The Random Effects Model in Discriminant Analysis

by

Libero Paul Fatti

A Thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg for the Degree of Doctor of *Philosophy*.

Johannesburg, March 1979.

Declaration

I, Libero Paul Fatti, hereby declare that this thesis is my own work and that it has not been presented to any other University for the purpose of obtaining a Degree.

L.P. Fatti
March, 1979.

Abstract

In this thesis the characteristics of discriminant analysis under the random effects model are investigated.

Assuming that the elements within any randomly selected population are normally distributed with mean vector μ and common covariance matrix Σ, and that over different populations μ has a normal distribution with mean vector ξ and covariance matrix T, the distributions of the population-based and sample-based Mahalanobis distances between two different populations are derived. From these, expressions and bounds are derived for the expected probabilities of mis- and correct classification under classical discriminant analysis, applied to two- and k-population problems respectively, when using either the population-based or sample-based linear discriminant functions.

The distributions and expected probabilities mentioned above are all expressed in terms of the eigenvalues of $T\Sigma^{-1}$, so the problems of hypothesis testing on, and more particularly, estimation of these eigenvalues are fully discussed.

Using the Predictive Bayesian Approach to Discriminant Analysis, expressions for the predictive density of an observation, given that it has come from a particular population, are derived under the random effects model. Brief consideration is also given to the empirical Bayes and semi-Bayes approaches to discriminant analysis under this model.

Finally, the results derived in this thesis are applied to a stratigraphic problem in underground mining.

Acknowledgements

I would like to express my gratitude to my supervisor Dr. D.M. Hawkins, until recently Professor of Mathematical Statistics at the University of the Witwatersrand, for his advice and encouragement throughout the period of research towards this thesis. I would also like to thank the following: Professors S. Geisser and D.J. de Waal and Dr. D. Bradu for useful discussions and suggestions; Dr. P.J.A. Nagel for allowing me the use of his programmes for the evaluation of Zonal polynomials and for his assistance in running them; the National Research Institute for Mathematical Sciences, CSIR, for providing me with research and computer facilities during part of 1978; and the University of South Africa for the use of their computer.

Finally, I would like to express my thanks to my first teacher in Statistics, Professor J.E. Kerrich, for sparking off in me an enduring love for Statistics.

Contents

Page

List of Tables and Figures

Chapter 1 Introduction

Suppose that $\pi_1, \pi_2, \ldots, \pi_k$ are k populations of p-component vectors. Let x be a vector known to have come from one of these populations. Discriminant analysis deals with the problem of identifying the population from which x was drawn.

The case covered most thoroughly in the literature is that in which the vectors from π_i follow a multivariate normal distribution with mean vector μ_i and a common covariance matrix Σ. (Anderson, 1958). Generally, it has been assumed that $\pi_1, \pi_2, \ldots, \pi_k$ are fixed populations predetermined by the problem faced.

This thesis deals with the case where the μ_i have been randomly selected from some population in advance of the experiment. Once the k mean vectors have been selected we are then faced with a conventional problem in discriminant analysis of classifying vectors into one of the k (now fixed) populations.

In different experiments, there are different sets of μ_i, in general with different numbers of elements k, all drawn independently from the same parent population.

The aim of this research is to investigate the characteristics of discriminant analysis under these circumstances. It will be assumed that the population from which the μ_i are drawn is multivariate normal with mean vector ξ and covariance matrix T.

This study was motivated by a stratigraphic problem in mining. (Hawkins and Rasmussen (1973), Hutchinson, Skinner and Bowes (1976)) In the Witwatersrand gold fields the gold bearing reef is one band (the "pay band") of a sedimentary succession and is usually visually unrecognisable. In badly faulted areas this pay band usually faults away, and the miner wishes to know the position in the sedimentary succession of the blank band facing him, from which he can deduce the new position of the pay band.

One method of identification is via trace element geochemistry of the bands. It is reasonable to suppose that the geochemistry of each band can be described by a (multivariate) statistical distribution. The mean of the distribution reflects the average conditions at the time of deposition of the band, while the spread reflects local variation in grade. Furthermore, the average conditions at different times and localities of deposition of the bands are themselves statistically variable, being themselves drawn from some parent population. Thus the bands intersected by any given cross-section will be fixed for the immediate classification problem and yet will follow some random effects model as we move from one area in the mine to another.

Another example of a random effects model in discriminant analysis occurs in anthropology (de Villiers, 1973, 1976). Here the problem is to classify an ancient skull from a certain period as having come from one of a number of tribes suspected to have lived in the locality in which the skull was found. The classification is based on various measurements (lengths and angles) made on the maxilla and/or mandible, and for any given tribe , sex and age-group these may be regarded as having a joint distribution with fixed mean vector and covariance matrix. Different tribes will, in general, have different mean vectors, and these may themselves be considered to have come from some multivariate distribution.

Another type of random effects model in discriminant analysis is considered by Geisser (1973), in the context of multiple birth discrimination. Supposing that a birth gives rise to t like-sexed offspring, the problem is to decide which of these offspring have come from the same eggs and which ones have come from different eggs. Assume that each offspring is characterised by a p-dimensional random variable x, where $x \sim N_p(\mu_i,\Sigma_W)$. Offspring from the same egg (monozygotes) have the same μ_i, whereas offspring from different eggs (heterozygotes) have different μ_i. Different μ_i are assumed to have been generated by a random effects model;

<u>i.e.</u> $$\mu_i \sim N_p(\mu,\Sigma_B), \quad \text{independently } \forall i \, .$$

Geisser considers the difference $z_r = x_t - x_r$ between the t^{th} and the r^{th} offspring. If t and r come from the same egg, then:

$$z_r \sim N_p(0,2\Sigma_W)$$

and if they are from different eggs, then

$$z_r \sim N_p(0,2\Sigma_W + 2\Sigma_B) \, .$$

The joint distribution of z_r and z_s is also multivariate normal with

$$\operatorname{cov}(z_r,z_s) = \begin{cases} \Sigma_W + \Sigma_B & \text{if t,r and s are all from different eggs,} \\ \Sigma_W + 2\Sigma_B & \text{if r and s are from the same egg but} \\ & \quad \text{t is from a different one,} \\ \Sigma_W & \text{otherwise} \end{cases}$$

Given the joint distribution of $z_1,\ldots,z_{t-1}$ for each of the various possible combinations of offspring and eggs, and the prior probabilities for each of these possible combinations, posterior probabilities can be calculated for each case, and the case for which this is a maximum is then chosen.

The situation discussed in this thesis is, however, entirely different from that just describeo. Here we assume that the i^{th} population is characterised by a $N_p(\mu_i,\Sigma)$ distribution and that different μ_i are independently distributed as $N_p(\xi,T)$. On the basis of these assumptions the characteristics of classification in this environment are then assessed.

i.e. Given an observation known to have come from one of k populations from the abovementioned random effects model, where the parameters of these populations are either known or estimated from training samples, how well are the classical procedures of discriminant analysis for classifying the observation into one of these populations likely to perform?

When it comes to the Predictive Bayesian Approach to discriminant analysis, the random effects model actually leads to a new procedure, for classifying the observation into one of the k populations.

1.1 The Scope of the research covered in this Thesis

As mentioned earlier, the aim of this thesis is to investigate the characteristics of discriminant analysis under the Random Effects model.

In order to do so, and to provide a framework within which to conduct the investigation, a summary of the theory of classical and Predictive Bayesian discriminant analysis is given in chapter 2. By the classical approach we mean that given by Anderson (1951,1958) and by the Predictive Bayesian approach we mean that of Geisser (1964,1966) and Dunsmore (1966).

Chapters 3 to 5 cover the classical approach. In chapter 3 the Random Effects model is set out in more detail, and then the distributions of the four quantities central to the classical approach are derived under this model. Chapter 4 uses the distributions derived in chapter 3 to evaluate the performance of classical discriminant analysis under the random effects model. Specifically, the probabilities of correct and misclassification are considered, separately for the two-group and multiple-group problems and for the two situations where the parameters are known and unknown.

All the results in chapters 3 and 4 are expressed in terms of $\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0$, the r nonzero eigenvalues of $T\Sigma^{-1}$ where T and Σ are the covariance matrices of the mean vector μ and observation vector X, respectively, so chapter 5 is devoted to the question of inference on these parameters. After a short review of hypothesis testing on the λ_i, the rest of the chapter addresses the question of their estimation, on the basis of "training samples" taken from a number of randomly selected populations.

Whereas the treatment of the classical approach is confined to an evaluation of the standard theory within the framework of the random effects model, the application of this model to the Predictive Bayesian approach results in a modification of the usual classification rule. Chapter 6 deals with this approach and in it the predictive density of a new observation, given the training samples and assuming that it comes from a specific group, is derived under the random effects model. A brief treatment of the Empirical Bayes and Semi-Bayes approaches completes this chapter.

In chapter 7 the theory of the preceding chapters is applied to some data obtained from underground mining, contrasting the results with those obtained by applying the usual fixed effects theory.

The thesis is concluded in chapter 8 with a discussion of various avenues for future research and with some comments on the applicability and usefulness of the theory developed here to the solution of practical problems in discriminant analysis.

Chapter 2 A Summary of the Classical and Bayesian approaches to Discriminant Analysis

In this chapter a brief summary is given of the theory of Discriminant Analysis under the Normal distribution.

The Classical approach, pioneered by Fisher (1936), Welch (1939), Wald (1944) and others is described by Anderson (1958), Lachenbruch (1975) and Giri (1977) so only a brief sketch of the basic theory will be given in section 2.1. The coverage is not complete, and prime emphasis will be given only to those aspects that will be of direct relevance to the treatment of the random effects model.

The Predictive Bayesian approach, pioneered by Geisser (1964), (1966) and Dunsmore (1966) is described in section 2.2. Once again, only a brief summary of the approach will be given, and only one main result, useful for comparison with the results derived in this thesis, will be given. A description of the approach is given in Press (1972).

A critical comparison of the Classical and Predictive Bayesian approaches, as well as a concise description of them that highlights the point of departure between the two is given by Aitchison, Habbema and Kay (1977). This paper comes out strongly in favour of the Bayesian approach, at least within the framework of the "fixed effects" (Classical approach) or "Diffuse prior" (Predictive Bayesian approach) model. It would be interesting to compare the relative efficacies of these two approaches within the random effects framework.

2.1 Classical Discriminant Analysis

Suppose we have a p-dimensional observation x known to have come from one of k populations $\pi_1, \pi_2, \ldots, \pi_k$.. Anderson (1958) proves that the Baye [illegible] fication procedure, that assigns x to one of the populations in [illegible] way that the expected loss from misclassification is minimised, is, under mild restrictions, an admissible procedure and that the class of Bayes procedures is minimal complete.

Assuming that the costs of misclassification from all k populations are equal, the Bayes procedure leads to the following simple classification rule:

Assign x to population π_i where,

$$q_i \ f_i(x) = \max_{j=1,\dots,k} q_j \ f_j(x) \qquad (2.1.1)$$

where q_j is the prior probability that x comes from π_j and $f_j(x)$ is the probability (density) function of x assuming that it has come from π_j.

The case considered most frequently in the literature and in practice is that in which observations from π_j follow a multivariate normal distribution with mean vector μ_j and common covariance matrix Σ. In this case,

$$q_j \ f_j(x) = q_j(2\pi)^{-p/2} \ |\Sigma|^{-\frac{1}{2}} \exp \{- \tfrac{1}{2}(x-\mu_j)' \ \Sigma^{-1}(x-\mu_j)\}$$

Taking logarithms and simplifying, rule (2.1.1) becomes:

Assign x to population π_i where,

$$\log q_i - \tfrac{1}{2}(x-\mu_i)' \ \Sigma^{-1}(x-\mu_i) = \max_{j=1,\dots,k} \{ \log q_j - \tfrac{1}{2}(x-\mu_j)' \Sigma^{-1}(x-\mu_j)\} \qquad (2.1.2)$$

or

$$(x - \tfrac{1}{2}(\mu_i + \mu_j))' \ \Sigma^{-1}(\mu_i - \mu_j) > \log \frac{q_j}{q_i} \quad \forall \ j=1,\dots,k; \ j \neq i \quad (2.1.3)$$

In the case where the prior probabilities q_j are all equal, rules (2.1.2) and (2.1.3) become, respectively:

Assign x to population π_i where,

$$(x - \mu_i)' \Sigma^{-1}(x - \mu_i) = \min_{j=1,\ldots,k} (x - \mu_j)' \Sigma^{-1}(x - \mu_j) \quad (2.1.4)$$

and

$$(x - \tfrac{1}{2}(\mu_i + \mu_j))' \Sigma^{-1}(\mu_i - \mu_j) > 0 \quad \forall\ j=1,\ldots,k;\ j \neq i\ . \quad (2.1.5)$$

From (2.1.4) it is clear that for equal prior probabilities the Bayesian classification rule is also a minimum distance rule in that x is classified into that population π_i to which it is closest as measured by the Mahalanobis distance from x to π_i :

$$\delta_i^2(x) = (x - \mu_i)' \Sigma^{-1}(x - \mu_i)$$

The Case k = 2

In this case rule (2.1.3) becomes:

Assign x to π_1 if:

$$u_{12}(x) = (x - \tfrac{1}{2}(\mu_1 + \mu_2))' \Sigma^{-1}(\mu_1 - \mu_2) > \log \frac{q_2}{q_1} \quad (2.1.6)$$

and to π_2 otherwise.

To obtain the probabilities of misclassification under rule (2.1.6) note that if we let X be the random vector corresponding to the observed x then, under the assumption that X is from π_1, $u_{12}(X)$ has a univariate normal distribution with mean:

$$E[u_{12}(X)|\pi_1] = \tfrac{1}{2}(\mu_1 - \mu_2)'\ \Sigma^{-1}(\mu_1 - \mu_2)$$
$$= \tfrac{1}{2}\ \delta^2_{12}$$

where δ^2_{12} denotes the Mahalanobis distance between π_1 and π_2, and variance:

$$\mathrm{Var}[u_{12}(X)|\pi_1] = E[(\mu_1 - \mu_2)'\Sigma^{-1}(X - \mu_1)(X - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)]$$
$$= (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$
$$= \delta^2_{12} \ .$$

So, given that X is from π_1,

$$u_{12}(X) \sim N(\tfrac{1}{2}\ \delta^2_{12}\ ,\delta^2_{12}) \qquad (2.1.7)$$

where $\delta^2_{12} = (\mu_1 - \mu_2)'\ \Sigma^{-1}(\mu_1 - \mu_2)$.

Similarly, it can be shown that if X is from π_2, then

$$u_{12}(X) \sim N(-\ \tfrac{1}{2}\ \delta^2_{12}\ ,\delta^2_{12}) \qquad (2.1.8)$$

Therefore,

$P_1 =$ P[Misclassify a random observation from π_1]

$= P[u_{12}(X) < c|\pi_1]$ where $c = \log \dfrac{q_1}{q_2}$

$$= \Phi\left(\frac{c - \frac{1}{2}\ \delta^2_{12}}{\delta_{12}}\right) \qquad (2.1.9)$$

where $\Phi(\cdot)$ denotes the standard normal distribution function, and

$$\begin{aligned} P_2 &= P[\text{Misclassify a random observation from } \pi_2] \\ &= P[u_{12}(X) > c|\pi_2] \\ &= \Phi\left(-\frac{c + \frac{1}{2}\delta_{12}^2}{\delta_{12}}\right) \end{aligned} \qquad (2.1.10)$$

For equal prior probabilities $q_1 = q_2 = \frac{1}{2}$, $c = 0$ and (2.1.9) and (2.1.10) become:

$$P_1 = P_2 = \Phi(-\tfrac{1}{2}\delta_{12}) \qquad (2.1.11)$$

<u>$k > 2$ populations</u>

This case has not received nearly as much attention as the two - population problem. Although there is not much increase in complexity at a conceptual level when moving from the two-to the multiple population problem, the evaluation of misclassification probabilities becomes considerably more complicated. To see this, note that if we use the notation:

$$u_{ij}(x) = (x - \tfrac{1}{2}(\mu_i + \mu_j))'\,\Sigma^{-1}(\mu_i - \mu_j) \qquad (2.1.12)$$

then classification rule (2.1.3) becomes:
Assign x to population π_i where,

$$u_{ij}(x) > \log\frac{q_j}{q_i} \qquad (2.1.13)$$

$$\forall\ j=1,\ldots,k;\ \ j\neq i$$

Letting X be the random vector corresponding to x, and assuming that X is from π_i we have, as in the case $k = 2$ populations:

$$E[u_{ij}(X)|\pi_i] = \tfrac{1}{2}\,\delta_{ij}^2$$

$$\mathrm{Var}[u_{ij}(X)|\pi_i] = \delta_{ij}^2$$

where $\delta_{ij}^2 = (\mu_i - \mu_j)'\,\Sigma^{-1}(\mu_i - \mu_j)$

and it is easy to show that

$$\mathrm{cov}[u_{ij}(X),\, u_{i\ell}(X)|\pi_i] = \delta_{ij\ell}$$

where $\delta_{ij\ell} = (\mu_i - \mu_j)'\,\Sigma^{-1}(\mu_i - \mu_\ell)$.

Using the notation:

$$u_{ij} = u_{ij}(X)$$

and noting that the k-1 random variables u_{ij}, $j=1,\ldots,k;\;\; j\neq i$ are all linear functions of the normally distributed random vector X we have that, given $X \in \pi_i$:

$$\underline{u}_i = (u_{i1},\ldots,u_{ii-1},\, u_{ii+1},\ldots,u_{ik})'$$

has a (k-1) - dimensional Normal distribution with mean vector:

$$\tfrac{1}{2}\,\underline{\delta}_i^2 = \tfrac{1}{2}(\delta_{i1}^2,\ldots,\delta_{ii-1}^2,\, \delta_{ii+1}^2,\ldots,\delta_{ik}^2)'$$

and covariance matrix:

$$\Delta_i = (\delta_{ij\ell}) \ , \ \ j,\ell=1,\ldots,k; \ \ j,\ell \neq i \qquad (2.1.14)$$

where we have used the notation:

$$\delta_{ijj} = \delta_{ij}^2$$

Remark 2.1.1 If $k-1 > p$ then $\underset{\sim}{u}_i$ will have a singular normal distribution with its mass concentrated on a p-dimensional subspace.

Therefore, the probability of correct classification, given $X \in \pi_i$ is:

$$P\left[\bigcap_{\substack{j=1\\j\neq i}}^{k} u_{ij} > c_{ji} \middle| \pi_i\right] = \int_{c_{1i}}^{\infty} \cdots\cdots \int_{c_{ki}}^{\infty} f_i(\underset{\sim}{u}_i) \prod_{\substack{j=1\\j\neq i}}^{k} du_{ij} \qquad (2.1.15)$$

where,

$$c_{ji} = \log \frac{q_j}{q_i} = \log q_j - \log q_i$$

and $f_i(\underset{\sim}{u}_i)$ is the density function of the (k-1)-dimensional Normal distribution given in (2.1.14).

Lachenbruch (1973) has evaluated the integral in (2.1.15) when the prior probabilites q_j are all equal (so that the lower limits of integration are all zero) for two particular configurations of the mean vectors μ_i. The two configurations that he considers are:

(a) the μ_i are collinear, with equal spacing of δ units between adjacent means,

and (b) the μ_i are placed at the vertices of a regular (k-1)-dimensional simplex with side of length δ units.

For configuration (a), with μ_1 and μ_k at the two extremes, (2.1.15) becomes

$$P[\text{correct classification}|\pi_i] = \Phi\left(\frac{\delta}{2}\right)-\Phi\left(-\frac{\delta}{2}\right) \quad \text{for } i=2,\ldots,k-1$$
$$= \Phi\left(\frac{\delta}{2}\right) \qquad \text{for } i=1 \text{ and } k$$

and for configuration (b) it becomes:

$$P[\text{correct classification}|\pi_i] = \int_{-\infty}^{\infty}\left[\Phi\left(\frac{\delta^2/2-x}{\sqrt{\delta^2/2}}\right)\right]^{k-1}\phi\left(\frac{x}{\sqrt{\delta^2/2}}\right)dx$$

where $\phi(\cdot)$ is the standard normal density function.

For a general configuration of mean vectors, however, tables of the (k-1)-dimensional normal distribution (or an algorithm to compute them) are required to evaluate (2.1.15).

The following lower bounds on the minimum probability P_o of correct classification when the prior probabilities q_j are all equal, that are far easier to compute than (2.1.15), have been given by Cacoullos (1973):

$$P_o \geq G_{k-1}\left(\frac{\delta^2}{4}\right) \qquad (2.1.16)$$

and

$$P_o \geq 1-(k-1)\ \Phi\left(-\frac{\delta}{2}\right) \qquad (2.1.17)$$

where,

$G_\nu(\cdot)$ is the distribution function of the chi-squared distribution on ν degrees of freedom,

and $\delta^2 = \min_{\forall i<j} \delta^2_{ij}$ is the minimum Mahalanobis distance between any two of the k populations.

For $k \leq 3$, (2.1.17), which is derived using Bonferroni's first inequality, gives a stronger bound than (2.1.16), whereas the opposite *is generally true for* $k > 3$.

2.1.1 Unknown Parameters

Thus far it has been assumed that all the parameters in the populations π_i, $i=1,\ldots,k$, are *known*. In most practical situations, however, these are not known, and have to be estimated from a "training sample" consisting of n_i observations x_{ij}, $j=1,\ldots,n_i$, known to have come from π_i , for each of the k populations π_i , $i=1,\ldots,k$.

Anderson (1951) proposed that the unknown parameters μ_i , $i=1,\ldots,k$ and Σ in (2.1.3) be replaced by their maximum likelihood estimators, the sample means,

$$x_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

and pooled sample covariance matrix, *respectively*

$$S = \frac{1}{\nu} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - x_{i.})(x_{ij} - x_{i.})'$$

where $\nu = \sum_{i=1}^{k} (n_i - 1)$. This gives the sample-based classification rule:

Assign x to population π_i where,

$$\log q_i - \tfrac{1}{2}(x - x_{i.})' S^{-1} (x - x_{i.}) = \max_{j=1,\ldots,k} \{\log q_j - \tfrac{1}{2}(x - x_{j.})' S^{-1} (x - x_{j.})\} \tag{2.1.18}$$

or

$$V_{ij} = V_{ij}(x) = (x - \tfrac{1}{2}(x_{i.} + x_{j.}))' S^{-1} (x_{i.} - x_{j.}) > \log \frac{q_j}{q_i} \tag{2.1.19}$$

$$\forall\ j=1,\ldots,k;\ j \neq i$$

This procedure of "plugging in" the sample estimates of the unknown parameters into the optimal Bayes classification rule (2.1.2) or (2.1.3) is essentially an Empirical Bayes procedure; see, for example, Maritz (1970). Aitchison, Habbema and Kay (1977) refer to it as an "estimative" method, in contrast to the "predictive" method used in the "pure" Bayesian approach of Geisser (1964) that will be described in section 2.2.

Anderson (1958) justifies the use of the sample-based discriminant function V_{ij} defined in (2.1.9) in the two-population case by pointing out that it can be written as:

$$V_{12} = x' \, S^{-1}(x_{1.} - x_{2.}) - \tfrac{1}{2}(x_{1.} + x_{2.})' \, S^{-1}(x_{1.} - x_{2.})$$

and that the first term ("Fisher's discriminant function") is the linear function of x that has the greatest "between group" variance relative to the "within group" variance. He also appeals to the fact that "*it seems intuitively reasonable*".

Geisser (1967) adds further justification by pointing out, in the two - population case, that the posterior mean of the population *discriminant* function u_{12}, defined in (2.1.6), given the training sample and assuming a noninformative prior distribution for μ_1, μ_2 and Σ, [illegible], for fixed x:

$$E[u_{12}|x, TS] = \frac{1}{2} \, p\left(n_2^{-1} - n_1^{-1}\right) + V_{12} \tag{2.1.20}$$

where TS denotes the training sample $\{x_{ij}; \; j=1,\ldots,n_i; \; i=1,2\}$.

Expression (2.1.20) derives from the fact that, under the abovementioned prior assumptions, the posterior mean of

$$\delta_i^2(x) = (x - \mu_i)' \, \Sigma^{-1}(x - \mu_i)$$

is:

$$E[\delta_i^2(x)|x, TS] = p\, n_i^{-1} + d_i^2(x) \qquad (2.1.21)$$

where

$$d_i^2(x) = (x - x_{i.})'\, S^{-1}(x - x_{i.})$$

This result is clearly not confined to the two-population case, and the bias in V_{ij} and $d_i^2(x)$ evident from (2.1.20) and (2.1.21) respectively, may be incorporated into classification rules (2.1.18) and (2.1.19) by substituting $\log q_j - \frac{1}{2}\, p\, n_j^{-1}$ for $\log q_j$, $j=1,\ldots,k$, in these two rules.

Remark 2.2.1 In the situation where the training samples from the different populations all have the same size,

i.e. $n_j = n$, $j=1,\ldots,k$

the bias $\frac{1}{2}\, p(n_j^{-1} - n_i^{-1})$ in V_{ij} vanishes, and that in $d_i^2(x)$ is a constant, $p\, n^{-1}$, and therefore does not affect rule (2.1.18).

As a final justification for using sample-based rules (2.1.18) and (2.1.19) Glick (1972) proves that, under very general conditions, sample-based classification rules are asymptotically optimal in the sense that they converge (almost surely) to their corresponding population-based optimal rules (2.1.1).

k = 2 populations

This is the case that has received the most attention in the literature. Conditional on $x_{1.}$, $x_{2.}$ and S, and letting X be the random vector corresponding to x, $V = V_{12}(X)$ has a normal distribution with mean

$$E[V|x_{1.}, x_{2.}, S; X \in \pi_1] = (\mu_1 - \tfrac{1}{2}(x_{1.} + x_{2.}))' \, S^{-1}(x_{1.} - x_{2.})$$

and variance

$$\mathrm{Var}[V|x_{1.}, x_{2.}, S; X \in \pi_1] = (x_{1.} - x_{2.})' \, S^{-1} \, \Sigma \, S^{-1}(x_{1.} - x_{2.})$$

Using rule (2.1.19) with k=2 and considering the case $q_1 = q_2 = \frac{1}{2}$, i.e.: Assign x to π_1 if

$$V > 0$$

and to π_2 otherwise, (2.1.22)

and arguing in a way similar to that leading to (2.1.11) we obtain the following expression for the conditional probability that a randomly chosen member of π_i will be misclassified:

$$P_i^C = P[\text{misclassification}|x_{1.} , x_{2.} , S; X \in \pi_i]$$

$$= \Phi\left(\frac{(-1)^i(\mu_i - \frac{1}{2}(x_{1.} + x_{2.}))' \, S^{-1}(x_{1.} - x_{2.})}{\sqrt{\{(x_{1.} - x_{2.})' \, S^{-1} \Sigma S^{-1}(x_{1.} - x_{2.})\}}}\right) \quad i=1,2 \qquad (2.1.23)$$

John (1961), Hills (1966), Lachenbruch and Mickey (1968), Dunn (1971) Sorum (1972a), McLachlan (1974a, b, c, 1975, 1976a, b) have studied the conditional error rates (2.1.23) (termed the "actual" error rate by Hills).

A simple estimator of P_i^C , i=1,2, is obtained by replacing μ_i and Σ respectively by $x_{i.}$ and S in (2.1.23). This yields:

$$\hat{p}_1^c = \hat{p}_2^c = \Phi(-\tfrac{d}{2}) \qquad , \qquad (2.1.24)$$

where $d^2 = d_{12}^2 = (x_{1.} - x_{2.})' \, S^{-1} (x_{1.} - x_{2.})$

Glick(1972) proves that this "apparent error-rate" $\Phi(-\frac{d}{2})$ converges uniformly to the "optimum" error rate $\Phi(-\frac{\delta}{2})$ given in (2.1.11) as the sample sizes n_1 and n_2 increase.

However, for moderate sample sizes (2.1.24) may be badly biased and give much too favourable an impression of the probability of error. Hills (1966) proves that:

$$E[\Phi(-\tfrac{d}{2})] < \Phi(-\tfrac{\delta}{2}) < E[P_1^c]$$

and Dunn and Varady (1966), Lachenbruch and Mickey (1968) and Dunn (1971) show empirically that this bias may indeed be substantial for moderate sample sizes.

McLachlan (1974c) gives the following estimator of P_1^c, with bias of order 3 with respect to $(n_1^{-1}, n_2^{-1}, \nu^{-1})$ where $\nu = n_1 + n_2 - 2$:

$$\hat{p}_1^c = \Phi(-\tfrac{d}{2}) + \phi(\tfrac{d}{2})\{ \frac{p-1}{n_1 d} + \frac{d}{32\nu} (4(4p-1)-d^2)\} + 0_2 \qquad (2.1.25)$$

(0_2 denotes the term of order 2 with respect to $(n_1^{-1}, n_2^{-1}, \nu^{-1})$; this is given explicitly in McLachlan (1976).)

While the conditional error rates are of interest in assessing the performance of a particular discriminant function, the unconditional or expected error rates, obtained by considering $x_{1.}$, $x_{2.}$ and S as random variables, are more appropriate when considering the expected performance of the sample discriminant function V when based on randomly chosen samples of sizes n_1 and n_2 from π_1 and π_2 , respectively.

Several authors, including Okamoto (1963, 1968) Hills (1966), Lachenbruch (1967, 1968), Lachenbruch and Mickey (1968), Dunn (1971), Sorum (1972b) and Anderson (1973a, 1973b) have studied the expected error rate when the sample-based classification rule (2.1.22) is used.

Okamoto (1963) obtained an asymptotic expansion for the distribution of the sample discriminant function V. Applying this to the classification rule (2.1.22) and assuming equal-sized training samples $n_1 = n_2 = n$, yields the following expression, to terms of order n^{-2}, for the expected probability of misclassification for a randomly chosen member of π_1:

$$P_1^e = P[\text{misclassification}|\pi_1] = \Phi\left(-\frac{\delta}{2}\right) + \phi\left(\frac{\delta}{2}\right)\frac{1}{\nu}\left(\frac{p-1}{\delta} + \frac{p}{4}\,\delta\right) + O(n^{-2}) \tag{2.1.26}$$

(Okamoto also gives a (very complicated) expression for the terms of order n^{-2}.)

Anderson (1973a, 1973b) derives an alternative asymptotic expansion for V in the "studentized" form which, for $n_1 = n_2 = n$ has the form:

$$P\left[\frac{V - \frac{1}{2}d^2}{d} \leq y|\pi_1\right] = \Phi(y) + \phi(y)\,\frac{1}{\nu}\left(\frac{2(p-1)}{\delta} - (p+\tfrac{1}{4})y - \tfrac{1}{4}\,y^3\right) + O(n^{-2}) \tag{2.1.27}$$

Expression (2.1.27) is useful when one wishes to choose the cut-off point for V for classifying x into π_1 so as to achieve a given probability of misclassification. (Anderson (1973b), McLachlan(1977))

Lachenbruch and Mickey (1968) use a simulation study to compare the performances of a number of estimators of P_i^c and P_i^e including Okamoto's expansion with two different estimators for δ, and a distribution-free method proposed by Lachenbruch (1967) based on a sample reuse approach.

k > 2 Populations

As in the case where the parameters are known, the multiple population problem has received far less attention than the two-population problem.

McKay (1977) has considered the problem of variable selection within the context of multiple population discriminant analysis, and Michaelis (1973) has performed simulation experiments to assess the error rate of the classification rule (2.1.19) based on the linear discriminant function V_{ij} in some multiple population situations. Glick (1972) proves that the "apparent non-error rate", obtained by replacing the parameters in (2.1.15) by sample-based (maximum likelihood) estimators, converges uniformly to the "optimum" probability of correct classification as the sample sizes increase.

Assuming equal prior probabilities $q_i = 1/k$, $i=1,\ldots,k$ for the k populations, classification rule (2.1.18) becomes:

Assign x to π_i where,

$$d_i^2(x) = \min_{j=1,\ldots,k} d_j^2(x)$$

where $$d_j^2(x) = (x - x_{j.})' S^{-1}(x - x_{j.}) \qquad (2.1.28)$$

If $x \in \pi_i$, letting X be the random variable corresponding to x and considering x_i and S as random variables,

$$\nu^{-1} n_i(n_i + 1)^{-1} d_i^2(X) \sim f_{p,\ \nu-p+1} \qquad (2.1.29)$$

and

$$\nu^{-1} n_j(n_j+1)^{-1} d_j^2(X) \sim f_{p,\ \nu-p+1}(\lambda_{ij}) \qquad (2.1.30)$$

where,

$f_{p,\ \nu-p+1}$ denotes the central, unnormed f-distribution with p and $\nu-p+1$ degrees of freedom,

$f_{p,\ \nu-p+1}(\lambda_{ij})$ denotes the corresponding noncentral distribution with noncentrality parameter

$$\lambda_{ij} = n_j(n_j+1)^{-1}\ \delta_{ij}^2$$

and $$\delta_{ij}^2 = (\mu_i - \mu_j)'\ \Sigma^{-1}(\mu_i - \mu_j)\ .$$

(See, for example, Giri (1977) chapter 7).

So the probability of <u>correct</u> classification using rule (2.1.28) and given $x \epsilon \pi_i$ and be written:

$$P[\text{correct classification}|x \epsilon \pi_i] = P[z_i < \min_{\substack{j=1,\ldots k \\ j\neq i}} z_j] \qquad (2.1.31)$$

where,

$$\nu^{-1}n_i(n_i+1)^{-1}\ z_i \sim f_{p,\ \nu-p+1}$$

$$\nu^{-1}n_j(n_j+1)^{-1}\ z_j \sim f_{p,\ \nu-p+1}(\lambda_{ij}) \quad j=1,\ldots,k;\ \ j\neq i$$

and the z_j, $j=1,\ldots,k$ are not independent random variables, due to the fact that X and S occur in all the $d_j^2(X)$, $j=1,\ldots,k$.

To evaluate the probability on the right-hand side of (2.1.31) requires the joint distribution of k correlated random variables, k-1 of which have noncentral f marginal distributions, the last one having a central f marginal distribution. This problem has received little, if any, attention to date.

Cacoullos (1973) gives the following lower bound on the minimum probability P_o of correct classification using rule (2.1.28):

$$P_o \geq \sum_{i=0}^{k} P[z_i \leq (\nu-p+1)n_i(16p\nu)^{-1}\ \delta^2] - k \qquad (2.1.32)$$

where,

z_i denotes a (normed) F - random variable with p and $(\nu - p + 1)$ degrees of freedom,

$\delta^2 = \min\limits_{\forall\, i<j} \delta^2{}_{ij}$,

and $n_0 = 1$.

2.2 The Predictive Bayesian Approach

Given the training sample $TS = \{x_{ij},\ j = 1,\dots,n_i;\ \ i = 1,\dots,k\}$ from k populations π_i, $i = 1,\dots,k$ and an observation x of unknown origin, the Predictive Bayesian approach consists in evaluating the posterior probability, given TS and the underlying model together with any known parameters, that x belongs to π_r for r = 1 to k, and then assigning x to that population for which this probability is the greatest.

More specifically, suppose that each π_r, $r = 1,\dots,k$ is specified by a probability density function $f(\cdot|\theta_r,\psi_r)$, where θ_r is the set of *unknown parameters and ψ_r the set of known parameters (if any)*. Let $\theta = \bigcup\limits_{r=1}^{k} \theta_r$ and $\psi = \bigcup\limits_{r=1}^{k} \psi_r$ be the sets of distinct unknown and known parameters, respectively, in the k populations. Denoting the joint prior distribution of θ given ψ by $g(\theta|\psi)$, then the predictive density of x given the training sample TS, ψ and assuming that x comes from π_r, is:

$$f(x|TS,\ \psi,\ \pi_r\ ) = \int_\theta f(x|\theta_r,\ \psi_r)P(\theta|TS,\ \psi)d\theta \qquad (2.2.1)$$

where $P(\theta|TS,\ \psi)$ is the posterior density of θ given the training sample and ψ, *and is given by:*

$$P(\theta|TS,\ \psi) \propto \ell(TS|\theta,\ \psi)\, g\,(\theta|\psi) \qquad (2.2.2)$$

where $\ell(TS|\theta, \psi)$ is the joint likelihood of the training sample.

