)}{10-5}\ln\frac{10}{5} = 10\ln 2$, $\frac{k}{\theta_1} = \frac{10\ln 2}{10} = \ln 2,\ \frac{k}{\theta_2} = 2\ln 2 = \ln 4, \mathrm{e}^{-\frac{k}{\theta_1}} = \mathrm{e}^{-\ln 2} = \frac{1}{2} = 0.5$, and $\mathrm{e}^{-\frac{k}{\theta_2}} = \mathrm{e}^{-\ln 4} = \frac{1}{4} = 0.25$. In (1): the observed value of $x = 6 < 10(\ln 2) = 10(0.69314718056) \approx 6.9315$. Accordingly, we classify x to M_2, that is, the manager M_2 was likely to be on duty. Thus,

$$\begin{aligned}
P(2|2, A) &= \text{The probability of making a correct decision}\\
&= \int_{x<k} P_2(x)\mathrm{d}x = \int_0^k P_2(x)\mathrm{d}x = \int_0^k \frac{1}{5}\mathrm{e}^{-\frac{x}{5}}\mathrm{d}x\\
&= 1 - \mathrm{e}^{-\ln 4} = 1 - \frac{1}{4} = 0.75;\\
P(2|1, A) &= \text{Probability of misclassification or making an incorrect decision}\\
&= \int_0^k P_1(x)\mathrm{d}x = \int_0^k \frac{1}{10}\mathrm{e}^{-\frac{x}{10}}\mathrm{d}x = 1 - \mathrm{e}^{-\frac{k}{10}} = 1 - \mathrm{e}^{-\ln 2} = \frac{1}{2} = 0.5.
\end{aligned}$$

Solution 12.3.2.(2). Here, $u = 6 + 6 + 8 = 20,\ n = 3$ and $k_1 = \frac{\theta_1\theta_2}{\theta_1-\theta_2}\, n\, \ln\frac{\theta_1}{\theta_2} = \frac{(10)(5)}{10-5}3\ln\frac{10}{5} = 30\ln 2$. Since $30\ln 2 \approx 20.795$ and the observed value of u is 20, $u < k_1$, and we assign the sample to π_2 or to $P_2(X)$ or M_2, with $\frac{k_1}{\theta_2} = \frac{30\ln 2}{5} = 6\ln 2$ and $\frac{k_1}{\theta_1} = \frac{30\ln 2}{10} = 3\ln 2$. Thus,

$$\begin{aligned}
P(2|2, A) &= \text{Probability of making a correct classification decision}\\
&= Pr\{u < k_1|P_2(X)\} = \int_0^{k_1} \frac{u^{n-1}}{\theta_2^n\Gamma(n)}\mathrm{e}^{-\frac{u}{\theta_2}}\mathrm{d}u\\
&= \int_0^{6\ln 2} \frac{v^2\mathrm{e}^{-v}}{\Gamma(3)}\mathrm{d}v, \quad \text{with } \Gamma(3) = 2! = 2.
\end{aligned}$$

Integrating by parts,

$$\int v^2 \mathrm{e}^{-v}\mathrm{d}v = -[v^2+2v+2]\mathrm{e}^{-v}.$$

Then,

$$\frac{1}{2}\int_0^{6\ln 2} v^2\mathrm{e}^{-v}\mathrm{d}v = -\{[2v^2/2+v+1]\mathrm{e}^{-v}\}_0^{6\ln 2} = 1-\frac{1}{64}[(6\ln 2)^2/2+(6\ln 2)+1]$$
$$\approx 1-\frac{1}{64}[13.797]\approx 0.785, \text{ and}$$

$$P(2|1,A) = \text{Probability of misclassification}$$
$$= \int_0^{k_1} P_1(X)\mathrm{d}X = \frac{1}{2}\int_0^{3\ln 2} v^2\mathrm{e}^{-v}\mathrm{d}v = -[\tfrac{1}{2}v^2+v+1]\mathrm{e}^{-v}\big|_0^{3\ln 2}$$
$$= 1-\frac{1}{2^3}[(3\ln 2)^2/2+(3\ln 2)+1]\approx 0.485.$$

Example 12.3.3. Let the two populations π_1 and π_2 be univariate normal with mean values μ_1 and μ_2, respectively, and the same variance σ^2, that is, $P_1(x) : N_1(\mu_1,\sigma^2)$ and $P_2(x) : N_1(\mu_2,\sigma^2)$. Let the prior probabilities of drawing an observation from these populations be $q_1 = \frac{1}{2}$ and $q_2 = \frac{1}{2}$, respectively, and the costs or loss involved with misclassification be $C(1|2) = C(2|1)$. Determine the regions of misclassification and the corresponding probabilities of misclassification if (1): a single observation x is available; (2): iid observations $x_1,\ldots,x_n$ are available, from π_1 or π_2.

Solution 12.3.3.(1). If one observation is available,

$$P_i(x) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu_i)^2}{2\sigma^2}},\ -\infty<x<\infty,\ -\infty<\mu_i<\infty,\ \sigma>0.$$

Consider regions

$$A_1 : \frac{P_1(x)}{P_2(x)} \geq \frac{C(1|2)\,q_2}{C(2|1)\,q_1} = 1 \Rightarrow \mathrm{e}^{-\frac{1}{2\sigma^2}[(x-\mu_1)^2-(x-\mu_2)^2]} \geq 1$$
$$\Rightarrow -\Big[\frac{1}{2\sigma^2}[(x-\mu_1)^2-(x-\mu_2)^2\Big] \geq 0.$$

Now, note that

$$-[(x-\mu_1)^2-(x-\mu_2)^2] = 2x(\mu_1-\mu_2)-(\mu_1^2-\mu_2^2)\geq 0 \Rightarrow$$
$$x \geq \frac{1}{2}\frac{(\mu_1^2-\mu_2^2)}{(\mu_1-\mu_2)} = \frac{1}{2}(\mu_1+\mu_2) \text{ for } \mu_1>\mu_2 \Rightarrow$$
$$A_1 : x \geq \frac{1}{2}(\mu_1+\mu_2) \text{ and } A_2 : x < \frac{1}{2}(\mu_1+\mu_2).$$

The probabilities of misclassification are the following for $k = \frac{1}{2}(\mu_1 + \mu_2)$:

$$P(2|1) = \int_{-\infty}^{k} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma^2}} \mathrm{d}x = \Phi\Big(\frac{k-\mu_1}{\sigma}\Big)$$
$$P(1|2) = \int_{k}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu_2)^2}{2\sigma^2}} \mathrm{d}x = 1 - \Phi\Big(\frac{k-\mu_2}{\sigma}\Big)$$

where $\Phi(\cdot)$ is the distribution function of a univariate standard normal density and $k = \frac{1}{2}(\mu_1 + \mu_2)$.

Solution 12.3.3.(2). In this case, $x_1, \ldots, x_n$ are iid and $X' = (x_1, \ldots, x_n)$. The multivariate densities are

$$P_i(X) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu_i)^2} = \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}(\sum_{j=1}^{n}(x_j-\bar{x})^2 + n(\bar{x}-\mu_i)^2)}}{\sigma^n(\sqrt{2\pi})^n}, \ i = 1, 2,$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$. Hence for $\mu_1 > \mu_2$,

$$A_1 : \frac{P_1(X)}{P_2(X)} \geq 1 \Rightarrow \mathrm{e}^{-\frac{n}{2\sigma^2}[(\bar{x}-\mu_1)^2 - (\bar{x}-\mu_2)^2]} \geq 1.$$

Taking logarithms and simplifying, we have

$$-\frac{n}{2\sigma^2}[(\bar{x}-\mu_1)^2 - (\bar{x}-\mu_2)^2] \geq 0 \Rightarrow$$
$$\bar{x} \geq \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{1}{2}(\mu_1 + \mu_2) \text{ for } \mu_1 > \mu_2$$

where

$$\bar{x} \sim N_1\Big(\mu_i, \frac{\sigma^2}{n}\Big), \ i = 1, 2.$$

Therefore the probabilities of misclassification are the following:

$$P(2|1) = \int_{-\infty}^{k} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{n}{2\sigma^2}(\bar{x}-\mu_1)^2} \mathrm{d}\bar{x} = \Phi\Big(\frac{\sqrt{n}(\mu_2-\mu_1)}{2\sigma}\Big)$$
$$P(1|2) = \int_{k}^{\infty} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{n}{2\sigma^2}(\bar{x}-\mu_2)^2} \mathrm{d}\bar{x} = 1 - \Phi\Big(\frac{\sqrt{n}(\mu_1-\mu_2)}{2\sigma}\Big)$$

where $k = \frac{1}{2}(\mu_1 + \mu_2)$ and $\Phi(\cdot)$ is the distribution function of a univariate standard normal random variable.

Example 12.3.4. Assume that no prior probabilities or costs are involved. A tuber crop called tapioca is planted by farmers. While farmer F_1 applies a standard fertilizer to the soil to enhance the growth of the tapioca plants, farmer F_2 does not apply any fertilizer and let the plants grow naturally. At harvest time, a tapioca plant is pulled up with all its tubers attached to the bottom of the stem. The upper part of the stem is cut off and the lower part with its tubers is put out for sale. Tuber yield per plant, x, is measured by weighing the lower part of the stem with the tubers attached. It is known from past experience that x is normally distributed with mean value $\mu_1 = 5$ and variance $\sigma^2 = 1$ for F_1 type farms, that is, $x \sim N_1(\mu_1 = 5, \sigma^2 = 1)|F_1$ and that for F_2 type farms, $x \sim N_1(\mu_2 = 3, \sigma^2 = 1)|F_2$, the weights being measured in kilograms. A road-side vendor is selling tapioca and his collection is either from F_1 type farms or F_2 type farms, but not both. A customer picked (1): one stem with its tubers attached weighing 4.2 kg (2) a random sample of four stems respectively weighing 6, 4, 3 and 5 kg. To which type of farms will you classify the observations in (1) and (2)?

Solution 12.3.4. (1). The decision is based on $k = \frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{2}(5 + 3) = 4$. In this case, the decision rule $A = (A_1, A_2)$ is such that $A_1 : x \geq k$ and $A_2 : x < k$ for $\mu_1 > \mu_2$. Note that $\frac{k-\mu_1}{\sigma} = k - \mu_1 = 4 - 5 = -1$ and $\frac{k-\mu_2}{\sigma} = (4 - 3) = 1$. As the observed x is $4.2 > 4 = k$, we classify x into $P_1(X) : N_1(\mu_1, 1)$. Moreover,

$$\begin{aligned} P(1|1, A) &= \text{Probability of making a correct classification decision} \\ &= Pr\{x \geq k|P_1(x)\} = \int_k^\infty \frac{\mathrm{e}^{-\frac{1}{2}(x-\mu_1)^2}}{\sqrt{(2\pi)}}\mathrm{d}x \\ &= \int_{-1}^\infty \frac{\mathrm{e}^{-\frac{1}{2}u^2}}{\sqrt{(2\pi)}} = 0.5 + \int_0^1 \frac{\mathrm{e}^{-\frac{1}{2}u^2}}{\sqrt{(2\pi)}}\mathrm{d}x \approx 0.84, \end{aligned}$$

and

$$\begin{aligned} P(1|2, A) &= \text{Probability of misclassification} \\ &= Pr\{x \geq k|P_2(x)\} = \int_k^\infty \frac{\mathrm{e}^{-\frac{1}{2}(x-\mu_2)^2}}{\sqrt{(2\pi)}}\mathrm{d}x = \int_1^\infty \frac{\mathrm{e}^{-\frac{1}{2}u^2}}{\sqrt{(2\pi)}}\mathrm{d}x \approx 0.16. \end{aligned}$$

Solution 12.3.4. (2). In this case, $\bar{x} = \frac{1}{4}(6+4+3+5) = 4.5$, $n = 4, \bar{x} \sim N(\mu_i, \frac{1}{n})$, $i = 1, 2$, $\frac{(k-\mu_1)}{\sigma/\sqrt{n}} = 2(4 - 5) = -2$ and $\frac{(k-\mu_2)}{\sigma/\sqrt{n}} = 2(4 - 3) = 2$. Since the observed $\bar{x}$ is $4.5 > 4 = k$, we assign the sample to $P_1(X) : N(\mu_1, 1)$, the criterion being $A_1 : \bar{x} \geq k$ and $A_2 : \bar{x} < k$. Additionally,

$$P(1|1,A) = \text{Probability of a correct classification}$$
$$= Pr\{\bar{x} \geq k|P_1(X)\} = \int_k^\infty \frac{\mathrm{e}^{-\frac{n}{2}(\bar{x}-\mu_1)^2}}{\sqrt{(2\pi)}}\mathrm{d}\bar{x} = \int_{-2}^\infty \frac{\mathrm{e}^{-\frac{1}{2}u^2}}{\sqrt{(2\pi)}}\mathrm{d}u$$
$$= 0.5 + \int_0^2 \frac{\mathrm{e}^{-\frac{1}{2}u^2}}{\sqrt{(2\pi)}}\mathrm{d}u \approx 0.98,$$

and

$$P(1|2,A) = \text{Probability of misclassification}$$
$$= Pr\{\bar{x} \geq k|P_2(X)\} = \int_k^\infty \frac{\mathrm{e}^{-\frac{n}{2}(\bar{x}-\mu_2)^2}}{\sqrt{(2\pi)}}\mathrm{d}\bar{x} = \int_2^\infty \frac{\mathrm{e}^{-\frac{1}{2}u^2}}{\sqrt{(2\pi)}}\mathrm{d}u \approx 0.023.$$

Example 12.3.5. Let π_1 and π_2 be two p-variate real nonsingular normal populations sharing the same covariance matrix, $\pi_1 : N_p(\mu^{(1)}, \Sigma)$, $\Sigma > O$, and $\pi_2 : N_p(\mu^{(2)}, \Sigma)$, $\Sigma > O$, whose mean values are such that $\mu^{(1)} \neq \mu^{(2)}$. Let the prior probabilities be $q_1 = q_2$ and the cost functions be $C(1|2) = C(2|1)$. Consider a single p-vector X to be classified into π_1 or π_2. Determine the regions of misclassification and the corresponding probabilities.

Solution 12.3.5. The p-variate real normal densities are the following:

$$P_i(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}(X-\mu^{(i)})'\Sigma^{-1}(X-\mu^{(i)})} \tag{i}$$

for $i = 1, 2,\ \Sigma > O,\ \mu^{(1)} \neq \mu^{(2)}$. Consider the inequality

$$\frac{P_1(X)}{P_2(X)} \geq \frac{C(1|2)\,q_2}{C(2|1)\,q_1} = 1 \Rightarrow$$
$$\mathrm{e}^{-\frac{1}{2}[(X-\mu^{(1)})'\Sigma^{-1}(X-\mu^{(1)})-(X-\mu^{(2)})'\Sigma^{-1}(X-\mu^{(2)})]} \geq 1.$$

Taking logarithms, we have

$$-\tfrac{1}{2}[(X-\mu^{(1)})'\Sigma^{-1}(X-\mu^{(1)}) - (X-\mu^{(2)})'\Sigma^{-1}(X-\mu^{(2)})] \geq 0 \Rightarrow$$
$$(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}X - \tfrac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)}) \geq 0.$$

Let

$$u = (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}X - \tfrac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)}). \tag{12.3.7}$$

Then, u has a univariate normal distribution since it is a linear function of the components of X, which is a p-variate normal. Thus,

$$\begin{aligned}\mathrm{Var}(u) &= \mathrm{Var}[(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}X] \\ &= (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}\mathrm{Cov}(X)\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \\ &= (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) = \Delta^2 \qquad (12.3.8)\end{aligned}$$

where Δ^2 is Mahalanobis' distance. The mean values of u under π_1 and π_2 are respectively,

$$\begin{aligned}E(u)|\pi_1 &= (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}E(X)|\pi_1 - \tfrac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)}) \\ &= \tfrac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) = \tfrac{1}{2}\Delta^2, \qquad (12.3.9) \\ E(u)|\pi_2 &= (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}\mu^{(2)} - \tfrac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)}) \\ &= \tfrac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(2)}-\mu^{(1)}) = -\tfrac{1}{2}\Delta^2, \qquad (12.3.10)\end{aligned}$$

so that

$$\begin{aligned}&u \sim N_1(\tfrac{1}{2}\Delta^2, \Delta^2) \text{ under } \pi_1, \\ &u \sim N_1(-\tfrac{1}{2}\Delta^2, \Delta^2) \text{ under } \pi_2. \qquad (12.3.11)\end{aligned}$$

Accordingly, the regions of misclassification are

$$A_2 : u < 0|\pi_1 : u \sim N_1(\tfrac{1}{2}\Delta^2, \Delta^2) \text{ and } A_1 : u \geq 0|\pi_2 : u \sim N_1(-\tfrac{1}{2}\Delta^2, \Delta^2), \quad (12.3.12)$$

and the probabilities of misclassification are as follows:

$$\begin{aligned}P(2|1) &= \int_{-\infty}^{0} \frac{1}{\Delta\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2\Delta^2}(u-\frac{1}{2}\Delta^2)^2}\mathrm{d}u \\ &= \int_{-\infty}^{\frac{0-\frac{1}{2}\Delta^2}{\Delta}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t = \Phi(-\tfrac{1}{2}\Delta) \qquad (ii) \\ P(1|2) &= \int_{0}^{\infty} \frac{1}{\Delta\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2\Delta^2}(u+\frac{1}{2}\Delta^2)}\mathrm{d}u \\ &= \int_{\frac{0+\frac{1}{2}\Delta^2}{\Delta}}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t = 1-\Phi(\tfrac{1}{2}\Delta) \qquad (iii)\end{aligned}$$

where $\Phi(\cdot)$ denotes the distribution function of a univariate standard normal variable.

Note 12.3.1. If no conditions are imposed on the prior probabilities, q_1 and q_2, or on the costs of misclassification, $C(2|1)$ and $C(1|2)$, then the regions are determined as $A_1 : u \geq k,\ k = \ln \frac{C(1|2)\, q_2}{C(2|1)\, q_1}$, and $A_2 : u < k$. In this case, the probabilities of misclassification will be $\Phi\big(\frac{k-\frac{1}{2}\Delta^2}{\Delta}\big)$ and $1 - \Phi\big(\frac{k+\frac{1}{2}\Delta^2}{\Delta}\big)$, respectively.

Note 12.3.2. If the prior probabilities q_1 and q_2 are not known, we may assume that the two populations π_1 and π_2 are equally likely to be chosen or equivalently that $q_1 = q_2 = \frac{1}{2}$, in which instance $k = \ln \frac{C(1|2)}{C(2|1)}$. Then, the correct decisions are to assign the vector X at hand to π_1 in the region A_1 and to π_2 in the region A_2, where $A_1 : u \geq k$ and $A_2 : u < k,\ k = \ln \frac{q_2\, C(1|2)}{q_1\, C(2|1)}$ with $q_1, q_2, C(2|1)$ and $C(1|2)$ assumed to be known and

$$u = (\mu^{(1)} - \mu^{(2)})' \Sigma^{-1} X - \tfrac{1}{2}(\mu^{(1)} - \mu^{(2)})' \Sigma^{-1}(\mu^{(1)} + \mu^{(2)})$$

whose first term, namely $(\mu^{(1)} - \mu^{(2)})' \Sigma^{-1} X$, is known as the linear discriminant function, which is utilized to discriminate or to separate two p-variate populations, not necessarily normally distributed, having mean value vectors $\mu^{(1)}$ and $\mu^{(2)}$ and sharing the same covariance matrix $\Sigma > O$.

Example 12.3.6. Assume that no prior probabilities or costs are involved. Applicants to a certain training program are given tests to evaluate their aptitude for languages and aptitude for science. Let the test scores be denoted by x_1 and x_2, respectively. Let X be the bivariate vector $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. After completing the training program, their aptitudes are tested again. Let $X^{(1)'} = [x_1^{(1)}, x_2^{(1)}]$ be the score vector in the group of successful trainees and let $X^{(2)'} = [x_1^{(2)}, x_2^{(2)}]$ be the score vector in the group of unsuccessful trainees. From previous experience of conducting such tests over the years, it is known that $X^{(1)} \sim N_2(\mu^{(1)}, \Sigma),\ \Sigma > O$, and $X^{(2)} \sim N_2(\mu^{(2)}, \Sigma),\ \Sigma > O$, where

$$\mu^{(1)} = \begin{bmatrix} 4 \\ 1 \end{bmatrix},\ \mu^{(2)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix},\ \Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \Rightarrow \Sigma^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}.$$

Then (1): one applicant taken at random before the training program started obtained the test scores $X_0 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$; (2): three applicants chosen at random before the training program started had the following scores:

$$\begin{bmatrix} 4 \\ 2 \end{bmatrix},\ \begin{bmatrix} 3 \\ 1 \end{bmatrix},\ \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

In (1), classify X_0 to π_1 or π_2 and in (2), classify the entire sample of three vectors into π_1 or π_2.

Solution 12.3.6. Let us compute certain quantities which are needed to answer the questions:

$$\frac{1}{2}(\mu^{(1)}+\mu^{(2)})=\frac{1}{2}\left(\begin{bmatrix}4\\1\end{bmatrix}+\begin{bmatrix}2\\1\end{bmatrix}\right)=\frac{1}{2}\begin{bmatrix}6\\2\end{bmatrix}=\begin{bmatrix}3\\1\end{bmatrix};$$

$$\mu^{(1)}-\mu^{(2)}=\begin{bmatrix}2\\0\end{bmatrix},\ (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}X=[2,-2]X=2x_1-2x_2;$$

$$\Delta^2=(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})=[2,0]\begin{bmatrix}1&-1\\-1&2\end{bmatrix}\begin{bmatrix}2\\0\end{bmatrix}=4;$$

$$(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)})=[2,0]\begin{bmatrix}1&-1\\-1&2\end{bmatrix}\begin{bmatrix}6\\2\end{bmatrix}=8.$$

Hence,

$$\begin{aligned}u&=(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}X-\frac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)})\\&=2x_1-2x_2-4;\end{aligned}$$

$$u|\pi_1\ \sim N_1(\tfrac{1}{2}\Delta^2,\Delta^2),\ u|\pi_2\ \sim N_1(-\tfrac{1}{2}\Delta^2,\Delta^2);$$

$$A_1: u\geq 0,\ A_2: u<0.$$

Since, in (1), the observed $X_0=\begin{bmatrix}4\\1\end{bmatrix}$, the observed u is $u=2x_1-2x_2-4=8-2-4=2>0$ and we classify the observed X_0 into $\pi_1: N_1(\frac{1}{2}\Delta^2,\Delta^2)$, the criterion being $A_1: u\geq 0$ and $A_2: u<0$. Thus,

$$\begin{aligned}P(1|1,A)&=\text{Probability of making a correct classification decision}\\&=Pr\{u\geq 0|\pi_1\}=\int_0^\infty \frac{\mathrm{e}^{-\frac{1}{2\Delta^2}(u-\frac{1}{2}\Delta^2)^2}}{\Delta\sqrt{(2\pi)}}\mathrm{d}u=\int_{-\frac{\Delta}{2}}^\infty\frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}}\mathrm{d}v\\&=\int_{-1}^\infty\frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}}\mathrm{d}v=0.5+\int_0^1\frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}}\mathrm{d}v\approx 0.841;\end{aligned}$$

$$\begin{aligned}P(1|2,A)&=\text{Probability of misclassification}\\&=\int_0^\infty\frac{\mathrm{e}^{-\frac{1}{2\Delta^2}(u+\frac{1}{2}\Delta^2)^2}}{\Delta\sqrt{(2\pi)}}\mathrm{d}u=\int_1^\infty\frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}}\mathrm{d}v\approx 0.159.\end{aligned}$$

When solving (2), the entire sample is to be classified. Proceeding as in the derivation of the criterion u in case (1), it is seen that for the problem at hand, X_0 will be replaced by $\bar{X}$,

the average of the sample vectors or the sample mean value vector, and then u will become $u_1 = 2\bar{x}_1 - 2\bar{x}_2 - 4$ where $\bar{X}' = [\bar{x}_1, \bar{x}_2]$. Thus, we require the sample average:

$$\begin{bmatrix}4\\2\end{bmatrix} + \begin{bmatrix}3\\1\end{bmatrix} + \begin{bmatrix}5\\1\end{bmatrix} = \begin{bmatrix}12\\4\end{bmatrix} \Rightarrow \text{ observed sample mean } = \frac{1}{3}\begin{bmatrix}12\\4\end{bmatrix}.$$

This means that $\bar{x}_1 = \frac{12}{3} = 4,\ \bar{x}_2 = \frac{4}{3}$, and the observed $u_1 = 2\bar{x}_1 - 2\bar{x}_2 - 4 = 8 - \frac{8}{3} - 4 > 0$. Hence, we classify the whole sample to π_1 as the criterion is $A_1 : u_1 \geq 0$ and $A_2 : u_1 < 0$. Since $\bar{X}$ is normally distributed with $E[\bar{X}] = \mu^{(i)}$ and $\text{Cov}(\bar{X}) = \frac{1}{n}\Sigma,\ i = 1, 2$, where n is the sample size, the densities of u_1 under π_1 and π_2 are the following:

$$u_1|\pi_1 \sim N_1(\tfrac{1}{2}\Delta^2, \tfrac{1}{3}\Delta^2),\ n = 3,$$
$$u_1|\pi_2 \sim N_1(-\tfrac{1}{2}\Delta^2, \tfrac{1}{3}\Delta^2).$$

Moreover,

$$\begin{aligned}P(1|1, A) &= \text{Probability of making a correct classification decision}\\ &= Pr\{u_1 \geq 0|\pi_1\} = \int_0^\infty \frac{\sqrt{3}}{\Delta\sqrt{(2\pi)}} \mathrm{e}^{-\frac{3}{2\Delta^2}(u_1 - \frac{1}{2}\Delta^2)^2} \mathrm{d}u_1\\ &= \int_{-\frac{\sqrt{3}\Delta}{2}}^\infty \frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}} \mathrm{d}v = \int_{-\sqrt{3}}^\infty \frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}} \mathrm{d}v = 0.5 + \int_0^{\sqrt{3}} \frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}} \mathrm{d}v \approx 0.958\end{aligned}$$

and

$$\begin{aligned}P(1|2, A) &= \text{Probability of misclassification}\\ &= Pr\{u_1 \geq 0|\pi_2\} = \int_0^\infty \frac{\sqrt{3}}{\Delta\sqrt{(2\pi)}} \mathrm{e}^{-\frac{3}{2\Delta^2}(u_1 + \frac{1}{2}\Delta^2)^2} \mathrm{d}u_1\\ &= \int_{\sqrt{3}}^\infty \frac{\mathrm{e}^{-\frac{1}{2}v^2}}{\sqrt{(2\pi)}} \approx 0.042.\end{aligned}$$

12.4. Linear Discriminant Function

Let X be a $p \times 1$ vector and B a $p \times 1$ arbitrary constant vector, $B' = (b_1, \ldots, b_p)$. Consider the arbitrary linear function $w = B'X$. Then, the mean value and variance of w are the following: $E(w) = B'E(X)$ and $\text{Var}(w) = \text{Var}(B'X) = B'\text{Cov}(X)B = B'\Sigma B$ where $\Sigma > O$ is the covariance matrix of X. Suppose that the X could be from a p-variate real population π_1 with mean value vector $\mu^{(1)}$ or from the p-variate real population π_2 with mean value vector $\mu^{(2)}$. Suppose that both the populations π_1 and π_2 have the same

covariance matrix $\Sigma > O$. Then, a measure of discrimination or separation between π_1 and π_2 is $|B'\mu^{(1)} - B'\mu^{(2)}|$ as measured in terms of the standard deviation $\sqrt{\mathrm{Var}(w)}$ for determining the best choice of B. Taking the squared distance, let

$$\delta = \frac{[B'\mu^{(1)} - B'\mu^{(2)}]^2}{B'\Sigma B} = \frac{[B'(\mu^{(1)} - \mu^{(2)}]^2}{B'\Sigma B} = \frac{B'(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})'B}{B'\Sigma B} \tag{12.4.1}$$

since the square of a scalar quantity is the scalar quantity times its transpose, $B'(\mu^{(1)} - \mu^{(2)})$ being a scalar quantity. Accordingly, we will maximize δ as specified in (12.4.1). This will be achieved by selecting a particular B in such a way that δ attains a maximum which corresponds to the maximum distance between π_1 and π_2. Without any loss of generality, we may assume that $B'\Sigma B = 1$, so that only the numerator in (12.4.1) need be maximized, subject to the condition $B'\Sigma B = 1$. Let λ denote a Lagrangian multiplier and

$$\eta = B'(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})'B - \lambda(B'\Sigma B - 1).$$

Let us take the partial derivative of η with respect to the vector B and equate the result to a null vector (the reader may refer to Chap. 1 for the derivative of a scalar variable with respect to a vector variable):

$$\begin{aligned} \frac{\partial \eta}{\partial B} = O &\Rightarrow 2(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})'B - 2\lambda\Sigma B = O \\ &\Rightarrow \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})'B = \lambda B. \end{aligned} \tag{i}$$

Note that $(\mu^{(1)} - \mu^{(2)})'B \equiv \alpha$ is a scalar quantity and B is a specific vector coming from *(i)* and hence we may write *(i)* as

$$B = \frac{\alpha}{\lambda}\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \equiv c\,\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \tag{ii}$$

where c is a real scalar quantity. Observe that δ as given in (12.4.1) will remain the same if B is multiplied by any scalar quantity. Thus, we may take $c = 1$ in *(ii)* without any loss of generality. The linear discriminant function then becomes

$$B'X = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}X, \tag{12.4.2}$$

and when $B'X$ is as given in (12.4.2), δ as defined in (12.4.1), can be expressed as follows:

$$\begin{aligned} \delta &= \frac{(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})}{(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})} \\ &= (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) = \Delta^2 \equiv \text{ Mahalanobis' distance} \\ &= \mathrm{Var}(w) = \text{ Variance of the discriminant function.} \end{aligned} \tag{12.4.3}$$

This δ is also the generalized squared distance between the vectors $\mu^{(1)}$ and $\mu^{(2)}$ or the squared distance between the vectors $\Sigma^{-\frac{1}{2}}\mu^{(1)}$ and $\Sigma^{-\frac{1}{2}}\mu^{(2)}$ in the mathematical sense (Euclidean distance). Hence Mahalanobis' distance between two p-variate populations with different mean value vectors and the same covariance matrix is a measure of discrimination or separation between the populations, and the linear discriminant function is given in (12.4.2). Hence for an observed value X, if $u = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}X > 0$ when $\mu^{(1)}$, $\mu^{(2)}$ and Σ are known, then we choose population π_1 with mean value $\mu^{(1)}$, and if $u < 0$, then we select population π_2 with mean value $\mu^{(2)}$. When $u = 0$, both π_1 and π_2 are equally favored.

Example 12.4.1. In a small township, there is only one grocery store. The town is laid out on the East and West sides of the sole main road. We will refer to the villagers as East-enders and West-enders. These townspeople shop only once a week for groceries. The grocery store owner found that the East-enders and West-enders have somewhat different buying habits. Consider the following items: $x_1 =$ grain items in kilograms, $x_2 =$ vegetable items in kilograms, $x_3 =$ dairy products in kilograms, and let $[x_1, x_2, x_3] = X'$ where X is the vector of weekly purchases. Then, the expected quantities bought by the East-enders and West-enders are $E(X) = \mu^{(1)}$ and $E(X) = \mu^{(2)}$, respectively, with the common covariance matrix $\Sigma > O$. From past history, the grocery store owner determined that

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},\ \mu^{(1)} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix},\ \mu^{(2)} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix},\ \Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Consider the following situations: (1) A customer walked in and bought $x_1 = 1$ kg of grain items, $x_2 = 2$ kg of vegetable items, and $x_3 = 1$ kg of dairy products. Is she likely to be an East-ender or West-ender? (2): Another customer bought the three types of items in the quantities (10, 1, 1), respectively. Is she more likely to be an East-ender than a West-ender?