When the x_{ij} in the training sample are random observations then $\ell(TS|\theta, \psi)$ becomes:

$$\ell(TS|\theta, \psi) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} f(x_{ij}|\theta_i, \psi_i) \quad . \tag{2.2.3}$$

Finally, given the set $q = \{q_i, i = 1,\ldots,k\}$ of prior probabilities that x belongs to π_i, $i = 1,\ldots,k$, we obtain the posterior probability that x belongs to π_r :

$$P[x \in \pi_r|TS, \psi, q] \propto q_r \, f(x|TS, \psi, \pi_r) \tag{2.2.4}$$

where the constant of proportionality is obtained from:

$$\sum_{r=1}^{k} P[x \in \pi_r|TS, \psi, q] = 1 \tag{2.2.5}$$

For the situation, considered in this thesis, where all the parameters are unknown à priori, all references to ψ_i and ψ are deleted from formulae (2.2.1) to (2.2.5).

Geisser (1964, 1966) gives formulae for the posterior probability given by (2.2.4) for the case where the π_i, $i = 1,\ldots,k$ are each characterised by a univariate or multivariate normal distribution and assuming a noninformative prior distribution for the unknown parameters . Different formulae are given for each of the various possible assumptions about the parameters of these distributions, such as whether they are known or unknown and whether or not some of them are equal for all k populations.

For the case of interest in this thesis, viz, unknown and different mean vectors, and unknown but common covariance matrix for the k populations, Geisser derives the following formulae for the posterior probability that x belongs to π_r, given the training sample TS.

For the univariate case:

$$P[x \in \pi_r | TS, q] = q_r \left(\frac{n_r}{n_r+1}\right)^{\frac{1}{2}} \left\{1 + \frac{n_r(x_{r.}-x)^2}{(n_r+1)(N-k)s^2}\right\}^{-\frac{1}{2}(N-k+1)} \qquad (2.2.6)$$

where,

s^2 is the pooled sample variance

and $N = \sum_{i=1}^{k} n_i$

and for the multivariate (p-dimensional) case:

$$P[x \in \pi_r | TS, q] \propto q_r \left(\frac{n_r}{n_r+1}\right)^{\frac{1}{2}p} \left\{1 + \frac{n_r(x_{r.}-x)' S^{-1}(x_{r.}-x)}{(n_r+1)(N-k)}\right\}^{-\frac{1}{2}(N-k+1)} \qquad (2.2.7)$$

where S is the pooled sample covariance matrix.

Remark 2.2.1 Factors of proportionality that do not affect the probabilities have been omitted from expressions (2.2.6) and (2.2.7) .

Chapter 3 Distribution Theory associated with Classical Discriminant Analysis under the Random Effects Model

In this chapter we consider some of the distributions that arise when applying the random effects model to the classical theory of discriminant analysis.

As mentioned earlier, our concern is to investigate the characteristics of discriminant analysis under the random effects model. In the classical approach this involves assessing the performance of the classification rules derived from this approach, as described in chapter 2, when applied to problems where the k populations have emanated from a random effects model. Thus we are concerned with the performance of future classification problems; once the populations have been chosen the problem becomes a more conventional one of classifying observations of unknown origin into one of k fixed populations.

The assumption underlying the random effects model is that the k populations in any particular application have, in fact, been drawn from the same parent population. If we know the parameters of the parent distribution then we should be able to assess the expected performance of any future classification problem involving k populations randomly chosen from it. (Clearly, k may vary from one application to the next).

As mentioned in chapter 1, we assume that observations from population i have a $N_p(\mu_i,\Sigma)$ distribution and that different μ_i are independent realizations from a $N_p(\xi,T)$ distribution. Intuitively speaking, if T is in some sense large compared to Σ, then we would expect discriminant analysis to perform well. If not, then we cannot expect very reliable classification.

More specifically, if T is large compared to Σ, then we would expect that the Mahalanobis distance:

$$\delta_{ij}^2 = (\mu_i - \mu_j)' \Sigma^{-1}(\mu_i - \mu_j) \tag{3.1}$$

between any two randomly selected populations π_i and π_j would be large. As pointed out by Das Gupta (1972), the probabilities of correct classification under a large class of classification rules (including those considered here), based either on known or estimated parameters, are monotonic increasing functions of the δ_{ij}^2, and so we would expect reliable classification under these circumstances.

This fact is also evident from the various expressions involving δ_{ij}^2 for the probabilities of mis - and correct classification under the classical approach, as given in chapter 2.

Under the random effects model δ_{ij}^2 is a random variable, and it is clear from the preceding discussion that its distribution is of central importance in understanding the characteristics of discriminant analysis under this model. The distribution of δ_{ij}^2 is therefore considered in section 3.1.

Another distance measure appearing in the classification rules described in chapter 2 is the Mahalanobis distance between a new observation x and the i^th population π_i:

$$\delta_i^2(x) = (x - \mu_i)' \Sigma^{-1}(x - \mu_i) \tag{3.2}$$

As mentioned there, the Bayesian classification procedure, when the parameters are known and prior probabilities are equal, is equivalent to classifying x into that population π_i to which it is closest in terms of $\delta_i^2(x)$. Although $\delta_i^2(x)$ does not appear in any of the formulae for the probabilities of mis-and correct classification, its distribution under the random effects model is of interest because of the insight it provides into the relationship between the parameter values and the likelihood of correct classification. The distribution of $\delta_i^2(X)$, where X is the random variable corresponding to x, is considered in Section 3.2.

The sample equivalents of δ_{ij}^2 and $\delta_i^2(x)$ are d_{ij}^2 and $d_i^2(x)$, respectively, where:

$$d_{ij}^2 = (x_{i.} - x_{j.})' \, S^{-1}(x_{i.} - x_{j.}) \qquad (3.3)$$

and

$$d_i^2(x) = (x - x_{i.})' \, S^{-1}(x - x_{i.}) \quad . \qquad (3.4)$$

These two quantities are important in the classical approach to discriminant analysis when the parameters Σ and μ_i, $i=1,\ldots,k$ are unknown and are estimated from training samples. Specifically, d_{ij}^2 appears in some of the expressions in chapter 2 for the probability of misclassification (conditional and unconditional) when the "plug-in" classification rules are used. In turn, these "plug-in" rules, when the prior probabilities are equal, are equivalent to a minimum distance classification rule in terms of the $d_i^2(x)$.

Under the random effects model both d_{ij}^2 and $d_i^2(x)$ are random variables, firstly because of their sampling distributions, and secondly because the underlying parameters μ_i, $i=1,\ldots,k$ in these sampling distributions are themselves random variables. Their distributions are considered in section 3.3 .

3.1 The Distribution of δ_{ij}^2

We now investigate the distribution of $\delta_{ij}^2 = (\mu_i - \mu_j)' \Sigma^{-1}(\mu_i - \mu_j)$ under the random effects model;

i.e. where μ_i and μ_j are independent realizations from a $N_p(\xi, T)$ distribution.

Because μ_i and μ_j are assumed to have been randomly selected from all possible combinations represented by the pair of indices (i,j), $j=1,\ldots,k$; $i\neq j$, the distribution of δ^2_{ij} will not depend on the values of i and j. In this section, therefore, the subscript ij will be omitted and the notation $\delta^2 = \delta^2_{ij}$ will be used.

It will be assumed that Σ is a symmetric positive definite matrix and that T is a symmetric positive definite or semidefinite matrix of rank $r \leq p$. The case where Σ is not of full rank will be given brief consideration.

The main result of this section is given in Theorem 3.1.1, in which the distribution of δ^2 is expressed as a sum of weighted chi-squared random variables. The remainder of the section will be devoted to the properties of this distribution, and in particular to obtaining expressions for the density - and distribution functions of δ^2.

Theorem 3.1.1

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ be the $r(\leq p)$ nonzero eigenvalues of $T\Sigma^{-1}$. Then δ^2 is distributed like:

$$2 \sum_{i=1}^{r} \lambda_i v_i$$

where the v_i are independent χ^2_1 random variables.

Remark 3.1.1 This theorem is an immediate consequence of a result given by Box (1954), a proof of which is given in Johnson and Kotz (1970b), pages 150-1. See also Ruben (1962). However, because of its importance in this thesis, another proof, slightly different from those mentioned above, is given here.

Proof: Let $X = \mu_i - \mu_j$. Then $X \sim N_p(0, 2T)$. Let $T = T_1 T_1'$ where T_1 is the $(p \times r)$ matrix whose columns are the r orthonormal eigenvectors corresponding to the r *nonzero* eigenvaluos of T multiplied by the square root of their respective eigenvalues, and let $X = \sqrt{2}\, T_1 Z$.

Then $Z \sim N_r(0, I_r)$, and

$$\delta^2 = X' \Sigma^{-1} X = 2\, Z' T_1' \Sigma^{-1} T_1 Z = 2\, Z' VZ$$

where $V = T_1' \Sigma^{-1} T_1$.

We can express the $(r \times r)$ symmetric matrix V in the canonical *form:*

$$V = P \Lambda P'$$

where Λ is the diagonal matrix whose diagonal elements are the eigenvalues of V, and P is the orthogonal matrix whose columns are the corresponding orthonormal eigenvectors of V.

Noting that:

$$\text{eigs } \{V\} = \text{eigs } \{T_1' \Sigma^{-1} T_1\} = \text{eigs } \{T_1 T_1' \Sigma^{-1}\} = \text{eigs } \{T \Sigma^{-1}\}$$

we have:

$$\delta^2 = 2\, Z' VZ = 2\, Z' P \Lambda P'Z = 2\, Y' \Lambda Y = 2 \sum_{i=1}^{r} \lambda_i y_i^2$$

where:

$$Y = (y_1, \ldots, y_r)' = P'Z \sim N_r(0, I)$$

and $\{\lambda_i;\ i=1,\ldots,r\}$ are the r nonzero eigenvalues of $T\Sigma^{-1}$. The result now follows from the fact that $v_i = y_i^2$, $i=1,\ldots,r$, are independently and identically *distributed* χ_1^2 random variables.

Remark 3.1.2 The result still holds if Σ is not of full rank, and Σ^{-1} denotes the Moore-Penrose unverse of Σ . (See for example, Graybill (1976).) In this case the summation goes to r_1 where $r_1 = \text{rank}(T\Sigma^{-1})$.

As an immediate result of Theorem 3.1.1, we obtain the following expression for the cumulants of δ^2 :

$$K_s = 2^{2s-1}(s-1)! \sum_{i=1}^{r} \lambda_i^s \qquad s = 1,2,\ldots \tag{3.1.1}$$

In particular, the mean and variance of δ^2 are, respectively:

$$E[\delta^2] = K_1 = 2\sum_{i=1}^{r} \lambda_i = 2\text{Tr } T\Sigma^{-1} \tag{3.1.2}$$

and

$$\text{Var } [\delta^2] = K_2 = 8 \sum_{i=1}^{r} \lambda_i^2 = 8\text{Tr}(T\Sigma^{-1})^2 \tag{3.1.3}$$

The distribution of the sum of weighted, independent chi-squared random variables has received considerable attention in the literature, and infinite series expansions for the density and distribution functions have been obtained in the following three forms:

(i) as Power series

(ii) as Laguerre series

and (iii) as mixtures of chi-squared distributions.

Good reviews of this work have been given by Kotz, Johnson and Boyd (1967a)(with derivations) and by Johnson and Kotz (1970b) chapter 29. In the special case where the eigenvalues are all of even multiplicity, finite series expansions have been obtained. (Robbins (1948) and Box (1954)). A recent article on the power series expansion has been written by Davis (1977).

The simplest approximation to the distribution of the sum of weighted, independent chi-squared random variables is the scaled chi-squared approximation proposed by Satterthwaite (Box, 1954). Other, more accurate approximations have been considered by various authors, a recent article on the subject being by Solomon and Stephens (1977). However, in view of the satisfactory computational experience with the evaluation of the exact distribution as a mixture of chi-squared distributions as reported later in this section, these approximations were not considered in this thesis.

Robbins and Pitman (1949) derive the distribution of the sum of weighted, independent chi-squared random variables as an infinite chi-squared series. Letting

$$Y = \sum_{i=1}^{r} \alpha_i V_i = \alpha_r \sum_{i=1}^{r} a_i V_i \tag{3.1.4}$$

where,

$$\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_r > 0 \;,$$

$$a_i = \alpha_i/\alpha_r \;, \quad i=1,\ldots,r \;, \quad a_r = 1$$

and $V_i \sim \chi^2_{\nu_i}$ independently, $i=1,\ldots,r$,

these authors show that the distribution function of Y can be expressed as:

$$F_Y(y) = \sum_{j=0}^{\infty} c_j^* \, G_{\nu+2j}(y/\alpha_r) \tag{3.1.5}$$

where,

$G_{\nu+2j}(\cdot)$ is the $\chi^2_{\nu+2j}$ distribution function, $\nu = \sum_{i=1}^{r} \nu_i$ and the constants c_j^* are defined by the identity:

$$\prod_{i=1}^{r-1} a_i^{-\frac{1}{2}\nu_i} (1-(1-a_i^{-1})z)^{-\frac{1}{2}\nu_i} = \sum_{j=0}^{\infty} c_j^* \, Z^j \tag{3.1.6}$$

They also provide convenient recursion formulae whereby the

c_j^* may be computed.

Ruben (1960), considering the case where $\nu_i = 1$, $i=1,\ldots,r$ (the case of interest here) derived the following generalization of (3.1.5):

$$F_Y(y) = \sum_{j=0}^{\infty} c_j \, G_{r+2j}(y/\beta) \qquad (3.1.7)$$

where β is an arbitrary positive oonstant and the constants c_j, as in (3.1. 6), are defined by the identity:

$$\prod_{i=1}^{r} (\beta/\alpha_i)^{\frac{1}{2}}(1-(1-\beta/\alpha_i)z)^{-\frac{1}{2}} = \sum_{j=0}^{\infty} c_j \, z^j \qquad (3.1.8)$$

The following recursion formulae for the c_j are also given:

$$c_0 = \prod_{i=1}^{r} (\beta/\alpha_i)^{\frac{1}{2}}$$

$$c_j = \frac{1}{2j} \sum_{i=0}^{j-1} h_{j-i} \, c_i \, , \quad j \geq 1$$

$$\text{where} \quad h_j = \sum_{i=1}^{r} (1-\beta/\alpha_i)^j \qquad (3.1.9)$$

Ruben (1960) proves that for any $\beta > 0$ the series (3.1.7) is unformly convergent in any bounded y-interval of $y > 0$, and uniformly convergent for all $y > 0$ if β is chosen so that $\max\limits_{j=1,\ldots,r} |1-\beta/\alpha_j| < 1$. He also suggests that the value:

$$\beta = 2\alpha_1 \, \alpha_r/(\alpha_1 + \alpha_r) \qquad (3.1.10)$$

may be close to the optimal choice of β as regards the rate of convergence of the infinite series (3.1.7).

Remark 3.1.3 For (3.1.7) to be a true mixture distribution the c_j must be nonnegative and $\sum_{j=0}^{\infty} c_j = 1$. Ruben (1960) shows that, for $0 < \beta \leq \alpha_r$ these criteria are satisfied, so that (3.1.5) is a mixture distribution. (Here $\beta = \alpha_r$). For the choice of β in (3.1.10), (3.1.7) may or may not be a mixture distribution, depending on the actual values of the α_i. If $\beta > r\ (\sum_{i=1}^{r} \lambda_i^{-1})^{-1}$ then (3.1.7) is not a mixture distribution.

The density function of Y is, from (3.1.7):

$$f_Y(y) = \beta^{-1} \sum_{j=0}^{\infty} c_j\ g_{r+2j}(y/\beta) \qquad (3.1.11)$$

where $g_{r+2j}(\cdot)$ is the χ^2_{r+2j} density function.

From Theorem 3.1.1 the distribution of δ^2 has

$$\alpha_i = 2\lambda_i \text{ and } \nu_i = 1, \quad i=1,\ldots,r \qquad (3.1.12)$$

so its distribution and density functions may be expressed as (3.1.7) (or as (3.1.5)) and (3.1.11), respectively.

A major simplification of the distribution of δ^2 results when $\lambda_i = \lambda$, $i=1,\ldots,r$. For then, by the additivity property of the chi-squared distribution:

$$\delta^2/2\lambda \sim \chi^2_r \qquad (3.1.13)$$

Since $\{\lambda_i\}$ = eigs $\{T\Sigma^{-1}\}$ = eigs $\{A^{-1}\ T\ A^{-1'}\}$ where $\Sigma = AA'$, and $A^{-1}\ T\ A^{-1'}$ is a nonnegative definite symmetric matrix, this could only occur when:

$$A^{-1}\ T\ A^{-1'} = \lambda B$$

$$\text{or} \qquad T = \lambda A B A' \qquad (3.1.14)$$

where B is a symmetric idempotent matrix of rank r. (See, for example, Graybill (1976), Theorem 1.7.2).

For r = p (i.e. T is of full rank) condition (3.1.14) implies that:

$$T = \lambda A I A' = \lambda \Sigma$$

i.e. that T is a scalar multiple of Σ.

As mentioned earlier, the probability of correct classification is a monotonic increasing function of δ^2. Therefore, for reliable classification we require the value of δ^2 to be as large as possible. In terms of the distribution of δ^2, this implies not only that the expectation of δ^2 should be large, but also that the probability of low values of δ^2 be low.

Therefore, using Chebychev's inequality, a criterion for establishing whether classification is likely to be reliable (in the sense that the probability of correct classification is large) could be based on the expectation and variance of δ^2; a high value of the former and a low value of the latter indicating the most favourable situation. From expressions (3.1.2) and (3.1.3) for the mean and variance of δ^2, respectively, it is clear that this situation is achieved when $\sum_{i=1}^{r} \lambda_i$ is large and, given $\sum_{i=1}^{r} \lambda_i$, $\sum_{i=1}^{r} \lambda_i^2$ is as small as possible.

So, given $\sum_{i=1}^{r} \lambda_i = \text{Tr } T\Sigma^{-1}$ and r = r(t), the best situation is when the λ_i are all equal, the worst being when one is very large and the rest small. *Furthermore, the greater the rank of T, the better.*

3.1.1 Computing the Density and Distribution functions of δ^2_{ij}

In order to have an idea of the form of the distribution of $\delta^2 = \delta^2_{ij}$, its density and distribution functions were computed using (3.1.7), (3.1.11) and (3.1.12) for particular sets of eigenvalues $\{\lambda_i\}$ of $T\Sigma^{-1}$.

To do this, two Fortran subroutines were written:

CONSTS computes the constants c_j using formulae (3.1.9),

and CHISER computes the chi-squared density and distribution functions, using formulae (2.3.1) and (2.3.2) in Johnson and Kotz (1970a) for the latter, for degrees of freedom starting from r and going up in steps of two for as many terms as necessary to obtain the density and distribution functions of δ^2 to the required level of accuracy. (See (3.1.7) and (3.1.11)).

Finally, using these two subroutines, the density and distribution functions of δ^2 were computed in a main program for values of δ^2 going up in equal steps from zero to an appropriate upper limit. Subroutines CONSTS and CHISER are given in Appendix 3.2.

Using r = 5, three different sets of eigenvalues, all with the same trace, were used, namely {11, 1, 1, 1, 1}, {3, 3, 3, 3, 3} and {5, 4, 3, 2, 1}, representing two extreme situations and one in the middle, respectively. Table 3.1.1 below gives the expected value and standard deviation of δ^2 for each of the three sets of eigenvalues.

Table 3.1.1

Case	Eigenvalues	$E[\delta^2]$	$\sqrt{Var[\delta^2]}$
(a)	11, 1, 1, 1, 1	30.0	31.6
(b)	3, 3, 3, 3, 3	30.0	19.0
(c)	5, 4, 3, 2, 1	30.0	21.0

Figures 3.1.1 and 3.1.2 give the density and distribution functions

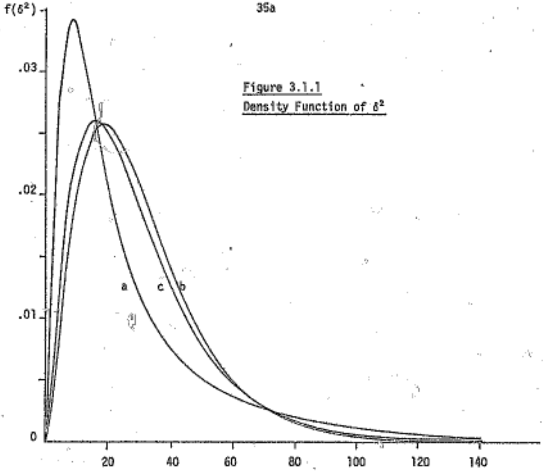


Figure 3.1.1
Density Function of δ²

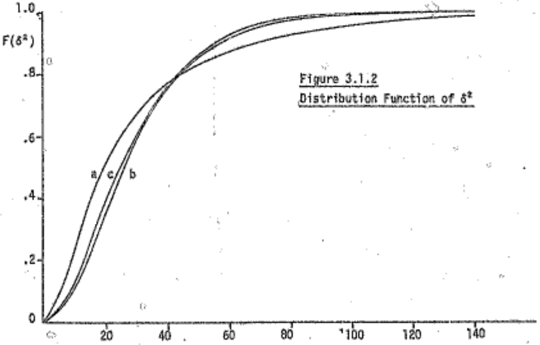


Figure 3.1.2
Distribution Function of δ²

of δ^2, respectively for each of the three cases (a), (b) and (c). From them we clearly see that the remarks concerning the relative magnitudes of the λ_i are borne out in practice.

For example, considering the two - group classification problem, we have from Chapter 2 in the case where the parameters μ_1, μ_2 and Σ will be known and the prior probabilities are equal, that:

$$P[\text{misclassification}] = \Phi(-\tfrac{1}{2}.\sqrt{\delta^2}).$$

Suppose now that we wish this probability to be less than .05. This means that $\frac{1}{2}\sqrt{\delta^2}$ must be greater than 1.64,

$$\text{i.e.: } \delta^2 > (2 \times 1.64)^2 = 10.76$$

From Figure 3.1.2 we see that the probabilities of this occurring in any future classification probability are 0.74, 0.88 and 0.86 respectively, for cases (a), (b) and (c).

3.2 The Distribution of $\delta_i^2(X)$

Using the distribution of δ_{ij}^2 obtained in Section 3.1, we now obtain the distribution of $\delta_i^2(X) = (X-\mu_i)'\,\Sigma^{-1}(X-\mu_i)$ under the assumptions given that section.

Clearly the distribution of $\delta_i^2(X)$ depends on which of the k populations X comes from, so we consider first the situation where X is from π_i.

It follows immediately from the properties of the multivariate normal distribution that in this case $\delta_i^2(X)$ has the central chi-squared distribution on p degrees of freedom.

$$\text{i.e. } \delta_i^2(X)\,|\,X \in \pi_i \sim \chi_p^2 \qquad (3.2.1)$$

When X comes from π_j, $j \neq i$ then, conditional on $\delta^2 = \delta_{ij}^2 = (\mu_i - \mu_j)' \Sigma^{-1}(\mu_i - \mu_j)$, $\delta_i^2(X)$ has a noncentral chi-squared distribution on p degrees of freedom, with noncentrality parameter δ^2.

$$\text{i.e.} \quad \delta_i^2(X) \,|\, X \in \pi_j,\ \delta^2 \sim \chi_p^2(\delta^2) \qquad (3.2.2)$$

Therefore, using the notation $Z = \delta_i^2(X)$, we have the following representation of the conditional density function of $\delta_i^2(X)$ as a mixture of central chi-squared densities:

$$f_{\delta_i^2(X)}(z|X \in \pi_j,\ \delta^2) = \sum_{s=0}^{\infty} \frac{(\frac{1}{2}\delta^2)^s}{s!} e^{-\frac{1}{2}\delta^2} g_{p+2s}(z) \qquad (3.2.3)$$

where $g_{p+2s}(z)$ is the density function of the χ_{p+2s}^2 distribution.

The unconditional distribution of z is now obtained by integrating $f_{\delta_i^2(X)}(z|X \in \pi_j,\ \delta^2)$, as given in (3.2.3), over the distribution of δ^2. This is done most conveniently by using the fact that conditional on δ^2 the distribution of z is a mixture of a central chi-square distributions with p + 2S degrees of freedom where the mixing is done over the variable S which, as is evident from (3.2.3), has a Poisson distribution with parameter $\frac{1}{2}\delta^2$.

Since only the distribution of S depends on δ^2, its unconditional distribution will first be obtained and this will then be substituted into (3.2.3) to give the unconditional distribution of z.

$$\text{So} \quad P[S = s] = \int_0^{\infty} P[S = s|\delta^2]\, f_{\delta^2}(\delta^2)d\delta^2 \qquad (3.2.4)$$

where,

$$P[S = s|\delta^2] = \frac{(\frac{1}{2}\delta^2)^s}{s!} e^{-\frac{1}{2}\delta^2}$$

and $f_{\delta^2}(\delta^2)$ is the density function of δ^2.

Using expressions (3.1.11) and (3.1.12), $f^{\cdot}_{\delta^2}(\delta^2)$ can be written in the following form:

$$f_{\delta^2}(\delta^2) = \beta^{-1} \sum_{j=0}^{\infty} c_j \; g_{r+2j}(\delta^2/\beta) \qquad (3.2.5)$$

where,

β is an arbitrary positive constant,

the c_j are given by formulae (3.1.9) with $\alpha_i = 2\lambda_i$, $i=1,\dots,r$

and $g_{r+2j}(\cdot)$ is the density function of the χ^2_{r+2j} distribution.

Substituting (3.2.5) into (3.2.4) and interchanging the order of summation and integration (this is justified by the uniform convergence of the series (3.2.5) for all $\delta^2 > 0$ when β is chosen appropriately - see the comment following (3.1.9))yields:

$$P[S = s] = \sum_{j=0}^{\infty} \frac{c_j \; \beta^{-(\frac{1}{2}r+j)}}{\Gamma(\frac{1}{2}r+j)2^{\frac{1}{2}r+j+s} \; s!} \int_0^{\infty} (\delta^2)^{\frac{1}{2}r+j+s-1} e^{-\frac{1}{2}\delta^2(1+\beta^{-1})} d\delta^2$$

The integral is readily evaluated as a gamma function, giving:

$$P[S = s] = (1+\beta^{-1})^{-s} \; (\Gamma(s+1))^{-1} \sum_{j=0}^{\infty} \frac{c_j \; \Gamma(\frac{1}{2}r+j+s)}{(1+\beta)^{\frac{1}{2}r+j} \; \Gamma(\frac{1}{2}r+j)} \qquad (3.2.6)$$

The unconditional density of $z = \delta_i^2(X)$ is now obtained by replacing the Poisson distribution by (3.2.6) as mixing distribution in (3.2.3), yielding:

$$f_{\delta_i^2(X)}(z|X \in \pi_j) = \sum_{s=0}^{\infty} a_s \; g_{p+2s}(z) \qquad (3.2.7)$$

where $a_s = P[S = s]$ as given in (3.2.6).

The mean and variance of $\delta_i^2(X)$ are most easily evaluated from expression (3.2.1) when $X \in \pi_i$, and from (3.2.7) when $X \in \pi_j$, $j \neq i$. For the first case we immediately get:

$$E[\delta_i^2(X) | X \in \pi_i] = E[\chi_p^2] = p \qquad (3.2.8)$$

and

$$\mathrm{Var}[\delta_i^2(X) | X \in \pi_i] = \mathrm{Var}[\chi_p^2] = 2p \qquad (3.2.9)$$

For $X \in \pi_j$, $j \neq i$, we use the following well-known results on conditional expectations:

$$E[\delta_i^2(X)] = E_s[E[\delta_i^2(X)|s]] \qquad (3.2.10)$$

and

$$\mathrm{Var}[\delta_i^2(X)] = E_s[\mathrm{Var}[\delta_i^2(X)|s]] + \mathrm{Var}_s[E[\delta_i^2(X)|s]] \qquad (3.2.11)$$

where $E_s[\cdot]$ and $\mathrm{Var}_s[\cdot]$ denote the expectation and variance, respectively, of $\cdot$, taken over the distribution of S. Now, from (3.2.7), conditional on $S = s$, $\delta_i^2(X)$ has a χ_{p+2s}^2 distribution, whence

$$E[\delta_i^2(x)|s] = p + 2s$$

$$\text{and} \qquad \mathrm{Var}[\delta_i^2(x)|s] = 2p + 4s.$$

Applying these to (3.2.10) and (3.2.11) we get:

$$E[\delta_i^2(X)] = E_s[p + 2s] = p + 2E_s[s] \qquad (3.2.12)$$

and $\quad \mathrm{Var}[\delta_i^2(X)] = E_s[2p+4s] + \mathrm{Var}_s[p+2s]$

$$= 2p+4\,E_s[s] + 4\,\mathrm{var}_s[s] \qquad (3.2.13)$$

Furthermore, conditional on $\delta^2 = \delta_{ij}^2$, S has a Poisson distribution with with parameter $\delta^2/2$, so using the above results on conditional expectations to find the mean and variance of S, we get

$$E[s] = E_{\delta^2}[E[s|\delta^2]] = E_{\delta^2}\left[\frac{\delta^2}{2}\right] = \frac{1}{2}\cdot 2\sum_{\ell=1}^{r}\lambda_\ell \quad \text{from (3.1.2)}$$

$$= \sum_{\ell=1}^{r}\lambda_\ell \qquad (3.2.14)$$

and $\quad \mathrm{Var}[s] = E_{\delta^2}[\mathrm{Var}[s|\delta^2]] + \mathrm{Var}_{\delta^2}[E[s|\delta^2]]$

$$= E_{\delta^2}\left[\frac{\delta^2}{2}\right] + \mathrm{Var}_{\delta^2}\left[\frac{\delta^2}{2}\right]$$

$$= \frac{1}{2}\cdot 2\sum_{\ell=1}^{r}\lambda_\ell + \frac{1}{4}\cdot 8\sum_{\ell=1}^{r}\lambda_\ell^2 \quad \text{from (3.1.2) and (3.1.3)}$$

$$= \sum_{\ell=1}^{r}\lambda_i + 2\sum_{\ell=1}^{r}\lambda_\ell^2 \qquad (3.2.15)$$

Finally, substituting (3.2.14) and (3.2.15) into (3.2.12) and (3.2.13) and simplifying, we get

$$E[\delta_i^2(X)|X\in\pi_j] = p + 2\sum_{\ell=1}^{r}\lambda_\ell \qquad (3.2.16)$$

and

$$\mathrm{Var}[\delta_i^2(X)|X\in\pi_j] = 2p + 8\left\{\sum_{\ell=1}^{r}\lambda_\ell + \sum_{\ell=1}^{r}\lambda_\ell^2\right\} \qquad (3.2.17)$$

Remark 3.2.1 Although the uniform convergence of expression (3.2.7) for the density of $\delta_i^2(X)|X\in\pi_j$ is difficult to establish directly, the existence of the (finite) expectation (3.2.16) implies it, by the Lebesgue Dominated Convergence Theorem. It is therefore permissible to integrate under the summation sign in (3.2.7), yielding the following expression for the distribution function of $\delta_i^2(X)|X \in \pi_j$:

$$P[\delta_i^2(X) \le z|X \in \pi_j] = \sum_{s=0}^{\infty} a_s \, G_{p+2s}(z) \qquad (3.2.18)$$

where,

$G_{p+2s}(z)$ is the χ^2_{p+2s} distribution function and $a_s = P[S = s]$ is given in (3.2.6) .

Remark 3.2.2 Comparing expressions (3.2.8) and (3.2.16) and recalling that x is classified into that population π_i for which $\delta_i^2(x)$ is a *minimum*, clearly demonstrates the importance, *for reliable classification*, of having $\sum_{\ell=1}^{r} \lambda_\ell = Tr(T\Sigma^{-1})$ *as large as possible*. Furthermore, as in the case with δ_{ij}^2 , expression (3.2.17) for the variance of $\delta_i^2(X)|X \in \pi_j$ shows that, for given $\sum_{\ell=1}^{r} \lambda_\ell$, $\sum_{\ell=1}^{r} \lambda_\ell^2$ should be as small as possible, i.e. the λ_i should all be equal and $r = r(T)$ should be as large as possible, for the most reliable classification.

3.2.1 Computing the Density and Distribution functions of δ_i^2 (X)

As in Section 3.1, the density and distribution functions of $\delta_i^2(X)$ were computed for particular sets of parameter values, using (3.2.1), (3.2.7) and (3.2.18). The constants a_s, given in (3.2.7) and (3.2.6) were computed using the Fortran subroutine CONST1, given in Appendix 3.2, and the chi-squared density and distribution functions were computed using the subroutine CHISER, described in Section 3.1 .

The same three sets of eigenvalues as used in Section 3.1 were used for the distribution of $\delta_i^2(X)|X \notin \pi_i$, and the distribution of $\delta_i^2(X)|X \in \pi_i$ was also computed. The expected value and standard deviation of $\delta_i^2(X)$ for each of these cases are given in Table 3.2.1 and the

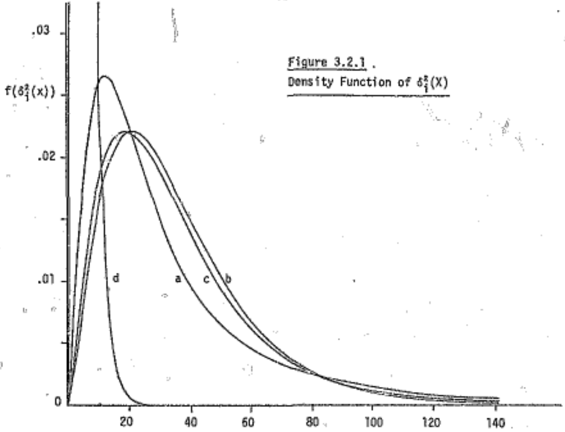


Figure 3.2.1 . Density Function of $\delta_1^2(X)$

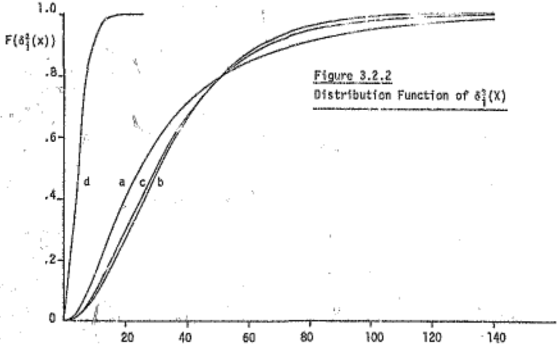


Figure 3.2.2 Distribution Function of $\delta_1^2(X)$

density and distribution functions are given in Figures 3.2.1 and 3.2.2, respectively.

Table 3.2.1

Case	Eigenvalues	$E[\delta_i^2(X)]$	$\sqrt{Var[\delta_i^2(X)]}$
(a) $X \notin \pi_i$	11, 1, 1, 1, 1	35.0	33.6
(b) $X \notin \pi_i$	3, 3, 3, 3, 3	35.0	22.1
(c) $X \notin \pi_i$	5, 4, 3, 2, 1	35.0	23.9
(d) $X \in \pi_i$		5.0	3.2

As in the previous section, these figures confirm the general remarks, made under Remark 3.2.2, regarding the desirability of having the λ_i as close together as possible.

§ 3.3 The distribution of d_{ij}^2 and $d_i^2(X)$

In this section we consider the distributions, under our random effects model, of the two statistics d_{ij}^2 and $d_i^2(X)$ of interest in discriminant analysis when the parameters μ_i, $i=1,\ldots,k$ and Σ are unknown and have to be estimated from a training sample.