Solution 12.4.1. The inverse of the covariance matrix, $\mu^{(1)} - \mu^{(2)}$, as well as other relevant quantities are the following:

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix},\ \mu^{(1)} - \mu^{(2)} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

$$(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1} = [1, 0, -1] \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} = [\tfrac{1}{3}, -1, -2].$$

In (1), $X' = (1, 2, 1)$ and since

$$(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}X = [\tfrac{1}{3}, -1, -2]\begin{bmatrix}1\\2\\1\end{bmatrix} < 0,$$

we classify this customer as a West-ender from her buying pattern. In (2),

$$(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}X = [\tfrac{1}{3}, -1, -2]\begin{bmatrix}10\\1\\1\end{bmatrix} > 0,$$

so that, given her purchases, this customer is classified as an East-ender.

12.5. Classification When the Population Parameters are Unknown

We now consider the classification problem involving two populations π_1 and π_2 for which the parameters of the corresponding densities are unknown. Since the structure of the parameters in these general densities $P_1(X)$ and $P_2(X)$ is not known, we will present a specific example: Consider the two p-variate normal populations of Example 12.3.3. Let $\pi_1 : N_p(\mu^{(1)}, \Sigma)$ and $\pi_2 : N_p(\mu^{(2)}, \Sigma)$, which share the same positive definite covariance matrix Σ. Suppose that we have a single observation vector X to be classified into π_1 or π_2. When the parameters $\mu^{(1)}$, $\mu^{(2)}$ and Σ are unknown, we will have to estimate them from some training samples. But, for a problem such as classifying skeletal remains, one does not have samples from the respective ancestral groups. Nevertheless, one can obtain training samples from living racial groups, and so, secure estimates of the parameters involved. Assume that we have simple random samples of sizes n_1 and n_2 from $N_p(\mu^{(1)}, \Sigma)$ and $N_p(\mu^{(2)}, \Sigma)$, respectively. Denote the sample values by $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$, and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$, and let $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ be the sample averages. That is,

$$X_j^{(1)} = \begin{bmatrix}x_{1j}^{(1)}\\ \vdots \\ x_{pj}^{(1)}\end{bmatrix},\ j = 1, \ldots, n_1; \quad X_j^{(2)} = \begin{bmatrix}x_{1j}^{(2)}\\ \vdots \\ x_{pj}^{(2)}\end{bmatrix},\ j = 1, \ldots, n_2;$$

$$\bar{X}^{(1)} = \begin{bmatrix}\bar{x}_1^{(1)}\\ \vdots \\ \bar{x}_p^{(1)}\end{bmatrix},\ \bar{x}_k^{(1)} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{kj}^{(1)}; \quad \bar{X}^{(2)} = \begin{bmatrix}\bar{x}_1^{(2)}\\ \vdots \\ \bar{x}_p^{(2)}\end{bmatrix},\ \bar{x}_k^{(2)} = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{kj}^{(2)}. \tag{12.5.1}$$

Let the sample matrices be denoted by bold-faced letters where the $p \times n_1$ matrix $\mathbf{X^{(1)}}$ and the $p \times n_2$ matrix $\mathbf{X^{(2)}}$ are the sample matrices and let $\bar{\mathbf{X}}^{(1)}$ and $\bar{\mathbf{X}}^{(2)}$ be the matrices of sample means. Thus, we have

$$\mathbf{X^{(1)}} = [X_1^{(1)}, \ldots, X_{n_1}^{(1)}] = \begin{bmatrix} x_{11}^{(1)} & \cdots & x_{1n_1}^{(1)} \\ \vdots & \ddots & \vdots \\ x_{p1}^{(1)} & \cdots & x_{pn_1}^{(1)} \end{bmatrix},$$

$$\bar{\mathbf{X}}^{\mathbf{(1)}} = [\bar{X}^{(1)}, \ldots, \bar{X}^{(1)}] = \begin{bmatrix} \bar{x}_1^{(1)} & \cdots & \bar{x}_1^{(1)} \\ \vdots & \ddots & \vdots \\ \bar{x}_p^{(1)} & \cdots & \bar{x}_p^{(1)} \end{bmatrix},$$

$$\mathbf{X^{(2)}} = [X_1^{(2)}, \ldots, X_{n_2}^{(2)}] = \begin{bmatrix} x_{11}^{(2)} & \cdots & x_{1n_2}^{(2)} \\ \vdots & \ddots & \vdots \\ x_{p1}^{(2)} & \cdots & x_{pn_2}^{(2)} \end{bmatrix},$$

$$\bar{\mathbf{X}}^{\mathbf{(2)}} = [\bar{X}^{(2)}, \ldots, \bar{X}^{(2)}] = \begin{bmatrix} \bar{x}_1^{(2)} & \cdots & \bar{x}_1^{(2)} \\ \vdots & \ddots & \vdots \\ \bar{x}_p^{(2)} & \cdots & \bar{x}_p^{(2)} \end{bmatrix}. \tag{12.5.2}$$

Then, the sample sum of products matrices are

$$S_i = (\mathbf{X^{(i)}} - \bar{\mathbf{X}}^{\mathbf{(i)}})(\mathbf{X^{(i)}} - \bar{\mathbf{X}}^{\mathbf{(i)}})', \ i = 1, 2;$$

$$S_m = (s_{ij}^{(m)}), \ s_{ij}^{(m)} = \sum_{k=1}^{n_m} (x_{ik}^{(m)} - \bar{x}_i^{(m)})(x_{jk}^{(m)} - \bar{x}_j^{(m)}), \ m = 1, 2, \ S = S_1 + S_2. \tag{12.5.3}$$

The unbiased estimators of $\mu^{(1)}, \mu^{(2)}$ and Σ are respectively $\bar{X}^{(1)}, \bar{X}^{(2)}$ and $\frac{S}{n_{(2)}} = \frac{S_1+S_2}{n_{(2)}}$, $n_{(2)} = n_1 + n_2 - 2$. The criteria for classification, the regions, the statistic, and so on, are available from Example 12.3.3. That is,

$$A_1 : u \geq k, \ A_2 : u < k, \ k = \ln \frac{C(1|2)q_2}{C(2|1)q_1},$$

where

$$u = X'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \frac{1}{2}(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} + \mu^{(2)}).$$

Note that q_1 and q_2 are the prior probabilities of selecting the populations π_1 and π_2 and $C(1|2)$ and $C(2|1)$ are the costs or loss associated with misclassification. We will assume that q_1, q_2, $C(1|2)$ and $C(2|1)$ are all known but the parameters $\mu^{(1)}$, $\mu^{(2)}$ and Σ are

estimated by their unbiased estimators. Denoting the estimator of u as v, we obtain the following criterion, assuming that we have one p-vector X to be classified into π_1 or π_2:

$$\begin{aligned} A_1 : v \geq k,\ A_2 : v < k,\ k &= \ln \frac{q_2 C(1|2)}{q_1 C(2|1)}, \\ v &= n_{(2)} X' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - n_{(2)} \tfrac{1}{2}(\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} + \bar{X}^{(2)}) \\ &= n_{(2)}[X - \tfrac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})]' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}). \end{aligned} \tag{12.5.4}$$

As it turns out, it already proves quite challenging to obtain the exact distribution of v as given in (12.5.4) where X is a single p-vector either from π_1 or from π_2.

12.5.1. Some asymptotic results

Before considering asymptotic properties of u and v as defined in Sect. 12.4, let us recall certain results obtained in earlier chapters. Let the $p \times 1$ vectors $Y_j,\ j = 1, \ldots, n$, be iid vectors from some population for which $E[Y_j] = \mu$ and $\mathrm{Cov}(Y_j) = \Sigma > O,\ j = 1, \ldots, n$. Let the sample matrix, the matrix of sample means wherein the sample mean $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$ and the sample sum of products matrix S be the as follows:

$$\begin{aligned} &\mathbf{Y} = [Y_1, \ldots, Y_n],\ \bar{\mathbf{Y}} = [\bar{Y}, \ldots, \bar{Y}],\ S = (s_{ij}),\ s_{ij} = \sum_{k=1}^{n} (y_{ik} - \bar{y}_i)(y_{jk} - \bar{y}_j), \\ &S = [\mathbf{Y} - \bar{\mathbf{Y}}][\mathbf{Y} - \bar{\mathbf{Y}}]' = \mathbf{Y}\,[\,I_n - JJ'/n\,]\,\mathbf{Y},\ Y_j' = [y_{1j}, y_{2j}, \ldots, y_{pj}], \end{aligned} \tag{i}$$

where J is a $n \times 1$ vector of unities. Since a matrix of the form $\mathbf{Y} - \bar{\mathbf{Y}}$ is present, we may let $\mu = O$ without any loss of generality in the following computations since $Y_j - \bar{Y} = (Y_j - \mu) - (\bar{Y} - \mu)$. Note that $B = B' = I_n - \frac{1}{n}JJ' = B^2$ and hence, B is idempotent and of rank $n - 1$. Since $B = B'$, there exists an orthonormal matrix Q such that $Q'BQ = \mathrm{diag}(1, \ldots, 1, 0) = D,\ QQ' = I,\ Q'Q = I$, the diagonal elements being 1's and 0 since $B = B^2$ and of rank $n - 1$. Then,

$$\begin{aligned} S &= \mathbf{Y}\,Q\,\mathrm{diag}(1, \ldots, 1, 0)\,Q'\,\mathbf{Y}' = \mathbf{Y}QDD'Q'\mathbf{Y}', \\ D &= \mathrm{diag}(1, \ldots, 1, 0). \end{aligned} \tag{ii}$$

Consider $\Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}}$. Let $U_j = \Sigma^{-\frac{1}{2}} Y_j,\ j = 1, \ldots, n$, where Y_j is the j-th column of $\mathbf{Y}$ and it is assumed that $\mu = O$. Observe that $E[U_j] = O, \mathrm{Cov}(U_j) = I_p,\ j = 1, \ldots, n$, and the U_j's are uncorrelated. Letting $\mathbf{U} = [U_1, \ldots, U_n]$, *(ii)* implies that

$$\Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}} = \mathbf{U}QDDQ'\mathbf{U}'. \tag{iii}$$

Denoting by $U_{(j)}$ the j-th row of $\mathbf{U}$, it follows that the elements of $U_{(j)}$ are iid uncorrelated real scalar variables with mean value zero and variance 1. Consider the transformation $V_{(j)} = U_{(j)}Q$; then $E[V_{(j)}] = O$ and $\mathrm{Cov}[V_{(j)}] = I_n,\ j = 1, \ldots, p$, the $V_{(j)}$'s being the uncorrelated. Let $\mathbf{V}$ be the $p \times n$ matrix whose rows are $V_{(j)},\ j = 1, \ldots, p$. Let the columns of $\mathbf{V}$ be $V_j,\ j = 1, \ldots, n$, that is, $\mathbf{V} = [V_1, \ldots, V_n]$. Then, *(iii)* implies the following:

$$\begin{aligned}
\Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}} &= \mathbf{V} D D' \mathbf{V}' = \{[V_1, \ldots, V_n]D\}\{[V_1, \ldots, V_n]D\}' \\
&= [V_1, \ldots, V_{n-1}, O][V_1, \ldots, V_{n-1}, O]' = V_1 V_1' + \cdots + V_{n-1} V_{n-1}' \Rightarrow \\
E[\Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}}] &= E[V_1 V_1'] + \cdots + E[V_{n-1} V_{n-1}'] = I_p + \cdots + I_p = (n-1) I_p \Rightarrow \\
E[S] &= (n-1)\Sigma \ \text{ or } \ E\Big[\frac{S}{n-1}\Big] = \Sigma. \qquad (iv)
\end{aligned}$$

Additionally,

$$\begin{aligned}
\mathrm{Cov}(\bar{Y}) &= \frac{1}{n^2}\mathrm{Cov}[Y_1 + \cdots + Y_n] = \frac{1}{n^2}[\mathrm{Cov}(Y_1) + \cdots + \mathrm{Cov}(Y_n)] \\
&= \frac{1}{n^2}[\Sigma + \cdots + \Sigma] = \frac{n}{n^2}\Sigma = \frac{\Sigma}{n} \to O \text{ as } n \to \infty, \qquad (v)
\end{aligned}$$

when Σ is finite with respect to any norm of Σ, namely $\|\Sigma\| < \infty$. Appealing to the extended Chebyshev inequality, this shows that the unbiased estimator of μ, namely $\bar{Y}$, converges to μ in probability, that is,

$$Pr(\bar{Y} \to \mu) \to 1 \text{ when } n \to \infty \text{ or } \lim_{n\to\infty} Pr(\bar{Y} \to \mu) = 1. \qquad (vi)$$

An unbiased estimator of Σ is $\hat{\Sigma} = \frac{S}{n-1}$ with $E[\hat{\Sigma}] = \Sigma$. Will $\hat{\Sigma}$ also converge to Σ in probability when $n \to \infty$? In order to establish this, we require the covariance structure of the elements in S. For arbitrary populations, it is somewhat difficult to verify this result; however, it is rather straightforward for normal populations. We will examine this aspect next.

12.5.2. Another method

Let the $p \times 1$ vectors $X_j,\ j = 1, \ldots, n$, be a simple random sample of size n from a population having a real $N_p(\mu, \Sigma),\ \Sigma > O$, distribution. Letting S denote the sample sum of products matrix, S will be distributed as a Wishart matrix with $m = n - 1$ degrees of freedom and $\Sigma > O$ as its parameter matrix, whose density is

$$f(S) = \frac{1}{2^{\frac{mp}{2}} |\Sigma|^{\frac{m}{2}} \Gamma_p(\frac{m}{2})} |S|^{\frac{m}{2} - \frac{p+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S)},\ S > O,\ m \geq p; \qquad (i)$$

the reader may also refer to real matrix-variate gamma density discussed in Chap. 5. This is usually written as $S \sim W_p(m, \Sigma),\ \Sigma > O$. Letting $S_{(*)} = \Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}}$, $S_{(*)} \sim W_p(m, I)$. Consider the transformation $S_{(*)} = TT'$ where $T = (t_{ij})$ is a lower triangular matrix whose diagonal elements are positive, that is, $t_{ij} = 0,\ i < j$, and $t_{jj} > 0,\ j = 1, \ldots, p$. It was explained in Chaps. 1 and 3 that the t_{ij}'s are mutually independently distributed with the t_{ij}'s such that $i > j$ distributed as standard normal variables and t_{jj}^2, as a chisquare variable having $m - (j-1)$ degrees of freedom. The j-th diagonal element of TT' is of the form $t_{j1}^2 + \cdots + t_{jj-1}^2 + t_{jj}^2$ where $t_{jk}^2 \sim \chi_1^2$, for $k = 1, \ldots, j-1$, that is, the square of a real standard normal variable. Thus, the j-th diagonal element is distributed as $\chi_1^2 + \cdots + \chi_1^2 + \chi_{m-(j-1)}^2 \sim \chi_m^2$ since all the individual chisquare variables are independently distributed, in which case the resulting number of degrees of freedom is the sum of the degrees of freedom of the chisquares. Now, noting that for a χ_ν^2,

$$E[\chi_\nu^2] = \nu \text{ and } \mathrm{Var}(\chi_\nu^2) = 2\nu, \tag{ii}$$

the expected value of each of the diagonal elements in TT', which are the diagonal elements in $S_{(*)}$, will be $m = n - 1$. The non-diagonal elements in TT' result from a sum of terms of the form $t_{ik}t_{ii},\ k < i$, whose expected value is $E[t_{ik}t_{ii}] = E[t_{ik}]E[t_{jj}]$; but since $E[t_{ik}] = 0,\ i > k$, all the non-diagonal elements will have zero as their expected values. Accordingly,

$$E[S_{(*)}] = \mathrm{diag}(m, \ldots, m) \Rightarrow E\Big[\frac{S_{(*)}}{m}\Big] = I \Rightarrow E\Big[\frac{S}{m}\Big] = \Sigma,\ m = n-1, \tag{iii}$$

and the estimator $\hat{\Sigma} = \frac{S}{m}$ is unbiased for Σ, m being equal to $n-1$. Now, let us examine the covariance structure of $S_{(*)}$. Let W denote a single vector comprising all the distinct elements of $S_{(*)} = TT'$ and consider its covariance structure. In this vector of order $\frac{p(p+1)}{2} \times 1$, convert all the original t_{ij}'s and t_{jj}'s in terms of standard normal and chisquare variables. Let $z_1, \ldots, z_{\frac{p(p-1)}{2}}$ be the standard normal variables and $y_1, \ldots, y_p$ denote the chisquare variables. Then, each element of $\mathrm{Cov}(W) = [W - E(W)][W - E(W)]'$ will be a sum of terms of the type

$$[\mathrm{Var}(y_k)][\mathrm{Var}(z_j)] = \mathrm{Var}(y_k) = [\text{twice the number of degrees of freedom of } y_k], \tag{iv}$$

which happens to be a linear function of m. Our estimator being $\hat{\Sigma} = \frac{S}{m} = \Sigma^{\frac{1}{2}} \frac{S_{(*)}}{m} \Sigma^{\frac{1}{2}}$, the covariance structure of $\frac{S_{(*)}}{m}$ which is $\frac{1}{m^2}\mathrm{Cov}(W)$ tends to O when $m \to \infty$, since each element of $\mathrm{Cov}(W)$ is of the form $a\,m + b$ where a and b are real scalars, so that $\frac{a\,m+b}{m^2} \to 0$ as $m \to \infty$, or equivalently, as $n \to \infty$ since $m = n - 1$. Thus, it follows from an extended version of Chebyshev's inequality that

$$Pr\Big(\frac{S}{m} \to \Sigma\Big) \to 1 \text{ as } m \to \infty \text{ or as } n \to \infty \text{ since } m = n-1. \tag{v}$$

These last two results are stated next as a theorem.

Theorem 12.5.1. *Let the $p \times 1$ vectors $X_j,\ j = 1, \ldots, n$, be iid with $E[X_j] = \mu$ and $\mathrm{Cov}(X_j) = \Sigma,\ j = 1, \ldots, n$. Assume that Σ is finite in the sense that $\|\Sigma\| < \infty$. Then, letting $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ denote the sample mean,*

$$Pr(\bar{X} \to \mu) \to 1 \text{ as } n \to \infty. \tag{12.5.5}$$

Further, letting $X_j \sim N_p(\mu, \Sigma),\ \Sigma > O$,

$$Pr\Big(\hat{\Sigma} = \frac{S}{m} \to \Sigma\Big) \to 1 \text{ as } m \to \infty \text{ or as } n \to \infty \text{ since } m = n-1. \tag{12.5.6}$$

Let us now examine the criterion in (12.5.4). In this case, we can obtain an asymptotic distribution of the criterion v for large $n_{(2)}$ or when $n_{(2)} \to \infty$ in the sense that $n_1 \to \infty$ and $n_2 \to \infty$. When $n_{(2)} \to \infty$, we have $\bar{X}^{(1)} \to \mu^{(1)},\ \bar{X}^{(2)} \to \mu^{(2)}$ and $\frac{S}{n_{(2)}} \to \Sigma$, so that the criterion v in (12.5.4) becomes

$$\begin{aligned} u &= (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}X - \tfrac{1}{2}(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} + \mu^{(2)}) \\ &= [X - \tfrac{1}{2}(\mu^{(1)} - \mu^{(2)})]'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}), \end{aligned} \tag{12.5.7}$$

which is nothing but u as specified in (12.3.7) with the densities $N_1(\frac{1}{2}\Delta^2, \Delta^2)$ in π_1 and $N_1(-\frac{1}{2}\Delta^2, \Delta^2)$ in π_2. Hence, the following result:

Theorem 12.5.2. *When $n_1 \to \infty$ and $n_2 \to \infty$, the criterion v provided in (12.5.4) becomes u as specified in (12.5.7) with the univariate normal densities $N_1(\frac{1}{2}\Delta^2, \Delta^2)$ in π_1 and $N_1(-\frac{1}{2}\Delta^2, \Delta^2)$ in π_2, where Δ^2 is Mahalanobis' distance given in (12.3.8). We classify X, the observation vector at hand, to π_1 when $X \in A_1$ and, to π_2 when $X \in A_2$ where $A_1 : u \geq k$ and $A_2 : u < k$ with $k = \ln \frac{C(1|2)\,q_2}{C(2|1)\,q_1}$, q_1 and q_2 being the prior probabilities of selecting the populations π_1 and π_2, respectively, and $C(2|1)$ and $C(1|2)$ denoting the costs or loss associated with misclassification.*

In a practical situation, when n_1 and n_2 are large, we may replace Δ^2 in Theorem 12.5.2 by the corresponding sample value $n_{(2)}(\bar{X}^{(1)} - \bar{X}^{(2)})'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$ where $S = S_1 + S_2$ and $n_{(2)} = n_1 + n_2 - 2$ and utilize the criterion u as specified in (12.5.7) to classify the given vector X into π_1 and π_2. It is assumed that $q_1,\ q_2,\ C(2|1)$ and $C(1|2)$ are available.

12.5.3. A new sample from π_1 or π_2

As in Examples 12.3.1 and 12.3.2, suppose that a simple random sample of size n_3 is available either from $\pi_1 : N_p(\mu^{(1)}, \Sigma)$ or from $\pi_2 : N_p(\mu^{(2)}, \Sigma)$, $\Sigma > O$. Letting the new sample be $X_1^{(3)}, \ldots, X_{n_3}^{(3)}$, the $p \times n_3$ sample matrix, the sample mean $\bar{X}^{(3)} = \frac{1}{n_3}\sum_{j=1}^{n_3} X_j^{(3)}$, the $p \times n_3$ matrix of sample means and the sample sum of products matrix are the following:

$$\begin{aligned}
\mathbf{X^{(3)}} &= [X_1^{(3)}, \ldots, X_{n_3}^{(3)}],\ \bar{\mathbf{X}}^{\mathbf{(3)}} = [\bar{X}^{(3)}, \bar{X}^{(3)}, \ldots, \bar{X}^{(3)}],\\
S_3 &= [\mathbf{X^{(3)}} - \bar{\mathbf{X}}^{\mathbf{(3)}}][\mathbf{X^{(3)}} - \bar{\mathbf{X}}^{\mathbf{(3)}}]' = (s_{ij}^{(3)}),\\
s_{ij}^{(3)} &= \sum_{k=1}^{n_3}(x_{ik}^{(3)} - \bar{x}_i^{(3)})(x_{jk}^{(3)} - \bar{x}_j^{(3)}).
\end{aligned} \tag{12.5.8}$$

An unbiased estimate from this third sample is $\hat{\Sigma} = \frac{S_3}{n_3 - 1}$, as $E[\hat{\Sigma}] = \Sigma$. A pooled estimate of Σ obtained from the three samples is

$$\frac{S_1 + S_2 + S_3}{n_1 + n_2 + n_3 - 3} \equiv \frac{S}{n_{(3)}},\ S = S_1 + S_2 + S_3,\ n_{(3)} = n_1 + n_2 + n_3 - 3. \tag{12.5.9}$$

Then, the criterion corresponding to (12.3.4) changes to:

$$A_1 : w \geq k \text{ and } A_2 : w < k,\ k = \ln \frac{C(1|2)\, q_2}{C(2|1)\, q_1}, \tag{12.5.10}$$

where

$$w = n_{(3)}[\bar{X}^{(3)} - \tfrac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})]' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) \tag{12.5.11}$$

with $S = S_1 + S_2 + S_3$, $n_{(3)} = n_1 + n_2 + n_3 - 3$ and $\bar{X}^{(3)}$ being the sample average from the third sample, which either comes from $\pi_1 : N_p(\mu^{(1)}, \Sigma)$ or $\pi_2 : N_p(\mu^{(2)}, \Sigma)$, $\Sigma > O$. Thus, the classification rule is the following:

$$A_1 : w \geq k \ \text{ and } \ A_2 : w < k,\ k = \ln \frac{C(1|2)\, q_2}{C(2|1)\, q_1}, \tag{12.5.12}$$

w being as defined in (12.5.11). That is, classify the new sample into π_1 if $w \geq k$ and, into π_2 if $w < k$.

As was explained in Sect. 12.5.2, as $n_j \to \infty$, $j = 1, 2$, $\bar{X}^{(i)} \to \mu^{(i)}$, $i = 1, 2$, and although n_3 usually remains finite, as $n_1 \to \infty$ and $n_2 \to \infty$, we have $n_{(3)} \to \infty$ and

$\frac{S}{n_{(3)}} \to \Sigma$. Accordingly, the criterion w as given in (12.5.11) converges to w_1 for large values of n_1 and n_2, where

$$w_1 = [\bar{X}^{(3)} - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})]' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}). \tag{12.5.13}$$

Compared to u as specified in (12.3.7), the only difference is that X associated with u is replaced by $\bar{X}^{(3)}$ in w_1. Hence, the variance in u will be multiplied by $\frac{1}{n_3}$, and the asymptotic distributions will be as follows:

$$w_1|\pi_1 \sim N_1\Big(\frac{1}{2}\Delta^2, \frac{1}{n_3}\Delta^2\Big) \text{ and } w_1|\pi_2 \sim N_1\Big(-\frac{1}{2}\Delta^2, \frac{1}{n_3}\Delta^2\Big), \tag{12.5.14}$$

as $n_1 \to \infty$ and $n_2 \to \infty$.

Theorem 12.5.3. *Consider two populations $\pi_1 : N_p(\mu^{(1)}, \Sigma)$ and $\pi_2 : N_p(\mu^{(2)}, \Sigma)$, $\Sigma > O$, and simple random samples of respective sizes n_1 and n_2 from these two populations. Suppose that a simple random sample of size n_3 is available, either from π_1 or π_2. For classifying the third sample into π_1 or π_2, the criterion to be utilized is w as given in (12.5.11). Then, the asymptotic distribution of w, when $n_i \to \infty$, $i = 1, 2$, is that of w_1 specified in (12.5.13) and the regions of classification are as given in (12.5.12).*

In a practical situation, when the sample sizes n_1 and n_2 are large, one may replace Δ^2 by its sample analogue, and then use (12.5.14) to reach a decision. As it turns out, it proves quite difficult to derive the exact density of w.

Example 12.5.1. A certain milk collection and distribution center collects and sells the milk supplied by local farmers to the community, the balance, if any, being dispatched to a nearby city. In that locality, there are two types of cows. Some farmers only keep Jersey cows and others, only Holstein cows. Samples of the same quantities of milk are taken and the following characteristics are evaluated: x_1, the fat content, x_2, the glucose content, and x_3, the protein content. It is known that $X' = (x_1, x_2, x_3)$ is normally distributed as $X \sim N_3(\mu^{(1)}, \Sigma)$, $\Sigma > O$, for Jersey cows, and $X \sim N_3(\mu^{(2)}, \Sigma)$, $\Sigma > O$, for Holstein cows, with $\mu^{(1)} \neq \mu^{(2)}$, the covariance matrices Σ being assumed identical. These parameters which are not known, are estimated on the basis of 100 milk samples from Jersey cows and 102 samples from Holstein cows, all the samples being of equal volume. The following are the summarized data with our standard notations, where S_1 and S_2 are the sample sums of products matrices:

$$\bar{X}^{(1)} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}, \ \bar{X}^{(2)} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \ S_1 = \begin{bmatrix} 50 & -50 & 50 \\ -50 & 100 & 0 \\ 50 & 0 & 150 \end{bmatrix}, \ S_2 = \begin{bmatrix} 150 & -150 & 150 \\ -150 & 300 & 0 \\ 150 & 0 & 450 \end{bmatrix}.$$

Three farmers just brought in their supply of milk and (1): a sample denoted by X_1 is collected from the first farmer's supply and evaluated; (2) a sample, X_2, is taken from a second farmer's supply and evaluated; (3) a set of 5 random samples are collected from a third farmer's supply, the sample average being $\bar{X}$. The data is

$$X_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix},\ X_2 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \text{ and } \bar{X} = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix},\ n = 5.$$

Classify, X_1, X_2 and the sample of size 5 to either coming from Jersey or Holstein cows.

Solution 12.5.1. The following preliminary calculations are needed:

$$\frac{S}{n_1+n_2-2} = \frac{S_1+S_2}{n_1+n_2-2} = \frac{S_1+S_2}{200} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 3 \end{bmatrix},$$

$$\Big(\frac{S}{200}\Big)^{-1} = \begin{bmatrix} 6 & 3 & -2 \\ 3 & 2 & -1 \\ -2 & -1 & 1 \end{bmatrix},\ \bar{X}^{(1)} - \bar{X}^{(2)} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix},\ \bar{X}^{(1)} + \bar{X}^{(2)} = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}.$$

Then,

$$(\bar{X}^{(1)} - \bar{X}^{(2)})'\Big(\frac{S}{200}\Big)^{-1} = [1, -1, 0]\begin{bmatrix} 6 & 3 & -2 \\ 3 & 2 & -1 \\ -2 & -1 & 1 \end{bmatrix} = [3, 1, -1],$$

$$\frac{1}{2}(\bar{X}^{(1)} - \bar{X}^{(2)})'\Big(\frac{S}{200}\Big)^{-1}(\bar{X}^{(1)} + \bar{X}^{(2)}) = \frac{1}{2}[3, 1, -1]\begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix} = 4,$$

$$(\bar{X}^{(1)} - \bar{X}^{(2)})'\Big(\frac{S}{200}\Big)^{-1}X = [3, 1, -1]X = 3x_1 + x_2 - x_3 \Rightarrow w = 3x_1 + x_2 - x_3 - 4$$

where the w is given in (12.5.11). For answering (1), we substitute X_1 to X in w. That is, w at X_1 is $3(2) + (1) - (1) - 4 = 2 > 0$. Hence, we assign X_1 to Jersey cows. For answering (2), we replace X in w by X_2, that is, $3(1) + (1) - (2) - 4 = -2 < 0$. Thus, we assign X_2 to Holstein cows. For answering (3), we replace X in w by $\bar{X}$. That is, $3(2) + (2) - (1) - 4 = 3 > 0$. Accordingly, we classify this sample as coming from Jersey cows.