Specifically, suppose we have the training sample:

$$x_{ij}\ , \quad j=1,\ldots,n_i \ ; \quad i=1,\ldots,k$$

from the k populations π_i, $i=1,\ldots,k$, where the x_{ij} are p-dimensional random vectors.

Under the assumptions enumerated earlier:

$$x_{ij} \sim N_p(\mu_i, \Sigma) \quad \text{independently,} \quad \forall i,j \ .$$

As usual, the maximum likelihood estimators are, for μ_i, i=1,...,k:

$$\hat{\mu}_i = x_{i.} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij} \quad i=1,\ldots,k \qquad (3.3.1)$$

and for Σ (corrected for bias):

$$\hat{\Sigma} = S = \nu^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - x_{i.})(x_{ij} - x_{i.})' \qquad (3.3.2)$$

$$\text{where} \quad \nu = \sum_{i=1}^{k} (n_i - 1)$$

and from standard multivariate normal theory we know that:

$$x_{i.} \sim N_p(\mu_i, n_i^{-1}\Sigma) \quad i=1,\ldots,k \quad \text{independently}$$

and (3.3.3)

$$\nu S \sim W_p(\Sigma, \nu) \quad \text{independently of the } x_{i.}$$

where $W_p(\Sigma, \nu)$ denotes the p-dimensional Wishart distribution with ν degrees of freedom and parameter matrix Σ .

The two statistics are defined as follows:

$$d_{ij}^2 = (x_{i.} - x_{j.})' S^{-1}(x_{i.} - x_{j.}) \quad i,j=1,\ldots,k; \quad i\neq j \quad (3.3.4)$$

and

$$d_i^2(X) = (X - x_{i.})' S^{-1}(X - x_{i.}) \quad i=1,\ldots,k \qquad (3.3.5)$$

where X is a random observation from one of the π_i, i=1,...,k .

We will first consider the distribution of d_{ij}^2. From (3.3.3) it follows immediately that:

$$x_{i.} - x_{j.} \sim N_p\left(\mu_i - \mu_j, \left(\frac{n_i + n_j}{n_i\, n_j}\right) \Sigma\right)$$

and therefore that, conditional on $\mu_i - \mu_j$, $\left(\frac{n_i\, n_j}{n_i + n_j}\right) d_{ij}^2$ follows a noncentral p-dimensional Hotelling's T^2 distribution with ν degrees of freedom. (See Anderson (1958), chapter 5 or Giri (1977), chapter 7). Therefore, conditional on α^2,

$$\left(\frac{n_i\, n_j}{n_i + n_j}\right) \frac{(\nu - p + 1)}{\nu p} d_{ij}^2 \sim F(p, \nu - p + 1; \alpha^2) \qquad (3.3.6)$$

where,

$$\alpha^2 = \left(\frac{n_i\, n_j}{n_i + n_j}\right) (\mu_i - \mu_j)' \Sigma^{-1}(\mu_i - \mu_j)$$

$$= \left(\frac{n_i\, n_j}{n_i + n_j}\right) \delta_{ij}^2$$

and $F(\nu_1, \nu_2; \alpha^2)$ denotes the noncentral F distribution with ν_1 and ν_2 degrees of freedom and noncentrality parameter α^2.

It will be more convenient in what follows to work with the unnormed noncentral f-distribution, $f(\nu_1, \nu_2; \alpha^2)$(see, for example C.R. Rao (1965), pp 175-6), so if we let

$$z = \left(\frac{n_i\, n_j}{n_i + n_j}\right) \nu^{-1}\, d_{ij}^2 \qquad (3.3.7)$$

then, conditional on α^2,

$z \sim f(p, \nu - p + 1; \alpha^2)$ and therefore has density function:

$$f_z(z|\alpha^2) = \sum_{s=0}^{\infty} \frac{(\frac{1}{2}\alpha^2)^s}{s!} e^{-\frac{1}{2}\alpha^2} g_{p+2s,\ \nu-p+1}(z) \qquad (3.3.8)$$

where

$$g_{p+2s,\ \nu-p+1}(z) = \frac{\Gamma(\frac{1}{2}(\nu+1)+s)}{\Gamma(\frac{1}{2}p+s)\Gamma(\frac{1}{2}(\nu-p+1))} \frac{z^{\frac{1}{2}p+s-1}}{(1+z)^{\frac{1}{2}(\nu+1)+s}}$$

$$(3.3.9)$$

is the density function of the central unnormed f-distribution with $p + 2s$ and $\nu - p + 1$ degrees of freedom, which we will denote by $f(p + 2s, \nu - p + 1)$.

To obtain the unconditional distribution of z we now integrate $f_z(z|\alpha^2)$ over the distribution of

$$= \left(\frac{n_i\ n_j}{n_i + n_j}\right) \delta^2_{ij}$$

where the distribution of $\delta^2 = \delta^2_{ij}$ is given in section 3.1 . As in section 3.2, we note from (3.3.8) that the conditional distribution of z is a mixture of unnormed f-distributions with $p + 2s$ and $\nu - p + 1$ degrees of freddom, where the mixing variable S has a Poisson distribution with parameter $\frac{1}{2}\alpha^2$. Noting that the density function of α^2 is, from (3.2.5);

$$f_{\alpha^2}(\alpha^2) = \left(\frac{n_i\ n_j}{n_i + n_j}\right)^{-1} f_{\delta^2}(\alpha^2 \left(\frac{n_i\ n_j}{n_i + n_j}\right)^{-1})$$

$$= \left(\frac{n_i\ n_j\ \beta}{n_i + n_j}\right)^{-1} \sum_{j=0}^{\infty} c_j\ g_{r+2j}(\alpha^2 \left(\frac{n_i\ n_j\ \beta}{n_i + n_j}\right)^{-1})$$

$$(3.3.10)$$

it is clear that the unconditional distribution of S is exactly the same as in Section 3.2, with β replaced by $\frac{n_i n_j \beta}{n_i + n_j}$. The unconditional density of z therefore becomes:

$$f_z(z) = \sum_{s=0}^{\infty} a_s^* g_{p+2s,\ \nu-p+1}(z) \tag{3.3.11}$$

where,

$g_{p+2s,\ \nu-p+1}(z)$ is the density function of the $f(p+2s,\ \nu-p+1)$ distribution given in (3.3.9),

$$a_s^* = \left[1 + \left(\frac{n_i n_j \beta}{n_i + n_j}\right)^{-1}\right]^{-s} (\Gamma(s+1))^{-1} \sum_{j=0}^{\infty} \frac{c_j \Gamma(\frac{1}{2} r + j + s)}{\left(1 + \frac{n_i n_j \beta}{n_i + n_j}\right)^{\frac{1}{2}r+j} \Gamma(\frac{1}{2}r+j)} \tag{3.3.12}$$

and the c_j are given by formulae (3.1.9) with $\alpha_i = 2\lambda_i$, $i=1,\ldots,r$.

Finally, transforming back to d_{ij}^2 using (3.3.7) we get the following expression for its density function:

$$f_{d_{ij}^2}(d_{ij}^2) = \left(\frac{n_i n_j}{n_i + n_j}\right) \nu^{-1} \sum_{s=0}^{\infty} a_s^* g_{p+2s,\ \nu-p+1}\left\{\left(\frac{n_i n_j}{n_i + n_j}\right) \nu^{-1} d_{ij}^2\right\} \tag{3.3.13}$$

The mean and variance of d_{ij}^2 are also most readily found in the manner of Section 3.2, the details of which may be found in Appendix 3.1, yielding:

$$E[d_{ij}^2] = \frac{\nu}{\nu - p - 1} \left[\left(\frac{n_i + n_j}{n_i n_j}\right) p + 2 \sum_{\ell=1}^{r} \lambda_\ell\right] \tag{3.3.14}$$

and

$$\mathrm{Var}[d_{ij}^2] = \frac{2\nu^2}{(\nu-p-1)^2(\nu-p-3)} \left\{ \left(\frac{n_i+n_j}{n_i\,n_j}\right)^2 (\nu-1)p + 4\left(\frac{n_i+n_j}{n_i\,n_j}\right)(\nu-1) \sum_{\ell=1}^{r} \lambda_\ell \right.$$

$$\left. + 4\left(\sum_{\ell=1}^{r} \lambda_\ell\right)^2 + 4(\nu-p-1) \sum_{\ell=1}^{r} \lambda_\ell^2 \right\} \qquad (3.3.15)$$

The existence of the (finite) mean of d_{ij}^2 permits integration under the summation sign in (3.3.13) (see Remark 3.2.1) yielding the following expression for the distribution function of d_{ij}^2 :

$$P[d_{ij}^2 \le z] = \sum_{s=0}^{\infty} a_s^* \, G_{p+2s,\ \nu-p+1} \left\{ \left(\frac{n_i\,n_j}{n_i+n_j}\right) \nu^{-1} z \right\} \qquad (3.3.16)$$

where $G_{p+2s,\ \nu-p-1}(\cdot)$ is the $f(p+2s,\ \nu-p+1)$ distribution function.

<u>Remark 3.3.1</u> For the balanced situation where the training sample contains the same number n from each of the k populations, all the relevant formulae of this section may be simplified by replacing $n_i\,n_j/(n_i+n_j)$ by $\frac{2}{n}$ wherever it appears. For example, the mean and variance of d_{ij}^2 become :

$$E[d^2] = E[d_{ij}^2] = \left(\frac{2\nu}{\nu-p-1}\right)\left(\frac{p}{n} + \sum_{\ell=1}^{r} \lambda_\ell\right) \qquad (3.3.17)$$

and

$$\mathrm{Var}[d^2] = \mathrm{Var}[d_{ij}^2] = \frac{8\nu^2}{(\nu-p-1)^2(\nu-p-3)} \left(\frac{(\nu-1)p}{n^2} + \frac{2(\nu-1)}{n} \sum_{\ell=1}^{r} \lambda_\ell \right.$$

$$\left. + \left(\sum_{\ell=1}^{r} \lambda_\ell\right)^2 + (\nu-p-1) \sum_{\ell=1}^{r} \lambda_\ell^2 \right\} \qquad (3.3.18)$$

Note further that for large ν and n expressions (3.3.17) and (3.3.18) tend to the corresponding expressions (3.1.2) and (3.1.3) for the mean and variance, respectively, of δ_{ij}^2 .

The distribution of $d_i^2(X)$ depends on which of the k populations X comes from. If X belongs to π_i, then it follows immediately from the definition (3.3.5) of $d_i^2(X)$ that $\left(\frac{n_i}{n_i+1}\right) d_i^2(X)$ follows a central p-dimensional Hotelling's T^2 distribution with ν degrees of freedom. Therefore:

$$\left(\frac{n_i}{n_i+1}\right) \frac{(\nu-p+1)}{\nu p} d_i^2(X) | X \in \pi_i \sim F(p, \nu-p+1) \qquad (3.3.19)$$

where $F(p, \nu-p+1)$ denotes the central (normed) F-distribution with p and $\nu-p+1$ degrees of freedom.

If X belongs to π_j, $j \neq i$, then from (3.3.4) and (3.3.5) it is clear that the distribution of $d_i^2(X)$ is the same as that for d_{ij}^2 with n_j equal to 1. Therefore, using expressions (3.3.13) and (3.3.16) we immediately obtain the following expressions for the density and distribution functions of $d_i^2(X)$:

$$f_{d_i^2(X)}(d_i^2(X) | X \notin \pi_i) = \left(\frac{n_i}{n_i+1}\right) \nu^{-1} \sum_{s=0}^{\infty} a_s^* \, g_{p+2s,\nu-p+1}\left[\left(\frac{n_i}{n_i+1}\right)\nu^{-1} \, d_i^2(X)\right] \qquad (3.3.20)$$

$$P[d_i^2(X) \le z | X \notin \pi_i] = \sum_{s=0}^{\infty} a_s^* \, G_{p+2s,\nu-p+1}\left(\left(\frac{n_i}{n_i+1}\right)\nu^{-1} \, z\right) \qquad (3.3.21)$$

where $g_{p+2s,\,\nu-p+1}(\cdot)$ and $G_{p+2s,\,\nu-p+1}(\cdot)$ are defined in (3.3.11) and (3.3.16) respectively, and a_s^* is defined in (3.3.12) with n_j equal to 1.

The mean and variance of $d_i^2(X)$ follow immediately from (3.3.19) for the case where $X \in \pi_i$:

$$E[d_i^2(X) | X \in \pi_i] = \left(\frac{n_i+1}{n_i}\right) \frac{\nu p}{\nu-p-1} \qquad (3.3.22)$$

and

$$\mathrm{Var}[d_i^2(X) \mid X \in \pi_i] = 2\left(\frac{n_i+1}{n_i}\right)^2 \frac{\nu^2(\nu-1)p}{(\nu-p-1)^2(\nu-p-3)} \tag{3.3.23}$$

and from (3.3.14) and (3.3.15) with $n_j = 1$ when $x \notin \pi_i$:

$$E[d_i^2(X) \mid X \notin \pi_i] = \frac{\nu}{\nu-p-1}\left\{\left[\frac{n_i+1}{n_i}\right]p + 2\sum_{\ell=1}^{r} \lambda_\ell\right\} \tag{3.3.24}$$

$$\mathrm{Var}[d_i^2(X) \mid X \notin \pi_i] = \frac{2\nu^2}{(\nu-p-1)^2(\nu-p-3)}\left\{\left(\frac{n_j+1}{n_i}\right)^2(\nu-1)p+4\left(\frac{n_i+1}{n_i}\right)(\nu-1)\sum_{\ell=1}^{r}\lambda_\ell + 4\left(\sum_{\ell=1}^{r}\lambda_\ell\right)^2 + 4(\nu-p-1)\sum_{\ell=1}^{r}\lambda_\ell^2\right\} \tag{3.3.25}$$

Remark 3.3.2 As in the case of d_{ij}^2 we note that for large ν and n_i the mean and variance of $d_i^2(X)$ tend to the corresponding expressions for $\delta_i^2(X)$ given in Section 3.2, both when $X \in \pi_i$ and when $X \notin \pi_i$. In view of this, the remarks concerning the magnitudes of $\sum_{\ell=1}^{r} \lambda_\ell$ and $\sum_{\ell=1}^{r} \lambda_\ell^2$ as related to the reliability of classification when the parameters are known, made in Sections3.1 and 3.2, also pertain to the situation when the classification rules are based on estimated parameters, discussed in this section.

Remark 3.3.3 The constants a_s^* in the distributions of d_{ij}^2 and $d_i^2(X) \mid X \notin \pi_i$ are the same as the constants a_s in Section 3.2, with the parameter β replaced by $\left(\frac{n_i\, n_j}{n_i+n_j}\right)\beta$ ($n_j = 1$ in the case of $d_i^2(X)$). Therefore the subroutine CONST1, used to compute the a_s may also be used for the a_s^*. So, as done in Sections 3.1 and 3.2, the density and distribution functions of $d_i^2(X)$ may be computed using a subroutine that computes sequences of density and distribution function values for the

$f(p+2s, \nu-p+1)$ distribution for values of s increasing from zero in steps of one, as done for the chi-squared distribution by the subroutine CHISER.

Appendix 3.1 Derivation of the Mean and Variance of d_{ij}^2

From (3.3.8) we have that, if

$$z = \left(\frac{n_i\, n_j}{n_i + n_j}\right)\nu^{-1}\, d_{ij}^2$$

and

$$\alpha^2 = \left(\frac{n_i\, n_j}{n_i + n_j}\right)\delta_{ij}^2$$

then, conditional on α^2, the distribution of z is a mixture of unnormed f-distributions with $p + 2S$ and $\nu - p + 1$ degrees of freedom, where S has a Poisson distribution with parameter $\frac{1}{2}\alpha^2$. Given $S = s$, therefore, z has the following conditional mean and variance (See, for example, Johnson and Kotz (1970b)):

$$E[z|s] = \frac{p+2s}{\nu-p-1} = \frac{p}{\nu-p-1} + \left(\frac{2}{\nu-p-1}\right)s \qquad \text{(A 3.1.1)}$$

and

$$\text{Var}[z|s] = \frac{2(p+2s)(\nu+2s-1)}{(\nu-p-1)^2(\nu-p-3)} = \frac{2}{(\nu-p-1)^2(\nu-p-3)}\,(p(\nu-1)+2(\nu+p-1)s+4s^2) \qquad \text{(A 3.1.2)}$$

Using (3.2.14), (3.2.15) and the relationship between α^2 and δ_{ij}^2 given above, we immediately get:

$$E[s] = \left(\frac{n_i\ n_j}{n_i + n_j}\right) \sum_{\ell=1}^{r} \lambda_\ell \qquad \text{(A 3.1.3)}$$

and

$$\mathrm{Var}[s] = \left(\frac{n_i\ n_j}{n_i + n_j}\right) \sum_{\ell=1}^{r} \lambda_\ell + 2\left(\frac{n_i\ n_j}{n_i + n_j}\right)^2 \sum_{\ell=1}^{r} \lambda_\ell^2 \qquad \text{(A 3.1.4)}$$

We now apply results (3.2.10) and (3.2.11) to (A 3.1.1) and (A 3.1.2) to obtain the unconditional mean and variance of z.

$$E[z] = E_s[E[z|s]] = \frac{p}{\nu-p-1} + \left(\frac{2}{\nu-p-1}\right) E[s]$$

$$\left(\frac{1}{\nu-p-1}\right)(p + 2\left(\frac{n_i\ n_j}{n_i\ n_j}\right) \sum_{\ell=1}^{r} \lambda_\ell\ ) \qquad \text{(A 3.1.5)}$$

$$\mathrm{Var}[z] = E_s[\mathrm{Var}[z|s]] + \mathrm{Var}_s[E[z|s]]$$

$$= \frac{2}{(\nu-p-1)^2(\nu-p-3)}\ (p(\nu-1) + 2(\nu+p-1)E[s] + 4E[s^2]\}$$

$$+ \left(\frac{2}{\nu-p-1}\right)^2 \mathrm{Var}[s]$$

Using (A 3.1.3) and (A 3.1.4) and the fact that $E[s^2] = \mathrm{Var}[s^2] + (E[s])^2$ we get, after a little simplification,

$$\mathrm{Var}[z] = \frac{2}{(\nu-p-1)^2(\nu-p-3)}\ \{p(\nu-1) + 4\left(\frac{n_i\ n_j}{n_i + n_j}\right)(\nu-1) \sum_{\ell=1}^{r} \lambda_\ell$$

$$+ 4\left(\frac{n_i\ n_j}{n_i + n_j}\right)^2 \left(\sum_{\ell=1}^{r} \lambda_\ell\right)^2 + 4\left(\frac{n_i\ n_j}{n_i + n_j}\right)^2 (\nu-p-1) \sum_{\ell=1}^{r} \lambda_\ell^2\ \}$$

(A 3.1.6)

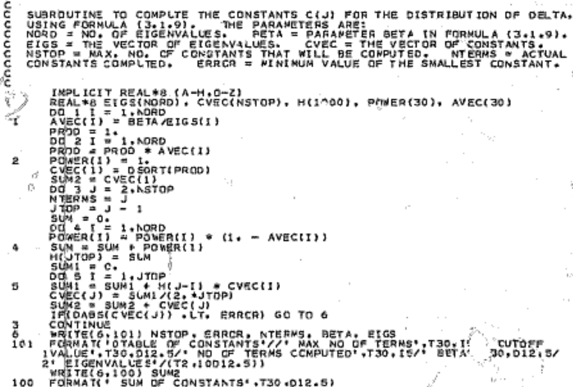
Finally, transforming back to d_{ij}^2 we get:

$$E[d_{ij}^2] = \left(\frac{n_i+n_j}{n_i\,n_j}\right)\nu\,E[z] = \left(\frac{\nu}{\nu-p-1}\right)\left\{\left(\frac{n_i+n_j}{n_i\,n_j}\right)p + 2\sum_{\ell=1}^{r}\lambda_\ell\right\} \qquad \text{(A 3.1.7)}$$

and

$$\begin{aligned}\mathrm{Var}[d_{ij}^2] &= \left(\frac{n_i+n_j}{n_i\,n_j}\right)^2 \nu^2\,\mathrm{Var}[z] \\ &= \frac{2\nu^2}{(\nu-p-1)^2(\nu-p-3)}\left\{\left(\frac{n_i+n_j}{n_i\,n_j}\right)^2 p(\nu-1) + 4\left(\frac{n_i+n_j}{n_i\,n_j}\right)(\nu-1)\sum_{\ell=1}^{r}\lambda_\ell \right. \\ &\qquad \left. + 4\left(\sum_{\ell=1}^{r}\lambda_\ell\right)^2 + 4(\nu-p-1)\sum_{\ell=1}^{r}\lambda_\ell\right\} \qquad \text{(A 3.1.8)}\end{aligned}$$

Appendix 3.2 Fortran Subroutines used in computing the Density and Distribution Functions of δ^2_{ij} and $\delta^2_i(X)$



```
      SUBROLTINE CENSTS( NORD ,BETA,EIGS,CVEC,NSTOP,NTERMS,ERROR)
C
C
C   SUBROUTINE TO COMPLTE THE CONSTANTS C(J) FOR THE DISTRIBUTION OF DELTA,
C   USING FORMULA (3.1.9).   THE PARAMETERS ARE:
C   NORD = NO. OF EIGENVALUES.   PETA = PARAMETER BETA IN FORMULA (3.1.9).
C   EIGS = THE VECTOR OF EIGENVALUES.   CVEC = THE VECTOR OF CONSTANTS.
C   NSTOP = MAX. NO. CF CONSTANTS THAT WILL BE COMPUTED.   NTERMS = ACTUAL NO.
C   CONSTANTS COMPLTED.   ERRCR = MINIMUM VALUE OF THE SMALLEST CONSTANT.
C
C
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 EIGS(NORD), CVEC(NSTOP), H(1000), POWER(30), AVEC(30)
      DO 1 I = 1,NORD
 1    AVEC(I) = BETA/EIGS(I)
      PROD = 1.
      DO 2 I = 1,NORD
      PROD = PROD * AVEC(I)
 2    POWER(I) = 1.
      CVEC(1) = DSORT(PROD)
      SUM2 = CVEC(1)
      DO 3 J = 2,NSTOP
      NTERMS = J
      JTOP = J - 1
      SUM = 0.
      DO 4 I = 1,NORD
      POWER(I) = POWER(I) * (1. - AVEC(I))
 4    SUM = SUM + POWER(I)
      H(JTOP) = SLM
      SUM1 = C.
      DO 5 I = 1,JTOP
 5    SUM1 = SUM1 + H(J-I) * CVEC(I)
      CVEC(J) = SUM1/(2.*JTOP)
      SUM2 = SUM2 + CVEC(J)
      IF(DABS(CVEC(J)) .LT. ERRCR) GO TO 6
 3    CONTINUE
 6    WRITE(6,101) NSTOP, ERRCR, NTERMS, BETA, EIGS
 101  FORMAT('0TABLE OF CONSTANTS'//' MAX NO OF TERMS',T30,I5/' CUTOFF
     1VALUE',T30,D12.5/' NO OF TERMS CCMPUTED',T30,I5/' BETA',T30,D12.5/
     2' EIGENVALUES'/(T2,10D12.5))
      WRITE(6,100) SUM2
 100  FORMAT(' SUM OF CONSTANTS',T30,D12.5)
      WRITE(6,102) (CVEC(I),I=1,NTERMS)
 102  FORMAT('CCONSTANTS'/(T2,10D12.5))
      RETURN
      END
      SUBROUTINE CONST1(NORD,BETA,FACT,CVEC,DVEC,NTERMS,NSTOP,NMAX,ERROR
     1,ERROR1)
C
C
C  SUBROUTINE TO COMPUTE THE CONSTANTS A(S) FOR THE DISTRIBUTION OF DELTA(X).
C  USING FORMULA (3.2.6), OR FOR THE DISTRIBUTION OF E OR D(X) USING FORMULA
C  (3.3.12).    PARAMETERS ARE:
C  NORD = NO. OF EIGENVALUES.    BETA = PARAMETER BETA IN THE FORMULAE.
C  FACT = 1. FOR (3.2.6) AND = N(I)*N(J)/(N(I)+N(J)) FOR (3.3.12).    CVEC =
C  VECTOR OF CONSTANTS C(J) FROM SUBROUTINE CONSTS.    NTERMS = NO. OF ELEMENT
C  IN CVEC.    NSTOP = MAX NO. OF CONSTANTS THAT WILL BE COMPUTED.
C  NMAX = ACTUAL NO. OF CONSTANTS COMPUTED.    ERROR = CUTOFF VALUE FOR
C  CALCULATING CONSTANTS.    ERROR1 = MINIMUM VALUE OF SMALLEST CONSTANT.
C
C
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 CVEC(NTERMS), DVEC(NSTOP), COEFFT(1000)
      INTEGER N(1000)
      BET  = BETA * FACT
      BP1INV = 1./(1.+BET)
      BINP1I = 1./(1. + 1./BET)
      AND2 = NORD/2.
      TERM = BP1INV**AND2
      SUM = 0.
      DO 1 J = 1,NTERMS
      COEFFT(J) = CVEC(J) * TERM
      SUM = SUM + COEFFT(J)
 1    TERM = TERM * BP1INV
      DVEC(1) = SUM * FACT
      SUM2 = DVEC(1)
      N(1) = NTERMS
      START = BINP1I
      DO 2 I = 2,NSTOP
      NMAX = I
      AI = I
      ITOP = I - 1
      SUM = 0.
      DO 3 J = 1,NTERMS
      N(I) = J
      AJ = J
      PROD = COEFFT(J) * START
      DO 4 K = 1,ITOP
 4    PROD = PROD * (AND2 + AJ + K - 2.)/K
      SUM = SUM + PROD
      IF(J .GT. 20 .AND. PROD .LT. ERROR)  GO TO 6
 3    CONTINUE
 6    DVEC(I) = SUM * FACT
      SUM2 = SUM2 + DVEC(I)
      IF (DVEC(I) .LT. ERROR1) GO TO 5
 2    START = START * BINP1I
 5    WRITE(6,101) NMAX,(DVEC(I),I=1,NMAX)
 101  FORMAT ('0CONSTANTS - FORMULA (3.2.6)'//' NO. OF TERMS COMPUTED',
     1T30,I5/' CONSTANTS'/(T2,10D12.5))
      WRITE(6,100) ERROR, ERROR1, SUM2
 100  FORMAT('0CUTOFF VALUE IN SUM',T30,D12.5/' CUTOFF VALUE IN CONSTS',
     1T30,D12.5/' SUM OF CONSTANTS',T30,D12.5)
      WRITE(6,102) (N(I),I = 1,NMAX)
 102  FORMAT ('0NO. OF TERMS IN EACH CONSTANT'/(T2,10I12))
      RETURN
      END
      SUBROUTINE CHISER (A ,X ,NSTART ,NSTOP ,CHI ,PDFCHI )
C
C
C     SUBROUTINE TO COMPUTE A SEQUENCE OF CHI-SQUARED CDF AND PDF TERMS FOR
C     DEGREES OF FREEDOM GOING UP IN STEPS OF TWO.   PARAMETERS ARE:
C     A = BETA IN FORMULAE (3.1.7), ETC..    X = X-VALUE FOR WHICH PDF AND CDF
C     TO BE COMPUTED.    NSTART = DEGREES OF FREEDOM FOR FIRST TERM.
C     NSTOP = NUMBER OF TERMS IN SEQUENCE.    CHI = VECTOR CDF VALUES.
C     PDFCHI = VECTOR OF PDF VALUES.
C
C
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 CHI(NSTOP) , PDFCHI(NSTOP)
      FACT =DEXP(-X/(2.*A))
      TERM = 1.
      IF (MOD(NSTART,2) .GT. 0) GO TO 1
      NS = NSTART/2
      SUM = TERM
      IF (NS .LT. 2) GO TO 7
      DO 3 I = 2,NS
      TERM = TERM*X/(2.*A*(I-1.))
 3    SUM = SUM + TERM
 7    CONTINUE
      PDFCHI(1) = TERM * FACT * .5
      CHI(1) = 1. - SUM * FACT
      DO 4 J = 2,NSTOP
      TERM = TERM*X/(2.*A*(NS + J - 2.))
      IF (TERM .LT. 1.D-20) TERM = 0.
      PDFCHI(J) = TERM * FACT * .5
      SUM = SUM + TERM
 4    CHI(J) = 1. - SUM * FACT
      GO TO 2
 1    CONTINUE
      NS = (NSTART-1)/2
      TERM = TERM/DSQRT (X*3.141592653589793/(2.*A))
      SUM = 0.
      IF (NS .LT. 1) GO TO 8
      DO 5 I = 1,NS
      TERM = TERM * X/(2.*A*(I - 0.5))
 5    SUM = SUM + TERM
 8    CONTINUE
      PDFCHI(1) = TERM * FACT * .5
      PHI = DERF(DSQRT(X/(2.*A)))
      CHI(1) = PHI - SUM * FACT
      DO 6 J = 2,NSTOP
      TERM = TERM * X/(2.*A*(NS + J - 1.5))
      IF (TERM .LT. 1.D-20) TERM = 0.
      PDFCHI(J) = TERM * FACT * .5
      SUM = SUM + TERM
 6    CHI(J) = PHI - SUM * FACT
 2    CONTINUE
      RETURN
      END
```

Chaper 4 Evaluating the Performance of Classical Discriminant Analysis under the Random Effects Model - Probabilities of Correct and Misclassification

In this chapter we apply the results of Chapter 3 to evaluate the probabilities of correct- and misclassification under the random effects model when the classical rules of discriminant analysis are used.

i.e. We are interested in the expected performance of these rules when applied to future classification problems where the k populations π_i, i = 1,...,k, will have arisen from the random effects model. Using the classification rule based on the parameters of these k populations, whether known at the time or estimated from a training sample, we will classify an observation of unknown origin into one of them. How well are we likely to perform? Or more specifically: What are the expected probabilities of correct- or misclassification?

This chapter attempts to answer these questions.

As in Chapter 2 we will first consider the situation where the parameters in the distributions of the k populations are known and the classification rules are expressed in terms of them. See Section 2.1. Thereafter we will discuss the more common situation where the parameters are unknown and the parameters in the abovementioned classification rules are replaced by their sample estimates, resulting in the "plug-in" rules discussed in Section 2.2.

In each of the above two situations separate consideration will be given to the case where k = 2, since the results in this case are more tractable than those for general k. Moreover, as is clear from Chapter 2, far more work has been done on this case, and consequently much more is known about it.

It is traditional in most of the literature to talk of the probabilities of misclassification in the case where k = 2 but of the probabilities of correct classification when k > 2. We will follow this tradition here.

As in Chapter 3, the results will all be expressed in terms of the eigenvalues $\{\lambda_i, i = 1,\ldots,k\}$ of $T\Sigma^{-1}$, either directly or in terms of quantities derived from them. In Chapter 5 we will address the question of estimating the λ_i when they are unknown.

4.1 Known Parameters

In this situation the Bayes classification rule, when the prior probabilities of each of the k populations are all equal, may be expressed either in terms of the Mahalanobis distance:
i.e. assign the new observation x to that population π_i for which

$$\delta_i^2(x) = \min_{j=1,\ldots,k} \delta_j^2(x) \qquad (4.1.1)$$

where $\delta_j^2(x) = (x-\mu_j)'\Sigma^{-1}(x-\mu_j)$

or in terms of the linear discriminant function:
i.e. assign x to π_i if

$$u_{ij}(x) > 0 \quad \forall j = 1,\ldots,k; \quad j \neq i \qquad (4.1.2)$$

where $u_{ij}(x) = (x - (\mu_i+\mu_j))'\Sigma^{-1}(\ -\mu_j).$

See Section 2.1.

The distribution of $\delta_i^2(x)$, under the assumption that x either belongs to, or does not belong to π_i was discussed in Section 3.2, giving a general insight into the expected probabilities of correct- and misclassification when using (4.1.1) or (4.1.2), as well as their relationship to the eigenvalues $\{\lambda_i, i = 1,\ldots,r\}$ of $T\Sigma^{-1}$. Expressions for these probabilities will now be derived for the specific case where there are two populations. We will consider only the situation where the prior probabilities q_i are all equal.

4.1.1 The case k = 2 Populations

When the prior probabilities q_i, i = 1,2 are equal, we have from (2.1.11) the following simple expression for the conditional probability of misclassification, given δ^2:

$$P(\delta^2) = P[\text{misclassification}|\delta^2] = \Phi(-\tfrac{1}{2}\delta) \qquad (4.1.3)$$

where, $\delta^2 = \delta_{12}^2 = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$

and $\Phi(\cdot)$ is the Standard Normal Distribution Function.

The unconditional probability of misclassification is therefore:

$$P = E[P(\delta^2)] = E[\Phi(-\tfrac{1}{2}\delta)] \qquad (4.1.4)$$

where the expectation is taken over th- -istribution of δ^2.

Now, from Section 3.1 we know that under the random effects model δ^2 is distributed as $2\sum_{i=1}^{r}\lambda_i v_i$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$ are the nonzero eigenvalues of $T\Sigma^{-1}$ and the v_i are independent χ_1^2 random variables.

An approximation to (4.1.4) may be obtained by approximating $\Phi(-\frac{\delta}{2})$ by the first three terms of its Taylor expansion about $E[\delta^2]$ and then taking expectations. For any twice- differentiable function $f(x)$ of a random variable this approximation takes the form:

$$E[f(X)] \doteq f(E[X]) + \frac{f''(E[X])}{2!} \mathrm{Var}[X] \tag{4.1.5}$$

where $f''(\cdot)$ denotes the second derivative of $f(\cdot)$.
So the approximation becomes:

$$P \doteq \Phi(-\tfrac{1}{2}\sqrt{E[\delta^2]}) + \tfrac{1}{2}\Phi''(-\tfrac{1}{2}\sqrt{E[\delta^2]})\mathrm{Var}[\delta^2]. \tag{4.1.6}$$

Now, from section 3.1 we have that

$$E[\delta^2] = 2\sum_{i=1}^{r} \lambda_i$$

and

$$\mathrm{Var}[\delta^2] = 8 \sum_{i=1}^{r} \lambda_i^2 . \tag{4.1.7}$$

Also,

$$\begin{aligned}\Phi''(\;Z) &= \frac{d}{dz}(-\tfrac{1}{4}z^{-\frac{1}{2}}\phi(-\tfrac{1}{2}\sqrt{z}))\\ &= \frac{d}{dz}(-\tfrac{1}{2}z^{-\frac{1}{2}}\frac{1}{\sqrt{2\pi}}e^{-z/8})\\ &= \frac{1}{8\sqrt{2\pi}}e^{-z/8}\,z^{-3/2}(1 + \text{[illegible]}) .\end{aligned} \tag{4.1.8}$$

Substituting (4.1.8) and (4.1.7) into (4.1.6) yields the following approximate expression for the probability of misclassification:

$$P \doteq \Phi(-\sqrt{\tfrac{1}{2}\sum_{i=1}^{r}\lambda_i}) + \tfrac{1}{2}\,\frac{1}{8\sqrt{2\pi}}\,e^{-2\sum_{i=1}^{r}\lambda_i/8}(2\sum_{i=1}^{r}\lambda_i)^{-3/2}(1+\frac{2\sum_{i=1}^{r}\lambda_i}{4})8\sum_{i=1}^{r}\lambda_i^2$$

$$= \Phi(-\sqrt{\tfrac{1}{2}\sum_{i=1}^{r}\lambda_i}) + \tfrac{1}{2}\phi(\sqrt{\tfrac{1}{2}\sum_{i=1}^{r}\lambda_i})\,\frac{(1+\frac{1}{2}\sum_{i=1}^{r}\lambda_i)\sum_{i=1}^{r}\lambda_i^2}{(2\sum_{i=1}^{r}\lambda_i)^{3/2}} \qquad (4.1.9)$$

where $\phi(\cdot)$ is the standard normal density function.

An exact expression for the probability of misclassification may be obtained by evaluating $E[\Phi(-\frac{1}{2}\delta)]$ in (4.1.4) directly. To do this we need the density function of $z = \delta^2$ which, from (3.1.11) may be expressed as:

$$f_{\delta^2}(z) = \frac{1}{\beta}\sum_{j=0}^{\infty} c_j\, g_{r+2j}(\tfrac{z}{\beta}) \qquad (4.1.10)$$

where β is an arbitrary positive constant, $g_{r+2j}(\cdot)$ is the $\chi^2{}_{r+2j}$ density function and the c_j are given by (3.1.9) and (3.1.12). Thus

$$P = \int_0^{\infty} \Phi(-\tfrac{1}{2}\sqrt{z})\,\frac{1}{\beta}\sum_{j=0}^{\infty} c_j\, g_{r+2j}(\tfrac{z}{\beta})dz$$

$$= \frac{1}{\beta}\sum_{j=0}^{\infty} c_j \int_0^{\infty} \Phi(-\tfrac{1}{2}\sqrt{z})g_{r+2j}(\tfrac{z}{\beta})dz \qquad (4.1.11)$$

where the exchange of the summation and integration operations is justified by the uniform convergence of (4.1.10). Note that

$$\Phi(-\tfrac{1}{2}\sqrt{z}) = P[X \le -\tfrac{1}{2}\sqrt{z}] \qquad \text{where } X \sim N(0,1)$$

$$= \tfrac{1}{2}P[X^2 \ge \tfrac{z}{4}]$$

$$= \tfrac{1}{2}(1 - G_1(\tfrac{z}{4})) \qquad (4.1.12)$$

where $G_r(\cdot)$ denotes the distribution function of the χ_r^2 distribution. Substituting this into (4.1.11) yields

$$P = \tfrac{1}{2} - \frac{1}{2\beta} \sum_{j=0}^{\infty} c_j \int_0^{\infty} G_1(\tfrac{z}{4}) g_{r+2j}(\tfrac{z}{\beta}) dz \qquad (4.1.13)$$

where we have assumed that (4.1.10) is a mixture distribution, so that $\sum_{j=0}^{\infty} c_j = 1$ (See Remark 3.1.3). *Denoting the integral in (4.1.13) by* I_j and making the transformation $y = \frac{z}{\beta}$ gives:

$$I_j = \beta \int_0^{\infty} G_1(\tfrac{\beta}{4}\, y) g_{r+2j}(y) dy \ .$$

Integrating by parts and simplifying yields:

$$I_j = \beta(1 - \tfrac{\beta}{4} \int_0^{\infty} g_1(\tfrac{\beta}{4}\, y) G_{r+2j}(y) dy).$$

Substituting I_j back into (4.1.12) yields:

$$P = \frac{\beta}{8} \sum_{j=0}^{\infty} c_j \int_0^{\infty} g_1(\tfrac{\beta}{4}\, y) G_{r+2j}(y) dy. \qquad (4.1.14)$$

The integral in (4.1.14) may be evaluated by using the following expressions for $G_{r+2j}(y)$, obtained by direct integration (See, for example, Johnson and Kotz (1970a) page 173)

$$G_{r+2j}(y) = \begin{cases} 1 - e^{-\frac{1}{2}y} \sum\limits_{i=0}^{\frac{1}{2}r+j-1} (\frac{y}{2})^i / i! & \text{for } r \text{ even} \\ 2\Phi(\sqrt{y}) - 1 - e^{-\frac{1}{2}y} \sum\limits_{i=0}^{\frac{1}{2}(r-1)+j-1} (\frac{y}{2})^{i+\frac{1}{2}} / \Gamma(i+\frac{3}{2}) & \\ & \text{for } r \text{ odd.} \end{cases} \qquad (4.1.15)$$

Considering first the case where r is even, using (4.1.14), (4.1.15) and the formula:

$$g_1(\tfrac{\beta}{4}y) = \frac{1}{\sqrt{2\pi}}(\tfrac{\beta}{4}y)^{-\frac{1}{2}}e^{-\beta y/8} ,$$

we get

$$P = \tfrac{1}{2} - \frac{1}{2}\sqrt{\frac{\beta}{\pi}}\sum_{j=0}^{\infty} c_j \sum_{i=0}^{\frac{1}{2}r+j-1} \frac{1}{i!}\int_0^{\infty} (\tfrac{y}{2})^{i-\frac{1}{2}} e^{-\frac{1}{2}y(1+\beta/4)}dy$$

$$= \tfrac{1}{2}\{1 - \tfrac{1}{2}\sqrt{\frac{\beta}{\pi}}\sum_{j=0}^{\infty} c_j \sum_{i=0}^{\frac{1}{2}r+j-1} \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+1)}(1+\tfrac{\beta}{4})^{-(i+\frac{1}{2})}\} \qquad (4.1.16)$$

Consider now the case where r is odd. Using the same approach as above, we get,

$$P = \frac{\sqrt{\beta}}{2\sqrt{2\pi}}\int_0^{\infty}\Phi(\sqrt{y})y^{-\frac{1}{2}}e^{-\beta\frac{y}{8}}dy - \tfrac{1}{2}(1+\tfrac{1}{2}\sqrt{\frac{\beta}{\pi}}\sum_{j=0}^{\infty}c_j\sum_{i=0}^{\frac{1}{2}(r-1)+j-1}\frac{\Gamma(i+1)}{\Gamma(i+\frac{3}{2})(1+\beta/4)^{i+1}}) .$$

(4.1.17)

Denoting the first term in (4.1.17) by I, we get after making the transformation $x = \sqrt{y}$:

$$I = 2\int_0^{\infty}\Phi(x)\frac{1}{\sqrt{2\pi}\sqrt{4/\beta}}e^{-\frac{1}{2}(\frac{x^2}{4/\beta})}dx . \qquad (4.1.18)$$

The above integral is a particular case of Hojo's integrals (see, for example Kendall and Stuart Volume 1 (1969) pages 326-7). F. Downton (1973) gives the following closely related result:

Given

$$X \sim N(\mu,\sigma^2)$$

$$Y \sim N(0,1) \quad \text{independently,}$$

then:

$$P[Y \leq X] \;=\; \int_{-\infty}^{\infty} \Phi(t)\,\frac{1}{\sqrt{2\pi}\sigma}\,e^{(t-\mu)^2/2\sigma^2}\,dt\ . \qquad (4.1.19)$$

By analogy with (4.1.19), (4.1.18) can be expressed as:

$$I = 2P[Y \leq X \cap X \geq 0]$$

where

$$X \sim N(0,4/\beta)$$

$$Y \sim N(0,1) \qquad \textit{independently}$$

or,

$$I = 2P[X - Y \geq 0 \cap X \geq 0]\ . \qquad (4.1.20)$$

To evaluate the joint probability in (4.1.20), we need the joint distribution of X-Y and X. Now, by independence, the joint probability density function of X and Y is:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{4/\beta}}\,e^{-\frac{1}{2}(y^2 + \beta\,\frac{x^2}{4})}\ .$$

Making the transformation:

$$T = X - Y$$

$$U = X$$

and noting that the Jacobian of the transformation is unity, we get the density of T and U as:

$$f_{T,U}(t,u) = \frac{1}{2\pi\sqrt{4/\beta}}\ e^{-\frac{1}{2}(u^2-2tu+t^2+\beta u^2/4)}$$

$$= \frac{1}{2\pi\sqrt{4/\beta}}\ e^{-\frac{1}{2}(u^2(1+\beta/4)-2tu+t^2)} .$$

So T and U have bivariate normal distribution with zero mean vector, variances σ_T^2 and σ_U^2 and correlation coefficient ρ, where the latter three parameters may be obtained from the following identities:

$$\sigma_T^2\ \sigma_U^2(1-\rho^2) = \frac{4}{\beta}$$
$$\sigma_U^2(1-\rho^2) = (1+\frac{\beta}{4})^{-1}$$
$$\sigma_T^2(1-\rho^2) = 1\ .$$

This yields:

$$\sigma_T^2 = 1 + \frac{4}{\beta}$$
$$\sigma_U^2 = \frac{4}{\beta}$$

and

$$\rho = (1+\frac{\beta}{4})^{-\frac{1}{2}}\ . \qquad (4.1.21)$$

Now, applying the result given in Anderson (1958) page 43, problem 43, viz: if

$$P[X \geq 0 \cap Y \geq 0] = \alpha$$

and

$$\binom{X}{Y} \sim N(\binom{0}{0},\ \begin{pmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2\end{pmatrix})$$

then

$$\rho = \cos(1-2\alpha)\pi$$

or

$$\alpha = \tfrac{1}{2}(1 - \tfrac{1}{\pi}\cos^{-1}\rho),$$

we get, from (4.1.20) and (4.1.21) that:

$$I = 2P[T \geq 0 \cap U \geq 0] = 1 - \tfrac{1}{\pi}\cos^{-1}((1 + \tfrac{\beta}{4})^{-\frac{1}{2}}). \qquad (4.1.22)$$

Substituting (4.1.22) back into (4.1.17) and simplifying, yields the following expression for the probability of misclassification when r is odd:

$$P = \tfrac{1}{2}\{1 - \tfrac{2}{\pi}\cos^{-1}((1+\tfrac{\beta}{4})^{-\frac{1}{2}}) - \tfrac{1}{2}\sqrt{\tfrac{\beta}{\pi}}\sum_{j=0}^{\infty} c_j \sum_{i=0}^{\frac{1}{2}(r-1)+j-1} \frac{\Gamma(i+1)}{\Gamma(i+1.5)}(1+\tfrac{\beta}{4})^{-(i+1)}\}$$

(4.1.23)

4.1.2 Evaluating the Probabilities of Misclassification for k = 2 populations

In order to evaluate formulae (4.1.16) and (4.1.23) for the probability of misclassification, the FORTRAN [illegible]utine PROBS, given in Appendix 4.3, was written. This was used to [illegible] probability of misclassification for the case r = 5 for the same th[illegible]ets of eigenvalues $\{\lambda_i\}$ that were used in Chapter 3, as well as for the corresponding three sets when the trace is halved. The results are given below in Table 4.1.1, together with those obtained from the approximate formula (4.1.9).

Table 4.1.1

Case	$\{\lambda_i\}$	Trace	Exact Probability of Misclassification	Approximate Probability of Misclassification
(a)	11.0,1.0,1.0,1.0,1.0	15.0	.0392	.0334
(b)	3.0,3.0,3.0,3.0,3.0	15.0	.0204	.0140
(c)	5.0,4.0,3.0,2.0,1.0	15.0	.0233	.0164
(d)	5.5,0.5,0.5,0.5,0.5	7.5	.0827	.1044
(e)	1.5,1.5,1.5,1.5,1.5	7.5	.0653	.0543
(f)	2.5,2.0,1.5,1.0,0.5	7.5	.0596	.0606

From Table 4.1.1 the relationship between the probability of misclassification and both the trace and relative sizes of the eigenvalues of $T\Sigma^{-1}$, that was predicted in Chapter 3, is clearly evident. However, the approximate formula (4.1.9), which is far easier to compute than the exact formulae and therefore useful for quick assessments of the probability of misclassification, is not very accurate.

4.1.3 The case k > 2 populations

From classification rule (4.1.1) the probability of correct classification, given $x \epsilon \pi_i$, becomes:

$$P[\text{correct classification}|x \epsilon \pi_i] = P[\delta_i^2(x) < \min_{\substack{j=1,\ldots,k \\ j \neq i}} \delta_j^2(x) | x \epsilon \pi_i] . \qquad (4.1.24)$$

Now, from Section 3.2 we have that, given $x \epsilon \pi_i$:

$$\delta_i^2(x) \sim \chi_p^2$$

and

$$\delta_j^2(x) \sim \chi_p^2(\delta_{ij}^2) \quad \text{conditionally on } \delta_{ij}. \qquad (4.1.25)$$

Unconditionslly $\delta_j^2(x)$ has the density given in (3.2.7):

$$f_{\delta_j^2(x)}(\delta_j^2(x)\,|\,x \in \pi_i) = \sum_{s=0}^{\infty} a_s\, g_{p+2s}(\delta_j^2(x)). \qquad (4.1.26)$$

where $g_{p+2s}(\cdot)$ denotes the χ_{p+2s}^2 density function and the coefficients a_s are given by (3.2.6). Moreover, the $\delta_j^2(x)$ are clearly not independent.

So, in order to evaluate (4.1.24) we need the joint distribution of the minimum of k - 1 correlated, identically distributed random variables $\delta_j^2(x)$ whose marginal densities are given by (4.1.26) and the chi-squared random variable $\delta_i^2(x)$ which is also correlated with the $\delta_i^2(x)$.

It is clear, therefore, that this approach to evaluating the probability of correct classification is not a promising one, and will not be pursued further here.

Another approach would be to use expression (2.1.15) for the probability of correct classification given $x \in \pi_i$, conditional on the values of $\{\delta_{ij\ell} = (\mu_i-\mu_j)'\Sigma^{-1}(\mu_i-\mu_\ell),\ j,\ell = 1,\ldots,k;\ \ j,\ell \neq i\}$ and then to obtain the unconditional probability by integrating it over the joint distribution of the $\delta_{ij\ell}$.

Since there is no analytic expression for (2.1.15), it would have to be evaluated numerically or by table look-up over a multidimensional grid of points defined by the $\delta_{ij\ell}$ and then integrated numerically over their joint distribution.

In addition to the complexity of the abovementioned operation, an expression for the joint distribution of the $\delta_{ij\ell}$ would have to be found. As in the previous approach, the marginal distributions of the $\delta_{ij\ell}$ are known. Viz: the $\delta_{ij}^2 = \delta_{ijj}$ have the distribution derived in Theorem 3.1.1

and the $\delta_{ij\ell}$, $j \neq \ell$ can, in a manner very similar to Theorem 3.1., be shown to be distributed as $\sum_{s=1}^{r} \lambda_s(v_s-w_s)$, where $\{\lambda_s\} = \text{Eigs}(T\Sigma^{-1})$ and the v_s and w_s are independent χ_1^2 random variables. It can also be shown that the correlation coefficient between δ_{ij}^2 and $\delta_{i\ell}^2$, $j \neq \ell$ is $\frac{1}{2}$.

However, the joint distribution of the $\delta_{ij\ell}$ is unknown, so this approach will also not be pursued any further.

This leaves only the lower bounds (2.1.16) and (2.1.17) on the probability of correct classification. However, these expressions give lower bounds on the <u>minimum</u> probability of correct classification. Stronger bounds than these may be obtained from Bonferroni's first inequality by noting that (4.1.24) can be written:

$$\begin{aligned} P[\text{correct classification}|x \in \pi_i] &= P[\bigcap_{\substack{j=1\\ j\neq i}}^{k} \delta_i^2(x) < \delta_j^2(x)|x \in \pi_i] \\ &\geq 1 - \sum_{\substack{j=1\\ j\neq i}}^{k} P[\delta_i^2(x) > \delta_j^2(x)|x \in \pi_i] \;. \end{aligned}$$

Now $P[\delta_i^2(x) > \delta_j^2(x)|x \in \pi_i]$ is just the probability of misclassification with two populations π_i and π_j, and is therefore equal to $\Phi(-\frac{1}{2}\delta_{ij}^2)$. So

$$P[\text{correct classification}|x \in \pi_i] \geq 1 - \sum_{\substack{j=1\\ j\neq i}}^{k} \Phi(-\tfrac{1}{2}\delta_{ij}). \tag{4.1.27}$$

Under the random effects model δ_{ij}^2 is a random variable, so (4.1.27) becomes:

$$\begin{aligned} P[\text{correct classification}|x \in \pi_i] &\geq 1 - \sum_{\substack{j=1\\ j\neq i}}^{k} E_{\delta_{ij}^2}[\Phi(-\tfrac{1}{2}\delta_{ij})] \\ &= 1 - (k-1)E_{\delta_{ij}^2}[\Phi(-\tfrac{1}{2}\delta_{ij})] \end{aligned} \tag{4.1.28}$$

since the δ_{ij}^2 are identically distributed.

Note that (4.1.28) does not depend on the particular population π_i from which x comes, so it is also the unconditional probability of correct classification. Finally, using results (4.1.16) and (4.1.23) of the previous sub-section in (4.1.28), we get

for r even:

$$P[\text{correct classification}] \geq 1 - \frac{k-1}{2}\{1 - \tfrac{1}{2}\sqrt{\frac{\beta}{\pi}} \sum_{j=0}^{\infty} c_j \sum_{i=0}^{\frac{1}{2}r+j-1} \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+1)}(1+\frac{\beta}{4})^{-(i+\frac{1}{2})}\} \tag{4.1.29}$$

for r odd:

$$P[\text{correct classification}] \geq 1 - \frac{k-1}{2}\{1 - \frac{2}{\pi}\cos^{-1}((1+\frac{\beta}{4})^{-\frac{1}{2}}) - \tfrac{1}{2}\sqrt{\frac{\beta}{\pi}} \sum_{j=0}^{\infty} c_j \sum_{i=0}^{\frac{1}{2}(r-1)+j-1} \frac{\Gamma(i+1)}{\Gamma(i+1.5)}(1+\frac{\beta}{4})^{-(i+1)}\}. \tag{4.1.30}$$

An upper bound on the probability of correct classification may also be obtained by using the fact that,

$$P[\text{misclassification} | x \in \pi_i] \geq P[\text{misclassification to } \pi_i\text{'s closest neighbour} | x \in \pi_i] = \Phi(-\tfrac{1}{2}\delta_i)$$

where $\delta_i^2 = \min_{\forall j \neq i} \delta_{ij}^2$.

So,

$$P[\text{correct classification} | x \in \pi_i] \leq 1 - \Phi(-\tfrac{1}{2}\delta_i) . \tag{4.1.31}$$

Under the random effects model, this becomes:

$$P[\text{correct classification}|x \in \pi_i] \le 1 - E_{\delta_i^2}[\Phi(-\tfrac{1}{2}\delta_i)] \ . \qquad (4.1.32)$$

To evaluate the expectation in (4.1.32) the distribution of $\delta_i^2 = \min\limits_{\forall j \neq i} \delta_{ij}^2$ is required. Unfortunately, although the δ_{ij}^2 have identical marginal distributions given by Theorem 3.1.1, and the correlation coefficient between δ_{ij}^2 and $\delta_{i\ell}^2$ is known, their joint distribution is unknown, and so the distribution of δ_i^2 cannot be found.

However, if we assume that μ_i is fixed, then it is possible to obtain the distribution of δ_i^2 and hence to evaluate the upper bound (4.1.32) on the probability of correct classification.

In what follows, we will therefore first obtain the distribution of δ_i^2, conditional on μ_i. Unfortunately it is not possible to obtain the unconditional distribution from it. This distribution will then be used to evaluate (4.1.32). Finally we shall show that a very similar expression for the upper bound is obtained if instead we ignore the intercorrelations between the δ_{ij}^2 and proceed as if they were independent. Under these circumstances it is not necessary to assume that μ_i is fixed.

The distribution of $\delta_i^2 = \min\limits_{\forall j \neq i} \ \delta_{ij}^2$, conditional on μ_i

We first consider the distribution of

$$\delta_{ij}^2 = (\mu_j - \mu_i)'\Sigma^{-1}(\mu_j - \mu_i)$$

conditional on μ_i, under the random effects model.

Under this model, the μ_j are independently and identically distributed $N_p(\xi,T)$ random variables. Therefore, conditionally on μ_i,

$$\mu_j - \mu_i \sim N_p(\xi-\mu_i,T) \quad \text{independently, } j = 1,\ldots,k; \ \ j \neq i \ . \qquad (4.1.33)$$

Theorem 4.1.1, given below, allows us to find the conditional distribution of δ_{ij}^2.

Theorem 4.1.1

Let $d^2 = X'\Sigma^{-1}X$, where $X \sim N_p(\mu, T)$. Then d^2 is distributed as $\sum_{i=1}^{r} \lambda_i v_i$ where the λ_i are the r ($\leq p$) nonzero eigenvalues of $T\Sigma^{-1}$ and the $v_i \sim \chi_1^2(\omega_i^2)$, independently. The square root ω_i of the noncentrality parameter of v_i is the i^{th} element of $P'\eta$ where P is the $(r\times r)$ orthogonal matrix whose i^{th} column is the eigenvector of $T_1'\Sigma^{-1}T_1$ corresponding to λ_i, $T = T_1T_1'$ and T_1 is a $p \times r$ matrix of rank $r = r(T)$, and η is the solution to $T_1\eta = \mu$.

The proof of this theorem, which is essentially a generalization of Theorem 3.1.1, is given in Appendix 4.1.

Applying Theorem 4.1.1 to (4.1.33) immediately yields the distribution of δ_{ij}^2, conditional on μ_i, in the following form:

$$\delta_{ij}^2 \sim \sum_{s=1}^{r} \lambda_s v_s, \quad \text{independently, } j = 1,\ldots,k; \;\; j \neq i \qquad (4.1.34)$$

where,

$\{\lambda_s\} = \text{eigs}(T\Sigma^{-1})$

$v_s \sim \chi_1^2(\omega_s^2)$ independently, $s = 1,\ldots,r$,

$(\omega_1,\omega_2,\ldots,\omega_r)' = P'\eta$,

P is the $(r\times r)$ orthogonal matrix defined in Theorem 4.1.1,

η is the solution to $T_1\eta = \xi - \mu_i$

and T_1 is the $(p\times r)$ matrix defined in Theorem 4.1.1.

Clearly, if T is of full rank, i.e. $r = p$, then $\eta = T_1^{-1}(\xi-\mu_i)$.

The mean and variance of v_s are, respectively (See, for example Johnson and Kotz (1970b) page 134):

$$E[v_s] = 1 + \omega_s^2$$
$$Var[v_s] = 2(1+2\omega_s^2)$$

and since the v_s are independent, we obtain the following expressions for the conditional mean and variance of δ_{ij}^2:

$$E[\delta_{ij}^2|\mu_i] = \sum_{s=1}^{r} \lambda_s(1+\omega_s^2) \qquad (4.1.35)$$

$$Var[\delta_{ij}^2|\mu_i] = 2\sum_{s=1}^{r} \lambda_s^2(1+2\omega_s^2). \qquad (4.1.36)$$

As in the case of the sum of weighted central chi-squared random variables, the distribution of the sum of weighted noncentral chi-squared random variables may also be expanded as an infinite series of central chi-squared distributions (See, for example, Ruben (1962), Press (1966), Kotz, Johnson and Boyd (1967b), Johnson and Kotz (1970b)). This yields the following expression for the distribution- and density functions, respectively, of δ_{ij}^2, conditional on μ_i. Letting $z = \delta_{ij}^2$:

$$F_{\delta_{ij}^2|\mu_i}(z) = \sum_{j=0}^{\infty} c_j' \, G_{r+2j}(\tfrac{z}{\beta})$$

and

$$f_{\delta_{ij}^2|\mu_i}(z) = \frac{1}{\beta}\sum_{j=0}^{\infty} c_j' \, g_{r+2j}(\tfrac{z}{\beta}) \qquad (4.1.37)$$

where β is an arbitrary positive constant, $G_{r+2j}(\cdot)$ and $g_{r+2j}(\cdot)$ are the distribution- and density functions, respectively, of the χ_{r+2j}^2 distribution and the constants c_j' are given by:

The mean and variance of v_s are, respectively (See, for example Johnson and Kotz (1970b) page 134):

$$E[v_s] = 1 + \omega_s^2$$
$$Var[v_s] = 2(1+2\omega_s^2)$$

and since the v_s are independent, we obtain the following expressions for the conditional mean and variance of δ_{ij}^2:

$$E[\delta_{ij}^2|\mu_i] = \sum_{s=1}^{r} \lambda_s(1+\omega_s^2) \qquad (4.1.35)$$

$$Var[\delta_{ij}^2|\mu_i] = 2\sum_{s=1}^{r} \lambda_s^2(1+2\omega_s^2). \qquad (4.1.36)$$

As in the case of the sum of weighted central chi-squared random variables, the distribution of the sum of weighted noncentral chi-squared random variables may also be expanded as an infinite series of central chi-squared distributions (See, for example, Ruben (1962), Press (1966), Kotz, Johnson and Boyd (1967b), Johnson and Kotz (1970b)). This yields the following expression for the distribution- and density functions, respectively, of δ_{ij}^2, conditional on μ_i. Letting $z = \delta_{ij}^2$:

$$F_{\delta_{ij}^2|\mu_i}(z) = \sum_{j=0}^{\infty} c_j' \, G_{r+2j}(\tfrac{z}{\beta})$$

and

$$f_{\delta_{ij}^2|\mu_i}(z) = \frac{1}{\beta}\sum_{j=0}^{\infty} c_j' \, g_{r+2j}(\tfrac{z}{\beta}) \qquad (4.1.37)$$

where β is an arbitrary positive constant, $G_{r+2j}(\cdot)$ and $g_{r+2j}(\cdot)$ are the distribution- and density functions, respectively, of the χ_{r+2j}^2 distribution and the constants c_j' are given by:

$$c_0' = e^{-\frac{1}{2}\sum_{s=1}^{r}\omega_s^2} \prod_{s=1}^{r} (\beta/\lambda_s)^{\frac{1}{2}}$$

$$c_j' = \frac{1}{2j}\sum_{i=0}^{j-1} h_{j-i}' c_i' \qquad j = 1,2,\ldots$$

where

$$h_j' = \sum_{s=1}^{r}(1-\beta/\lambda_s)^j + j\beta \sum_{s=1}^{r}(\omega_s^2/\lambda_s)(1-\beta/\lambda_s)^{j-1}$$

Ruben (1962) shows that for $0 < \beta \le \alpha_r$ (4.1.32) is a mixture distribution (it may or may not be for other values of β) and that it converges uniformly in any bounded z-interval of $z > 0$ for any β, and converges uniformly for all $z > 0$ if β is chosen so that $\max_j|1 - \frac{\beta}{\lambda_j}| < 1$.

Remembering that, conditionally on μ_i, *the* δ_{ij}^2, $j = 1,\ldots,k;\ j \ne i$, are independently distributed, all with distribution given by (4.1.37) we immediately get the distribution and density functions of $\delta_i^2 = \min_{\forall j \ne i} \delta_{ij}^2$ in the following form (See, for example Gibbons (1971)),

$$F_{\delta_i^2|\mu_i}(z) = 1 - (1 - F_{\delta_{ij}^2|\mu_i}(z))^{k-1}$$

$$f_{\delta_i^2|\mu_i}(z) = (k-1)(1 - F_{\delta_{ij}^2|\mu_i}(z))^{k-2} f_{\delta_{ij}^2|\mu_i}(z) \quad (4.1.38)$$

where $F_{\delta_{ij}^2|\mu_i}(z)$ and $f_{\delta_{ij}^2|\mu_i}(z)$ are given in (4.1.37).

Using (4.1.38), the upper bound (4.1.32) on the probability of correct classification under the random effects model, given $x \in \pi_i$, can be evaluated conditionally on μ_i. Using the notation

$$P_{\mu_i} = P[\text{correct classification}|x \in \pi_i, \mu_i]$$

we therefore have

$$P_{\mu_i} \leq 1 - \int_0^\infty \Phi(-\tfrac{1}{2}\sqrt{z}) f_{\delta_i^2|\mu_i}(z)dz$$

$$= \tfrac{1}{2}(1 + \int_0^\infty G_1(\tfrac{z}{4}) f_{\delta_i^2|\mu_i}(z)dz)$$

using result (4.1.12), where $G_1(\cdot)$ is the χ_1^2 distribution function. Integrating by parts yields,

$$P_{\mu_i} \leq \tfrac{1}{2}(2 - \tfrac{1}{4} \int_0^\infty g_1(\tfrac{z}{4}) F_{\delta_i^2|\mu_i}(z)dz$$

where $g_1(\cdot)$ is the χ_1^2 density function

$$= \tfrac{1}{2}(1 + \tfrac{1}{4} \int_0^\infty g_1(\tfrac{z}{4})(1 - \sum_{j=0}^{\infty} c_j' \, G_{r+2j}(\tfrac{z}{\beta}))^{k-1}dz)$$

from (3.1.37) and (3.1.38)

$$= \tfrac{1}{2}(1 + \tfrac{\beta}{4} \int_0^\infty g_1(\tfrac{\beta}{4}y)(1 - \sum_{j=0}^{\infty} c_j' \, G_{r+2j}(y))^{k-1}dy)$$

making the transformation $y = \frac{z}{\beta}$.

Now, using expressions (4.1.15) for $G_{r+2j}(y)$ and considering the case where <u>r is even</u>, we get

$$P_{\mu_i} \leq \tfrac{1}{2}(1 + \tfrac{\beta}{4} \int_0^\infty \frac{(\frac{\beta}{4}y)^{-\frac{1}{2}}}{\sqrt{2\pi}} e^{-\beta y/8}\Big[\sum_{j=0}^{\infty} c_j' \sum_{i=0}^{\frac{1}{2}r+j-1} (\tfrac{y}{2})^i/i!\Big]^{k-1} e^{-(k-1)y/2}dy)$$

where we have assumed that (4.1.37) is a mixture distribution so that $\sum_{j=0}^{\infty} c_j' = 1$. From the identity:

$$\left(\sum_{j=0}^{\infty} c_j' \sum_{i=0}^{\frac{1}{2}r+j-1} \left(\tfrac{y}{2}\right)^i / i!\right)^{k-1} \equiv \sum_{j=0}^{\infty} a_j y^j \tag{4.1.39}$$

where the a_j are obtained by equating coefficients of y^j on the left- and right-hand sides (See Appendix 4.2 for their values) we obtain:

$$P_{\mu_i} \leq \tfrac{1}{2}\{1 + \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\beta}{4}} \sum_{j=0}^{\infty} a_j \int_0^{\infty} y^{j-\frac{1}{2}} e^{-y(k-1+\beta/4)/2} \, dy$$

where the interchange of summation and integration operations is justified by the uniform convergence of (4.1.37) and hence of (4.1.39). Evaluating the above integral as a gamma function finally yields after some simplification,

$$P_{\mu_i} \leq \tfrac{1}{2}\{1 + \sqrt{\frac{\beta/4}{k+\beta/4-1}} \sum_{j=0}^{\infty} a_j \left(\frac{2}{k+\beta/4-1}\right)^j \left(\tfrac{1}{2}\right)^{[j]}\} \tag{4.1.40}$$

where $(a)^{[j]} = a(a+1).....(a+j-1)$.

Unfortunately, the case where r is odd is so complicated that it is not considered here.

Remark 4.1.1 The drawback to expression (4.1.40) is that it refers to the conditional probability of correct classification and requires μ_i to be given before it can be used.

An approach that gives an unconditional but approximate upper bound is to ignore the intercorrelations between the δ_{ij}^2, $j = 1,\ldots,k$; $j \neq i$ and to proceed as if they were independent. Therefore, instead of using the conditional distribution (4.1.37) in expression (4.1.38) for the distribution of $\delta_i^2 = \min_{\forall j \neq i} \delta_{ij}^2$, we use the unconditional distribution (4.1.1) for δ_{ij}^2 that was derived in Chapter 3. Noting that (4.1.10) and (4.1.37) differ only in respect of their constants c_j and c_j', respectively, it is clear that the arguments go through exactly as for the conditional case with c_j' replaced by c_j. So expression (4.1.40) can also be used as an approximate upper bound on the unconditional probability of correct classification if c_j' is replaced by c_j in definition (4.1.39) of the a_j.

Another link-up between the upper boun on the conditional probability of correct classification and the approximate upper bound on the unconditional probability is achieved if μ_i is fixed at the value $\mu_i = \xi$ in the former. For then it is clear from (4.1.34) that, conditionally on $\mu_i = \xi$

$$\delta_{ij}^2 \sim \sum_{s=1}^{r} \lambda_s \, v_s$$

where now the v_s are central χ_1^2 random variables. Comparing this with the unconditional distribution of δ_{ij}^2 derived in Theorem 3.1.1:

$$\delta_{ij}^2 \sim 2 \sum_{s=1}^{r} \lambda_s \, v_s$$

where the v_s are also central χ_1^2 random variables, we see that for a given set of eigenvalues $\{\lambda_s\}$, the values of the upper bound (4.1.40) for the probability of correct classification conditional on $\mu_i = \xi$, will be equal to that of the corresponding approximate bound on the unconditional probability for the case when the eigenvalues are all half as large.

This is intuitively reasonable, as one would expect poorer classification from populations situated near the mean of their distribution.

4.1.4 Evaluating the bounds on the probabilities of correct classification for $k > 2$ populations

Expressions (4.1.29) and (4.1.30) for the lower bound on the probability of correct classification have been derived directly from the two-population case, and they are also computed by the subroutine PROBS given in Appendix 4.3. Table 4.1.2 gives the values of the lower bound for the same three sets of eigenvalues $\{\lambda_i\}$, all with a trace of 15, that were used in earlier examples, and for $k = 5$ populations. Values for $k = 5$, $r = 4$ and a similar three sets of $\{\lambda_i\}$, all with trace 10, are also given, for comparison with the upper bounds discussed below.

Expression (4.1.40) for the upper bound on the conditional probability of correct classification is not evaluated as easily because of the increasing complexity of the formulae for the constants a_j appearing in it for values of j greater than $\frac{r}{2}$. See Appendix 4.2.

However, for the specific case where the eigenvalues $\{\lambda_s\}$ of $T\Sigma^{-1}$ are all equal, say $\lambda_s = \lambda$, $s = 1,\ldots,r$, and μ_i is fixed at the value $\mu_i =$ [illegible] it is clear from Remark 4.1.1 above and from definition (3.1.9) for the c_j that if $\beta = \lambda$ then $c'_0 = 1$ and $c'_j = 0$, $\forall j > 0$, and that if $\beta = 2\lambda$ then $c_0 = 1$ and $c_j = 0$, $\forall j > 0$. (This is also an immediate consequence of the fact that when the λ_s are all equal then δ^2_{ij} is proportional to a χ^2_r random variable. See (3.1.13)).

Under these circumstances (4.1.39) becomes:

$$\Big\{\sum_{i=0}^{\frac{1}{2}r-1} (\tfrac{y}{2})^i/i!\Big\}^{k-1} \equiv \sum_{j=0}^{\infty} a_j\, y^j \qquad (4.1.41)$$

so that the sequence of nonzero a_j terminates after a finite number of terms and they are readily computed, especially for low values of r.

For example, for the case $r = 4$ and $k = 5$ populations, (recall that formulae (4.1.39) and (4.1.40) are valid only for r even), using either (4.1.41) or the formulae derived in Appendix 4.2, we get the following values for the a_j:

$$a_o = 1,\ a_1 = 2,\ a_2 = \frac{3}{2},\ a_3 = \frac{1}{2},\ a_4 = \frac{1}{16}$$

and $a_j = 0$, $\forall j > 4$.

Using these values for the a_j, the upper bound (4.1.40) on the conditional probability of correct classification with $\mu_i = \xi$, as well as the approximate upper bound on the unconditional probability (see Remark 4.1.2) were computed for the case where $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 2.5$. For a given value of the trace of $T\Sigma^{-1}$, the case where the λ_i are all equal gives the best classification, so these upper bounds are also valid for the other cases with $r = 4$ given in Table 4.1.2.

Table 4.1.2

Bounds on the probabilities of correct classification for $k = 5$ populations

Case		Lower bound	Upper bound on conditional prob. evaluated at $\lambda_i = \xi$	Approximate Upper bound
(a)	11,1,1,1,1	.8433	-	-
(b)	3,3,3,3,3	.9183	-	-
(c)	5,4,3,2,1	.9068	-	-
(d)	7,1,1,1	.7517	.7970	.8694
(e)	2.5,2.5,2.5,2.5	.8220	.7970	.8694
(f)	4,3,2,1	.8063	.7970	.8694

As remarked at the end of the previous sub-section, classification tends to be poorer when the new observation comes from a population whose mean is situated at the centre of its distribution, than when it is situated elsewhere. This is reflected by the low value of the upper bound on the conditional probability evaluated at $\mu_i = \xi$ given in Table 4.1.2, which is in fact lower than the corresponding lower bound in two out of the three cases (d) to (f). Thus it would appear that the upper bound on the conditional probability is of limited use in practice, and that the approximate upper bound, obtained by assuming that the δ_{ij}, $j = 1,\ldots,k;\ j \neq i$, are independent, is far more useful.