12.6. Maximum Likelihood Method of Classification

As before, let π_1 be the p-variate real normal population $N_p(\mu^{(1)}, \Sigma)$, $\Sigma > O$, with the simple random sample $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ of size n_1 drawn from that population,

and $\pi_2 : N_p(\mu^{(2)}, \Sigma),\ \Sigma > O$, with the simple random sample $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ of size n_2 so distributed. A p-vector X at hand is to be classified into π_1 or π_2. Let the sample means and the sample sums of products matrices be $\bar{X}^{(1)},\ \bar{X}^{(2)},\ S_1$ and S_2. Then, the problem of classification of X into π_1 or π_2 can be stated in terms of testing a hypothesis of the following type: X and $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ are from $N_p(\mu^{(1)}, \Sigma)$ and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ are from π_2 constitutes the null hypothesis, versus, the alternative X and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ are from $N_p(\mu^{(2)}, \Sigma)$ and $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ are from $N_p(\mu^{(1)}, \Sigma)$. Let the likelihood functions under the null and alternative hypotheses be denoted as L_0 and L_1, respectively, where

$$L_0 = \Big\{\prod_{j=1}^{n_1} \frac{\mathrm{e}^{-\frac{1}{2}(X_j-\mu^{(1)})'\Sigma^{-1}(X_j-\mu^{(1)})}}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\Big\}\frac{\mathrm{e}^{-\frac{1}{2}(X-\mu^{(1)})'\Sigma^{-1}(X-\mu^{(1)})}}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}$$
$$\times \Big\{\prod_{j=1}^{n_2} \frac{\mathrm{e}^{-\frac{1}{2}(X_j-\mu^{(2)})'\Sigma^{-1}(X_j-\mu^{(2)})}}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\Big\},$$

$$L_0 = \frac{\mathrm{e}^{-\frac{1}{2}\rho_1}}{(2\pi)^{\frac{(n_1+n_2+1)p}{2}}|\Sigma|^{\frac{n_1+n_2+1}{2}}},\ \rho_1 = \nu + (X-\mu^{(1)})'\Sigma^{-1}(X-\mu^{(1)}), \qquad (i)$$

$$L_1 = \frac{\mathrm{e}^{-\frac{1}{2}\rho_2}}{(2\pi)^{\frac{(n_1+n_2+1)p}{2}}|\Sigma|^{\frac{n_1+n_2+1}{2}}},\ \rho_2 = \nu + (X-\mu^{(2)})\Sigma^{-1}(X-\mu^{(2)}), \qquad (ii)$$

where

$$\nu = \mathrm{tr}(\Sigma^{-1}S_1) + \frac{n_1}{2}(\bar{X}^{(1)}-\mu^{(1)})'\Sigma^{-1}(\bar{X}^{(1)}-\mu^{(1)})$$
$$+ \mathrm{tr}(\Sigma^{-1}S_2) + \frac{n_2}{2}(\bar{X}^{(2)}-\mu^{(2)})'\Sigma^{-1}(\bar{X}^{(2)}-\mu^{(2)}) \qquad (iii)$$

and S_1 and S_2 are the sample sums of products matrices from the samples $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$, respectively. Referring to Chaps. 1 and 3 for vector/matrix derivatives and the maximum likelihood estimators (MLE's) of the parameters of normal populations, the MLE's obtained from *(i)* are the following, denoting the estimators/estimates with a hat: The MLE's under L_0 are the following:

$$\hat{\mu}^{(1)} = \frac{n_1\bar{X}^{(1)}+X}{n_1+1},\ \hat{\mu}^{(2)} = \bar{X}^{(2)},\ \hat{\Sigma} = \frac{S_1+S_2+S_3^{(1)}}{n_1+n_2+1} \equiv \hat{\Sigma}_1,$$
$$S_3^{(1)} = (X-\hat{\mu}^{(1)})(X-\hat{\mu}^{(1)})' = \Big(X - \frac{n_1\bar{X}^{(1)}+X}{n_1+1}\Big)\Big(X - \frac{n_1\bar{X}^{(1)}+X}{n_1+1}\Big)'$$
$$= \Big(\frac{n_1}{n_1+1}\Big)^2(X-\bar{X}^{(1)})(X-\bar{X}^{(1)})', \qquad (12.6.1)$$

observing that the scalar quantity

$$(X-\hat{\mu}^{(1)})'\Sigma^{-1}(X-\hat{\mu}^{(1)}) = \mathrm{tr}(X-\hat{\mu}^{(1)})'\Sigma^{-1}(X-\hat{\mu}^{(1)}) = \mathrm{tr}(\Sigma^{-1}S_3^{(1)}).$$

By substituting the MLE's in L_0, we obtain the maximum of L_0:

$$\max L_0 = \frac{\mathrm{e}^{-\frac{(n_1+n_2+1)p}{2}}}{(2\pi)^{\frac{(n_1+n_2+1)p}{2}}|\hat{\Sigma}_1|^{\frac{(n_1+n_2+1)}{2}}},$$
$$\hat{\Sigma}_1 = \frac{S_1+S_2+(\frac{n_1}{n_1+1})^2(X-\bar{X}^{(1)})(X-\bar{X}^{(1)})'}{n_1+n_2+1}. \tag{12.6.2}$$

Under L_1, the MLE's are

$$\hat{\mu}^{(2)} = \frac{n_2\bar{X}^{(2)}+X}{n_2+1},\ \hat{\mu}^{(1)} = \bar{X}^{(1)},\ \hat{\Sigma} = \frac{S_1+S_2+S_3^{(2)}}{n_1+n_2+1} \equiv \hat{\Sigma}_2,$$
$$S_3^{(2)} = \Big(\frac{n_2}{n_2+1}\Big)^2(X-\bar{X}^{(2)})(X-\bar{X}^{(2)})'. \tag{12.6.3}$$

Thus,

$$\max L_1 = \frac{\mathrm{e}^{-\frac{(n_1+n_2+1)p}{2}}}{(2\pi)^{\frac{(n_1+n_2+1)p}{2}}|\hat{\Sigma}_2|^{\frac{n_1+n_2+1}{2}}},$$
$$\hat{\Sigma}_2 = \frac{1}{n_1+n_2+1}\Big[S_1+S_2+\Big(\frac{n_2}{n_2+1}\Big)^2(X-\bar{X}^{(2)})(X-\bar{X}^{(2)})'\Big].$$

Hence,

$$\lambda_1 = \frac{\max L_0}{\max L_1} = \Big(\frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|}\Big)^{\frac{n_1+n_2+1}{2}} \Rightarrow\ \lambda_1^{\frac{2}{n_1+n_2+1}} = z_1 = \frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|},\ \text{ so that}$$
$$z_1 = \frac{|S_1+S_2+(\frac{n_2}{n_2+1})^2(X-\bar{X}^{(2)})(X-\bar{X}^{(2)})'|}{|S_1+S_2+(\frac{n_1}{n_1+1})^2(X-\bar{X}^{(1)})(X-\bar{X}^{(1)})'|}. \tag{12.6.4}$$

If $z_1 \geq 1$, then $\max L_0 \geq \max L_1$, which means that the likelihood of X coming from π_1 is greater than or equal to the likelihood of X originating from π_2. Hence, we may classify X into π_1 if $z_1 \geq 1$ and classify X into π_2 if $z_1 < 1$. In other words,

$$A_1 : z_1 \geq 1 \ \text{ and } \ A_2 : z_1 < 1. \tag{iv}$$

If we let $S = S_1 + S_2$, then $z_1 \geq 1 \Rightarrow$

$$|S + \Big(\frac{n_2}{n_2+1}\Big)^2 (X - \bar{X}^{(2)})(X - \bar{X}^{(2)})'|$$
$$\geq |S + \Big(\frac{n_1}{n_1+1}\Big)^2 (X - \bar{X}^{(1)})(X - \bar{X}^{(1)})'|. \tag{v}$$

We can re-express this last inequality in a more convenient form. Expanding the following partitioned determinant in two different ways, we have the following, where S is $p \times p$ and Y is $p \times 1$:

$$\begin{vmatrix} S & -Y \\ Y' & 1 \end{vmatrix} = |S + YY'| = |S|\,|1 + Y'S^{-1}Y|$$
$$= |S|[1 + Y'S^{-1}Y], \tag{vi}$$

observing that $1 + Y'S^{-1}Y$ is a scalar quantity. Accordingly, $z_1 \geq 1$ means that

$$1 + \Big(\frac{n_2}{n_2+1}\Big)^2 (X - \bar{X}^{(2)})'S^{-1}(X - \bar{X}^{(2)}) \geq 1 + \Big(\frac{n_1}{n_1+1}\Big)^2 (X - \bar{X}^{(1)})'S^{-1}(X - \bar{X}^{(1)}).$$

That is,

$$z_2 = \Big(\frac{n_2}{n_2+1}\Big)^2 (X - \bar{X}^{(2)})'S^{-1}(X - \bar{X}^{(2)})$$
$$- \Big(\frac{n_1}{n_1+1}\Big)^2 (X - \bar{X}^{(1)})S^{-1}(X - \bar{X}^{(1)}) \geq 0 \ \Rightarrow$$
$$z_3 = \Big(\frac{n_2}{n_2+1}\Big)^2 (X - \bar{X}^{(2)})'\Big(\frac{S}{n_1+n_2-2}\Big)^{-1}(X - \bar{X}^{(2)})$$
$$- \Big(\frac{n_1}{n_1+1}\Big)^2 (X - \bar{X}^{(1)})'\Big(\frac{S}{n_1+n_2-2}\Big)^{-1}(X - \bar{X}^{(1)}) \geq 0. \tag{12.6.5}$$

Hence, the regions of classification are the following:

$$A_1 : z_3 \geq 0 \ \text{ and } \ A_2 : z_3 < 0. \tag{vii}$$

Thus, classify X into π_1 when $z_3 \geq 0$ and, X into π_2 when $z_3 < 0$. For large n_1 and n_2, some interesting results ensue. When $n_1 \to \infty$ and $n_2 \to \infty$, we have $\frac{n_i}{n_i+1} \to 1,\ i = 1, 2,\ \bar{X}^{(i)} \to \mu^{(i)},\ i = 1, 2$, and $\frac{S}{n_1+n_2-2} \to \Sigma$. Then, z_3 converges to z_4 where

$$z_4 = \tfrac{1}{2}(X - \mu^{(2)})'\Sigma^{-1}(X - \mu^{(2)}) - (X - \mu^{(1)})'\Sigma^{-1}(X - \mu^{(1)}) \geq 0 \tag{viii}$$
$$\Rightarrow [X - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})]'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \geq 0 \ \Rightarrow \ u \geq 0$$

where u is the same criterion u as that specified in (12.5.7). Hence, we have the following result:

Theorem 12.6.1. *Let $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ be a simple random sample of size n_1 from $\pi_1 : N_p(\mu^{(1)}, \Sigma),\ \Sigma > O$ and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ be a simple random sample of size n_2 from the population $\pi_2 : N_p(\mu^{(2)}, \Sigma),\ \Sigma > O$. Letting X be a vector at hand to be classified into π_1 or π_2, when $n_1 \to \infty$ and $n_2 \to \infty$, the likelihood ratio criterion for classification is the following: Classify X into π_1 if $u \geq 0$ and, X into π_2 if $u < 0$ or equivalently, $A_1 : u \geq 0$ and $A_2 : u < 0$ where $u = [X - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})]'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$ whose density is $u \sim N_1(\frac{1}{2}\Delta^2, \Delta^2)$ when X is assigned to π_1 and $u \sim N_1(-\frac{1}{2}\Delta^2, \Delta^2)$ when X is assigned to π_2, with $\Delta^2 = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$ denoting Mahalanobis' distance.*

The likelihood ratio criterion for classification specified in (12.6.5) can also be given the following interpretation: For large values of n_1 and n_2, the criterion reduces to the following: $(X - \mu^{(2)})'\Sigma^{-1}(X - \mu^{(2)}) - (X - \mu^{(1)})'\Sigma^{-1}(X - \mu^{(1)}) \geq 0$ where $(X - \mu^{(2)})'\Sigma^{-1}(X - \mu^{(2)})$ is the generalized distance between X and $\mu^{(2)}$, and $(X - \mu^{(1)})'\Sigma^{-1}(X - \mu^{(1)})$ is the generalized distance between X and $\mu^{(1)}$, which means that the generalized distance between X and $\mu^{(2)}$ is larger than the generalized distance between X and $\mu^{(1)}$ when $u > 0$. That is, X is closer to $\mu^{(1)}$ than $\mu^{(2)}$ and accordingly, we classify X into π_1, which is the case $u > 0$. Similarly, if X is closer to $\mu^{(2)}$ when compared to the distance to $\mu^{(1)}$, we assign X to π_2, which is the case $u < 0$. The case $u = 0$ is also included in the first inequality, but only for convenience. However, when $Pr\{u = 0|\pi_i,\ i = 1, 2\} = 0$, replacing $u > 0$ by $u \geq 0$ is fully justified.

Note 12.6.1. The reader may refer to Example 12.3.3 for an illustration of the computations involved in connection with the probabilities of misclassification. For large values of n_1 and n_2, one has the z_4 of *(viii)* as an approximation to the u appearing in the same equation as well as the u of (12.5.7) or that of Example 12.3.3. In order to apply Theorem 12.6.1, one needs to know the parameters $\mu^{(1)},\ \mu^{(2)}$ and Σ. When they are not available, one may substitute to them the corresponding estimates $\bar{X}^{(1)},\ \bar{X}^{(2)}$ and $\hat{\Sigma} = \frac{S_1+S_2}{n_1+n_2-2}$ when n_1 and n_2 are large. Then, the approximate probabilities of misclassification can be determined.

Example 12.6.1. Redo the problem considered in Example 12.5.1 by making use of the maximum likelihood procedure.

Solution 12.6.1. In order to answer the questions, we need to compute

$$z_4 = \Big(\frac{n_2}{n_2+1}\Big)^2 (X - \bar{X}^{(2)})'\Big(\frac{S}{n_1+n_2-1}\Big)^{-1}(X - \bar{X}^{(2)}) - \Big(\frac{n_1}{n_1+1}\Big)^2 (X - \bar{X}^{(1)})'\Big(\frac{S}{n_1+n_2-2}\Big)^2 (X - \bar{X}^{(1)}).$$

In this case, $\frac{n_1}{n_1+1} = \frac{100}{101} \approx 1$ and $\frac{n_2}{n_2+1} = \frac{102}{103} \approx 1$ and hence, the criterion z_4 is the same as w of (12.5.4) and the decisions arrived at an Example 12.5.1 will remain unchanged in this example. Since n_1 and n_2 are large, we have reasonably accurate approximations of the parameters as

$$\bar{X}^{(1)} \to \mu^{(1)},\ \bar{X}^{(2)} \to \mu^{(2)} \text{ and } \frac{S}{n_1+n_2-2} \to \Sigma,$$

so that the probabilities of misclassification can be evaluated by using their estimates. The approximate distributions are then given by

$$w|\pi_1 \sim N_1(\tfrac{1}{2}\hat{\Delta}^2, \hat{\Delta}^2) \text{ and } w|\pi_2 \sim N_1(-\tfrac{1}{2}\hat{\Delta}^2, \hat{\Delta}^2)$$

where $\hat{\Delta}^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})'(\frac{S}{n_1+n_2-2})^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$. From the computations done in Example 12.5.1, we have

$$(\bar{X}^{(1)} - \bar{X}^{(2)})' = [1, -1, 0],\ (\bar{X}^{(1)} - \bar{X}^{(2)})'\Big(\frac{S}{n_1+n_2-2}\Big)^{-1} = [3, 1, -1],$$

$$\hat{\Delta}^2 = [3, 1, -1]\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = 2$$

$$\Rightarrow w|\pi_1 \sim N_1(1, 2) \text{ and } w|\pi_2 \sim N_1(-1, 2), \text{ approximately.}$$

As well, $A_1 : w \geq 0$ and $A_2 : w < 0$. For the data pertaining to (1) of Example 12.5.1, we have $w > 0$ and X_1 is assigned to π_1. Observing that $w \to u$ of (12.5.7),

$$\begin{aligned} P(1|1, A) &= \text{Probability of arriving at a correct decision} \\ &= Pr\{u > 0|\pi_1\} = \int_0^\infty \frac{1}{\sqrt{2}\sqrt{(2\pi)}} \mathrm{e}^{-\frac{1}{4}(u-1)^2} \mathrm{d}u \\ &= \int_{\frac{0-1}{\sqrt{2}}}^\infty \frac{1}{\sqrt{(2\pi)}} \mathrm{e}^{-\frac{1}{2}v^2} \mathrm{d}v \approx 0.76; \\ P(1|2, A) &= \text{Probability of misclassification} \\ &= Pr\{u > 0|\pi_2\} = \int_0^\infty \frac{1}{\sqrt{2}\sqrt{(2\pi)}} \mathrm{e}^{-\frac{1}{4}(u+1)^2} \mathrm{d}u \\ &= \int_{\frac{1}{\sqrt{2}}}^\infty \frac{1}{\sqrt{(2\pi)}} \mathrm{e}^{-\frac{1}{2}v^2} \mathrm{d}v \approx 0.24. \end{aligned}$$

In Example 12.5.1, the observed vector provided for (2) is classified into π_2 since $w < 0$. Thus, the probability of making the right decision is $P(2|2, A) = Pr\{u < 0|\pi_2\} \approx 0.76$

and the probability of misclassification is $P(2|1, A) = Pr\{u < 0|\pi_1\} \approx 0.24$. Given the data related to (3) of Example 12.5.1, the only difference is that the distributions in π_1 and π_2 will be slightly different, the mean values remaining the same but the variance $\hat{\Delta}^2$ being replaced by $\hat{\Delta}^2/n$ where $n = 5$. The computations are similar to those provided for (1), the sample mean being assigned to π_1 in this case.