4.2 Unknown Parameters

In this section we consider the probabilities of correct- and mis-classification when the sample-based classification rule, with equal prior probabilities for each of the k populations, is used. viz: Assign new observation x to that population π_i for which,

$$d_i^2(x) = \min_{j=1,\ldots,k} d_j^2(x) \tag{4.2.1}$$

where

$$d_j^2(x) = (x-x_{j.})'S^{-1}(x-x_{j.}) \ ,$$

$x_{j.}$ is the mean of the training sample of size n_j from population π_j, and S is the pooled sample covariance matrix based on ν degrees of freedom, or equivalently, assign x to π_i if

$$V_{ij}(x) > 0 \qquad \forall j = 1,\ldots,k;\ j \neq i \tag{4.2.2}$$

where

$$V_{ij}(x) = (x - (x_{i.}+x_{j.}))'S^{-1}(x_{i.}-x_{j.}).$$

As described in Section 2.2, two types of misclassification probability may be defined when the sample-based classification rule is used. (Although we refer to the misclassification probability, the remarks hold equally well for the probability of correct classification). They are the conditional probability of misclassification, P_i^c given a particular training sample and that $x \in \pi_i$, and the expected probability of misclassification P_i given $x \in \pi_i$, when the classification rule is based on training samples of size n_j, $j = 1,\ldots,k$.

Both these probabilities may be expressed in terms of the population means μ_i (or functions of them) which, under the random effects model, are random variables. Under this model, therefore, we are interested in the expectations of P_i^c and P_i^e over the distribution of the μ_i.

Interpreted in a Bayesian sense, taking the expectation of P_i^c over the distribution of the μ_i gives the posterior probability of misclassification, given the training sample. As shall be seen in the case of $k = 2$ populations this leads to results that are not very useful from a practical point of view, so the great majority of this section will be devoted to obtaining expressions for the expected probabilities of correct and misclassification under the random effects model when the classification rules (4.2.1) and (4.2.2) are based on training samples of size n_j, $j = 1,\ldots,k$.

4.2.1 The case k = 2 populations

The conditional probability of misclassification, using the classification rules (4.2.1) or (4.2.2) based on training samples yielding $x_{1.}, x_{2.}$ and S, is given in Section 2.1, equation (2.1.23). Thus,

$$P_i^c(\mu_i) = P[\text{misclassification} | x_{1.}, x_{2.}, S, \mu_i; x \in \pi_i]$$

$$= \Phi\left\{\frac{(-1)^i(\mu_i - \frac{1}{2}(x_{1.} + x_{2.}))' S^{-1}(x_{1.} - x_{2.})}{\sqrt{(x_{1.} - x_{2.})' S^{-1} \Sigma S^{-1}(x_{1.} - x_{2.})}}\right\}$$

$$= \Phi((-1)^i \quad (\mu_i - a)'b/c) \tag{4.2.3}$$

where,

$$a = \tfrac{1}{2}(x_{1.} + x_{2.})$$

$$b = S^{-1}(x_{1.} - x_{2.})$$

and $\quad c = \sqrt{b'\Sigma b}$.

Under the random effects model $\mu_i \sim N(\xi, T)$, independently, so considering the case $X \in \pi_1$ and taking expectations over the distribution of μ_1 yields:

$$P_1^c = P[\text{misclassification}] x_{1.}, x_{2.}, S; X \in \pi_1] = E_{\mu_1}[\Phi(-\frac{(\mu_1 - a)'b}{c})] .$$

Letting $y = \dfrac{(\mu_1 - a)'b}{c}$, we have that, under the random effects model,

$$y \sim N(\frac{(\xi - a)'b}{c}, \frac{b'Tb}{c^2})$$

so,

$$P_1^c = \int_{-\infty}^{\infty} \Phi(-y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(y-\eta)^2/\sigma^2} \, dy$$

where,

$$\eta = (\xi - a)'b/c$$

and

$$\sigma^2 = b'Tb/c^2.$$

So,

$$P_1^C = \int_{-\infty}^{\infty} \Phi(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(y+\eta)^2/\sigma^2} \, dy.$$

This integral may be evaluated using the result in Downton (1973) referred to in expression (4.1.19) in Section 4.1. This immediately yields:

$$P_1^C = \Phi(1 - \frac{\eta}{\sqrt{1+\sigma^2}})$$

$$= \Phi\left(- \frac{(\xi - \frac{1}{2}(x_{1.}+x_{2.}))'S^{-1}(x_{1.}-x_{2.})}{\sqrt{(x_{1.}-x_{2.})'S^{-1}(\Sigma+T)S^{-1}(x_{1.}-x_{2.})}}\right) . \qquad (4.2.4)$$

Similarly,

$$P_2^C = P[\text{misclassification}|x_{1.},x_{2.},S;x \in \pi_2]$$

$$= \Phi\left(\frac{(\xi - \frac{1}{2}(x_{1.}+x_{2.}))'S^{-1}(x_{1.}-x_{2.})}{\sqrt{(x_{1.}-x_{2.})'S^{-1}(\Sigma+T)S^{-1}(x_{1.}-x_{2.})}}\right) . \qquad (4.2.5)$$

<u>Remark 4.2.1</u> Although results (4.2.4) and (4.2.5) are elegant mathematically, they are not very useful from a practical point of view. This is highlighted by the fact that since the prior probabilities q_1 and q_2 of π_1 and π_2, respectively, have been assumed equal, the average posterior probability of misclassification becomes, using (4.2.4) and (4.2.5):

$$P[\text{misclassification}|x_{1.}x_{2.},S] = \tfrac{1}{2}(P_1^C + P_2^C) = \tfrac{1}{2} \qquad (4.2.6)$$

independently of the values of $x_{1.}, x_{2.}$ and S.

The reason for this anomaly is that once $x_{1.}$ and $x_{2.}$ are given, the populations π_1 and π_2, and hence μ_1 and μ_2 are no longer randomly chosen but are fixed for the present problem. Therefore it is not meaningful to take the expectation of the conditional probability of misclassification, given the training sample, over the distribution of μ_i

From Remark 4.2.1 above it is clear that there is no further need for considering the conditional probability of misclassification under the random effects model.

The most useful result on the expected probability of misclassification for the two- population problem is that of Okamoto (1963), given in expression (2.1.26) of Section 2.1 for the case of equal-sized training samples $n_1 = n_2 = n$ from π_1 and π_2:

$$\begin{aligned} P_i^e(\delta^2) &= P[\text{misclassification}\,|\,n,\nu,\delta^2; x \in \pi_i] \\ &= \Phi(-\tfrac{\delta}{2}) + \tfrac{1}{\nu}\,\phi(\tfrac{\delta}{2})(\tfrac{p-1}{\delta} + \tfrac{p\delta}{4}) + O(n^{-2}) \qquad (4.2.7) \end{aligned}$$

where,

$$\delta^2 = \delta_{12}^2 = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\,,$$

ν is the degrees of freedom of S and $\phi(\cdot)$ is the standard normal density function.

The expected probability of misclassification under the random effects model may therefore be obtained by taking the expectation of (4.2.7) over the distribution of δ^2. Since there is no difference in (4.2.7) for $x \in \pi_1$ or $x \in \pi_2$ (this is not the case if $n_1 \neq n_2$) the subscript i will be dropped from $P_i^e(\delta^2)$. So,

$$P^e = P[\text{misclassification}|n,\nu] = E_{\delta^2}[P^e(\delta^2)]$$

$$= E_{\delta^2}[\Phi(-\tfrac{\delta}{2}) + \tfrac{1}{\nu}\,\phi(\tfrac{\delta}{2})(\tfrac{p-1}{\delta} + \tfrac{p\delta}{4})] + O(n^{-2}) \;. \qquad (4.2.8)$$

As in the case where the parameters are known, we may approximate (4.2.8) usihg the approximation (4.1.5). The first term in (4.2.8) is just the probability of misclassification when the parameters are known, and its appro.imation is given in (4.1.9), so we need look only at the second term. As before, we need the second derivative of this term with respect to δ^2. Some straightforward calculations yield, letting $z = \delta^2$:

$$\frac{d^2}{dz^2}\{\tfrac{1}{\nu}\,\phi(\tfrac{\sqrt{z}}{2})(\tfrac{p-1}{\sqrt{z}} + \tfrac{p\sqrt{z}}{4})\}$$

$$= \frac{z^{-\frac{5}{2}}}{4\nu}\,\phi(\tfrac{\sqrt{z}}{2})\{3(p-1) + \tfrac{p-2}{4}\,z - (\tfrac{p+1}{16})z^2 + \tfrac{p}{64}\,z^3\} \;. \qquad (4.2.9)$$

Applying (4.1.5), (4.1.7), (4.1.9) and (4.2.9) to (4.2.8), we get:

$$P^e \doteq \Phi\left(-\sqrt{\sum_{i=1}^{r}\lambda_i/2}\right) + \tfrac{1}{2}\phi\left(\sqrt{\sum_{i=1}^{r}\lambda_i/2}\right)\frac{(1 + \sum_{i=1}^{r}\lambda_i/2)\sum_{i=1}\lambda_i^2}{(2\sum_{i=1}^{r}\lambda_i)^{3/2}}$$

$$+ \frac{1}{\nu}\,\phi\left(\sqrt{\sum_{i=1}^{r}\lambda_i/2}\right)\left\{\frac{p-1}{\sqrt{2\sum_{i=1}^{r}\lambda_i}} + \frac{p}{4}\sqrt{2\sum_{i=1}^{r}\lambda_i}\right\}$$

$$+ \frac{(2\sum_{i=1}^{r}\lambda_i)^{-5/2}}{8\nu}\,\phi\left(\sqrt{\sum_{i=1}^{r}\lambda_i/2}\right)\{3(p-1) + (\tfrac{p-2}{4})2\sum_{i=1}^{r}\lambda_i - (\tfrac{p+1}{16})(2\sum_{i=1}^{r}\lambda_i)^2$$

$$+ \tfrac{p}{64}\,(2\sum_{i=1}^{r}\lambda_i)^3\}\;8\sum_{i=1}^{r}\lambda_i^2$$

i.e. $$P^e \doteq \phi\left(-\sqrt{\tfrac{1}{2}\sum_{i=1}^{r}\lambda_i}\right) + \frac{1}{\nu}\,\phi\left(\sqrt{\tfrac{1}{2}\sum_{i=1}^{r}\lambda_i}\right)\left\{\left(2\sum_{i=1}^{r}\lambda_i\right)^{-\frac{3}{2}}\left(p-1+\frac{p}{2}\sum_{i=1}^{r}\lambda_i\right)\right.$$

$$+\sum_{i=1}^{r}\lambda_i^2\left(2\sum_{i=1}^{r}\lambda_i\right)^{-5/2}\left(3(p-1)+\left(\frac{2\nu+p-2}{2}\right)\sum_{i=1}^{r}\lambda_i+\left(\frac{2\nu-p-1}{4}\right)\left(\sum_{i=1}^{r}\lambda_i\right)^2+\right.$$

$$+\sum_{i=1}^{r}\lambda_i^2\left(2\sum_{i=1}^{r}\lambda_i\right)\quad\left(3(p-1)+\left(\frac{\quad}{2}\right)\sum_{i=1}^{r}\lambda_i+\left(\frac{\quad}{4}\right)\left(\sum_{i=1}^{r}\lambda_i\right)^2+\right.$$

$$\left.\left.+\frac{p}{8}\left(\sum_{i=1}^{r}\lambda_i\right)^3\right)\right\}\,. \qquad (4.2.10)$$

A more accurate expression for P^e may be obtained by evaluating (4.2.8) exactly, using expressi n (4.1.10) for the density of δ^2.

Letting $z = \delta^2$ as before, this becomes:

$$P^e = \int_0^\infty\left(\phi\left(-\frac{\sqrt{z}}{2}\right)+\frac{1}{\nu}\,\phi\left(\frac{\sqrt{z}}{2}\right)\left((p-1)z^{-\frac{1}{2}}+\frac{p}{4}z^{\frac{1}{2}}\right)\right)\frac{1}{\beta}\sum_{j=0}^{\infty}c_j\,g_{r+2j}\left(\frac{z}{\beta}\right)dz + O(n^{-2})\,. \qquad (4.2.11)$$

The first term in the above integral is just the probability of misclassification in the case where the parameters are known, and is given in (4.1.16) and (4.1.23) for r even and odd, respectively. The second term may be evaluated, after interchanging the summation and integration operations, in terms of gamma functions. After some simplification, this yields,

<u>for r even</u>:

$$P^e = \tfrac{1}{2}\left\{1-\tfrac{1}{2}\sqrt{\frac{\beta}{\pi}}\sum_{j=0}^{\infty}c_j\left[\sum_{i=0}^{\frac{1}{2}r+j-1}\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+1)}\left(1+\frac{\beta}{4}\right)^{-(i+\frac{1}{2})}-\frac{\Gamma(\frac{1}{2}r+j-\frac{1}{2})}{\nu\Gamma(\frac{1}{2}r+j)}\left(1+\frac{\beta}{4}\right)^{-(\frac{1}{2}r+j-\frac{1}{2})}\right.\right.$$

$$\left.\left.\times\left(\frac{2(p-1)}{\beta}+\frac{p(\frac{1}{2}r+j-\frac{1}{2})}{(1+\beta/4)}\right)\right]\right\}+O(n^{-2})\,. \qquad (4.2.12)$$

for r odd:

$$P^e = \tfrac{1}{2}\{1 - \frac{2}{\pi}\cos^{-1}(\frac{1}{\sqrt{1+\beta/4}}) - \tfrac{1}{2}\sqrt{\frac{\beta}{\pi}}\sum_{j=0}^{\infty} c_j[\sum_{i=0}^{\frac{1}{2}(r-1)+j-1} \frac{\Gamma(i+1)}{\Gamma(i+1.5)}(1+\frac{\beta}{4})^{-(i+1)}$$

$$- \frac{\Gamma(\frac{1}{2}r+j-\frac{1}{2})}{\sqrt{\pi}\,\Gamma(\frac{1}{2}r+j)}(1+\frac{\beta}{4})^{-(\frac{1}{2}r+j-\frac{1}{2})}(\frac{2(p-1)}{\beta} + \frac{p(\frac{1}{2}r+j-\frac{1}{2})}{(1+\beta/4)})]\} + 0(n^{-2}).$$

(4.2.13)

4.2.2 Evaluating the Probabilities of Misclassification for k = 2 populations

FORTRAN subroutine PROB1, given in Appendix 4.3, evaluates formulae (4.2.12) and (4.2.13) for the probability of misclassification when the parameters are unknown. Table 4.2.1 gives the probabilities of misclassification for the case r = 5 for the same three sets of eigenvalues $\{\lambda_i\}$, all with a trace of 15, that were used in the earlier examples, and two values of ν, together with the corresponding approximate probabilities obtained from formula (4.2.10).

Table 4.2.1

case	$\{\lambda_i\}$	ν	Probability of Misclassification correct to $0(n^{-2})$	Approximate Probability of Misclassification
(a)	11,1,1,1,1	20	.0570	.0585
(b)	3,3,3,3,3	20	.0315	.0253
(c)	5,4,3,2,1	20	.0354	.0295
(d)	11,1,1,1,1	40	.0481	.0460
(e)	3,3,3,3,3	40	.0260	.0197
(f)	5,4,3,2,1	40	.0294	.0230

Comparing the probabilities of misclassification for the cases ν = 20 and ν = 40 with each other and with the corresponding probabilities in Table 4.1.1, which represent the case where ν → ∞, clearly indicates the effect that sample size has on them. Moreover, as in the case where the parameters are known, the approximation to the probability provided by formula (4.2.10) is only correct to about two decimal places.

4.2.3 The case k > 2 populations

Using classification rule (4.2.1), the probability of correct classification, given $x \in \pi_i$ becomes:

$$P[\text{correct classification}|x \in \pi_i] = P[d_i^2(x) \leq \underset{\substack{j=1,\ldots,k\\ j\neq i}}{\text{Min}} d_j^2(x)|x \in \pi_i] \tag{4.2.14}$$

Now, given that $x \in \pi_i$, the marginal distribution of $d_i^2(x)$ is proportional to the central $F(p,\nu-p+1)$ distribution, and is given by expression (3.3.19). On the other hand, the marginal distribution of $d_j^2(x)$, $j \neq i$, is, conditionally on δ_{ij}^2, proportional to the noncentral $F(p,\nu-p+1)$ distribution with noncentrality parameter proportional to δ_{ij}^2. See (3.3.6). Its unconditional distribution is given by (3.3.20) and (3.3.21). However, the joint distribution of the $d_j^2(x)$, $j = 1,\ldots,k$, is unknown, so that expression (4.2.14) cannot be evaluated.

Using classification rule (4.2.2), the probability of correct classification, given $x \in \pi_i$, is:

$$P[\text{correct classification}|x \in \pi_i] = P[V_{ij}(x) > 0, \forall j=1,\ldots,k;\ j \neq i|x \in \pi_i]. \tag{4.2.15}$$

As in the above case the marginal distribution of $V_{ij}(x)$, conditional on δ_{ij}^2, is known (Okamoto, 1963) and the unconditional distribution can, in principle, be obtained by integrating over the distribution of δ_{ij}^2. However, the joint distribution of the $V_{ij}(x)$ is again unknown, so that expression (4.2.15) can also not be evaluated.

As in the case where the parameters are known, we therefore consider bounds on the probability of correct classification. As before, Cacoullos' lower bound (2.1.32) refers to the <u>minimum</u> probability of correct classification and we can improve on them by using Bonferroni's first inequality. Using the analogous argument as that leading up to expression (4.1.28) in the case where the parameters are known, and using Okamoto's (1963) expression (4.2.7) for the probability of misclassification for two populations together with the assumption that the training samples from each of the k populations are all the same size n, yields the following lower bound on the probability of correct classification under the random effects model:

$$\text{P[correct classification]} \geq 1 - (k-1)E_{\delta_{ij}^2}[\Phi(-\tfrac{1}{2}\delta_{ij}) + \frac{1}{\nu}\phi(\tfrac{1}{2}\delta_{ij})$$

$$\times \{\frac{p-1}{\delta_{ij}} + \frac{p\delta_{ij}}{4}\}] + O(n^{-2}). \qquad (4.2.16)$$

Finally, substituting expressions (4.2.12) and (4.2.13) for the expectation in (4.2.16), yields,

<u>for r even</u>:

$$\text{P[correct classification]} \geq 1 - \frac{k-1}{2}(1 - \tfrac{1}{2}\sqrt{\frac{\beta}{\pi}}\sum_{j=0}^{\infty} c_j[\sum_{i=0}^{\frac{r}{2}+j-1} \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+1)}(1+\frac{\beta}{4})^{-(i+\frac{1}{2})}$$

$$- \frac{\Gamma(\frac{1}{2}r+j-\frac{1}{2})}{\nu\Gamma(\frac{1}{2}r+j)}(1+\frac{\beta}{4})^{-(\frac{1}{2}r+j-\frac{1}{2})}(\frac{2(p-1)}{\beta} + \frac{p(\frac{1}{2}r+j-\frac{1}{2})}{(1+\beta/4)})]) + O(n^{-2}) \quad (4.2.17)$$

<u>for r odd</u>:

$$\text{P[correct classification]} \geq 1 - \frac{k-1}{2}(1 - \frac{2}{\pi}\cos^{-1}(\frac{1}{\sqrt{1+\beta/4}}) - \tfrac{1}{2}\sqrt{\frac{\beta}{\pi}}\sum_{j=0}^{\infty} c_j$$

$$\times [\sum^{\frac{1}{2}(r-1)+j-1} \frac{\Gamma(i+1)}{\Gamma(i+1.5)}(1+\frac{\beta}{4})^{-(i+1)} - \frac{\Gamma(\frac{1}{2}r+j-\frac{1}{2})}{\nu\Gamma(\frac{1}{2}r+j)}(1+\frac{\beta}{4})^{-(\frac{1}{2}r+j-\frac{1}{2})}(\frac{2(p-1)}{\beta} + \frac{p(\frac{1}{2}r+j-\frac{1}{2})}{(1+\beta/4)})])$$

$$+ O(n^{-2}). \qquad (4.2.18)$$

We can also obtain an upper bound on the probability of correct classification in a manner similar to that used when the parameters are known. Using Okamoto's (1963) exprecsion (4.2.7) and assuming training samples of equal size n, yields the expression analogous to (4.1.32):

$$P[\text{correct classification} \mid x \in \pi_i] \leq 1 - E_{\delta_i^2}[\Phi(-\tfrac{1}{2}\delta_i)$$

$$+ \frac{1}{\nu}\,\phi(\tfrac{1}{2}\delta_i)\{\frac{p-1}{\delta_i} + \frac{p\delta_i}{4}\}] + O(n^{-2}) \qquad (4.2.19)$$

where

$$\delta_i^2 = \min_{\forall j \neq i} \delta_{ij}^2 .$$

The first term inside the expectation was evaluated in the case where the parameters are known, conditionally on μ_i. The second term is, using the distribution (4.1.38) of δ_i^2, conditionally on μ_j:

$$I = \int_0^\infty \frac{1}{\nu}\,\phi(\tfrac{1}{2}\sqrt{z})(\frac{p-1}{\sqrt{z}} + \frac{p\sqrt{z}}{4}) f_{\delta_i^2|\mu_i}(z)dy$$

$$= \frac{(k+1)}{\sqrt{2\pi}\nu\beta}\int_0^\infty e^{-z/8}(\frac{p-1}{\sqrt{z}} + \frac{p\sqrt{z}}{4})(1 - \sum_{j=0}^{\infty} c_j'\, G_{r+2j}(\frac{z}{\beta}))^{(k-2)} \sum_{j=0}^{\infty} c_j'\, g_{r+2j}(\frac{z}{\beta})dz$$

$$= \frac{(k-1)}{\sqrt{2\pi}\nu}\int_0^\infty e^{-\beta y/8}(\frac{p-1}{\sqrt{\beta y}} + \frac{p\sqrt{\beta y}}{4})e^{-(k-2)y/2}\sum_{s=0}^{\infty} a_s'\, y^s \sum_{j=0}^{\infty} \frac{c_j'\, y^{\frac{1}{2}r+j-1}}{2^{\frac{1}{2}r+j}\Gamma(\frac{1}{2}r+j)}\, e^{-\frac{1}{2}y}\, dy$$

for the case when r is even, where the a_s' are defined in (4.1.39) with $(k-1)$ replaced by $(k-2)$ and the c_j' are defined in (4.1.37). Interchanging the order of integration and evaluating the resulting integral yields:

$$I_{*}=\frac{k-1}{\sqrt{2\pi\nu}}\sum_{s=0}^{\infty}a'_s\sum_{j=0}^{\infty}\frac{c'_j}{2^{\frac{1}{2}r+j}\Gamma(\frac{1}{2}r+j)}(\frac{p-1}{\sqrt{\beta}}(\frac{2}{k+\beta/4-1})^{\frac{1}{2}r+s+j-\frac{1}{2}}\Gamma(\frac{1}{2}r+s+j-\frac{1}{2})$$

$$\frac{p\sqrt{\beta}}{4}(\frac{2}{k+\beta/4-1})^{\frac{1}{2}r+s+j+\frac{1}{2}}\Gamma(\frac{1}{2}r+s+j+\frac{1}{2})\}. \qquad (4.2.20)$$

Substituting (4.2.20) and (4.1.40) into (4.2.19) and simplifying, gives the following upper bound on the conditional probability of correct classification, given μ_i, when <u>r is even</u>:

$$P[\text{correct classification}|x \in \pi_i;\mu_i] \leq \tfrac{1}{2}\{1+\sqrt{\frac{\beta/4}{k+\beta/4-1}}\sum_{j=0}^{\infty}a_j(\frac{2}{k+\beta/4-1})^j(\tfrac{1}{2})^{[j]}\}$$

$$-\frac{k-1}{\sqrt{2\pi\nu}}\sum_{s=0}^{\infty}a'_s\sum_{j=0}^{\infty}\frac{c'_j\Gamma(\frac{1}{2}r+s+j-\frac{1}{2})}{2^{\frac{1}{2}r+j}\Gamma(\frac{1}{2}r+j)}(\frac{2}{k+\beta/4-1})^{\frac{1}{2}r+s+j-\frac{1}{2}}\{\frac{p-1}{\sqrt{\beta}}+\frac{p\sqrt{\beta}}{2}(\frac{\frac{1}{2}r+s+j-\frac{1}{2}}{k+\beta/4-1})\}$$

$$(4.2.21)$$

where,

the a_j are defined in (4.1.39) and evaluated in Appendix 4.2,
the a'_s are similarly defined, but with (k-1) replaced by (k-2) and
the c'_j are defined in (4.1.37).

<u>Remark 4.2.2</u> As in the case where the parameters are known, an approximate upper bound on the <u>unconditional</u> probability of correct classification with k populations may be obtained by ignoring the intercorrelations between the δ^2_{ij}, $j = 1,\ldots,k$; $j \neq i$, and proceeding as if they were independent. Arguing in exactly the same way as in Remark 4.1.2, we conclude that (4.2.21) is also an approximate upper bound on the unconditional probability if the c'_j are replaced by c_j (defined in (3.1.9)) in this expression and in the definition (4.1.39) of the a_j and a'_s. Furthermore, for a

given set of eigenvalues $\{\lambda_i\}$ the *upper* bound on the conditional probability of correct classification evaluated at $\mu_i' = \xi$ is *exactly equal to the approximate* bound on the unconditional probability for the case where the eigenvalues are all halved.

4.2.4 Evaluating the bounds on the probabilities of correct classification for $k > 2$ populations

Expressions (4.2.17) and (4.2.18) for the *lower bound* on the probability of correct classification are also computed by subroutine PROB1 given in Appendix 4.3. Table 4.2.2 gives the values of this bound for the same six sets of eigenvalues that were used in Table 4.1.2 for the case when the parameters are known, and for k = 5 populations. *The degrees of freedom* ν were taken to be 20.

Upper bound (4.2.21) on *the* conditional probability of correct classification, given $\mu_i = \xi$ was computed for the special *case where* the eigenvalues are equal, as was the corresponding approximate bound on the unconditional probability. See Sub-Section 4.1.4 for the details and for the values of the a_j when r = 4. The corresponding values for the a_s' are:

$$a_0' = 1,\ a_1' = \frac{3}{2},\ a_2' = \frac{3}{4},\ a_3' = \frac{1}{8} \text{ and } a_s' = 0,\ \forall s > 3.$$

For the same reason given in Sub-Section 4.1.4, the upper bounds computed for the case of equal eigenvalues are also valid for other sets of eigenvalues with the same trace.

Table 4.2.2

Bounds on the probabilities of correct classification for $k = 5$ populations and degrees of freedom $\nu = 20$

Case	$\{\lambda_i\}$	Lower bound	Upper bound on conditional prob. evaluated at $\mu_i = \xi$	Approximate upper bound
(a)	11,1,1,1,1	.7719	-	-
(b)	3,3,3,3,3	.8739	-	-
(c)	5,4,3,2,1	.8582	-	-
(d)	7,1,1,1	.6713	.7416	.8325
(e)	2.5,2.5,2.5,2.5	.7579	.7416	.8325
(f)	4,3,2,1	.7386	.7416	.8325

As in the case where the parameters are known, the upper bound on the conditional probability of correct classification, evaluated at $\mu_i = \xi$, tends to be unrealistically low, and is in fact lower than the lower bound in one case. For practical purposes, the approximate upper bound on the unconditional probability is therefore generally more useful.

Appendix 4.1

Proof of Theorem 4.1.1

Since T is a nonnegative definite symmetric matrix of rank r, we may as in Theorem 3.1.1 let $T = T_1 T_1'$, where T_1 is a p × r matrix of rank r. Making the transformation

$$X = T_1 Z$$

we immediately have that

$$Z \sim N_p(\eta, I)$$

where η is the solution to $T_1 \eta = \mu$.
Therefore $d^2 = X' \Sigma^{-1} X = Z' T_1' \Sigma^{-1} T_1 Z = Z' V Z$, where $V = T_1' \Sigma^{-1} T_1$ is an (r × r) positive definite symmetric matrix. Now V can be expressed in the canonical form:

$$V = P \Lambda P'$$

where $\Lambda = \mathrm{diag}\{\lambda_i\}$ and $\{\lambda_i\} = \mathrm{eigs}\{T_1' \Sigma^{-1} T_1\} = \mathrm{eigs}\{T \Sigma^{-1}\}$ and P is the orthogonal matrix whose i^{th} column is the eigenvector of V corresponding to λ_i.
Therefore d^2 becomes:

$$d^2 = Z' P \Lambda P' Z = Y' \Lambda Y = \sum_{i=1}^{r} \lambda_i y_i^2$$

where $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = P' Z \sim N_r(P'\eta, I)$

So $y_i^2 \sim \chi_1^2(\omega_i^2)$, independently, where ω_i is the i^{th} element of $P'\eta$.

Appendix 4.2

Evaluating the coefficients a_j in identity (4.1.39):

$$\Big(\sum_{j=0}^{\infty} c_j' \sum_{i=0}^{\frac{1}{2}r+j-1} (\tfrac{y}{2})^i/i!\Big)^{k-1} \equiv \sum_{j=0}^{\infty} a_j\, y^j$$

Theorem A 4.2.1

$$a_j = \frac{(k-1)^j}{2^j\, j!} \quad \text{for} \quad j = 0,1,\ldots \tfrac{1}{2}r-1$$

Proof The left hand side of (4.1.39) may be written:

$$\Big(\sum_{j=0}^{\infty} c_j' \sum_{i=0}^{\frac{1}{2}r+j-1} (\tfrac{y}{2})^i/i!\Big)^{k-1} = \{c_0'\Big(1+\frac{y}{2\times 1!}+\ldots+\frac{y^{\frac{1}{2}r-1}}{2^{\frac{1}{2}r-1}(\frac{1}{2}r-1)!}\Big)$$

$$+c_1'\Big(1+\frac{y}{2\times 1!}+\ldots+\frac{y^{\frac{1}{2}r}}{2^{\frac{1}{2}r}(\frac{1}{2}r)!}\Big)+\ldots+c_j'\Big(1+\frac{y}{2\times 1!}+\ldots+\frac{y^{\frac{1}{2}r+j-1}}{2^{\frac{1}{2}r+j-1}(\frac{1}{2}r+j-1)!}\Big)$$

$$+\ldots\}^{k-1}$$

$$= \{1+\frac{y}{2\times 1!}+\ldots+\frac{y^{\frac{1}{2}r-1}}{2^{\frac{1}{2}r-1}(\frac{1}{2}r-1)} + (1-c_0')\frac{y^{\frac{1}{2}r}}{2^{\frac{1}{2}r}(\frac{1}{2}r)!}+\ldots.$$

$$+\Big(1-\sum_{i=0}^{j-1} c_i'\Big)\frac{y^{\frac{1}{2}r+j-1}}{2^{\frac{1}{2}r+j-1}(\frac{1}{2}r+j-1)!}+\ldots.\}^{k-1} \qquad (A\ 4.2.1)$$

where we have assumed that (4.1.37) is a mixture distribution, so that $\sum_{j=0}^{\infty} c_j' = 1$

$$= \Big(\sum_{s=0}^{\infty} b_s\, y^s\Big)^{k-1} \qquad (A\ 4.2.2)$$

where $b_s = \dfrac{1}{2^s\, s!}$ for $s = 0,1,\ldots \tfrac{1}{2}r-1$

$$= \Big(1-\sum_{i=0}^{j-1} c_i'\Big)/2^s\, s! \quad \text{for} \quad s = \tfrac{1}{2}r+j-1;\ j = 1,2,\ldots..$$

Using the multinomial theorem to evaluate (A 4.2.2) and substituting this into identity (4.1.39) immediately yields:

$$a_j = \Sigma \frac{(k-1)!}{\ell_0!\,\ell_1!\ldots\ell_j!}\, b_1^{\ell_1}\, b_2^{\ell_2} \ldots b_j^{\ell_j} \qquad \text{(A 4.2.3)}$$

where the summation is taken over all partitions $\ell_0, \ell_1, \ldots \ell_j$ of k-1 for which:

$$\sum_{i=1}^{j} i\,\ell_i = j \qquad \text{(A 4.2.4)}$$

Subsituting the values of b_s given in (A 4.2.1) into (A 4.2.3) and using (A 4.2.4) gives, <u>for $j < \frac{1}{2}r-1$</u>:

$$a_j = \frac{1}{2^j}\, \Sigma \frac{(k-1)!}{\ell_0!\,\ell_1!\ldots\ell_j!}\, \left(\frac{1}{1!}\right)^{\ell_1} \left(\frac{1}{2!}\right)^{\ell_2} \ldots \left(\frac{1}{j!}\right)^{\ell_j} \qquad \text{(A 4.2.5)}$$

The first few coefficients are,from (A 4.2.5):

$$a_0 = 1$$

$$a_1 = \frac{1}{2}\left\{\frac{(k-1)!}{(k-2)!\,1!}\left(\frac{1}{1!}\right)\right\} = \frac{k-1}{2 \times 1!}$$

$$a_2 = \frac{1}{2^2}\left\{\frac{(k-1)!}{(k-2)!\,1!}\,\frac{1}{2!} + \frac{(k-1)!}{(k-3)!\,2!}\left(\frac{1}{1!}\right)^2\right\}$$

$$= \frac{k-1}{2^2}\left\{\frac{1}{2!} + \frac{k-2}{2!}\right\} = \frac{(k-1)^2}{2^2 \times 2!}$$

and so on.

The rest of the proof follows by induction. Assume that the result is true for all $j \le i$ and all k, where $i < \frac{1}{2}r-1$.

i.e. $a_j^{(k-1)} = \frac{(k-1)^j}{2^j\ j!}$ for $j = 0,1,\ldots,i$ (A 4.2.6)

where the superscript in $a_j^{(k-1)}$ indicates its dependence on k-1.

Now,

$$(\sum_{s=0}^{\infty} b_s\ y_s)^{k-1} \equiv (\sum_{s=0}^{\infty} b_s\ y^s)(\sum_{s=0}^{\infty} b_s\ y^s)^{k-2}$$

$$\text{i.e.} \quad \sum_{j=0}^{\infty} a_j^{(k-1)}\ y^j = (\sum_{s=0}^{\infty} b_s\ y^s)(\sum_{j=0}^{\infty} a_j^{(k-2)}\ y^j) \qquad \text{(A 4.2.7)}$$

Equating coefficients of y^{i+1} on both sides of (A 4.2.7) yields:

$$\begin{aligned} a_{i+1}^{(k-1)} &= b_0\ a_{i+1}^{(k-2)} + b_1\ a_i^{(k-2)} + b_2\ a_{i-1}^{(k-2)} + \ldots + b_{i+1}\ a_0^{(k-2)} \\ &= a_{i+1}^{(k-2)} + \frac{1}{2\times 1!}\frac{(k-1)^i}{2^i\ i!} + \frac{1}{2^2\times 2!}\frac{(k-1)^{i-1}}{2^{i-1}(i-1)!} + \ldots + \frac{1}{2^{i+1}(i+1)!} \end{aligned}$$

by assumption (A 4.2.6) .

Therefore,

$$\begin{aligned} a_{i+1}^{(k-1)} - a_{i+1}^{(k-2)} &= \frac{1}{2^{i+1}(i+1)!} \sum_{j=0}^{i} \binom{i+1}{j}(k-2)^j \\ &= \frac{1}{2^{i+1}(i+1)!}((k-2+1)^{i+1} - (k-2)^{i+1}) \\ &= \frac{(k-1)^{i+1}}{2^{i+1}(i+1)!} - \frac{(k-2)^{i+1}}{2^{i+1}(i+1)!} \qquad \text{(A 4.2.8)} \end{aligned}$$

and since (A 4.2.8) holds identically for all k it immediately follows that:

$$a_{i+1}^{(k-1)} = \frac{(k-1)^{i+1}}{2^{i+1}(i+1)!} \quad \text{for all k} \qquad \text{(A 4.2.9)}$$

Finally, as the theorem has already been shown to be true for all $j \le 3$, it is true for all $j \le \frac{1}{2}r-1$ by induction.