12.7. Classification Involving k Populations

Consider the p-variate populations $\pi_1, \ldots, \pi_k$ and let X be a p-vector at hand to be classified into one of these k populations. Let $q_1, \ldots, q_k$ be the prior probabilities of selecting these populations, $q_j > 0,\ j = 1, \ldots, k$, with $q_1 + \cdots + q_k = 1$. Let the cost of misclassification of a p-vector belonging to π_i being improperly classified into π_j be $C(j|i)$ for $i \neq j$ so that $C(i|i) = 0,\ i = 1, \ldots, k$. A decision rule $A = (A_1, \ldots, A_k)$ determines subspaces $A_j \subset R_p,\ j = 1, \ldots, k$, with $A_i \cap A_j = \phi$ (the empty set) for all $i \neq j$. Let the probability/density functions associated with the k populations be $P_j(X),\ j = 1, \ldots, k$, respectively. Let $P(j|i, A) = Pr\{X \in A_j|\pi_i : P_i(X), A\}$ = probability of an observation coming from or belonging to the population π_i or originating from the probability/density function $P_i(X)$, being improperly assigned to π_j or misclassified as coming from $P_j(X)$, and the cost associated with this misclassification be denoted by $C(j|i)$. Under the rule $A = (A_1, \ldots, A_k)$, the probabilities of correctly classifying and misclassifying an observed vector are the following, assuming that the $P_j(X)'$s, $j = 1, \ldots, k$, are densities:

$$P(i|i, A) = \int_{A_i} P_i(X)\mathrm{d}X \ \text{ and } \ P(j|i, A) = \int_{A_j} P_i(X)\mathrm{d}X,\ i, j = 1, \ldots, k, \qquad (i)$$

where $P(i|i, A)$ is a probability of achieving a correct classification, that is, of assigning an observation X to π_i when the population is actually π_i, and $P(j|i, A)$ is the probability of an observation X coming from π_i being misclassified as originating from π_j. Consider a p-vector X at hand. What is then the probability that this X came from $P_i(X)$, given that X is an observation vector from one of the populations $\pi_1, \ldots, \pi_k$? This is in fact a conditional statement involving

$$\frac{q_i P_i(X)}{q_1 P_1(X) + q_2 P_2(X) + \cdots + q_k P_k(X)}.$$

Suppose that for specific i and j, the conditional probability

$$\frac{q_i P_i(X)}{q_1 P_1(X) + \cdots + q_k P_k(X)} \geq \frac{q_j P_j(X)}{q_1 P_1(X) + \cdots + q_k P_k(X)}. \qquad (ii)$$

This is tantamount to presuming that the likeliness of X originating from $P_i(X)$ is greater than or equal to that of X coming from $P_j(X)$. In this case, we would like to assign X to

π_i rather than π_j. If *(ii)* holds for all $j = 1, \ldots, k,\ j \neq i$, then we classify X into π_i. Equation *(ii)* for $j = 1, \ldots, k,\ j \neq i$, implies that

$$q_i P_i(X) \geq q_j P_j(X) \Rightarrow \frac{P_i(X)}{P_j(X)} \geq \frac{q_j}{q_i},\ j = 1, \ldots, k,\ j \neq i. \tag{12.7.1}$$

Accordingly, we adopt (12.7.1) as a decision rule $A = (A_1, \ldots, A_k)$. This decision rule corresponds to the following: When $X \in A_1 \subset R_p$ or X falls in A_1, then X is classified into π_1, when $X \in A_2$, then X is assigned to π_2, and so on. What is the expected cost of an X belonging to π_i being misclassified into π_j under some decision rule $B = (B_1, \ldots, B_k),\ B_j \subset R_p,\ j = 1, \ldots, k,\ B_i \cap B_j = O,\ i \neq j$, for all i and j? This is $q_i P_i(X) C(j|i) \equiv E_i(B)$. The expected cost of an X belonging to π_j being misclassified into π_i under the same decision rule B is $E_j(B) = q_j P_j(X) C(i|j)$. If $E_i(B) < E_j(B)$, then we favor $P_i(X)$ over $P_j(X)$ as it is always desirable to minimize the expected cost in any procedure or decision. If $E_i(B) < E_j(B)$ for all $j = 1, \ldots, k,\ j \neq i$, then $P_i(X)$ or π_i is preferred over all other populations to which X could be assigned. Note that

$$E_i(B) < E_j(B) \Rightarrow q_i P_i(X) C(j|i) < q_j P_j(X) C(i|j) \Rightarrow \frac{P_i(X)}{P_j(X)} < \frac{q_j\, C(i|j)}{q_i\, C(j|i)}, \tag{iii}$$

for $j = 1, \ldots, k,\ j \neq i$, so that *(iii)* is the situation resulting from the following misclassification rule: if

$$\frac{P_i(X)}{P_j(X)} \geq \frac{q_j\, C(i|j)}{q_i\, C(j|i)},\ j = 1, \ldots, k,\ j \neq i, \tag{12.7.2}$$

we classify X into π_i or equivalently, $X \in A_i$, which is the decision rule $A = (A_1, \ldots, A_k)$. Thus, the decision rule B in *(iii)* is identical to A. Observing that when $C(i|j) = C(j|i)$, (12.7.2) reduces to (12.7.1); the decision rule $A = (A_1, \ldots, A_k)$ in (12.7.1) is seen to yield the maximum probability of assigning an observation X at hand to π_i compared to the probability of assigning X to any other $\pi_j,\ j = 1, \ldots, k,\ j \neq i$, when the costs of misclassification are equal. As well, it follows from (12.7.2) that the decision rule $A = (A_1, \ldots, A_k)$ gives the minimum expected cost associated with assigning the observation X at hand to π_i compared to assigning X to any other population $\pi_j,\ j = 1, \ldots, k\ , j \neq i$.

12.7.1. Classification when the populations are real Gaussian

Let the populations be p-variate real normal, that is, $\pi_j \sim N_p(\mu^{(j)}, \Sigma),\ \Sigma > O,\ j = 1, \ldots, k$, with different mean value vectors but the same covariance matrix $\Sigma > O$. Let the density of π_j be denoted by $P_j(X) \simeq N_p(\mu^{(j)}, \Sigma),\ \Sigma > O$. A vector X at hand is to be assigned to one of the π_i's, $i = 1, \ldots, k$. In Sect. 12.3 or Example 12.3.3, the decision

rule involves two populations. Letting the two populations be $\pi_i : P_i(X)$ and $\pi_j : P_j(X)$ for specific i and j, it was determined that the decision rule consists of classifying X into π_i if $\ln \frac{P_i(X)}{P_j(X)} \geq \ln \rho$, $\rho = \frac{q_j C(i|j)}{q_i C(j|i)}$, with $\rho = 1$ so that $\ln \rho = 0$ whenever $C(i|j) = C(j|i)$ and $q_i = q_j$. When $\ln \rho = 0$, we have seen that the decision rule is to classify the p-vector X into π_i or $P_i(X)$ if $u_{ij}(X) \geq 0$ and to assign X to $P_j(X)$ or π_j if $u_{ij}(X) < 0$, where

$$\begin{aligned} u_{ij}(X) &= (\mu^{(i)} - \mu^{(j)})' \Sigma^{-1} X - \tfrac{1}{2}(\mu^{(i)} - \mu^{(j)})' \Sigma^{-1}(\mu^{(i)} + \mu^{(j)}) \\ &= [X - \tfrac{1}{2}(\mu^{(i)} + \mu^{(j)})]' \Sigma^{-1}(\mu^{(i)} - \mu^{(j)}). \end{aligned} \tag{iv}$$

Now, on applying the result obtained in *(iv)* to (12.7.1) and (12.7.2), one arrives at the following decision rule:

$$A_i : u_{ij}(X) \geq 0 \text{ or } A_i : u_{ij}(X) \geq \ln k,\ k = \frac{q_j\, C(i|j)}{q_i\, C(j|i)},\ j = 1, \ldots, k,\ j \neq i, \tag{12.7.3}$$

with $\ln \rho = 0$ occurring when $q_i = q_j$ and $C(i|j) = C(j|i)$.

Note 12.7.1. What will interchanging i and j in $u_{ij}(X)$ entail? Note that, as defined, $u_{ij}(X)$ involves the terms $(\mu^{(i)} - \mu^{(j)}) = -(\mu^{(j)} - \mu^{(i)})$ and $(\mu^{(i)} + \mu^{(j)})$, the latter being unaffected by the interchange of $\mu^{(i)}$ and $\mu^{(j)}$. Hence, for all i and j,

$$u_{ij}(X) = -u_{ji}(X),\ i \neq j. \tag{12.7.4}$$

When the underlying population is $X \sim N_p(\mu^{(i)}, \Sigma)$, $E[u_{ij}(X)|\pi_i] = \frac{1}{2}\Delta_{ij}^2$, which implies that $E[u_{ji}|\pi_i] = -\frac{1}{2}\Delta_{ij}^2 = -E[u_{ij}(X)|\pi_i]$ where $\Delta_{ij}^2 = (\mu^{(i)} - \mu^{(j)})' \Sigma^{-1}(\mu^{(i)} - \mu^{(j)})$.

Note 12.7.2. For computing the probabilities of correctly classifying and misclassifying an observed vector, certain assumptions regarding the distributions associated with the populations $\pi_j,\ j = 1, \ldots, k$, are needed, the normality assumption being the most convenient one.

Example 12.7.1. A certain milk collection and distribution center collects and sells the milk supplied by local farmers to the community, the balance, if any, being dispatched to a nearby city. In that locality, there are three dairy cattle breeds, namely, Jersey, Holstein and Guernsey, and each farmer only keeps one type of cows. Samples are taken and the following characteristics are evaluated in grams per liter: x_1, the fat content, x_2, the glucose content, and x_3, the protein content. It has been determined that $X' = (x_1, x_2, x_3)$

is normally distributed as $X \sim N_3(\mu^{(1)}, \Sigma)$ for Jersey cows, $X \sim N_3(\mu^{(2)}, \Sigma)$ for Holstein cows and $X \sim N_3(\mu^{(3)}, \Sigma)$ for Guernsey cows, with a common covariance matrix $\Sigma > O$, where

$$\mu^{(1)} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix},\ \mu^{(2)} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix},\ \mu^{(3)} = \begin{bmatrix} 2 \\ 3 \\ 3 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

(1): A farmer brought in his supply of milk from which one liter was collected. The three variables were evaluated, the result being $X_0' = (2, 3, 4)$. (2): Another one liter sample was taken from a second farmer's supply and it was determined that the vector of the resulting measurements was $X_1' = (2, 2, 2)$. No prior probabilities or costs are involved. Which breed of dairy cattle is each of these farmers likely to own?

Solution 12.7.1. Our criterion is based on $u_{ij}(X)$ where

$$u_{ij}(X) = (\mu^{(i)} - \mu^{(j)})'\Sigma^{-1}X - \frac{1}{2}(\mu^{(i)} - \mu^{(j)})'\Sigma^{-1}(\mu^{(i)} + \mu^{(j)}).$$

Let us evaluate the various quantities of interest:

$$\mu^{(1)} - \mu^{(2)} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},\ \mu^{(1)} - \mu^{(3)} = \begin{bmatrix} 0 \\ 0 \\ -2 \end{bmatrix},\ \mu^{(2)} - \mu^{(3)} = \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix},$$
$$\mu^{(1)} + \mu^{(2)} = \begin{bmatrix} 3 \\ 6 \\ 3 \end{bmatrix},\ \mu^{(1)} + \mu^{(3)} = \begin{bmatrix} 4 \\ 6 \\ 4 \end{bmatrix},\ \mu^{(2)} + \mu^{(3)} = \begin{bmatrix} 3 \\ 6 \\ 5 \end{bmatrix};$$

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix};$$

$$A_1 : \{u_{12}(X) \geq 0, u_{13}(X) \geq 0\},\ A_2 : \{u_{21}(X) \geq 0, u_{23}(X) \geq 0\},$$
$$A_3 : \{u_{31}(X) \geq 0, u_{32}(X) \geq 0\};$$

$$(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}X = (1, 0, -1)\Sigma^{-1}X = (\tfrac{1}{3}, -1, -2)X = \tfrac{1}{3}x_1 - x_2 - 2x_3$$
$$(\mu^{(1)} - \mu^{(3)})'\Sigma^{-1}X = (0, 0, -2)\Sigma^{-1}X = (0, -2, -4)X = -2x_2 - 4x_3$$
$$(\mu^{(2)} - \mu^{(3)})'\Sigma^{-1}X = (-1, 0, -1)\Sigma^{-1}X = (-\tfrac{1}{3}, -1, -2)X = -\tfrac{1}{3}x_1 - x_2 - 2x_3;$$

$$\frac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)})=\frac{1}{2}[\tfrac{1}{3},-1,-2]\begin{bmatrix}3\\6\\3\end{bmatrix}=-\frac{11}{2}$$

$$\frac{1}{2}(\mu^{(1)}-\mu^{(3)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(3)})=\frac{1}{2}[0,-2,-4]\begin{bmatrix}4\\6\\4\end{bmatrix}=-14$$

$$\frac{1}{2}(\mu^{(2)}-\mu^{(3)})'\Sigma^{-1}(\mu^{(2)}+\mu^{(3)})=\frac{1}{2}[-\tfrac{1}{3},-1,-2]\begin{bmatrix}3\\6\\5\end{bmatrix}=-\frac{17}{2}.$$

Hence,

$$u_{12}(X)=\tfrac{1}{3}x_1-x_2-2x_3+\tfrac{11}{2};\quad u_{13}(X)=-2x_2-4x_3+14;$$
$$u_{21}(X)=-\tfrac{1}{3}x_1+x_2+2x_3-\tfrac{11}{2};\quad u_{23}(X)=-\tfrac{1}{3}x_1-x_2-2x_3+\tfrac{17}{2};$$
$$u_{31}(X)=2x_2+4x_3-14;\quad u_{32}(X)=\tfrac{1}{3}x_1+x_2+2x_3-\tfrac{17}{2}.$$

In order to answer (1), we substitute X_0 to X and first, evaluate $u_{12}(X_0)$ and $u_{13}(X_0)$ to determine whether they are ≥ 0. Since $u_{12}(X_0)=\frac{1}{3}(2)-(3)-2(4)+\frac{11}{2}<0$, the condition is violated and hence we need not check for $u_{13}(X_0)\geq 0$. Thus, X_0 is not in A_1. Now, consider $u_{21}(X_0)=-\frac{1}{3}(2)+3+2(4)-\frac{11}{2}>0$ and $u_{23}(X_0)=-\frac{1}{3}(2)-(3)-2(4)-\frac{17}{2}<0$; again the condition is violated and we deduce that X_0 is not in A_2. Finally, we verify A_3: $u_{31}(X_0)=2(3)+2(4)-14=0$ and $u_{32}(X_0)=\frac{1}{3}(2)+(3)+2(4)-\frac{17}{2}>0$. Thus, $X_0\in A_3$, that is, we conclude that the sample milk came from Guernsey cows.

For answering (2), we substitute X_1 to X in $u_{ij}(X)$. Noting that $u_{12}(X_1)=\frac{1}{3}(2)-(2)-2(2)+\frac{11}{2}>0$ and $u_{13}(X_1)=-2(2)-4(2)+14>0$, we can surmise that $X_1\in A_1$, that is, the sample milk came from Jersey cows. Let us verify A_2 and A_3 to ascertain that no mistake has been made in the calculations. Since $u_{21}(X_1)<0$, X_1 is not in A_2, and since $u_{31}(X_0)<0$, X_1 is not in A_3. This completes the computations.