Remark A 4.2.1 The coefficients a_j for $j \ge \frac{1}{2}r$ are most readily calculated from (A 4.2.1) with the help of Theorem A 4.2.1. Unfortunately no general result is available for them. Writing (A 4.2.1) as:

$$\left(\sum_{j=0}^{\infty} c_j' \sum_{i=0}^{\frac{1}{2}r+j-1} \frac{y^i}{2^i\ i!} \right)^{k-1} = \left(\sum_{j=0}^{\infty} \frac{y^j}{2^i\ i!} - \sum_{j=\frac{1}{2}r}^{\infty} \left(\sum_{i=0}^{j-\frac{1}{2}r} c_j' \right) \frac{y^j}{2^j\ j!} \right)^{k-1} \qquad \text{(A 4.2.10)}$$

and using the following obvious generalization of Theorem A 4.2.1:

$$\left(\sum_{j=0}^{\infty} \frac{y^j}{2^j\ j!} \right)^{k-1} = \sum_{j=0}^{\infty} \frac{(k-1)^j}{2^j\ j!} y^j \qquad \text{(A 4.2.11)}$$

we obtain the first few higher coefficients as follows:

$$a_{\frac{1}{2}r} = \frac{(k-1)^{\frac{1}{2}r}}{2^{\frac{1}{2}r}(\frac{1}{2}r)!} - \binom{k-1}{1} \frac{c_0'}{2^{\frac{1}{2}r}(\frac{1}{2}r)!} = \frac{k-1}{2^{\frac{1}{2}r}(\frac{1}{2}r)!} \left((k-1)^{\frac{1}{2}r-1} - c_0'\right) \qquad \text{(A 4.2.12)}$$

$$a_{\frac{1}{2}r+1} = \frac{(k-1)^{\frac{1}{2}r+1}}{2^{\frac{1}{2}r+1}(\frac{1}{2}r+1)!} - \binom{k-1}{1} \left\{ \frac{(c_0' + c_1')}{2^{\frac{1}{2}r+1}(\frac{1}{2}r+1)!} + \frac{c_0'}{2^{\frac{1}{2}r}(\frac{1}{2}r)!} \frac{(k-2)}{2 \times 1!} \right\}$$

$$= \frac{k-1}{2^{\frac{1}{2}r+1}(\frac{1}{2}r+1)!} \left\{(k-1)^{\frac{1}{2}r} - c_0'(1 + (k-2)(\tfrac{1}{2}r+1)) - c_1'\right\} \qquad \text{(A 4.2.13)}$$

$$a_{\frac{1}{2}r+2} = \frac{(k-1)^{\frac{1}{2}r+2}}{2^{\frac{1}{2}r+2}(\frac{1}{2}r+2)!} - \binom{k-1}{1} \left\{ \frac{(c_0' + c_1' + c_2')}{2^{\frac{1}{2}r+2}(\frac{1}{2}r+2)!} + \frac{c_0' + c_1'}{2^{\frac{1}{2}r+1}(\frac{1}{2}r+1)!} \cdot \frac{(k-2)}{2 \times 1!} \right.$$

$$\left. + \frac{c_0'}{2^{\frac{1}{2}r}(\frac{1}{2}r)!} \cdot \frac{(k-2)^2}{2^2 \times 2!} \right\}$$

$$= \frac{(k-1)}{2^{\frac{1}{2}r+2}(\frac{1}{2}r+2)!} \{(k-1)^{\frac{1}{2}r+1} - c_0'(1 + (k-2)(\tfrac{1}{2}r+2)(1+\tfrac{1}{2}(k-2)(\tfrac{1}{2}r+1)))$$

$$- c_1'(1 + (k-2)(\tfrac{1}{2}r+2)) - c_2'\} \qquad \text{(A 4.2.14)}$$

and so on.

Result (A 4.2.13) only holds for $\frac{1}{2}r > 1$ and (A 4.2.14) only for $\frac{1}{2}r > 2$.

For $\underline{\frac{1}{2}r = 2}$, i.e. $r = 4$, (A 4.2.14) becomes, instead:

$$a_4 = \frac{k-1}{2^4 \times 4!} \{(k-1)^3 - c_0'(1 + 4(k-2)(1 + \tfrac{3}{2}(k-2)) - c_0'\ 3(k-2))$$

$$- c_1'(1 + 4(k-2)) - c_2'\} \qquad \text{(A 4.2.15)}$$

and for $\underline{\frac{1}{2}r = 1}$, i.e. $r = 2$, (A 4.2.13) and (A 4.2.14) become, respectively:

$$a_2 = \frac{k-1}{2^2 \times 2!} \{ k - 1 - c_0'(1 + (k-2)(2 - c_0')) - c_1'\} \qquad \text{(A 4.2.16)}$$

$$a_3 = \frac{k-1}{2^3 \times 3!} \{(k-1)^2 - c_0'(1 + 3(k-2)(k - 1 - c_0(k-2) - c_1 + c_0^2(\tfrac{k}{3} - 1))$$

$$- c_1'(1 + 3(k-2)) - c_2'\} \qquad \text{(A 4.2.17)}$$

Appendix 4.3 FORTRAN Subroutines for computing probabilities of correct- and misclassification

```
      SUBROUTINE PROBS(NORD,BETA,CVEC,NTERMS,ERROR,NTERM1,PROB2,NGPS,
     1PROBK)
C
C
C     PROGRAM TO COMPUTE PROBABILITIES OF MISCLASSIFICATION.  KNOWN PARAMETERS
C     THE PARAMETERS ARE:
C     NORD = NO. OF EIGENVALUES.    BETA = THE PARAMETER BETA.
C     CVEC = THE VECTOR OF CONSTANTS C(J),    NTERMS = LENGTH OF VECTOR CVEC.
C     ERROR = MAXIMUM VALUE OF THE LAST TERM IN THE INFINITE SUM IN THE FORMUL/
C     NTERM1 = NO. OF TERMS IN SUMMATION ACTUALLY COMPUTED.
C     PROB2 = PROBABILITY OF MISCLASSIFICATION WITH TWO GROUPS.
C     NGPS = NO. OF GROUPS.    PROBK = LOWER BOUND ON THE PROBABILITY OF CORRECT
C     CLASSIFICATION WITH 'NGPS' GROUPS.
C
C
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 CVEC(NTERMS)
      BETIN = 1./(1.+.25*BETA)
      PI = 3.14159265358979З
      SQTPI = DSQRT(PI)
      DF = NDF
      IF(MOD(NORD,2) .GT. 0) GO TO 10
      ITOP = NORD/2
      TERM = DSQRT(BETIN) * SQTPI
      SUM = TERM
      IF(ITOP .LE. 1) GO TO 2
      DO 1 I = 2,ITOP
      AI = I
      TERM = TERM*(AI-1.5)*BETIN/(AI-1.)
    1 SUM = SUM + TERM
    2 CONTINUE
      SUM1 = CVEC(1) * SUM
      NTERM1 = 1
      IF(NTERMS .LE. 1) GO TO 4
      DO 3 J = 2,NTERMS
      AJ = ITOP + J - 1.
      TERM = TERM*(AJ-1.5)*BETIN/(AJ-1.)
      SUM = SUM + TERM
      TERM1 = CVEC(J) * SUM
      IF(TERM1 .LT. ERROR) GO TO 4
      NTERM1 = J
    3 SUM1 = SUM1 + TERM1
    4 PROB2 = .5*(1.-.5*DSQRT(BETA)/SQTPI*SUM1)
      PROBK = 1. - (NGPS-1.)*PROB2
      GO TO 20
   10 ITOP = (NORD-1)/2
      TERM = 2.*BETIN/SQTPI
      SUM = TERM
      IF(ITOP .LE. 0) SUM = 0.
      IF(ITOP .LE. 1) GO TO 12
      DO 11 I = 2,ITOP
      AI = I
      TERM = TERM*(AI-1.)*BETIN/(AI-.5)
   11 SUM = SUM + TERM
   12 CONTINUE
      SUM1 = CVEC(1) * SUM
      NTERM1 = 1
      IF(NTERMS .LE. 1) GO TO 14
      DO 13 J = 2,NTERMS
      AJ = ITOP + J - 1.
      TERM = TERM*(AJ-1.)*BETIN/(AJ-.5)
      SUM = SUM + TERM
      TERM1 = CVEC(J) * SUM
      IF(TERM1 .LT. ERROR) GO TO 14
      NTERM1 = J
   13 SUM1 = SUM1 + TERM1
   14 PROB2 = .5 - DARCOS(DSQRT(BETIN))/PI - DSQRT(BETA)/(4.*SQTPI)*SUM1
      PROBK = 1. - (NGPS-1.)*PROB2
   20 CONTINUE
      WRITE(6,101) ERROR, NTERM1, PROB2, NGPS, PROBK
  101 FORMAT('0PROBABILITY OF MISCLASSIFICATION.   KNOWN PARAMETERS.'
     1/'0CUTOFF VALUE',T30,D12.6/' NO. OF TERMS COMPUTED',T30,I5/
     2' TWO-GROUP PROBABILITY',T30,D12.6/' LOWER BOUND ON PROBABILITY OF
     3 CORRECT CLASSIFICATION WITH',I3,' GROUPS',T75,D12.6)
      RETURN
      END
      SUBROUTINE PROB1(NORD,NORD1,NDF,BETA,CVEC,NTERMS,ERROR,NTERM1,
     1PROB2,NGPS,PROBK)
C
C
C     PROGRAM TO COMPUTE PROBABILITIES OF MISCLASSIFICATION.   UNKNOWN PARAMETER
C     THE PARAMETERS ARE:
C     NORD = NO. OF EIGENVALUES.   NORD1 = THE DIMENSION OF THE PROBLEM,
C     NDF = DEGREES OF FREEDOM OF COVARIANCE MATRIX.   BETA = PARAMETER BETA.
C     CVEC = THE VECTOR OF CONSTANTS C(J).   NTERMS = LENGTH OF VECTOR CVEC.
C     ERROR = MAXIMUM VALUE OF THE LAST TERM IN THE INFINITE SUM IN THE FORMULAE
C     NTERM1 = NO. OF TERMS IN SUMMATION ACTUALLY COMPUTED.
C     PROB2 = PROBABILITY OF MISCLASSIFICATION WITH TWO GROUPS.
C     NGPS = NO. OF GROUPS.   PROBK = LOWER BOUND ON THE PROBABILITY OF CORRECT
C     CLASSIFICATION WITH 'NGPS' GROUPS.
C
C
C
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 CVEC(NTERMS)
      BETIN = 1./(1.+.25*BETA)
      PI = 3.141592653589793
      SQTPI = DSQRT(PI)
      DF = NDF
      IF(MOD(NORD,2) .GT. 0) GO TO 10
      ITOP = NORD/2
      TERM = DSQRT(BETIN) * SQTPI
      SUM = TERM
      IF(ITOP .LE. 1) GO TO 2
      DO 1 I = 2,ITOP
      AI = I
      TERM = TERM*(AI-1.5)*BETIN/(AI-1.)
1     SUM = SUM + TERM
2     CONTINUE
      SUM1  = CVEC(1) * (SUM - TERM/DF*(2.*(NORD1-1.)/BETA + NORD1*
     1(AI-.5)*BETIN))
      NTERM1 = 1
      IF(NTERMS .LE. 1) GO TO 4
      DO 3 J = 2,NTERMS
      AJ = ITOP + J - 1.
      TERM = TERM*(AJ-1.5)*BETIN/(AJ-1.)
      SUM = SUM + TERM
      TERM1 = CVEC(J) * (SUM - TERM/DF*(2.*(NORD1-1.)/BETA + NORD1*
     1(AJ-.5)*BETIN))
      IF(TERM1 .LT. ERROR) GO TO 4
      NTERM1 = J
3     SUM1 = SUM1 + TERM1
4     PROB2 = .5*(1.-.5*DSQRT(BETA)/SQTPI*SUM1)
      PROBK = 1. - (NGPS-1.)*PROB2
      GO TO 20
10    ITOP = (NORD-1)/2
      TERM = 2.*BETIN/SQTPI
      SUM = TERM
      IF(ITOP .LE. 0) SUM = 0.
      IF(ITOP .LE. 1) GO TO 12
      DO 11 I = 2,ITOP
      AI = I
      TERM = TERM*(AI-1.)*BETIN/(AI-.5)
11    SUM = SUM + TERM
12    CONTINUE
      SUM1  = CVEC(1) * (SUM - TERM/DF*(2.*(NORD1-1.)/BETA + NORD1*
     1AI*BETIN))
      NTERM1 = 1
      IF(NTERMS .LE. 1) GO TO 14
      DO 13 J = 2,NTERMS
      AJ = ITOP + J - 1.
      TERM = TERM*(AJ-1.)*BETIN/(AJ-.5)
      SUM = SUM + TERM
      TERM1 = CVEC(J) * (SUM - TERM/DF*(2.*(NORD1-1.)/BETA + NORD1*
     1AJ*BETIN))
      IF(TERM1 .LT. ERROR) GO TO 14
      NTERM1 = J
13    SUM1 = SUM1 + TERM1
14    PROB2 = .5 - DARCOS(DSQRT(BETIN))/PI - DSQRT(BETA)/(4.*SQTPI)*SUM1
      PROBK = 1. - (NGPS-1.)*PROB2
20    CONTINUE
      WRITE(5,101) ERROR, NTERM1, NORD1, NDF, PROB2, NGPS, PROBK
101   FORMAT('0PROBABILITY OF MISCLASSIFICATION,   ESTIMATED PARAMETERS'
     1/'0CUTOFF VALUE',T30,D12.6/' NO. OF TERMS COMPUTED',T30,I5/
     2' DIMENSION OF PROBLEM',T30,I5/' DEGREES OF FREEDOM',T30,I5/
     3' TWO-GROUP PROBABILITY',T30,D12.6/' LOWER BOUND ON PROBABILITY OF
     4 CORRECT CLASSIFICATION WITH',I3,' GROUPS',T75,D12.6)
      RETURN
      END
```

Chaper 5 Hypothesis Testing on and Estimation of the Eigenvalues of $T\Sigma^{-1}$

5.1 Introduction

The results derived in chapter 3 and 4 are all expressed in terms of $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$, the r nonzero eigenvalues of $T\Sigma^{-1}$, either explicitly as in the expressions of the means and variances, or implicitly through the constants c_j appearing in all the density and distribution functions as well as in the probabilities of correct- and misclassification.

It is clear, therefore, that in any practical implementation of these results, sample-based estimates of these quantities will be required.

Since we are only concerned with the nonzero eigenvalues of $T\Sigma^{-1}$, the logical first step is to test the hypotheses that some of the smaller eigenvalues are in fact zero. (They cannot be negative).

In this chapter, therefore, we will consider the two questions of hypothesis testing on and estimation of these eigenvalues.

Section 5.2 will be devoted to the first of these two questions. None of the results given in this section are new, so only the formulae for the various tests will be given, together with a discussion on their applicability to our problem.

In the remaining sections of this chapter the less understood question of estimation of the eigenvalues will be considered. Various estimators will be proposed, and in Section 5.5 they will be compared by means of a simulation experiment.

As in Section 3.3, we will assume that we have a training sample of random observations from each of k populations. Furthermore, because of the inherent problems associated with estimation in random effects models when the samples are unbalanced (see, for example Johnson and Leone

Vol II (1964) page 13) it will be assumed that the sample sizes from each of the k populations are the same.

Therefore, our sample will consist of p-dimensional random vectors,

$$x_{ij}; \quad j = 1,\ldots,n; \quad i = 1,\ldots,k \qquad . (5.1.1)$$

where, under our random effects model,

$$x_{ij} \sim N_p(\mu_i,\Sigma) \quad , \quad \text{independently}$$

and $$\mu_i \sim N_p(\xi,T) \quad , \text{independently.}$$

Let

$$x_{i.} = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \qquad i = 1,\ldots,k$$

and

$$x_{..} = \frac{1}{k}\sum_{i=1}^{k} x_{i.} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}$$

where

$$N = kn \quad .$$

From the data we can construct the following MANOVA table:

Table 5.1.1

Source of Variation	Sums of Squares	Degrees of freedom	Mean Squares	Expected Mean Squares
Between groups	$A_1 = n\sum_{i=1}^{k}(x_{i.}-x_{..})(x_{i.}-x_{..})'$	$\nu_1=k-1$	$S_1=\frac{A_1}{\nu_1}$	$\Sigma+nT$
Within groups	$A_2=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-x_{i.})(x_{ij}-x_{i.})'$	$\nu_2=N-k$	$S_2=\frac{A_2}{\nu_2}$	Σ

Defining,

$$\Sigma_1 = \Sigma + nT \qquad (5.1.2)$$

we have, under the random effects model:

$$\begin{aligned} & A_1 \sim W_p(\nu_1, \Sigma_1) \\ \text{and} \quad & A_2 \sim W_p(\nu_2, \Sigma), \text{ independently} \end{aligned} \qquad (5.1.3)$$

where $W_p(\nu,\Sigma)$ denotes the p-dimensional Wishart distribution with ν degrees of freedom and parameter matrix Σ.

5.2 Hypothesis testing on the λ_i

In this section we discuss the problem of testing whether some, or all of the eigenvalues $\{\lambda_i\}$ of $T\Sigma^{-1}$ are equal to zero.

Log[illegible]e first hypothesis to test is $H_0 : T = 0$, for if it were tru[illegible] 0, $\forall_i$ which would imply that the k populations we[illegible] identical and [illegible] it would be fruitless to continue with the discrim[illegible] analysis.

From (5.1.2) this null hypothesis becomes,

$$\begin{aligned} & H_0 : \Sigma_1 = \Sigma \\ & \text{with alternative,} \\ & H_1 : \Sigma_1 > \Sigma \, . \end{aligned} \qquad (5.2.1)$$

Clearly H_1 would imply that $r(T) > 0$.

The usual MANOVA tests using the statistics A_1 and A_2 defined in Table 5.1.1 are based on the fixed effects model. See for example, de Waal (1976). Under the null hypothesis, however, the distributions of these two statistics are not affected if instead the random effects model pertains, so the abovementioned tests are also appropriate for our situation. On the other hand, under the alternative hypothesis, A_1 has the noncentral Wishart distribution $W_p(\nu_1,\Sigma_1,\Omega)$ with noncentrality parameter Ω when the fixed effects model pertains, as opposed to the distribution given in (5.1.3) for the random effects model. So the power functions of these tests will be different and will have different interpretations under the two models.

All the invariant tests of hypotheses (5.2.1) are based on

$$\{g_1 \geq g_2 \geq \ldots\ldots\ldots \geq g_p\} = eigs(A_1A_2^{-1}). \qquad (5.2.2)$$

Two frequently applied test statistics are:

(i) The likelihood ratio statistic (Wilk's criterion)

$$T_1 = \log(|A_2|/|A_1+A_2|) = \sum_{i=1}^{p} \log(1+g_i) \qquad (5.2.3)$$

(ii) *Hotelling's T_0^2 statistic:*

$$T_2 = \nu_2 T_0^2 = \mathrm{tr}\, A_1A_2^{-1} = \sum_{i=1}^{p} g_i. \qquad (5.2.4)$$

Remark 5.2.1 Two further test statistics due to Roy and Pillai respectively, also appear frequently in the literature, but they won't be considered here. The reason for mentioning Hotelling's T_0^2 statistic is that it is considered again in Sub-Section 5.4.2 where its distribution under the random effects model is discussed.

Anderson (1958), using results from Box (1949), shows that the asymptotic null distribution of T_1 can be written:

$$P[m_1T_1 \le z] = G_{p\nu_1}(z) + \frac{\gamma_2}{m_1{}^2}(G_{p\nu_1+4}(z) - G_{p\nu_1}(z))$$

$$+ \frac{1}{m_1^4}\{\gamma_4(G_{p\nu_1+8}(z) - G_{p\nu_1}(z)) - \gamma_2^2(G_{p\nu_1+4}(z) - G_{p\nu_1}(z))\} + O(N^{-6}) \tag{5.2.5}$$

where,

$$m_1 = \nu_2 + \tfrac{1}{2}(\nu_1-p-1)$$

$$\gamma_2 = p\nu_1(p^2+\nu_1^2-5)/48$$

$$\gamma_4 = \tfrac{1}{2}\gamma_2^2 + \frac{p\nu_1}{1920}(3(p^4+\nu_1^4) + 10p^2\nu_1^2 + 50(p^2+\nu_1^2) + 159)$$

and $G_\nu(\cdot)$ is the χ_ν^2 distribution function. As a rough rule, Anderson (1958) suggests that accuracy to three decimal places may be achieved using the first term only in the above expression if $p^2 + \nu_1^2 \le m_1/3$.

The asymptotic null distribution of T_2 is given by Fujikoshi (1977) in the following form:

$$P[m_2T_2 \le z] = G_{p\nu_1}(z) + \frac{p\nu_1(p+\nu_1+1)}{4m_2}(G_{p\nu_1}(z)$$

$$-2G_{p\nu_1+2}(z) + G_{p\nu_1+4}(z)) + O(m_2^{-2}) \tag{5.2.6}$$

where

$$m_2 = \nu_2 - p - 1.$$

If H_0 is rejected, the next test of interest is whether any subset of the λ_i could all be zero. If true, then the distribution of δ^2, the Mahabanobis distance between any two populations, under the random effects model could be expressed in terms of the remaining non-zero λ_i's only. See Theorem 3.1.1. The null hypothesis of this test is,

$$H_{01} : \lambda_{r+1} = \lambda_{r+2} = \ldots\ldots = \lambda_p = 0$$

where $0 < r < p$.

Fujikoshi (1977) discusses tests for dimensionality of the noncentrality parameter Ω under the fixed effects MANOVA model. That these tests are appropriate for testing H_{01} can be seen by the following argument..

Conditionally on $\mu_1, \mu_2, \ldots, \mu_k$ we have a fixed effects model, in which case A_1 has the noncentral Wishart distribution $W_p(\nu_1, \Sigma, \Omega)$ with noncentrality parameter,

$$\Omega = \tfrac{1}{2}n\Sigma^{-1}\sum_{i=1}^{k}(\mu_i-\mu_.)(\mu_i-\mu_.)' \qquad (5.2.7)$$

where $\mu_. = \frac{1}{k}\sum_{i=1}^{k}\mu_i$. Now, under the random effects model,

$$\mu_i \sim N_p(\xi, T) \quad \text{independently, } \forall_i \qquad (5.2.8)$$

so that,

$$\sum_{i=1}^{k}(\mu_i-\mu_.)(\mu_i-\mu_.)' \sim W_p(\nu_1, T). \qquad (5.2.9)$$

Clearly, from (5.2.7), $r(\Omega) = r(\sum_{i=1}^{k}(\mu_i-\mu_.)(\mu_i-\mu_.)')$, and as long as $k > r(T)$, then from (5.2.9), with probability 1, $r(\sum_{i=1}^{k}(\mu_i-\mu_.)(\mu_i-\mu_.)') = r(T)$. So, for $k > r(T)$,

$$r(\Omega) = r(T) = r(T\Sigma^{-1}) \qquad (5.2.10)$$

and therefore any test for dimensionality of Ω will also be a test of $r(T\Sigma^{-1})$. Finally, since $r(T\Sigma^{-1})$ is equal to the number of non-zero λ_i, testing H_{01} is equivalent to testing the hypothesis $r(\Omega) = r$ against the alternative $r(\Omega) > r$.

The two test statistics, corresponding to T_1 and T_2, for testing H_{01} are:

$$T_{11} = \sum_{i=r+1}^{p} \log(1+g_i) \tag{5.2.11}$$

and

$$T_{21} = \sum_{i=r+1}^{p} g_i . \tag{5.2.12}$$

Fujikoshi (1977) gives the following results on the asymptotic null distributions of T_{11} and T_{21}.

$$P[m_{11}T_{11} \leq z] = G_f(z) + O(m_{11}^{-2}) \tag{5.2.13}$$

where,

$$f = (p-r)(\nu_1-r)$$

and $m_{11} = \nu_2 + \frac{1}{2}(\nu_1-p-1) + \sum_{i=1}^{r} \lambda_i^{-1}$.

$$P[m_{21}T_{21} \leq z] = G_f(z) + \frac{f(p+\nu_1-2r+1)}{4m_{21}} (G_f(z) - 2G_{f+2}(z)$$

$$+ G_{f+4}(z)) + O(m_{21}^{-1}) \tag{5.2.14}$$

where $m_{21} = \nu_2 - p - 1 + r + \sum_{i=1}^{r} \lambda_i^{-1}$.

To apply these tests we clearly need to know λ_i^{-1}, $i = 1,\ldots,r$ appearing in m_{11} and m_{21}. A simple expedient is to replace λ_i^{-1} by $\hat{\lambda}_i^{-1}$ where $\hat{\lambda}_i$ is one of the estimators of λ_i discussed in the remainder of this chapter.

5.3. Estimation of $\{\lambda_i\} = \mathrm{Eigs}(T\Sigma^{-1})$

From Table 5.1.1, expressions (5.1.2) and (5.1.3) and the usual theory associated with the Multivariate Normal distribution it is clear that $S_1 = \nu^{-1}A_1$ and $S_2 = \nu_2^{-1}A_2$ are maximum likelihood point estimators (corrected for bias) of $\Sigma_1 = \Sigma + nT$ and Σ, respectively.

Thus we have the following maximum likelihood estimators for Σ and T:

$$\hat{\Sigma} = S_2$$

and

$$\hat{T} = \frac{1}{n}(S_1 - S_2) \tag{5.3.1}$$

since the transformation is one-to-one. (See, for example, Anderson (1958) page 48).

Moreover, as long as the λ_i are distinct, the eigenvalues of $\hat{T}\hat{\Sigma}^{-1}$ will be the maximum likelihood estimators of the corresponding eigenvalues of $T\Sigma^{-1}$ (See, for example, Anderson (1958) pages 279-80). Therefore, noting that:

$$\begin{aligned}\hat{T}\hat{\Sigma}^{-1} &= \frac{1}{n}(S_1 - S_2)S_2^{-1} \\ &= \frac{1}{n}(S_1S_2^{-1} - I)\end{aligned} \tag{5.3.2}$$

where I is the identity matrix, we have the following maximum likelihood estmators of the λ_i, as long as they are distinct:

$$\hat{\lambda}_i = \frac{1}{n}(\ell_i - 1) \qquad (5.3.3)$$

where $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$ are the eigenvalues of $S_1S_2^{-1}$.

Remark 5.3.1 Note that $\{\ell_i\} = \mathrm{Eigs}\{S_1S_2^{-1}\} = \mathrm{eigs}\{\frac{\nu_2}{\nu_1} A_1A_2^{-1}\} = \{\frac{\nu_2}{\nu_1} g_i\}$. Girshick (1939) proves that the eigenvalues of a sample covariance matrix from a Normal sample are asymptotically independent, unbiased and normally distributed estimators of the corresponding population eigenvalues as long as they are distinct. Using the multivariate analogue of the argument used to prove that the F-distribution tends to the chi-square distribution as the denominator degrees of freedom get large (see for *example Wilks (1962) page 191) it can be shown that the above asymptotic* result also holds for the eigenvalues of $S_1S_2^{-1}$ as both numerator and denominator degrees of freedom get large.

However, as will become clear from the results of the simulation experiment described in section 5.5, very large sample sizes are necessary before these results can be assumed to hold to any reasonable degree of accuracy.

For moderate values of ν_1 and ν_2 the situation is not so simple. Khatri (1967) obtains the joint density function of the eigenvalues $g_1 > g_2 > \ldots > g_p > 0$ of $A_1A_2^{-1}$ which can be expressed in the following form:

$$f_{g_1,\ldots,g_p}(g_1,\ldots,g_p) = c \prod_{i=1}^{p} \gamma_i^{-\frac{1}{2}\nu_1} g_i^{\frac{1}{2}(\nu_1-p-1)} (1+\gamma g_i)^{-\frac{1}{2}(\nu_1+\nu_2)} (\prod_{i<j}(g_i-g_j))$$

$$\times\ {}_1F_0(\tfrac{1}{2}(\nu_1+\nu_2);\ \gamma I - \Gamma^{-1},\ G(I+\gamma G)^{-1}) \qquad (5.3.4)$$

where,

$\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_p > 0$ are the eigenvalues of $\Sigma_1\Sigma^{-1}$

γ is an arbitrary non-negative real number

$\prod_{i<j}$ denotes the product $\prod_{i=1}^{p}\prod_{j=i+1}^{p}$

$\Gamma = \mathrm{diag}\{\gamma_i\}$

$G = \mathrm{diag}\{g_i\}$

${}_1F_0(\nu;\ A,B)$ denotes a generalized hypergeometric function with matrix arguments. (See, for example, Johnson and Kotz (1972) equation (3.1.2))

and c is a constant.

Remark 5.3.2 Since $\Sigma_1 = \Sigma + nT$ we have the following relationship between the γ_i and the λ_i:

$$\{\gamma_i\} = \mathrm{eigs}\{(\Sigma+nT)\Sigma^{-1}\} = \mathrm{eigs}\{I+nT\Sigma^{-1}\} = \{1+n\lambda_i\}.$$

Therefore, estimators of the γ_i would also produce estimators of the corresponding λ_i.

As it stands, formula (5.3.4) is not very useful for obtaining estimators of the γ_i (and hence of the λ_i), but Chang (1970) shows that when $\nu_1 + \nu_2$ is large and the γ_i are distinct then the following expression for the limiting joint density of the g_i may be derived from (5.3.4):

$$f_{g_1,\ldots,g_p}(g_1,\ldots,g_p) = c \prod_{i<j}^{p}\left(\frac{g_i-g_j}{\gamma_j^{-1}-\gamma_i^{-1}}\right)^{\frac{1}{2}} \prod_{i=1}^{p} \frac{g_i^{\frac{1}{2}(\nu_1-p-1)}}{\gamma_i^{\frac{1}{2}\nu_1}(1+g_i\gamma_i^{-1})^{\frac{1}{2}(\nu_1+\nu_2-p+1)}} \tag{5.3.5}$$

where $\gamma_1 > \gamma_2 > \ldots > \gamma_p$ are the eigenvalues of $\Sigma_1\Sigma^{-1}$,

$$c = \left(\frac{2\pi}{\nu_1+\nu_2}\right)^{p(p-1)/4} \frac{\Gamma_p(\frac{1}{2}(\nu_1+\nu_2))}{\Gamma_p(\frac{1}{2}\nu_1)\Gamma_p(\frac{1}{2}\nu_2)}$$

and $\Gamma_p(\frac{1}{2}\nu) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(\frac{1}{2}(\nu-j+1))$ is the multivariate gamma function.

As a check on formula (5.3.5) we evaluate it for the case $p = 1$:

$$f_{g_1}(g_1) = \frac{\Gamma(\frac{1}{2}(\nu_1+\nu_2))}{\Gamma(\frac{1}{2}\nu_1)\Gamma(\frac{1}{2}\nu_2)} \gamma_1^{-1}\left(\frac{g_1}{\gamma_1}\right)^{\frac{1}{2}\nu_1-1} \bigg/ \left(1+\frac{g_1}{\nu_1}\right)^{\frac{1}{2}(\nu_1+\nu_2)} \qquad (5.3.6)$$

so that g_1/γ_1 has an (unnormed) f-distribution. So Chang's limiting distribution (5.3.5) is exact in the one-dimensional case, with expected value,

$$E[\frac{g_1}{\gamma_1}] = \frac{\nu_1}{\nu_2-2} .$$

Thus $\ell_1 = \frac{\nu_2}{\nu_1} g_1$ has expected value $(\frac{\nu_2}{\nu_2-2}) \gamma_1$ from which the following unbiased estimator of γ_1 results:

$$\hat{\gamma}_1 = \left(\frac{\nu_2-2}{\nu_2}\right) \ell_1. \qquad (5.3.7)$$

For higher dimensions, however, the calculation of expected values from (5.3.5) becomes intractable analytically.

Remark 5.3.3 *In a very recent paper,* Khatri and Srivastava (1978) give the following asymptotic expansion for the joint density function for $g_1 > g_2 > \ldots > g_p > 0$ when the γ_i are distinct:

$$f^+_{g_1,\ldots,g_p}(g_1,\ldots,g_p) = f_{g_1,\ldots,g_p}(g_1,\ldots,g_p)\{1+\frac{1}{2(\nu_1+\nu_2)}\left(\sum_{i<j} c_{ij}^{-1} + \frac{p(p-1)(4p+1)}{12}\right]$$

$$+ O((\nu_1+\nu_2)^{-2}\} \qquad (5.3.8)$$

where,

$f_{g_1,\ldots,g_p}(g_1,\ldots,g_p)$ is Chang's expression (5.3.5)

$$c_{ij} = (\gamma_j^{-1}-\gamma_i^{-1})(g_i-g_j)(1+g_i\gamma_i^{-1})^{-1}(1+g_j\gamma_j^{-1})^{-1}$$

and $\sum_{i<j}$ denotes the double sum $\sum_{i=1}^{p}\sum_{j=i+1}^{p}$.

For the situation where only the first q γ_i are distinct and the last (p-q) are equal they give a similar, but more complicated expression for the joint density of the g_i.

Unfortunately the abovementioned paper appeared in print after the research in this chapter had been completed, so that expression (5.3.8) was not used to obtain maximum likelihood estimators of the γ_i. However, since $\nu_1 + \nu_2 = kn-1$ and k must be greater than r(T) (which usually equals p) to ensure that $r(\hat{T}) = r(T)$, where $\hat{T}$ is given in 5.3.1, $\nu_1 + \nu_2$ will tend to be large in most practical applications. Thus the correction factor in (5.3.8) will be small in practice.

Nevertheless, it would be a relatively straightforward but lengthy matter to obtain unrestricted and restricted maximum marginal likelihood estimators of the γ_i from (5.3.8) corresponding to those obtained from (5.3.5) described in the remainder of this section and in the next. It would then be interesting to compare these two additional estimators of the γ_i with those proposed below, by repeating the simulation experiments described in Section 5.5.

5.3.1 Maximum Marginal Likelihood Estimators of $\{\gamma_i\} = \text{Eigs}(\Sigma_1\Sigma^{-1})$

James (1966), considering the eigenvalues of a simple Wishart matrix, argues that although the sample eigenvalues and eigenvectors are jointly maximum likelihood estimators of their population counterparts, the sample eigenvalues do not maximise the likelihood function of their marginal distribution. He then goes on to solve the maximum likelihood equations obtained from the limiting marginal distribution of the sample eigenvalues to give estimators (to $O(\nu^{-2})$) of the population eigenvalues. *It is interesting to note that* Lawley (1956) *obtains the identical* estimators using a quite different approach. We now apply the same approach as James (1966) to Chang's formula (5.3.5) for the limiting density of $\{g_i\} = \text{eigs}(A_1A_2^{-1})$:

Starting with the log likelihood of the γ_i,

$$L = L(\underset{\sim}{\gamma}|\underset{\sim}{g}) = \log c + \tfrac{1}{2}(\nu_1-p-1)\sum_{i=1}^{p} \log g_i - \tfrac{1}{2}\nu_1 \sum_{i=1}^{p} \log \gamma_i$$

$$- \tfrac{1}{2}(\nu_1+\nu_2-p+1)\sum_{i=1}^{p} \log(1+\frac{g_i}{\gamma_i}) + \tfrac{1}{2}\sum\sum_{i<j} \log (g_i-g_j)$$

$$-\tfrac{1}{2}\sum\sum_{i<j} \log(\gamma_j^{-1}-\gamma_i^{-1}) \qquad (5.3.9)$$

differentiating with respect to γ_i and simplifying yields:

$$\frac{\partial L}{\partial \gamma_i} = \frac{1}{2\gamma_i}(-\nu_1 + (\nu_1+\nu_2-p+1)\frac{g_i}{\gamma_i+g_i} + \sum_{j\neq i} \frac{\gamma_j}{\gamma_j-\gamma_i}) \qquad i = 1,\ldots,p$$

$$(5.3.10)$$

where $\sum\limits_{j\neq i}$ *denotes the single sum from* $j = 1$ to p excluding the term where $j = i$.