12.7.2. Some distributional aspects

For computing the probabilities of correctly classifying and misclassifying an observation, we require the distributions of our criterion $u_{ij}(X)$. Let the populations be normally distributed, that is, $\pi_j\sim N_p(\mu^{(j)},\Sigma)$, $\Sigma>O$, with the same covariance matrix Σ for all k populations, $j=1,\ldots,k$. Then, the probability of achieving a correct classification when X is assigned to π_i is the following under the decision rule $A=(A_1,\ldots,A_k)$:

$$P(i|i,A)=\int_{A_i}P_i(X)\mathrm{d}X \tag{12.7.5}$$

where $\mathrm{d}X = \mathrm{d}x_1 \wedge \ldots \wedge \mathrm{d}x_p$ and the integral is actually a multiple integral. But A_i is defined by the inequalities $u_{i1}(X) \geq 0,\ u_{i2}(X) \geq 0, \ldots, u_{ik}(X) \geq 0$, where $u_{ii}(X)$ is excluded. This is the case when no prior probabilities and costs are involved or when the prior probabilities are equal and the cost functions are identical. Otherwise, the region is $\{A_i : u_{ij}(X) \geq \ln k_{ij},\ k_{ij} = \frac{q_j C(i|j)}{q_i C(j|i)},\ j = 1, \ldots, k,\ j \neq i\}$. Integrating (12.7.5) is challenging as the region is determined by $k-1$ inequalities.

When the parameters $\mu^{(j)},\ j = 1, \ldots, k$, and Σ are known, we can evaluate the joint distributions of $u_{ij}(X),\ j = 1, \ldots, k,\ j \neq i$, under the normality assumption for $\pi_j,\ j = 1, \ldots, k$. Let us examine the distributions of $u_{ij}(X)$ for normally distributed $\pi_i : P_i(X), i = 1, \ldots, k$. In this instance, $E[X]|\pi_i = \mu^{(i)}$, and under π_i,

$$\begin{aligned} E[u_{ij}(X)]|\pi_i &= (\mu^{(i)} - \mu^{(j)})' \Sigma^{-1} \mu^{(i)} - \tfrac{1}{2}(\mu^{(i)} - \mu^{(j)})' \Sigma^{-1}(\mu^{(i)} + \mu^{(j)}) \\ &= \tfrac{1}{2}(\mu^{(i)} - \mu^{(j)})' \Sigma^{-1}(\mu^{(i)} - \mu^{(j)}) = \tfrac{1}{2}\Delta_{ij}^2; \\ \mathrm{Var}(u_{ij}(X))|\pi_i &= \mathrm{Var}[(\mu^{(i)} - \mu^{(j)})' \Sigma^{-1} X] = \Delta_{ij}^2. \end{aligned}$$

Since $u_{ij}(X)$ is a linear function of the vector normal variable X, it is normal and the distribution of $u_{ij}(X)|\pi_i$ is

$$u_{ij}(X) \sim N_1(\tfrac{1}{2}\Delta_{ij}^2, \Delta_{ij}^2),\ j = 1, \ldots, k,\ j \neq i. \tag{12.7.6}$$

This normality holds for each $j,\ j = 1, \ldots, k,\ j \neq i$, and for a fixed i. Then, we can evaluate the joint density of $u_{i1}(X), u_{i2}(X), \ldots, u_{ik}(X)$, excluding $u_{ii}(X)$, and we can evaluate $P(i|i, A)$ from this joint density. Observe that for $j = 1, \ldots, k,\ j \neq i$, the $u_{ij}(X)$'s are linear functions of the same vector normal variable X and hence, they have a joint normal distribution. In that case, the mean value vector is a $(k-1)$-vector, denoted by $\mu_{(ii)}$, whose elements are $\frac{1}{2}\Delta_{ij}^2,\ j = 1, \ldots, k,\ j \neq i$, for a fixed i, or equivalently,

$$\mu_{(ii)}' = [\tfrac{1}{2}\Delta_{i1}^2, \ldots, \tfrac{1}{2}\Delta_{ik}^2] = E[U_{ii}'] \ \text{ with } \ U_{ii}' = [u_{i1}(X), \ldots, u_{ik}(X)],$$

excluding the elements $u_{ii}(X)$ and $\Delta_{ii}^2 = 0$. The subscript ii in U_{ii} indicates the region A_i and the original population $P_i(X)$. The covariance matrix of U_{ii}, denoted by Σ_{ii}, will be a $(k-1) \times (k-1)$ matrix of the form $\Sigma_{ii} = [\mathrm{Cov}(u_{ir}, u_{it})] = (c_{rt}),\ c_{rt} = \mathrm{Cov}(u_{ir}(X), u_{it}(X))$. The subscript ii in Σ_{ii} indicates the region A_i and the original population $P_i(X)$. Observe that for two linear functions $t_1 = C'X = c_1 x_1 + \cdots + c_p x_p$ and $t_2 = B'X = b_1 x_1 + \cdots + b_p x_p$, having a common covariance matrix $\mathrm{Cov}(X) = \Sigma$, we have $\mathrm{Var}(t_1) = C'\Sigma C$, $\mathrm{Var}(t_2) = B'\Sigma B$ and $\mathrm{Cov}(t_1, t_2) = C'\Sigma B = B'\Sigma C$. Therefore,

$$c_{rt} = (\mu^{(i)} - \mu^{(r)})' \Sigma^{-1}(\mu^{(i)} - \mu^{(t)}),\ i \neq r, t;\ \ \Sigma_{ii} = (c_{rt}).$$

Let the vector U_{ii} be such that $U_{ii}' = (u_{i1}(X), \ldots, u_{ik}(X))$, excluding $u_{ii}(X)$. Thus, for a specific i,

$$U_{ii} \sim N_{k-1}(\mu_{(ii)}, \Sigma_{ii}),\ \Sigma_{ii} > O,$$

and its density function, denoted by $g_{ii}(U_{ii})$, is

$$g_{ii}(U_{ii}) = \frac{1}{(2\pi)^{\frac{k-1}{2}}|\Sigma_{ii}|^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}(U_{ii}-\mu_{(ii)})'\Sigma_{ii}^{-1}(U_{ii}-\mu_{(ii)})}.$$

Then,

$$\begin{aligned} P(i|i, A) &= \int_{u_{ij}(X)\geq 0,\ j=1,\ldots,k,\ j\neq i} g_{ii}(U_{ii})\mathrm{d}U_{ii} \\ &= \int_{u_{i1}(X)=0}^{\infty}\cdots\int_{u_{ik}(X)=0}^{\infty} g_{ii}(U_{ii})\mathrm{d}u_{i1}(X)\wedge\ldots\wedge\mathrm{d}u_{ik}(X), \end{aligned} \tag{12.7.7}$$

the differential $\mathrm{d}u_{ii}$ being absent from $\mathrm{d}U_{ii}$, which is also the case for $u_{ii}(X) \geq 0$ in the integral. If prior probabilities and cost functions are involved, then replace $u_{ij}(X) \geq 0$ in the integral (12.7.7) by $u_{ij}(X) \geq \ln k_{ij}$, $k_{ij} = \frac{q_j C(i|j)}{q_i C(j|i)}$. Thus, the problem reduces to determining the joint density $g_{ii}(U_{ii})$ and then evaluating the multiple integrals appearing in (12.7.7). In order to compute the probability specified in (12.7.7), we standardize the normal density by letting $V_{ii} = \Sigma_{ii}^{-\frac{1}{2}}U_{ii}$ where $V_{ii} \sim N_{k-1}(O, I)$, and with the help of this standard normal, we may compute this probability through V_{ii}. Note that (12.7.7) holds for each i, $i = 1, \ldots, k$, and thus, the probabilities of achieving a correct classification, $P(i|i, A)$ for $i = 1, \ldots, k$, are available from (12.7.7).

For computing probabilities of misclassification of the type $P(i|j, A)$, we can proceed as follows: In this context, the basic population is $\pi_j : P_j(X) \sim N_p(-\frac{1}{2}\Delta_{ij}^2, \Delta_{ij}^2)$, the region of integration being $A_i : \{u_{i1}(X) \geq 0, \ldots, u_{ik}(X) \geq 0\}$, excluding the element $u_{ii}(X) \geq 0$. Consider the vector U_{ij} corresponding to the vector U_{ii}. In U_{ij}, i stands for the region A_i and j, for the original population $P_j(X)$. The elements of U_{ij} are the same as those of U_{ii}, that is, $U_{ij}' = (u_{i1}(X), \ldots, u_{ik}(X))$, excluding $u_{ii}(X)$. We then proceed as before and compute the covariance matrix Σ_{ij} of U_{ij} in the original population $P_j(X)$. The variances of $u_{im}(X)$, $m = 1, \ldots, k$, $m \neq i$, will remain the same but the covariances will be different since they depend on the mean values. Thus, $U_{ij} \sim N_{k-1}(\mu_{(ij)}, \Sigma_{ij})$, and on standardizing, one has $V_{ij} \sim N_{k-1}(O, I)$, so that the required probability $P(i|j, A)$ can be computed from the elements of V_{ij}. Note that when the prior probabilities and costs are equal,

$$P(i|j,A) = \int_{u_{i1}(X)\geq 0,\ldots,u_{ik}(X)\geq 0} g_{ij}(U_{ij})\,\mathrm{d}u_{i1}(X)\wedge\ldots\wedge\mathrm{d}u_{ik}(X)$$
$$= \int_{u_{i1}(X)=0}^{\infty}\cdots\int_{u_{ik}(X)=0}^{\infty} g_{ij}(U_{ij})\,\mathrm{d}U_{ij}, \tag{12.7.8}$$

excluding $u_{ii}(X)$ in the integral as well as the differential $\mathrm{d}u_{ii}(X)$. Thus, $\mathrm{d}U_{ij} = \mathrm{d}u_{i1}(X)\wedge\ldots\wedge\mathrm{d}u_{ik}(X)$, excluding $\mathrm{d}u_{ii}(X)$.

Example 12.7.2. Given the data provided in Example 12.7.1, what is the probability of correctly assigning X to π_1? That is, compute the probability $P(1|1,A)$.

Solution 12.7.2. Observe that the joint density of $u_{12}(X)$ and $u_{13}(X)$ is that of a bivariate normal distribution since $u_{12}(X)$ and $u_{13}(X)$ are linear functions of the same vector X where X has a multivariate normal distribution. In order to compute the joint bivariate normal density, we need $E[u_{1j}(X)]$, $\mathrm{Var}(u_{1j}(X))$, $j=2,3$ and $\mathrm{Cov}(u_{12}(X),u_{13}(X))$. The following quantities are evaluated from the data given in Example 12.7.1:

$$\begin{aligned}\mathrm{Var}(u_{12}(X)) &= \mathrm{Var}[(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}X - \tfrac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}+\mu^{(2)})]\\ &= \mathrm{Var}[(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}X] = (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})\\ &= [\tfrac{1}{3},-1,-2]\begin{bmatrix}1\\0\\-1\end{bmatrix} = \frac{7}{3}\Rightarrow E[u_{12}(X)] = \frac{7}{6};\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(u_{13}(X)) &= (\mu^{(1)}-\mu^{(3)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(3)})\\ &= [0,-2,-4]\begin{bmatrix}0\\0\\-2\end{bmatrix} = 8\Rightarrow E[u_{13}(X)] = 4;\end{aligned}$$

$$\mathrm{Cov}(u_{12}(X),u_{13}(X)) = (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(3)}) = [\tfrac{1}{3},-1,-2]\begin{bmatrix}0\\0\\-2\end{bmatrix} = 4.$$

Hence, the covariance matrix of $U_{11} = \begin{bmatrix}u_{12}(X)\\u_{13}(X)\end{bmatrix}$, denoted by Σ_{11}, is the following:

$$\Sigma_{11} = \begin{bmatrix}\frac{7}{3} & 4\\ 4 & 8\end{bmatrix}\Rightarrow |\Sigma_{11}| = \frac{8}{3}$$
$$\Sigma_{11}^{-1} = \frac{3}{8}\begin{bmatrix}8 & -4\\ -4 & \frac{7}{3}\end{bmatrix} = \frac{1}{8}\begin{bmatrix}24 & -12\\ -12 & 7\end{bmatrix},$$

where

$$\frac{1}{8}\begin{bmatrix} 24 & -12 \\ -12 & 7 \end{bmatrix} = \frac{1}{8}\begin{bmatrix} 2\sqrt{6} & 0 \\ -\sqrt{6} & 1 \end{bmatrix}\begin{bmatrix} 2\sqrt{6} & -\sqrt{6} \\ 0 & 1 \end{bmatrix} = B'B \text{ with}$$

$$B = \frac{1}{\sqrt{8}}\begin{bmatrix} 2\sqrt{6} & -\sqrt{6} \\ 0 & 1 \end{bmatrix} \Rightarrow |B| = |B'| = |\Sigma_{11}|^{-\frac{1}{2}} = \frac{2\sqrt{6}}{8}.$$

The bivariate normal density of U_{11} is the following:

$$U_{11} = \begin{bmatrix} u_{12}(X) \\ u_{13}(X) \end{bmatrix} \sim N_2(\mu_{(1)}, \Sigma_{11}),\ \mu_{(1)} = \begin{bmatrix} 7/6 \\ 4 \end{bmatrix}, \tag{12.7.9}$$

with Σ_{11} and $\Sigma_{11}^{-1} = B'B$ as previously specified. Letting $Y = B(U_{11} - E[U_{11}])$, $Y \sim N_2(O, I)$. Note that

$$Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},\ U_{11} - E[U_{11}] = \begin{bmatrix} u_{12}(X) - E[u_{12}(X)] \\ u_{13}(X) - E[u_{13}(X)] \end{bmatrix} = \begin{bmatrix} u_{12}(X) - 7/6 \\ u_{13}(X) - 4 \end{bmatrix},$$

$$y_1 = \frac{2\sqrt{6}}{\sqrt{8}}(u_{12}(X) - 7/6) - \frac{\sqrt{6}}{\sqrt{8}}(u_{13}(X) - 4),$$

$$y_2 = \frac{\sqrt{6}}{\sqrt{8}}(u_{13}(X) - 4).$$

Then,

$$B^{-1} = \frac{\sqrt{8}}{2\sqrt{6}}\begin{bmatrix} 1 & \sqrt{6} \\ 0 & 2\sqrt{6} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \sqrt{2} \\ 0 & 2\sqrt{2} \end{bmatrix},$$

and we have

$$\begin{bmatrix} u_{12}(X) - 7/6 \\ u_{13}(X) - 4 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \sqrt{2} \\ 0 & 2\sqrt{2} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

which yields $u_{12}(X) = \frac{7}{6} + \frac{1}{\sqrt{3}}\, y_1 + \sqrt{2}\, y_2$ and $u_{13}(X) = 4 + 2\sqrt{2}\, y_2$. The intersection of the two lines corresponding to $u_{12}(X) = 0$ and $u_{13}(X) = 0$ is the point $(y_1, y_2) = (\sqrt{3}(\frac{5}{6}), -\sqrt{2})$. Thus, $u_{12}(X) \geq 0$ and $u_{13}(X) \geq 0$ give $y_2 \geq -\frac{4}{2\sqrt{2}} = -\sqrt{2}$ and $\frac{7}{6} + \frac{1}{\sqrt{3}}\, y_1 + \sqrt{2}\, y_2 \geq 0$. We can express the resulting probability as $\rho_1 - \rho_2$ where

$$\rho_1 = \int_{y_2=-\sqrt{2}}^{\infty}\int_{y_1=-\infty}^{\infty} \frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(y_1^2+y_2^2)}\mathrm{d}y_1 \wedge \mathrm{d}y_2 = 1 - \Phi(-\sqrt{2}), \tag{12.7.10}$$

which is explicitly available, where $\Phi(\cdot)$ denotes the distribution function of a standard normal variable, and

$$\begin{aligned}\rho_2 &= \int_{y_1=-\infty}^{\sqrt{3}(\frac{5}{6})} \int_{y_2=-\sqrt{2}}^{\frac{1}{\sqrt{2}}(\frac{7}{6}+\frac{1}{\sqrt{3}}y_1)} \tfrac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(y_1^2+y_2^2)}\mathrm{d}y_1 \wedge \mathrm{d}y_2 \\ &= \int_{y_1=-\infty}^{\sqrt{3}(\frac{5}{6})} \tfrac{1}{\sqrt{(2\pi)}}\mathrm{e}^{-\frac{1}{2}y_1^2}[\Phi(-\tfrac{1}{\sqrt{2}}(\tfrac{7}{6}+\tfrac{1}{\sqrt{3}}y_1)) - \Phi(-\sqrt{2})]\mathrm{d}y_1 \\ &= \int_{y_1=-\infty}^{\sqrt{3}(\frac{5}{6})} \tfrac{1}{\sqrt{(2\pi)}}\mathrm{e}^{-\frac{1}{2}y_1^2}\Phi(-\tfrac{1}{\sqrt{2}}(\tfrac{7}{6}+\tfrac{1}{\sqrt{3}}y_1))\mathrm{d}y_1 - \Phi(\sqrt{3}(\tfrac{5}{6}))\Phi(-\sqrt{2}). \qquad (12.7.11)\end{aligned}$$

Therefore, the required probability is

$$\begin{aligned}\rho_1 - \rho_2 &= 1 - \Phi(-\sqrt{2}) + \Phi(-\sqrt{2})\Phi(\sqrt{3}(\tfrac{5}{6})) \\ &\quad - \int_{y_1=-\infty}^{\sqrt{3}(\frac{5}{6})} \tfrac{1}{\sqrt{(2\pi)}}\mathrm{e}^{-\frac{1}{2}y_1^2}\,\Phi(-\tfrac{1}{\sqrt{2}}(\tfrac{7}{6}+\tfrac{1}{\sqrt{3}}y_1))\,\mathrm{d}y_1. \qquad (12.7.12)\end{aligned}$$

Note that all quantities, except the integral, are explicitly available from standard normal tables. The integral part can be read from a bivariate normal table. If a bivariate normal table is used, then one can approximate the required probability from (12.7.9). Alternatively, once evaluated numerically, the integral is found to be equal to 0.2182 which subtracted from 0.9941, yields a probability of 0.7759 for $P(1|1, A)$.

12.7.3. Classification when the population parameters are unknown

When training samples are available from the populations $\pi_i,\ i = 1, \ldots, k$, we can estimate the parameters and proceed with the classification. Let $X_j^{(i)},\ j = 1, \ldots, n_i$, be a simple random sample of size n_i from the i-th population π_i. Then, the sample average is $\bar{X}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_j^{(i)}$, and with our usual notations, the sample matrix, the matrix of sample means and sample sum of products matrix are the following:

$$\begin{aligned}\mathbf{X^{(i)}} &= [X_1^{(i)}, \ldots, X_{n_i}^{(i)}],\ \bar{\mathbf{X}}^{\mathbf{(i)}} = [\bar{X}^{(i)}, \ldots, \bar{X}^{(i)}], \\ S_i &= [\mathbf{X^{(i)}} - \bar{\mathbf{X}}^{\mathbf{(i)}}][\mathbf{X^{(i)}} - \bar{\mathbf{X}}^{\mathbf{(i)}}]',\ i = 1, \ldots, k,\end{aligned}$$

where

$$\mathbf{X^{(i)}} = \begin{bmatrix} x_{11}^{(i)} & \ldots & x_{1n_i}^{(i)} \\ \vdots & \ddots & \vdots \\ x_{p1}^{(i)} & \ldots & x_{pn_i}^{(i)} \end{bmatrix},\ S_i = (s_{rt}^{(i)}),\ s_{rt}^{(i)} = \sum_{m=1}^{n_i}(x_{rm}^{(i)} - \bar{x}_r^{(i)})(x_{tm}^{(i)} - \bar{x}_t^{(i)}),\ i = 1, \ldots, k.$$

Note that $\mathbf{X^{(i)}}$ and $\bar{\mathbf{X}}^{\mathbf{(i)}}$ are $p\times n_i$ matrices and $X_j^{(i)}$ is a $p\times 1$ vector for each $j=1,\ldots,n_i$, and $i=1,\ldots,k$. Let the population mean value vectors and the common covariance matrix be $\mu^{(1)},\ldots,\mu^{(k)}$, and $\Sigma>O$, respectively. Then, the unbiased estimators for these parameters are the following, identifying the estimators/estimates by a hat: $\hat{\mu}_j^{(i)}=\bar{X}^{(i)},\ i=1,\ldots,k$, and $\hat{\Sigma}=\frac{S}{n_1+\cdots+n_k-k},\ S=S_1+\cdots+S_k$. On replacing the population parameters by their unbiased estimators, the classification criteria $u_{ij}(X),\ j=1,\ldots,k,\ j\neq i$, become the following: Classify an observation vector X into π_i if $\hat{u}_{ij}(X)\geq \ln k_{ij},\ k_{ij}=\frac{q_jC(i|j)}{q_iC(j|i)},\ j=1,\ldots,k,\ j\neq i$, or $\hat{u}_{ij}\geq 0,\ j=1,\ldots,k,\ j\neq i$, if $q_1=\cdots=q_k$, and the $C(i|j)$'s are equal $j=1,\ldots,k,\ j\neq i$, where

$$\hat{u}_{ij}(X)=(\bar{X}^{(i)}-\bar{X}^{(j)})'\hat{\Sigma}^{-1}X-\tfrac{1}{2}(\bar{X}^{(i)}-\bar{X}^{(j)})'\hat{\Sigma}^{-1}(\bar{X}^{(i)}+\bar{X}^{(j)}) \tag{12.7.13}$$

for $j=1,\ldots,\ k,\ j\neq i$. Unfortunately, the exact distribution of $\hat{u}_{ij}(X)$ is difficult to obtain even when the populations π_i's have p-variate normal distributions. However, when $n_j\to\infty,\ \bar{X}^{(j)}\to\mu^{(j)},\ j=1,\ldots,k$, and when $n_j\to\infty,\ j=1,\ldots,k,\ \hat{\Sigma}\to\Sigma$. Then, asymptotically, that is, when $n_j\to\infty,\ j=1,\ldots,k,\ \hat{u}_{ij}(X)\to u_{ij}(X)$, so that the theory discussed in the previous sections is applicable. As well, the classification probabilities can then be evaluated as illustrated in Example 12.7.2.

12.8. The Maximum Likelihood Method when the Population Covariances Are Equal

Consider k real normal populations $\pi_i:\ P_i(X)\simeq N_p(\mu^{(i)},\Sigma),\ \Sigma>O,\ i=1,\ldots,k$, having the same covariance matrix but different mean value vectors $\mu^{(i)},\ i=1,\ldots,k$. A p-vector X at hand is to be classified into one of these populations $\pi_j,\ j=1,\ldots,k$. Consider a simple random sample $X_1^{(i)},X_2^{(i)},\ldots,X_{n_i}^{(i)}$ of sizes n_i from π_i for $i=1,\ldots,k$. Employing our usual notations, the sample means, sample matrices, matrices of sample means and the sample sum of products matrices are as follows:

$$\bar{X}^{(i)}=\frac{1}{n_i}\sum_{j=1}^{n_i}X_j^{(i)},\ \mathbf{X^{(i)}}=[X_1^{(i)},\ldots,X_{n_i}^{(i)}]=\begin{bmatrix}x_{11}^{(i)} & \cdots & x_{1n_i}^{(i)}\\ \vdots & \ddots & \vdots\\ x_{p1}^{(i)} & \cdots & x_{pn_i}^{(i)}\end{bmatrix},$$
$$\bar{\mathbf{X}}^{\mathbf{(i)}}=[\bar{X}^{(i)},\ldots,\bar{X}^{(i)}],\ S^{(i)}=[\mathbf{X^{(i)}}-\bar{\mathbf{X}}^{\mathbf{(i)}}][\mathbf{X^{(i)}}-\bar{\mathbf{X}}^{\mathbf{(i)}}]',$$
$$S^{(i)}=(s_{rt}^{(i)}),\ s_{rt}^{(i)}=\sum_{m=1}^{n_i}(x_{rm}^{(i)}-\bar{x}_r^{(i)})(x_{tm}^{(i)}-\bar{x}_t^{(i)}),\ S=S^{(1)}+S^{(2)}+\cdots+S^{(k)}. \tag{12.8.1}$$

Then, the unbiased estimators of the population parameters, denoted with a hat, are

$$\hat{\mu}^{(i)} = \bar{X}^{(i)},\ i = 1, \ldots, k, \quad \text{and} \quad \hat{\Sigma} = \frac{S}{n_1 + n_2 + \cdots + n_k - k}. \tag{12.8.2}$$

The null hypothesis can be taken as $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ and X originating from π_i and $X_1^{(j)}, \ldots, X_{n_j}^{(j)}$ coming from $\pi_j,\ j = 1, \ldots, k,\ j \neq i$, the alternative hypothesis being: X and $X_1^{(j)}, \ldots, X_{n_j}^{(j)}$ coming from π_j for $j = 1, \ldots, k,\ j \neq i$, and $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ originating from π_i. On proceeding as in Sect. 12.6, when the prior probabilities are equal and the cost functions are identical, the criterion for classification of the observed vector X to π_i for a specific i is

$$\begin{aligned} A_i : &\Big(\frac{n_j}{n_j + 1}\Big)^2 (X - \bar{X}^{(j)})' \Big(\frac{S}{n_{(k)}}\Big)^{-1} (X - \bar{X}^{(j)}) \\ &- \Big(\frac{n_i}{n_i + 1}\Big)^2 (X - \bar{X}^{(i)})' \Big(\frac{S}{n_{(k)}}\Big)^{-1} (X - \bar{X}^{(i)}) \geq 0 \end{aligned} \tag{12.8.3}$$

for $j = 1, \ldots, k,\ j \neq i$, where the decision rule is $A = (A_1, \ldots, A_k)$, $S = S^{(1)} + \cdots + S^{(k)}$ and $n_{(k)} = n_1 + n_2 + \cdots + n_k - k$. Note that (12.8.3) holds for each $i, i = 1, \ldots, k$, and hence, $A_1, \ldots, A_k$ are available from (12.8.3). Thus, the vector X at hand is classified into A_i, that is, assigned to the population π_i, if the inequalities in (12.8.3) are satisfied. This statement holds for each $i, i = 1, \ldots, k$. The exact distribution of the criterion in (12.8.3) is difficult to establish but the probabilities of classification can be computed from the asymptotic theory discussed in Sect. 12.7 by observing the following:

When $n_i \to \infty$, $\bar{X}^{(i)} \to \mu^{(i)}$, $i = 1, \ldots, k$, and when $n_1 \to \infty, \ldots, n_k \to \infty$, $\hat{\Sigma} \to \Sigma$. Thus, asymptotically, when $n_i \to \infty,\ i = 1, \ldots, k$, the criterion specified in (12.8.3) reduces to the criterion (12.7.3) of Sect. 12.7. Accordingly, when $n_i \to \infty$ or for very large n_i's, $i = 1, \ldots, k$, one may utilize (12.7.3) for computing the probabilities of classification, which was illustrated in Examples 12.7.1 and 12.7.2.

12.9. Maximum Likelihood Method and Unequal Covariance Matrices

The likelihood procedure can also provide a classification rule when the normal population covariance matrices are different. For example, let $\pi_1 : P_1(X) \simeq N_p(\mu^{(1)}, \Sigma_1)$, $\Sigma_1 > O$, and $\pi_2 : P_2(X) \simeq N_p(\mu^{(2)}, \Sigma_2),\ \Sigma_2 > O$, where $\mu^{(1)} \neq \mu^{(2)}$ and $\Sigma_1 \neq \Sigma_2$. Let a simple random sample $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ of size n_1 from π_1 and a simple random sample $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ of size n_2 from π_2 be available. Let $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ be the sample averages and S_1 and S_2 be the sample sum of products matrices, respectively. In classification problems, there is an additional vector X which comes from π_1 under the null hypothesis and

from π_2 under the alternative. Then, the maximum likelihood estimators, denoted by a hat, will be the following:

$$\hat{\mu}^{(1)} = \bar{X}^{(1)},\ \hat{\mu}^{(2)} = \bar{X}^{(2)},\ \hat{\Sigma}_1 = \frac{S_1}{n_1} \text{ and } \hat{\Sigma}_2 = \frac{S_2}{n_2}, \tag{i}$$

respectively, when no additional vector is involved. However, these estimators will change in the presence of the additional vector X, where X is the vector at hand to be assigned to π_1 or π_2. When X originates from π_1 or π_2, $\mu^{(1)}$ and $\mu^{(2)}$ are respectively estimated as follows:

$$\hat{\mu}_*^{(1)} = \frac{n_1\bar{X}_1 + X}{n_1 + 1} \text{ and } \hat{\mu}_*^{(2)} = \frac{n_2\bar{X}_2 + X}{n_2 + 1}, \tag{ii}$$

and when X comes from π_1 or π_2, Σ_1 and Σ_2 are estimated by

$$\hat{\Sigma}_{1*} = \frac{S_1 + S_3^{(1)}}{n_1 + 1} \text{ and } \hat{\Sigma}_{2*} = \frac{S_2 + S_3^{(2)}}{n_2 + 1} \tag{iii}$$

where

$$\begin{aligned} S_3^{(1)} &= (X - \hat{\mu}_*^{(1)})(X - \hat{\mu}_*^{(1)})' = \Big(\frac{n_1}{n_1 + 1}\Big)^2 (X - \bar{X}_1)(X - \bar{X}_1)' \\ S_3^{(2)} &= (X - \hat{\mu}_*^{(2)})(X - \hat{\mu}_*^{(2)})' = \Big(\frac{n_2}{n_2 + 1}\Big)^2 (X - \bar{X}_2)(X - \bar{X}_2)', \end{aligned} \tag{iv}$$

referring to the derivations provided in Sect. 12.6 when discussing maximum likelihood procedures. Thus, the null hypothesis can be X and $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ are from π_1 and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ are from π_2, versus the alternative: X and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ being from π_2 and $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$, from π_1. Let L_0 and L_1 denote the likelihood functions under the null and alternative hypotheses, respectively. Observe that under the null hypothesis, Σ_1 is estimated by $\hat{\Sigma}_{1*}$ of *(iii)* and Σ_2 is estimated by $\hat{\Sigma}$ of *(i)*, respectively, so that the likelihood ratio criterion λ is given by

$$\lambda = \frac{\max L_0}{\max L_1} = \frac{|\hat{\Sigma}_{2*}|^{\frac{n_2+1}{2}} |\hat{\Sigma}_1|^{\frac{n_1}{2}}}{|\hat{\Sigma}_{1*}|^{\frac{n_1+1}{2}} |\hat{\Sigma}_2|^{\frac{n_2}{2}}}. \tag{12.9.1}$$

The determinants in (12.9.1) can be represented as follows, referring to the simplifications discussed in Sect. 12.6:

$$\lambda = \frac{(n_1 + 1)^{\frac{p(n_1+1)}{2}}}{(n_2 + 1)^{\frac{p(n_2+1)}{2}}} \frac{|S_2|^{\frac{n_2+1}{2}}}{|S_1|^{\frac{n_1+1}{2}}} \frac{[1 + (\frac{n_2}{n_2+1})^2 (X - \bar{X}_2)' S_2^{-1} (X - \bar{X}_2)]^{\frac{n_2+1}{2}} |\hat{\Sigma}_1|^{\frac{n_1}{2}}}{[1 + (\frac{n_1}{n_1+1})^2 (X - \bar{X}_1)' S_1^{-1} (X - \bar{X}_1)]^{\frac{n_1+1}{2}} |\hat{\Sigma}_2|^{\frac{n_2}{2}}}. \tag{12.9.2}$$

The classification rule then consists of assigning the observed vector X to π_1 if $\lambda \geq 1$ and, to π_2 if $\lambda < 1$. We could have expressed the criterion in terms of $\lambda_1 = \lambda^{\frac{2}{n}}$ if $n_1 = n_2 = n$, which would have simplified the expressions appearing in (12.9.2).