Equating this to zero gives:

$$(\nu_1+\nu_2-p+1)\frac{g_i}{\gamma_i+g_i} + \sum_{j\neq i} \frac{\gamma_j}{\gamma_j-\gamma_i} = \nu_1 \qquad i = 1,\ldots,p. \qquad (5.3.11)$$

Before attempting to solve equations (5.3.11) for the γ_i, let us first check whether they do, in fact, give a maximum for the log likelihood (5.3.9). Taking second derivatives of L:

$$\frac{\partial^2 L}{\partial \gamma_i^2} = \frac{1}{2\gamma_i}\left(-(\nu_1+\nu_2-p+1)\frac{g_i}{(\gamma_i+g_i)^2} + \sum_{j\neq i}\frac{\gamma_j}{(\gamma_j-\gamma_i)^2}\right)$$

$$-\frac{1}{2\gamma_i^2}\left(-\nu_1 + (\nu_1+\nu_2-p+1)\frac{g_i}{\gamma_i+g_i} + \sum_{j\neq i}\frac{\gamma_j}{(\gamma_j-\gamma_i)}\right)$$

$$= \frac{1}{2\gamma_i}\left(-(\nu_1+\nu_2-p+1)\frac{g_i}{(\gamma_i+g_i)^2} + \sum_{j\neq i}\frac{\gamma_j}{(\gamma_j-\gamma_i)^2}\right) \qquad (5.3.12)$$

at the stationary point given by (5.3.11). *Clearly* $\frac{\partial^2 L}{\partial\gamma_i^2} < 0$ *for* $\nu_1 + \nu_2$ sufficiently large.

Similarly,

$$\frac{\partial^2 L}{\partial\gamma_i \partial\gamma_j} = -\frac{1}{2(\gamma_j-\gamma_i)^2} < 0. \qquad (5.3.13)$$

Using the criterion (see, for example Brand (1960) page 188)

$$H_{i,j} = \frac{\partial^2 L}{\partial\gamma_i^2}\,\frac{\partial^2 L}{\partial\gamma_j^2} - \left(\frac{\partial^2 L}{\partial\gamma_i\partial\gamma_j}\right)^2 \qquad (5.3.14)$$

we see that, for $\nu_1 + \nu_2$ sufficiently large $H_{i,j} > 0$, $\forall i,j$, at the stationary point, implying that (5.3.11) gives a maximum.

Going back to equations (5.3.11) it is obviously no straightforward matter to solve these in terms of the γ_i, $i = 1,\ldots p$. However, solving them *in terms of the* g_i (which gives the modal value of their distribution) yields:

$$g_i = \gamma_i \left(\frac{\nu_1 - \sum\limits_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i}}{\nu_2 - p + 1 + \sum\limits_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i}} \right) \quad i = 1, \ldots, p. \qquad (5.3.15)$$

At this stage, it is convenient to return to the

$$\{\ell_i\} = \mathrm{eigs}(S_1 S_2^{-1}) = \mathrm{eigs}(\frac{\nu_2}{\nu_1} A_1 A_2^{-1}) = \{\frac{\nu_2}{\nu_1} g_i\}.$$

The modal values of the ℓ_i are, from (5.3.15):

$$\ell_i = \gamma_i \left(\frac{1 - \frac{1}{\nu_1} \sum\limits_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i}}{1 - \frac{p-1}{\nu_2} + \frac{1}{\nu_2} \sum\limits_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i}} \right) \quad i = 1, \ldots, p. \qquad (5.3.16)$$

As a first check of the correctness of formula (5.3.16), note that, modal $\ell_i \to \gamma_i$ as ν_1 and ν_2 get large.

Further checks on (5.3.16) can be made by noting that, as $\nu_2 \to \infty$ the ℓ_i become the eigenvalues of the single (normed) Wishart Matrix $S_1 \Sigma^{-1}$, where $\nu_1 S_1 \Sigma^{-1} \sim W(\Sigma_1 \Sigma^{-1}, \nu_1)$. Formula (5.3.16) then reduces to:

$$\ell_i = \gamma_i (1 - \frac{1}{\nu_1} \sum_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i}) \qquad (5.3.17)$$

which is equivalent to James' (1966) equation (8.1) for the limiting maximum marginal likelihood estimators of the population eigenvalues of a Wishart matrix (he uses the notation $\alpha_i = \gamma_i^{-1}$). Formula (5.3.17) is also equivalent (to $O(\nu_1^{-2})$) to Lawley's (1956) expression for $E[\ell_i]$ obtained by using a perturbation argument.

5.3.2 Approximate solution of the Maximum Likelihood Equations

To obtain the maximum likelihood estimators of the γ_i from (5.3.11), note that from (5.3.16) we have:

$$\gamma_i = \ell_i \left[\frac{1 - \frac{p-1}{\nu_2} + \frac{1}{\nu_2} \sum\limits_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i}}{1 - \frac{1}{\nu_1} \sum\limits_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i}} \right] \tag{5.3.18}$$

and, for ν_2 large this becomes:

$$\begin{aligned} \gamma_i &\doteq \ell_i \left(1 - \frac{1}{\nu_1} \sum_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i} \right)^{-1} \\ &= \ell_i \left(1 + \frac{1}{\nu_1} \sum_{j \neq i} \frac{\gamma_j}{\gamma_j - \gamma_i} + O(\nu_1^{-2}) \right) \\ &= \ell_i + O(\nu_1^{-1}). \end{aligned} \tag{5.3.19}$$

"Plugging" (5.3.19) into the right hand side of equation (5.3.18) yields the following approximate formula for the asymptotic maximum marginal likelihood estimators for the γ_i:

$$\hat{\gamma}_i = \ell_i \left[\frac{1 - \frac{p-1}{\nu_2} + \frac{1}{\nu_2} \sum\limits_{j \neq i} \frac{\ell_j}{\ell_j - \ell_i}}{1 - \frac{1}{\nu_1} \sum\limits_{j \neq i} \frac{\ell_j}{\ell_j - \ell_i}} \right] + O(\nu_1^{-2}). \tag{5.3.20}$$

It may also be noted in passing that the method of successive approximations (see, for example McCracken and Dorn (1964)) for solving (5.3.18), considered as the system of equations,

$$\underset{\sim}{\gamma} = \underset{\sim}{f}(\underset{\sim}{\gamma}),$$

yields (5.3.20) in its first step if the initial values $\gamma_i = \ell_i$ are used.

As a check on formula (5.3.20), note again that, as $\nu_2 \to \infty$ we get

$$\hat{\gamma}_i = \ell_i\left(1 - \frac{1}{\nu_1}\sum_{j\neq i}\frac{\ell_j}{\ell_j-\ell_i}\right)^{-1} + O(\nu_1^{-2})$$

$$= \ell_i\left(1 + \frac{1}{\nu_1}\sum_{j\neq i}\frac{\ell_j}{\ell_j-\ell_i}\right) + O(\nu_1^{-2}) \qquad (5.3.21)$$

which is the same as formula (8.2) of James (1966) for the maximum marginal likelihood estimator, as well as Lawley's (1956) formula for the estimator with bias of order ν_1^{-2}, of the i^{th} population eigenvalue of a single Wishart matrix.

5.3.3 Numerical Solution of the Maximum Likelihood Equations

Since there is no exact analytic solution to the maximum likelihood equations (5.3.11), we now consider their numerical solution.

From expression (5.3.9) it is evident that the limiting log likelihood function of $\{\gamma_i;\ i = 1,\ldots,p\}$ tends to infinity whenever any two of the γ_i's are equal. However, since Chang's formula (5.3.5) is valid only for distinct population eigenvalues, these singularities in the log likelihood occur at inadmissible values of the γ_i. Nevertheless these "inadmissible singularities" could cause considerable difficulties when trying to solve the maximum likelihood equations (5.3.11) numerically.

To get around this problem, we consider the following reparameterisation of the problem:

Let

$$\frac{1}{\gamma_1} = e^{\delta_1} + c_1$$

and

$$\frac{1}{\gamma_i} - \frac{1}{\gamma_{i-1}} = e^{\delta_i} + \varepsilon_i \qquad i = 2,\ldots,p \qquad (5.3.22)$$

where the ε_i, $i = 1,\ldots,p$ are preassigned small positive quantities. The reasons for choosing this reparameterisation is as follows:

(a) it ensures that $\gamma_1 > \gamma_2 > \ldots > \gamma_p > 0$,

(b) the new parameters $\{\delta_i ;\ i = 1,\ldots,p\}$ are unconstrained in value , and

(c) the γ_i appear only in the forms $\frac{1}{\gamma_i}$ and $\frac{1}{\gamma_j} - \frac{1}{\gamma_i}$, $j > i$, in the density function (5.3.5) of the g_i (considered as a likelihood function) and both these forms can be expressed simply in terms of the new parameters.

Viz:

$$\frac{1}{\gamma_i} = \sum_{k=1}^{i} (e^{\delta_k} + \varepsilon_k)$$

and

$$\frac{1}{\gamma_j} - \frac{1}{\gamma_i} = \sum_{k=i+1}^{j} (e^{\delta_k} + \varepsilon_k) \qquad i,j = 1,\ldots,p; \quad j > i. \qquad (5.3.23)$$

A drawback to this reparameterization is that it entails preassigning values for the ε_i. In practice this presents no difficulty; a practical rule is to let ε_i be some small fraction of $\frac{1}{\hat{\gamma}_i^o} - \frac{1}{\hat{\gamma}_{i-1}^o}$ for $i = 2,\ldots,p$ and of $\frac{1}{\hat{\gamma}_1^o}$ for $i = 1$, where the $\hat{\gamma}_i^o$ are initial estimators of the γ_i.

In terms of the new parameters the log likelihood becomes:

$$L = L(\underset{\sim}{\delta}|g,\varepsilon) = \log c + \tfrac{1}{2}(\nu_1-p-1)\sum_{j=1}^{p} \log g_j \quad .$$

$$+ \tfrac{1}{2}\nu_1 \sum_{j=1}^{p} \log(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}) - \tfrac{1}{2}(\nu_1+\nu_2-p+1)\sum_{j=1}^{p} \log(1+g_j(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}))$$

$$+ \tfrac{1}{2} \sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \log(g_i-g_j) - \tfrac{1}{2}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \log(\sum_{j=i+1}^{j} e^{\delta_k+\varepsilon_k})$$

$$= f(\underset{\sim}{g}) + \tfrac{1}{2}\nu_1\sum_{j=1}^{p} \log(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}) - \tfrac{1}{2}(\nu_1+\nu_2-p+1)\sum_{j=1}^{p}\log(1+g_j(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}))$$

$$- \tfrac{1}{2}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \log(\sum_{k=i+1}^{j} e^{\delta_k+\varepsilon_k}) \qquad (5.3.24)$$

where $f(\underset{\sim}{g})$ is a function of the g_i only. Differentiating L with respect to the δ's and simplifying yields the new maximum likelihood equations:

$$\frac{\partial L}{\partial \delta_\ell} = \tfrac{1}{2}e^{\delta_\ell}\{\nu_1\sum_{j=\ell}^{p}(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k})^{-1} - (\nu_1+\nu_2-p+1)\sum_{j=\ell}^{p} g_j(1+g_j(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}))^{-1}$$

$$-\sum_{i=1}^{\ell-1}\sum_{j=\ell}^{p}(\sum_{k=i+1}^{j} e^{\delta_k+\varepsilon_k})^{-1}\} = 0 \qquad \ell = 2,\ldots,p$$

and

$$\frac{\partial L}{\partial \delta_1} = \tfrac{1}{2}e^{\delta_1}\{\nu_1\sum_{j=1}^{p}(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k})^{-1} - (\nu_1+\nu_2-p+1)\sum_{j=1}^{p} g_j(1+g_j(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}))^{-1}\}$$

$$= 0. \qquad (5.3.25)$$

A standard numerical technique for solving the maximum likelihood equations for the maximum likelihood estimator $\hat{\underset{\sim}{\delta}} = (\hat{\delta}_1,\ldots,\hat{\delta}_p)'$ is the Newton-Raphson iterative procedure. Defining the (p×1) vector of first derivatives $D_{\delta}(L(\underset{\sim}{\delta}))$, whose ℓ^{th} element is $\frac{\partial L}{\partial \delta_\ell}$ and the (p×p) Hessian matrix $D^2_{\underset{\sim}{\delta}}(L(\underset{\sim}{\delta}))$, whose $(\ell,m)^{th}$ element is $\frac{\partial^2 L}{\partial \delta_\ell \partial \delta_m}$, the Newton-Raphson iterative method can be written (See, for example, Silvey (1975) or Cox and Hinkley (1974)),

$$\hat{\underset{\sim}{\delta}}^{(r+1)} = \hat{\underset{\sim}{\delta}}^{(r)} - (D^2_{\underset{\sim}{\delta}}(L(\hat{\underset{\sim}{\delta}}^{(r)})))^{-1} D_{\underset{\sim}{\delta}}(L(\hat{\underset{\sim}{\delta}}^{(r)})) \qquad (5.3.26)$$

Given an initial approximation $\hat{\underset{\sim}{\delta}}^{(0)}$ to δ, successive approximations $\hat{\underset{\sim}{\delta}}^{(1)}, \hat{\underset{\sim}{\delta}}^{(2)},\ldots$ are obtained from (5.3.26) which hopefully converge to $\hat{\underset{\sim}{\delta}}$.

As an initial approximation we may let $\hat{\underset{\sim}{\gamma}}^{(0)} = \underset{\sim}{\ell} = (\ell_1,\ldots,\ell_p)'$ and then obtain $\hat{\underset{\sim}{\delta}}^{(0)}$ from (5.3.22).

viz:

$$\hat{\delta}_i^{(0)} = \begin{cases} \text{Log}_e\left(\frac{1}{\hat{\gamma}_i^{(0)}} - \frac{1}{\hat{\gamma}_{i-1}^{(0)}} - \varepsilon_i\right) & i = 2,\ldots,p \\ \text{Log}_e\left(\frac{1}{\hat{\gamma}_1^{(0)}} - \varepsilon_1\right) & i = 1 . \end{cases} \qquad (5.3.27)$$

Another, possibly better, initial approximation may be obtained by using the approximate maximum likelihood formula (5.3.20) for $\hat{\gamma}^{(0)}$.

Differentiating (5.3.25) with respect to δ_m yields the elements of the Hessian matrix:

$$\frac{\partial^2 L}{\partial \delta_\ell \partial \delta_m} = \tfrac{1}{2} e^{\delta_\ell + \delta_m} \{- \nu_1 \sum_{j=Low}^{p} (\sum_{k=1}^{j} e^{\delta_k + \varepsilon_k})^{-2}$$

$$+ (\nu_1+\nu_2-p+1) \sum_{j=Low}^{p} g_j^2 (1+g_j(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}))^{-2} + \sum_{i=1}^{Top} \sum_{j=Low}^{p} (\sum_{k=i+1}^{j} e^{\delta_k+\varepsilon_k})^{-2}\}$$

$$\ell ,m = 2,\ldots,p; \quad \ell \neq m. \qquad (5.3.28)$$

where Top = min(ℓ,m) - 1, Low = max(ℓ,m). For $\ell = 1$ or $m = 1$ the last term in (5.3.28) is dropped.

$$\frac{\partial^2 L}{\partial \delta_\ell^2} = \left.\frac{\partial^2 L}{\partial \delta_\ell \partial \delta_m}\right|_{m=\ell} + \frac{\partial L}{\partial \delta_\ell} \qquad \ell = 2,\ldots,p . \qquad (5.3.29)$$

For $\ell = 1$, *drop the* last term in $\left.\frac{\partial^2 L}{\partial \delta_\ell \partial \delta_m}\right|_{m=\ell}$ in (5.3.29).

Finally, as the transformation from $\underset{\sim}{\delta}$ to $\underset{\sim}{\gamma}$ is one-to-one the *maximum* likelihood estimator $\hat{\underset{\sim}{\gamma}}$ of $\underset{\sim}{\gamma}$ *may be obtained from* $\hat{\underset{\sim}{\delta}}$ by merely transforming *back via* (5.3.22).

5.3.4 Large Sample Distribution of the Maximum Marginal Likelihood Estimators $\{\hat{\gamma}_i\}$

It is well known (see, for example Silvey (1975), or Cox and Hinkley (1974)) that under certain regularity conditions that are usually satisfied in practice (and are satisfied here) the *maximum likelihood estimators* $\hat{\underset{\sim}{\gamma}} = (\hat{\gamma}_1,\ldots,\hat{\gamma}_p)'$ *are asymptotically efficient* and approximately normally distributed with mean vector $\underset{\sim}{\gamma} = (\gamma_1,\ldots,\gamma_p)'$ and covariance matrix $B_{\underset{\sim}{\gamma}}^{-1}$, where $B_{\underset{\sim}{\gamma}}$ is Fisher's Information matrix given by:

$$B_{\underset{\sim}{\gamma}} = (b_{ij}) = - E_{\underset{\sim}{g}}[D_{\underset{\sim}{\gamma}}^2(L(\underset{\sim}{\gamma}|\underset{\sim}{g}))] \qquad (5.3.30)$$

$L = L(\underset{\sim}{\gamma}|\underset{\sim}{g})$ is the log likelihood of $\underset{\sim}{\gamma}$ given in (5.3.9) and $D_{\underset{\sim}{\gamma}}^2(L(\underset{\sim}{\gamma}|\underset{\sim}{g}))$ is the Hessian matrix whose $(i,j)^{th}$ element is $\frac{\partial^2 L}{\partial \gamma_i \partial \gamma_j}$. The expectation in (5.3.30) is taken over the distribution of $\underset{\sim}{g} = (g_1,\ldots,g_p)'$.

Differentiating $\frac{\partial L}{\partial \gamma_i}$ given in (5.3.10) with respect to γ_j yields:

$$\frac{\partial^2 L}{\partial \gamma_j \partial \gamma_j} = - \frac{1}{2(\gamma_j - \gamma_i)^2} \quad , \quad \forall j \neq i \tag{5.3.31}$$

and with respect to γ_i:

$$\frac{\partial^2 L}{\partial \gamma_i^2} = \frac{1}{2\gamma_i}(-(\nu_1+\nu_2-p+1)\frac{g_i}{(\gamma_i+g_i)^2} + \sum_{j\neq i} \frac{\gamma_j}{(\gamma_j-\gamma_i)^2}$$

$$- \frac{1}{2\gamma_i^2}\left(-\nu_1 + (\nu_1+\nu_2-p+1)\frac{g_i}{\gamma_i+g_i} + \sum_{j\neq i}\frac{\gamma_j}{\gamma_j-\gamma_i}\right)$$

$$= \frac{1}{2\gamma_i^2}\{\nu_1-p+1+(\nu_1+\nu_2-p+1)(\left(\frac{\gamma_i}{\gamma_i+g_i}\right)^2 - 1) + \gamma_i^2 \sum_{j\neq i}\frac{1}{(\gamma_j-\gamma_i)^2}\} \tag{5.3.32}$$

The off-diagonal elements *of $D^2_{\underset{\sim}{\gamma}}(L)$ given* in (5.3.31) do not depend on $\underset{\sim}{g}$, so we have immediately, from (5.3.30)

$$b_{ij} = \frac{1}{2(\gamma_j-\gamma_i)^2} \quad , \quad i \neq j . \tag{5.3.33}$$

The diagonal elements b_{ii} are given by:

$$b_{ii} = - \frac{1}{2\gamma_i^2}\Big(\nu_1-p+1+(\nu_1+\nu_2-p+1)(E_{\underset{\sim}{g}}[\left(\frac{\gamma_i}{\gamma_i+g_i}\right)^2] - 1)$$

$$+ \gamma_i^2 \sum_{j\neq i}\frac{1}{(\gamma_j-\gamma_i)^2}\Big\}. \tag{5.3.34}$$

Now

$$E_{\underset{\sim}{g}}[\left(\frac{\gamma_i}{\gamma_i+g_i}\right)^2] = E_{\underset{\sim}{g}}[(1+\frac{g_i}{\gamma_i})^{-2}] . \tag{5.3.35}$$

As noted earlier, the evaluation of the expected values of the g_i using Chang's asymptotic expression (5.3.6) for their joint density is intractable analytically for $p > 1$, and so, *a fortiori*, is that *of* $(1 + \frac{g_i}{\gamma_i})^{-2}$.

[illegible], if we make the transformation:

$$u_i = \frac{g_i}{\gamma_i}\ ,\ i = 1,\dots,p$$

in (5.3.5), we get the limiting joint density of the u_i as:

$$f_{\underset{\sim}{u}}(u_1,\dots,u_p) = K \prod_{i<j}^{p} \left(\frac{\gamma_i u_i - \gamma_j u_j}{\gamma_j^{-1} - \gamma_i^{-1}}\right)^{\frac{1}{2}} \prod_{i=1}^{p} \frac{u_i^{\frac{1}{2}(\nu_1-p+1)-1}}{(1+u_i)^{(\nu_2+\nu_1-p+1)/2}} \tag{5.3.36}$$

where $K = c \prod_{i=1}^{p} \gamma_i^{-(\frac{p-1}{2})}$ and c is defined in (5.3.5).

. Anderson (1965) *has shown that if* ℓ_i, $i = 1,\dots,p$ *are the eigenvalues* of a single (normed) Wishart matrix, and γ_i, $i = 1,\dots,p$ are their *corresponding population values, then the "linkage factor"*

$$\prod_{i<j} \left(\frac{\ell_i - \ell_j}{\gamma_j^{-1} - \gamma_i^{-1}}\right)^{\frac{1}{2}}$$

tends to 1 with probability 1 as the sample size $n \to \infty$.

Now, in our case, the "linkage factor" is:

$$\prod_{i<j} \left(\frac{\gamma_i u_i - \gamma_j u_j}{\gamma_j^{-1} - \gamma_i^{-1}}\right)^{\frac{1}{2}} = \left(\frac{\nu_1}{\nu_2}\right)^{\frac{p(p-1)}{4}} \prod_{i<j} \left(\frac{\ell_i - \ell_j}{\gamma_j^{-1} - \gamma_i^{-1}}\right)^{\frac{1}{2}}$$

where the $\{\ell_i\} = \mathrm{eigs}(S_1 S_2^{-1})$. By the same argument used earlier, as $\nu_2 \to \infty$, the ℓ_i become eigenvalues of a single (normed) Wishart matrix, and so by Anderson's result our "linkage factor" tends to 1 with probability 1 as ν_1 and $\nu_2 \to \infty$.

Using the above result in (5.3.36) it is clear that, for large ν_1 and ν_2, the u_i are approximately independently distributed as (unnormed) f-random variables on $(\nu_1 - p+1)$ and ν_2 degrees of freedom. Hence, transforming to beta random variables:

$$x_i = \frac{u_i}{1+u_i} \qquad i = 1,\ldots,p$$

we have:

$$E\left[\left(1 + \frac{g_i}{\gamma_i}\right)^{-2}\right] = E[(1+u_i)^{-2}] = E[(1-x_i)^2]$$

$$= 1 - 2E[x_i] + E[x_i^2]$$

where, for large ν_1 and ν_2, x_i has, approximately, a beta *distribution* with parameters $n_1 = \frac{1}{2}(\nu_1-p+1)$ and $n_2 = \frac{1}{2}\nu_2$. So

$$E\left[\left(1 + \frac{g_i}{\gamma_i}\right)^{-2}\right] \doteq 1 - 2\left(\frac{n_1}{n_1+n_2}\right) + \frac{n_1(n_1+1)}{(n_1+n_2)(n_1+n_2+1)}$$

$$= 1 - \frac{n_1(n_1+2n_2+1)}{(n_1+n_2)(n_1+n_2+1)}$$

$$= 1 - \frac{(\nu_1-p+1)(\nu_1+2\nu_2-p+3)}{(\nu_1+\nu_2-p+1)(\text{[illegible]}+3)} \; . \qquad (5.3.37)$$

Substituting this result back into (5.[illegible] and (5.3.34) gives:

$$b_{ii} \doteq \frac{1}{2\gamma_i^2}\left\{\frac{(\nu_1-p+1)(\nu_1+2\nu_2-p+3)}{(\nu_1+\nu_2-p+3)} - \nu_1 + p - 1 - \gamma_i^2 \sum_{j\neq i} \frac{1}{(\gamma_j-\gamma_i)^2}\right\}$$

$$i = 1,\ldots,p. \qquad (5.3.38)$$

Finally, substituting (5.3.38) and (5.3.33) into (5.3.30) gives the approximate large sample distribution of the maximum marginal likelihood estimator $\hat{\underset{\sim}{\gamma}}$ of $\underset{\sim}{\gamma}$.

Example 5.3.1 To test how good this approximation is, the approximate means, standard deviations and correlation coefficients of the $\hat{\gamma}_i$ were calculated from the above formulae for the case $p = 3$, using the two sets of eigenvalues and three of the sample sizes, each represented by a pair of values for ν_1 and ν_2, that were used in the simulation experiments described in Section 5.5. In the first set the eigenvalues are equally spaced whereas in the second the spacing between γ_1 and γ_2 is much larger than that between γ_2 and γ_3. The three sample sizes represent, roughly, "medium sized", "large" and "very large" samples, respectively. The results are given in Table 5.3.1 below, together with the corresponding values obtained from the simulation experiments. (Because of the frequent failure, especially in the smaller sample sizes, of the maximum likelihood estimator described in Sub-section 5.3.3 to produce meaningful results, the results from the simulations on the approximate maximum likelihood estimators given in expression (5.3.20) are used. Admittedly (5.3.20) sometimes also produces meaningless results, but its alternative, the "hybrid" estimator described in Section 5.5 that always gives meaningful results, is not a maximum likelihood estimator. See Section 5.5 for a full discussion of these points.)

Table 5.3.1

Approximate Means, Standard Deviations and Correlation Coefficients of the Maximum Likelihood Estimators of the $\{\gamma_i\}$ for $p = 3$ dimensions

Notation: (i) Denotes the values obtained from the formulae

(ii) Denotes the values obtained from the simulation experiments.

A. Degrees of Freedom $\nu_1 = 15$, $\nu_2 = 64$

Means		Standard Deviations		Pair.	Correlation Coefficients	
(i)	(ii)	(i)	(ii)	(i,j)	(i)	(iii)
6	6.70	-	2.94	(1,2)	-	-.082
4	4.12	-	2.44	(1,3)	-	-.098
2	1.85	-	0.86	(2,3)	-	-.049
16	16.71	8.31	8.13	(1,2)	-.061	-.122
4	4.30	2.31	2.21	(1,3)	-.003	-.135
2	1.87	0.96	0.83	(2,3)	-.255	.006

B. Degrees of Freedom $\nu_1 = 30$, $\nu_2 = 124$

Means		Standard Deviations		Pair.	Correlation Coefficients	
(i)	(ii)	(i)	(ii)	(i,j)	(i)	(ii)
6	6.09	2.86	1.71	(1,2)	-.484	-.067
4	4.31	1.69	2.13	(1,3)	.005	.001
2	1.90	0.62	0.68	(2,3)	.107	-.051
16	15.56	5.13	4.63	(1,2)	-.023	.045
4	4.33	1.32	1.36	(1,3)	-.006	.009
2	1.94	0.61	0.81	(2,3)	-.100	-.168

C. Degrees of Freedom $\nu_1 = 60$, $\nu_2 = 244$

Mean		Standard Deviations		Pair	Correlation Coefficients	
(i)	(ii)	(i)	(ii)	(i,j)	(i)	(ii)
5	6.10	1.45	1.42	(1,2)	-.162	-.154
4	4.33	0.92	1.05	(1,3)	-.011	-.077
2	2.00	0.42	0.45	(2,3)	.045	.013
16	16.26	3.43	3.63	(1,2)	-.010	.018
4	4.26	0.87	0.89	(1,3)	-.003	-.068
2	2.01	0.42	0.46	(2,3)	-.046	-.061

The missing values in part A of Table 5.3.1 indicate that formulae (5.3.30), (5.3.33) and (5.3.38) broke down in that they produced negative variances. (This also occurred in both cases when the formulae were applied to the "small" sample size with $\nu_1 = 6$ and $\nu_2 = 28$.)

Looking at means and standard deviations alone, the agreement between the approximate and simulation results in the case where the spacings between the γ_i increase with their values is excellent, even for the "medium sized" samples. In the case where the spacings are equal, the agreement between the standard deviations is not quite so good for the "large" samples but is again excellent for the "very large" samples.

Looking at the correlation coefficients, the picture is not so rosy, although there is reasonable agreement for the "very large" samples. This, however, could as much be a result of the occasional breakdown in the simulation experiments of the approximate formula (5.3.20) for the maximum likelihood estimators, as of the poor performance of the approximate formula for their covariance matrix. It is well known that even a small fraction of outliers where the orderings of the variables are permuted, can have a drastic effect on the sample correlation coefficient. This fact is evidenced by the very large differences between the correlation coefficient in Table 5.3.1 and the corresponding coefficients in Table 5.5.5 where only "well-behaved" estimates have been included in the sample.

In summary, the formulae for the approximate mean vector and covariance matrix of the maximum marginal likelihood estimators $\{\hat{\gamma}_i\}$ derived in this sub-section would appear to be fairly good for large samples (as defined here and in Section 5.5) and gets better (and becomes applicable to smaller samples) as the differences between adjacent eigenvalues increase.

5.4 Additional Information on $\{\gamma_i\} = \text{Eigs}(\Sigma_1\Sigma^{-1})$

The *maximum likelihood estimators of the* γ_i, $i = 1,\ldots,p$ *obtained* in Section 5.3 are based on Chang's expression (5.3.5) for the limiting density of $\{g_i\} = \text{Eigs}(A_1A_2^{-1})$, where,

$$\begin{aligned} A_1 &\sim W_p(\nu_1,\Sigma_1) \\ A_2 &\sim W_p(\nu_2,\Sigma\ ) \qquad \text{independently.} \end{aligned}$$

In this section some exact results on the expected values of functions of the g_i are derived. Those will then be used to obtain moment estimators for the means and variances of the four quantities: δ^2_{ij}, $\delta^2_i(x)$, d^2_{ij} and $d^2_i(x)$ whose distributions under the random effects model are discussed in Chapter 3, as well as for the approximate probabilities of misclassification derived in Chapter 4. In addition, some of these exact results will be used to improve the estimators of the γ_i obtained in Section 5.3.

Specifically, in Sub-section 5.4.1, well-known results on the moments of the generalised variance from a multivariate normal distribution will *be used to obtain an exact moment estimator of* $\prod_{i=1}^{p} \gamma_i$. In Sub-sections

5.4.2 and 5.4.3 new results on the *distribution* of $\mathrm{Tr}(A_1A_2^{-1})$ lead to exact expressions for the mean and variance *of* $\sum_{i=1}^{p} g_i$ in terms of $\sum_{i=1}^{p} \gamma_i$ and $\sum_{i=1}^{p} \gamma_i^2$. These results are used in Sub-section 5.4.4 to obtain moment estimators for the means and variances of the four quantities and for the approximate probabilities of misclassification mentioned above. Finally, the combination of the various pieces of information to obtain improved estimators of the γ_i, either exactly or by means of the technique of resticted maximum likelihood estimation, is discussed in Sub-sections 5.4.5 and 5.4.6.

5.4.1 Moments of the Generalised Variance

The h^{th} moment of $|A|$, where $A \sim W_p(\nu,\Sigma)$, for h an integer greater than $-\frac{1}{2}(\nu-p+1)$, is given by:

$$\mu'_h(|A|) = |\Sigma|^h 2^{hp} \frac{\Gamma_p(\frac{1}{2}\nu+h)}{\Gamma_p(\frac{1}{2}\nu)} \qquad (5.4.1)$$

where $\Gamma_p(\frac{1}{2}\nu)$ is the *multivariate* gamma function defined in (5.3.5) (See, for example Johnson and Kotz (1972)). Therefore, since A_1 and A_2 are independent Wishart matrices,

$$\mu'_h(|A_1A_2^{-1}|) = \mu'_h(|A_1|)\mu'_h(|A_2|^{-1}) = \mu'_h(|A_1|)\mu'_{-h}(|A_2|)$$

$$= |\Sigma_1|^h 2^{hp} \frac{\Gamma_p(\frac{1}{2}\nu_1+h)}{\Gamma_p(\frac{1}{2}\nu_1)} |\Sigma|^{-h} 2^{-hp} \frac{\Gamma_p(\frac{1}{2}\nu_2-h)}{\Gamma_p(\frac{1}{2}\nu_2)}$$

for $\frac{1}{2}(\nu_2-p+1) > h > -\frac{1}{2}(\nu_1-p+1)$

$$= |\Sigma_1\Sigma^{-1}|^h \frac{\Gamma_p(\frac{1}{2}\nu_1+h)\Gamma_p(\frac{1}{2}\nu_2-h)}{\Gamma_p(\frac{1}{2}\nu_1)\Gamma_p(\frac{1}{2}\nu_2)} . \qquad (5.4.2)$$

Noting that $|A_1A_2^{-1}| = \prod_{i=1}^{p} g_i$ and $|\Sigma_1\Sigma^{-1}|^h = \prod_{i=1}^{p} \gamma_i^h$, and considering the case h = 1, we obtain:

$$E[\prod_{i=1}^{p} g_i] = \prod_{i=1}^{p} \gamma_i \frac{\Gamma_p(\frac{1}{2}\nu_1+1)\Gamma_p(\frac{1}{2}\nu_2-1)}{\Gamma_p(\frac{1}{2}\nu_1)\Gamma_p(\frac{1}{2}\nu_2)}$$

and, using the definition of $\Gamma_p(\frac{1}{2}\nu)$ given in (5.3.5) this reduces to:

$$E[\prod_{i=1}^{p} g_i] = \prod_{i=1}^{p} \gamma_i\left(\frac{\nu_1-i+1}{\nu_2-i-1}\right). \tag{5.4.3}$$

From (5.4.3) we immediately obtain the following moment estimator of $\prod_{i=1}^{p} \gamma_i$:

$$\widehat{\prod_{i=1}^{p} \gamma_i} = \prod_{i=1}^{p} g_i\left(\frac{\nu_2-i-1}{\nu_1-i+1}\right). \tag{5.4.4}$$

In terms of the $\{\ell_i\} = \text{Eigs}(S_1S_2^{-1})$ *this becomes:*

$$\widehat{\prod_{i=1}^{p} \gamma_i} = \prod_{i=1}^{p} (\frac{\nu_1}{\nu_2}\,\ell_i)\left(\frac{\nu_2-i-1}{\nu_1-i+1}\right)$$

$$= \left(\frac{\nu_1}{\nu_2}\right)^p \prod_{i=1}^{p} \ell_i\left(\frac{\nu_2-i-1}{\nu_1-i+1}\right) . \tag{5.4.5}$$

In a similar manner, exact moment estimators of $\prod_{i=1}^{p} \gamma_i^h$ may be obtained from (5.4.2), for $h = 2,3,\ldots,\frac{1}{2}(\nu_2-p+1) - 1$.

In Sub-section 5.4.5 the exact moment estimator (5.4.4) of $\prod_{i=1}^{p} \gamma_i$ will be used as a constraint on the values of the estimators of the γ_i, in order to obtain what will hopefully be improved estimators, through the method of restricted maximum likelihood estimation. A second constraint on the $\hat{\gamma}_i$, based on the expectation of $\text{Tr}(A_1A_2^{-1})$ derived in the next two sub-sections, will also be used in the restricted maximum likelihood estimation of the γ_i in Sub-section 5.4.5.

5.4.2 *On the Distribution of* $Tr(A_1A_2^{-1})$

In this sub-section the distribution of $Tr(A_1A_2^{-1}) = \sum_{i=1}^{p} g_i$ is investigated, and an expression for it as a sum of weighted, correlated f-random variables is derived. This will be used in Sub-section 5.4.3 to derive the expectation and variance of $Tr(A_1A_2^{-1})$ which will, in turn, be used to obtain estimators for the means and variances of the four quantities δ_{ij}^2, $\delta_i^2(x)$, d_{ij}^2 and $d_i^2(x)$ whose distributions are discussed in Chaper 3, as well as for the approximate probabilities of misclassification derived in Chapter 4. As mentioned earlier, the expectation of $Tr(A_1A_2^{-1})$ will also be used in Sub-section 5.4.5 as a constraint in the restricted maximum likelihood estimation of the γ_i.

To recap,

$$A_1 \sim W_p(\nu_1,\Sigma_1) \;,$$

$$A_2 \sim W_p(\nu_2,\Sigma) \text{ independently,}$$

$$\{g_i\} = \text{Eigs}(A_1A_2^{-1})$$

and

$$\{\gamma_i\} = \text{Eigs}(\Sigma_1\Sigma^{-1}).$$

Remark 5.4.1 Clearly (see expression (5.2.4)) $Tr(A_1A_2^{-1})$ is a multiple of Hotelling's T_0^2 statistic. For the central ($\Sigma_1 = \Sigma$) and noncentral cases ($A_1 \sim W_p(\nu_1,\Sigma,\Omega)$) a considerable amount of work has been done on the distribution of T_0^2. See, for example, Johnson and Kotz (1972) and Fujikoshi (1977). However, we have not been able to find any publications on the distribution of T_0^2 under the situation of interest here, where A_1

and A_2 both have central Wishart distributions but with different parameter matrices Σ_1 and Σ.

Now, (see, for example Bellman (1970)) it is possible to reduce Σ_1 and Σ to diagonal form simultaneously,

i.e. There exists a nonsingular matrix V such that,

$$\begin{aligned} & V\Sigma V' = I \\ \text{and} \quad & V\Sigma_1 V' = \Delta = \text{diag}(\gamma_i) \ . \end{aligned}$$

Therefore, making the transformation,

$$\begin{aligned} & A_1^* = VA_1V' \\ \text{and} \quad & A_2^* = VA_2V' \end{aligned}$$

we immediately have that,

$$\begin{aligned} & A_1^* \sim W_p(\nu_1,\Delta) \\ \text{and} \quad & A_2^* \sim W_p(\nu_2,I) \qquad \text{independently.} \end{aligned}$$

Furthermore,

$$\begin{aligned} \text{Tr}(A_1^*A_2^{*-1}) &= \text{Tr}(VA_1V'(VA_2V')^{-1}) \\ &= \text{Tr}(A_1A_2^{-1}) \end{aligned}$$

so it is clear that $\text{Tr}(A_1A_2^{-1})$ is invariant under this transformation. We will therefore assume in the rest of this section that

$$A_1 \sim W_p(\nu_1, \Delta)$$

and $$A_2 \sim W_p(\nu_1, I)$$

where $$\Delta = \text{diag}\{\gamma_i\} \; . \qquad (5.4.6)$$

<u>Remark 5.4.2</u> For the case where some of the γ_i are zero, we reduce the dimension p appropriately.

It is well known (see, for example, Anderson (1958), Theorem 3.3.2) that A_1 can be written as

$$A_1 = \sum_{i=1}^{\nu_1} Y_i Y_i' \qquad (5.4.7)$$

where $\qquad Y_i \sim N_p(0, \Delta)$ independently, $i = 1, \ldots, \nu_1$.

So,

$$\begin{aligned} Tr(A_1 A_2^{-1}) &= Tr\left(\sum_{i=1}^{\nu_1} Y_i Y_i' A_2^{-1}\right) \\ &= \sum_{i=1}^{\nu_1} Tr(Y_i' A_2^{-1} Y_i) \\ &= \sum_{i=1}^{\nu_1} Y_i' A_2^{-1} Y_i \\ &= \frac{1}{\nu_2} \sum_{i=1}^{\nu_1} D_i^2 \qquad (5.4.8) \end{aligned}$$

where

$$D_i^2 = Y_i' \, S_2^{-1} \, Y_i$$

and $$S_2 = \frac{1}{\nu_2} A_2 \; .$$

Clearly D_i^2 can be considered as a sample-based Mahabanobis distance between Y_i and the origin with the difference that S_2 is a sample covariance matrix corresponding to a population covariance matrix that is different from that in the distribution of Y_i.

We how consider the distribution of $\frac{1}{\nu_2} D_i^2 = Y_i' A_2^{-1} Y_i$. Our argument follows the same lines as those used by A.H. Bowker in deriving the distribution of Hotelling's T^2 statistic. See, for example, Anderson (1958) or Giri (1977).

Define a (p×p) random orthogonal matrix Q_i whose first row is $Y_i'(Y_i'Y_i)^{-\frac{1}{2}}$ and whose remaining p-1 rows are defined arbitrarily, and let

$$Z_i = Q_i Y_i$$

and

$$B_i = Q_i A_2 Q_i' .$$

The first element z_{i1} of Z_i is, from the definition of the first row of Q_i,

$$z_{i1} = Y_i'(Y_i'Y_i)^{-\frac{1}{2}} Y_i = (Y_i'Y_i)^{\frac{1}{2}}$$

whereas the other elements of Z_i are all identically zero, by the orthogonality of Q_i. Therefore

$$Y_i' A_2^{-1} Y_i = Z_i' B_i^{-1} Z_i = z_{i1}^2 \; b_i^{11}$$

where b_i^{11} is the $(1,1)^{th}$ element of B_i^{-1}. Now

$$b_i^{11} = (b_{i11} - \underline{b}_{i(1)}' B_{i22}^{-1} \underline{b}_{i(1)})^{-1} = b_{i11.2}^{-1}$$

where $$B_i = \begin{pmatrix} b_{i11} & \underset{\sim}{b}'_{i(1)} \\ \underset{\sim}{b}_{i(1)} & B_{i22} \end{pmatrix}$$

so we get

$$Y_i' A_2^{-1} Y_i = Y_i'Y_i/b_{i11.2} \quad . \qquad (5.4.9)$$

To obtain the distribution of $b_{i11.2}$, note that, conditionally on Q, B_i has a $W_p(\nu_2,I)$ distribution. Therefore, conditionally on Q, $b_{i11.2}$ as a $W_1(\nu_2-p+1,1)$ distribution (see, for example Giri (1977) Theorem 6.4.1)

i.e.

$$b_{i11.2} \sim \chi^2_{\nu_2-p+1}$$

and since this distribution does not depend on Q_i, it is also the unconditional distribution of $b_{i11.2}$. Therefore, using the notation $u_i = b_{i11.2}$ we have that,

$$Y_i' A_2^{-1} Y_i = Y_i' Y_i/u_i \qquad (5.4.10)$$

where $u_i \sim \chi^2_{\nu_2-p+1}$ independently of Y_i.

To find the distribution of $Y_i'Y_i$, make the transformation

$$X_i = (x_{i1},\ldots,x_{ip})' = \Delta^{-\frac{1}{2}}Y_i$$

where $$\Delta^{-\frac{1}{2}} = \mathrm{diag}(\gamma_i^{-\frac{1}{2}}) \ .$$

Therefore, from (5.4.7), $X_i \sim N_p(0,I)$, independently, so that,

$$Y_i'Y_i = X_i' \Lambda X_i = \sum_{j=1}^{p} \gamma_j x_{ij}^2 = \sum_{j=1}^{p} \gamma_j v_{ij} \tag{5.4.11}$$

where $v_{ij} \sim \chi_1^2$, independently $\forall i,j$. Substituting (5.4.11) into (5.4.10) we get

$$Y_i' A_2^{-1} Y_i = \frac{1}{\nu_2} D_i^2 = \sum_{j=1}^{p} \gamma_j \frac{v_{ij}}{u_i} \tag{5.4.12}$$

and substituting (5.4.12) into (5.4.8) in turn, yields

$$\mathrm{Tr}(A_1 A_2^{-1}) = \sum_{i=1}^{\nu_1} \sum_{j=1}^{p} \gamma_j \frac{v_{ij}}{u_i} \tag{5.4.13}$$

where

$$v_{ij} \sim \chi_1^2 \quad \text{independently, } i = 1,\ldots,\nu_1; \; j = 1,\ldots,p$$

and $u_i \sim \chi^2_{\nu_2-p+1}$ independently of the v_{ij}. However, the u_i are <u>not</u> mutually independent for different i. (For p = 1 it is easy to show that the u_i are all identical.)

Expression (5.4.13) can also be written as:

$$\mathrm{Tr}(A_1 A_2^{-1}) = \sum_{j=1}^{p} \gamma_j \sum_{i=1}^{\nu_1} f_{ij} \tag{5.4.14}$$

where the f_{ij} have an unnormed $f(1,\nu_2-p+1)$ distribution, independently for different j but not for different i.

For the case where the (nonzero) eigenvalues γ_j are all equal, say $\gamma_j = \gamma \ \forall_j$, expression (5.4.13) reduces to:

$$Tr(A_1A_2^{-1}) = \gamma \sum_{i=1}^{\nu_1} f_i \quad . \qquad (5.4.15)$$

where the f_i are dependent $f(p,\nu_2-p+1)$ random variables.

Equation (5.4.15) leads naturally to the scaled F-approximations to the distribution of Hotelling's T_0^2 statistic in the central case (γ=1), proposed by Pillai and Samson (1959), Hughes and Saw (1972) and McKeon (1974). For the case where the γ_j are unequal (i.e. Σ_1 is not proportional to Σ) a scaled chi-squared approximation (Box, 1954) to $\sum_{j=1}^{p} \gamma_j \, v_{ij}$ in (5.4.13) leads to an approximate expression for the distribution of $Tr(A_1A_2^{-1})$ in the form (5.4.15). So a scaled F-approximation such as any of those proposed by the abovementioned authors should again be appropriate here.

5.4.3 The Mean and Variance of $Tr(A_1A_2^{-1})$

We now use the distribution of $Tr(A_1A_2^{-1})$obtained in the previous sub-section to find its mean and variance.

Expression (5.4.14) immediately leads to the expected value :

$$\begin{aligned} E[Tr(A_1A_2^{-1})] &= \sum_{j=1}^{p} \gamma_j \sum_{i=1}^{\nu_1} E[f_{ij}] \\ &= \sum_{j=1}^{p} \gamma_j \sum_{i=1}^{\nu_1} \frac{1}{\nu_2-p-1} \\ &= \left(\frac{\nu_1}{\nu_2-p-1}\right) \sum_{j=1}^{p} \gamma_j \ . \qquad (5.4.16) \end{aligned}$$

Remark 5.4.3 The result (5.4.16) can be confirmed by the following direct derivation of the expectation:

$$E[\mathrm{Tr}(A_1A_2^{-1})] = \mathrm{Tr}(E[A_1]E[A_2^{-1}]) \; .$$

(This step is justified by the independence of A_1 and A_2 and because the trace operation consists only of multiplications and additions of their elements)

$$= \mathrm{Tr}(\nu_1\Sigma_1(\nu_2-p-1)^{-1}\Sigma^{-1})$$

from the properties of the Wishart and Inverse Wishart distributions (See, for example, Johnson and Kotz, 1972).

$$= \left(\frac{\nu_1}{\nu_2-p-1}\right)\mathrm{Tr}(\Sigma_1 \; \Sigma^{-1})$$

$$= \left(\frac{\nu_1}{\nu_2-p-1}\right) \sum_{j=1}^{p} \gamma_j \; .$$

The variance of $\mathrm{Tr}(A_1A_2^{-1})$ does not follow in such a straightforward manner, but is most readily obtained from expression (5.4.8):

$$\mathrm{Tr}(A_1A_2^{-1}) = \frac{1}{\nu_2}\sum_{i=1}^{\nu_1} D_i^2$$

where $D_i^2 = Y_i' \; S_2^{-1} \; Y_i$. Therefore,

$$\mathrm{Var}[\mathrm{Tr}(A_1A_2^{-1})] = \frac{1}{\nu_2^2}\left(\sum_{i=1}^{\nu_1} \mathrm{Var}[D_i^2] + 2\sum_{i<j} \mathrm{Cov}[D_i^2,D_j^2] \right). \qquad (5.4.17)$$

Using (5.4.12) we obtain

$$E[D_i^2] = \nu_2 \sum_{j=1}^{p} \gamma_j \, E[v_{ij}]E[u_i^{-1}]$$

where

$$v_{ij} \sim \chi_1^2 \qquad \text{independently } \forall j = 1,..,p$$

and

$$u_i \sim \chi^2_{\nu_2-p+1} \qquad \text{independently.}$$

So,

$$E[D_i^2] = \nu_2 \sum_{j=1}^{p} \gamma_j(\nu_2-p-1)^{-1} = \left(\frac{\nu_2}{\nu_2-p-1}\right) \sum_{j=1}^{p} \gamma_j \qquad (5.4.18)$$

using the fact that the r^{th} moment of the χ^2_ν distribution is

$$\mu_r' = \frac{\Gamma(\frac{1}{2}\nu+r)}{\Gamma(\frac{1}{2}\nu)} 2^r \qquad \forall r > -\tfrac{1}{2}\nu.$$

Similarly,

$$E[(D_i^2)^2] = \nu_2^2 E[u_i^{-2}]\Big(\sum_{j=1}^{p} \gamma_j^2 \, E[v_{ij}^2] + 2 \sum_{j<\ell} \gamma_j \, \gamma_\ell \, E[v_{ij}]E[v_{i\ell}]\Big)$$

$$= \nu_2^2\{(\nu_2-p-1)(\nu_2-p-3)\}^{-1}\Big(\sum_{j=1}^{p} \gamma_j^2 \, 3 + 2 \sum_{j<\ell} \gamma_j \, \gamma_\ell\Big)$$

so

$$\mathrm{Var}[D_i^2] = E[(D_i^2)^2] - (E[D_i^2])^2$$

$$= \frac{\nu_2^2}{(\nu_2-p-1)(\nu_2-p-3)}\Big(3\sum_{j=1}^{p} \gamma_j^2 + 2\sum_{j<\ell} \gamma_j \gamma_\ell\Big) - \frac{\nu_2^2}{(\nu_2-p-1)^2}\Big(\sum_{j=1}^{p} \gamma_j\Big)^2$$

$$= \frac{2\nu_2^2}{(\nu_2-p-1)^2(\nu_2-p-3)}\Big[\Big(\sum_{j=1}^{p} \gamma_j\Big)^2 + (\nu_2-p-1)\sum_{j=1}^{p} \gamma_j^2\Big) . \qquad (5.4.19)$$

To obtain $\mathrm{cov}[D_i^2, D_j^2]$ note that, from (5.4.8)

$$D_i^2 = Y_i' S_2^{-1} Y_i$$

where,

$$Y_i \sim N_p(0, \Delta) \quad \text{independently}, \forall i = 1, \ldots, p$$

$$\nu_2 S_2 \sim W_p(\nu_2, I)$$

and $\Delta = \mathrm{diag}\{\gamma_i\}$.

Using Theorem 3.1.1, with slight modification, it immediately follows that, conditionally on S_2,

$$D_i^2 = \sum_{\ell=1}^{p} \alpha_\ell \, v_{\ell i} \quad \forall i, j \qquad (5.4.20)$$

where

$$\{\alpha_\ell\} = \mathrm{Eigs}\{\Delta S_2^{-1}\}$$

and

$$v_{\ell i} \sim \chi_1^2 \qquad \text{independently}, \ \forall \ell = 1, \ldots, p.$$

Furthermore,

$$\mathrm{Cov}[D_i^2 D_j^2] = E[D_i^2 D_j^2] - E[D_i^2]E[D_j^2]$$

$$= E_{S_2}[E[D_i^2 D_j^2|S_2]] - E_{S_2}[E[D_i^2|S_2]]E_{S_2}[E[D_j^2|S_2]] \qquad (5.4.21)$$

where $E_{S_2}[\cdot]$ denotes the expection over the distribution of S_2. The conditional expectations in (5.4.21) follow immediately from (5.4.20):

$$E[D_i^2|S_2] = \sum_{\ell=1}^{p} \alpha_\ell E[v_{\ell i}] = \sum_{\ell=1}^{p} \alpha_\ell = \mathrm{Tr}(\Delta S_2^{-1})$$

and

$$E[D_i^2 \; D_j^2| S_2] = E[\sum_{\ell=1}^{p} \alpha_\ell v_{\ell i}]E[\sum_{\ell=1}^{p} \alpha_\ell v_{\ell j}]$$

by the independance of the $v_{\ell i}$

$$= (\sum_{\ell=1}^{p} \alpha_\ell)^2 = (\mathrm{Tr}(\Delta S_2^{-1}))^2$$

so,

$$\mathrm{Cov}[D_i^2 D_j^2] = E_{S_2}[(\mathrm{Tr}(\Delta S_2^{-1}))^2] -(E_{S_2}[\mathrm{Tr}(\Delta S_2^{-1})])^2$$

$$= \mathrm{Var}_{S_2}[\mathrm{Tr}(\Delta S_2^{-1})] \qquad (5.4.22)$$

where $\mathrm{Var}_{S_2}[\cdot]$ denotes the variance over the distribution of S_2. Now

$$\mathrm{Tr}(\Delta S_2^{-1}) = \mathrm{Tr}(\Delta^{\frac{1}{2}} S_2^{-1} \Delta^{\frac{1}{2}}) = \nu_2 \mathrm{Tr}(\Delta^{-\frac{1}{2}} A_2 \Delta^{-\frac{1}{2}})^{-1}$$

where

$$\Delta^{-\frac{1}{2}} = \text{diag}\{\gamma_i^{-\frac{1}{2}}\}$$

and

$$A_2 = \nu_2 S_2 \sim W_p(\nu_2, I).$$

Therefore,

$$\Delta^{-\frac{1}{2}} A_2 \Delta^{-\frac{1}{2}} \sim W_p(\nu_2, \Delta^{-\frac{1}{2}} I \Delta^{-\frac{1}{2}})$$

$$\sim W_p(\nu_2, \Delta^{-1})$$

so that $(\Delta^{-\frac{1}{2}} A_2 \Delta^{-\frac{1}{2}})^{-1}$ follows the inverted Wishart distribution $W_p^{-1}(\nu_2+p+1, \Delta)$ (See, for example Press, 1972). So,

$$\text{Tr}(\Delta S_2^{-1}) = \nu_2 \text{Tr}(W) \qquad (5.4.23)$$

where

$$W = (w_{ij}) \sim W_p^{-1}(\nu_2+p+1, \Delta) .$$

Furthermore,

$$\text{Var}[\text{Tr}(W)] = \text{Var}[\sum_{i=1}^{p} w_{ii}] = \sum_{i=1}^{p} \text{var}[w_{ii}] + 2\sum_{i<j} \text{Cov}[w_{ii}, w_{jj}]$$

These variances and covariances are given in Press (1972) on page 112, so substituting them into the above and remembering that $\Delta = \text{diag}\{\gamma_i\}$ we get, after some simplification,

$$\text{Var}[\text{Tr}(W)] = \sum_{i=1}^{p} \frac{2\gamma_i^2}{(\nu_2-p-1)^2(\nu_2-p-3)}$$

$$+ 2\sum_{i<j} \frac{2\gamma_i \gamma_j}{(\nu_2-p)(\nu_2-p-1)^2(\nu_2-p-3)} . \qquad (5.4.24)$$

Substituting (5.4.24) and (5.4.23) into (5.4.22) yields,

$$\mathrm{Cov}[D_i^2,D_j^2] = \frac{2\nu_2^2}{(\nu_2-p)(\nu_2-p-1)^2(\nu_2-p-3)}\{(\nu_2-p)\sum_{k=1}^{p}\gamma_k^2 + 2\sum_{k<\ell}\gamma_k\,\gamma_\ell\} \ . \tag{5.4.25}$$

Finally, substituting (5.4.25) and (5.4.19) into (5.4.17) yields, after some simplification:

$$\mathrm{Var}[\mathrm{Tr}(A_1A_2^{-1})] = \frac{{}_1(\nu_1+\nu_2-p-1)}{(\nu_2-p)(\nu_2-p-1)^2(\nu_2-p-3)}\{(\sum_{j=1}^{p}\gamma_j)^2 + (\nu_2-p-1)\sum_{j=1}^{p}\gamma_j^2\}$$

$$= \frac{2\nu_1(\nu_1+\nu_2-p-1)}{(\nu_2-p)(\nu_2-p-1)^2(\nu_2-p-3)}\{(\mathrm{Tr}(\Sigma_1\Sigma^{-1}))^2$$

$$+ (\nu_2-p-1)\mathrm{Tr}(\Sigma_1\Sigma^{-1})^2\} \ . \tag{5.4.26}$$

As a test for the correctness of formulae (5.4.16) and (5.4.26) for the mean and variance, respectively, of $\mathrm{Tr}(A_1A_2^{-1})$ we consider the case where $\Sigma_1 = \Sigma$, i.e. $\gamma_i = 1$, $i = 1,\ldots,p$. The formulae then reduce to:

$$E[\mathrm{Tr}(A_1A_2^{-1})] = \frac{\nu_1\ p}{\nu_2-p-1}$$

and

$$\mathrm{Var}[\mathrm{Tr}(A_1A_2^{-1})] = \frac{2p\nu_1(\nu_2-1)(\nu_1+\nu_2-p-1)}{(\nu_2-p)(\nu_2-p-1)^2(\nu_2-p-3)} \tag{5.4.27}$$

which agree with those given by Pillai and Samson (1959) as well as by Hughes and Saw (1972). (The formulae given by McKeon (1974) both appear to require the factor $\nu_2(\nu_2-p-1)^{-1}$.)

Using similar techniques to those used above it is clear that with increasing amounts of algebra the higher moments of $Tr(A_1A_2^{-1})$ may be obtained.

Formulae (5.4.16) and (5.4.26) will now be used to obtain moment estimators of $\sum_{i=1}^{p} \lambda_i$ and $\sum_{i=1}^{p} \lambda_i^2$ where $\{\lambda_i\} = Eigs(T\Sigma^{-1})$, which may in turn be used to estimate the means and variances of the four distance variables whose distributions were discussed in Chapter 3, as well as the approximate probabilities of misclassification derived in Chapter 4.

5.4.4 Moment Estimators for $\sum_{i=1}^{p} \lambda_i$ and $\sum_{i=1}^{p} \lambda_i^2$

The formulae for the means and variances of the four distance variables δ_{ij}^2, $\delta_i^2(x)$, d_{ij}^2 and $d_i^2(x)$ derived in Chapter 3, as well as those for the approximate probabilities (4.1.9) and (4.2.10) of misclassification derived in Chapter 4, are all expressed in terms of the two quantities:

$$\sum_{i=1}^{p} \lambda_i = Tr(T\Sigma^{-1})$$

and

$$\sum_{i=1}^{p} \lambda^2 = Tr(T\Sigma^{-1})^2 \ .$$

In this sub-section, moment estimators for these two quantities will be obtained in terms of the expectation and variance of $Tr(S_1S_2^{-1}) = \frac{\nu_2}{\nu_1} Tr(A_1A_2)$ derived in the previous sub-section. These may then be substituted into the abovementioned formulae to obtain estimators for the means and variances of the four distance variables and for the approximate probabilities of misclassification.

Substituting the expression given in Remark 5.3.2 for the relationship between the $\{\lambda_i\}$ and the $\{\gamma_i\}$:

$$\gamma_i = 1 + n\lambda_i$$

into expressions (5.4.16) and (5.4.26) for the mean and variance of $\mathrm{Tr}(A_1A_2^{-1})$, transforming to $\mathrm{Tr}(S_1S_2^{-1})$ and simplifying, *yields:*

$$E[\mathrm{Tr}\ S_1S_2^{-1}] = \frac{\nu_2}{\nu_2-p-1}(p + n \sum_{i=1}^{p} \lambda_i) \qquad (5.4.28)$$

and

$$\mathrm{Var}[\mathrm{Tr}(S_1S_2^{-1})] = C(n^2(\nu_2-p-1) \sum_{i=1}^{p} \lambda_i^2 + n^2(\sum_{i=1}^{p} \lambda_i)^2$$

$$+ 2n(\nu_2-1)\sum_{i=1}^{p} \lambda_i + p(\nu_2-1)) \qquad (5.4.29)$$

where

$$C = \frac{2\nu_2^2(\nu_1+\nu_2-p-1)}{\nu_1(\nu_2-p)(\nu_2-p-1)^2(\nu_2-p-3)} \quad .$$

So it follows immediately that the moment estimators for $\sum_{i=1}^{p} \lambda_i$ and $\sum_{i=1}^{p} \lambda_i^2$ are respectively,

$$\widehat{\sum_{i=1}^{p} \lambda_i} = \frac{(\nu_2-p-1)}{\nu_2 n} \hat{E}[\mathrm{Tr}(S_1S_2^{-1})] - \frac{p}{n} \qquad (5.4.30)$$

and

$$\widehat{\sum_{i=1}^{p} \lambda_i^2} = \frac{1}{n^2(\nu_2-p-1)} (C^{-1}\widehat{\mathrm{Var}}[\mathrm{Tr}(S_1S_2^{-1})] - n^2(\widehat{\sum_{i=1}^{p} \lambda_i})^2$$

$$- 2n(\nu_2-1)\widehat{\sum_{i=1}^{p} \lambda_i} - p(\nu_2-1)) \qquad (5.4.31)$$

where $\hat{E}[\mathrm{Tr}(S_1S_2^{-1})]$ and $\widehat{\mathrm{Var}}[\mathrm{Tr}(S_1S_2^{-1})]$ are sample-based estimators for the mean and variance of $\mathrm{Tr}(S_1S_2^{-1})$.

Now, the obvious estimator for $E[\mathrm{Tr}(S_1S_2^{-1})]$ from the training sample is

$$\hat{E}[\mathrm{Tr}(S_1S_2^{-1})] = \mathrm{Tr}(S_1S_2^{-1}) \qquad (5.4.32)$$

but there is no corresponding simple estimator for $\mathrm{Var}[\mathrm{Tr}(S_1S_2^{-1})]$. However, the Jackknife technique, originally proposed by Quenouille (1956) provides an attractive, if computationally lengthy, method for obtaining an estimator for the latter.

The Jackknife Technique Good descriptions of the technique are given by Gray and Schucany, (1972) Miller (1974) and Bissel and Ferguson (1975), so a brief summary here will suffice.

Given an unknown parameter θ for which a (possibly) biased estimator $\hat{\theta}$ is available from a random sample, suppose that the expected value of $\hat{\theta}$ may be written,

$$E[\hat{\theta}] = \theta + O(n^{-1}) \qquad (5.4.33)$$

where n is the sample size. The Jackknife technique for reducing this bias to $O(n^{-2})$ and at the same time producing an estimate of the variance of $\hat{\theta}$ proceeds as follows. Divide the sample into r subgroups each of size h (r=n and h=1 in most applications). Removing each subgroup from the sample in turn, and re-estimating θ from the remainder of the sample in each case, produces r "partial estimates" $\hat{\theta}_{-j}$, $j = 1,\ldots,r$, each based on a sample of size $h(r-1)$. Now combine these partial estimates with the whole-sample estimate to form r "pseudo-values" $\hat{\theta}_{*j}$:

$$\hat{\theta}_{*j} = r\hat{\theta} - (r-1)\hat{\theta}_{-j} \qquad j = 1,\ldots,r. \tag{5.4.34}$$

The Jackknife estimator of θ is the average of the $\hat{\theta}_{*j}$:

$$\hat{\theta}_{*} = \frac{1}{r}\sum_{j=1}^{r} \hat{\theta}_{*j} = r\hat{\theta} - (r-1)\hat{\theta}_{-.} \tag{5.4.35}$$

where

$$\hat{\theta}_{-.} = \frac{1}{r}\sum_{j=1}^{r} \hat{\theta}_{-j}$$

and it can easily be shown that $\hat{\theta}_{*}$ has a (possible) bias of order n^{-2}

i.e. $\qquad E[\hat{\theta}_{*}] = \theta + O(n^{-2})$.

Quenouille (1956) shows that, to order n^{-1}, the variance of $\hat{\theta}_{*}$ is the same as that of $\hat{\theta}$ for a wide class of estimators, and Tukey (1958) proposed the following estimator for $\mathrm{Var}[\hat{\theta}]$ or $\mathrm{Var}[\hat{\theta}_{*}]$:

$$S_T^2 = \frac{1}{r(r-1)}\sum_{j=1}^{r} (\hat{\theta}_{*j} - \hat{\theta}_{*})^2$$

$$= \frac{r-1}{r}\sum_{j=1}^{r} (\hat{\theta}_{-j} - \hat{\theta}_{-.})^2 \; ; \tag{5.4.36}$$

Tukey (1958) also suggested that a confidence interval for θ may be obtained by assuming that $t_r = (\hat{\theta}_{*}-\theta)/S_T$ has, approximately, a t-distribution on r-1 degrees of freedom.

Going back to formula (5.4.16) we have:

$$E[\mathrm{Tr}(S_1S_2^{-1})] = \frac{\nu_2}{\nu_1}\,E[\mathrm{Tr}(A_1A_2^{-1})]$$

$$= \frac{\nu_2}{\nu_2-p-1}\,\mathrm{Tr}(\Sigma_1\Sigma^{-1})$$

$$= \left(1 - \frac{p+1}{\nu_2}\right)^{-1}\mathrm{Tr}(\Sigma_1\Sigma^{-1})$$

$$= \mathrm{Tr}(\Sigma_1\Sigma^{-1}) + O(\nu_2^{-1}) \qquad (5.4.37)$$

which is clearly of the form (5.4.33), so it would appear that the Jackknife technique can provide an estimator for $\mathrm{Var}[\mathrm{Tr}(S_1S_2^{-1})]$ via (5.4.36).

Jackknife Estimation of $\mathrm{Var}[\mathrm{Tr}(S_1S_2^{-1})]$. As mentioned earlier, a drawback to the Jackknife technique is the fact that the amount of computation required can become very lengthy, especially when the training sample is large and $h = 1$, as is usually recommended. However, the computation can be reduced considerably in the case of $\mathrm{Tr}(S_1S_2^{-1})$ with $h = 1$ by using the following theorem.

Theorem 5.4.1

Let A_1 and A_2 be the $(p\times p)$ "Between groups" and "within groups" sum of squares matrices based on k groups and n observations per group, as defined in the MANOVA table 5.1.1. Let $T_{-(i,j)}$ denote the value of a statistic T computed from the MANOVA sample with observation x_{ij} removed from the i^{th} group. Then, using the notation of Section 5.1,

$$\mathrm{Tr}(A_1A_2^{-1})_{-(i,j)} = \mathrm{Tr}(A_1A_2^{-1}) + \mathrm{Tr}(A_1A_2^{-1}E) + \mathrm{Tr}(GA_2^{-1}F)$$

where,

$$E = n(n-1)ee'A_2^{-1}/(1-n(n-1)eA_2^{-1}e') \ ,$$

$$F = I + E \ ,$$

$$G = (N-n)ff' + (n-1)(e-f)(e-f)' - Nfg' - g((n-1)e+f)'$$

and $e = \frac{x_{ij}-x_{i.}}{n-1}$, $f = \frac{x_{ij}-x_{..}}{N-1}$, $g = x_{i.} - x_{..}$. The proof is given in Appendix 5.1.

From Theorem 5.4.1 *only a single matrix inversion*, that of A_2, is *required* for the computation of all N partial estimates $Tr(A_1A_2^{-1})_{-(i,j)}$, $\forall i,j$, and since the other formulae are all of a simple nature the total computation time on a modern computer is very small, even for large values of N and moderate values of p.

Note that, since $S_1 = \nu_1^{-1}A_1$ and $S_2 = \nu_2^{-1}A_2$

$$Tr(S_1S_2^{-1})_{-(i,j)} = \frac{\nu_2-1}{\nu_1} Tr(A_1A_2^{-1})_{-(i,j)}. \qquad (5.4.38)$$

Therefore, using $h = 1$ and $r = N$ in (3.4.36) we obtain the following estimator for $Var[Tr(S_1S_2^{-1})]$ from the jackknife method

$$\hat{Var}[Tr(S_1S_2^{-1})] = \frac{N-1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n} (Tr(S_1S_2^{-1})_{-(i,j)} - Tr(S_1S_2^{-1})_{-(\cdot,\cdot)})^2 . \qquad (5.4.39)$$

Substituting (5.4.39) and (5.4.32) into (5.4.30) and (5.4.31) yields moment estimators for $\sum_{i=1}^{p} \lambda_i$ and $\sum_{i=1}^{p} \lambda_i^2$, respectively, which can in turn be substituted into the relevant formulae to obtain estimators for the means and variances of $\hat{\delta}_{ij}^2$, $\delta_i^2(x)$, d_{ij}^2 and $d_i^2(x)$ as well as for the approximate probabilities of misclassification under the random effects model.

5.4.5 Restricted Maximum Likelihood Estimators of the $\{\gamma_i\}$

In this sub-section we investigate the use of the exact results on the moments of $|A_1A_2^{-1}|$ and $Tr(A_1A_2^{-1})$ obtained in sections 5.4.1 and 5.4.3 respectively, to improve our maximum likelihood estimators of the $\{\gamma_i\} = Eigs(\Sigma_1\Sigma^{-1})$ based on Chang's (1970) expression for the limiting density of the $\{g_i\} = Eigs(A_1A_2^{-1})$.

But firstly we investgate the special cases p = 1 and p = 2.

p = 1. In this case Chang's (1970) formula, $|A_1A_2^{-1}|$ and $Tr(A_1A_2^{-1})$ all lead to the same result, viz:

$$\left(\frac{g_1}{\gamma_1}\right) \sim f(\nu_1,\nu_2)$$

where $f(\nu_1,\nu_2)$denotes the unnormed f-distribution on ν_1 and ν_2 degrees of freedom (See (5.3.6) and (5.4.13)). Therefore, using any one of expressions (5.3.7),(5.4.4)or (5.4.16), we obtain the following unbiased moment estimator of γ_1:

$$\hat{\gamma}_1^* = \frac{\nu_2-2}{\nu_1} g_1 = \frac{\nu_2-2}{\nu_2} \ell_1 \qquad (5.4.40)$$

where $\{\ell_i\} = Eigs(S_1S_2^{-1})$ or $\ell_1 = S_1/S_2$ in this case.

The maximum likelihood estimator is given by:

$$\hat{\gamma}_1 = \frac{\nu_2}{\nu_1} g_1 = \ell_1 \qquad (5.4.41)$$

which clearly has a slight bias.

p = 2. In this case we can solve the moment estimators for $\prod_{i=1}^{p} \gamma_i$ and $\sum_{i=1}^{p} \gamma_i$ obtained from the exact first moments of $|A_1A_2^{-1}|$ and $Tr(A_1A_2^{-1})$, respectively, for γ_1 and γ_2. From (5.4.5) we have:

$$\widehat{\prod_{i=1}^{p} \gamma_i} = \left(\frac{\nu_1}{\nu_2}\right)^p \prod_{i=1}^{p} \left(\frac{\nu_2-i-1}{\nu_1-i+1}\right) \ell_i = a, \text{ say} \tag{5.4.42}$$

and from (5.4.30) and (5.4.32), remembering that $\sum_{i=1}^{p} \gamma_i = p + n \sum_{i=1}^{p} \lambda_i$, we have

$$\widehat{\sum_{i=1}^{p} \gamma_i} = \frac{\nu_2-p-1}{\nu_2} \sum_{i=1}^{p} \ell_i = b, \text{ say.} \tag{5.4.43}$$

Letting the estimators $\hat{\gamma}_1$ and $\hat{\gamma}_2$ satisfy the relationships:

$$\hat{\gamma}_1 \hat{\gamma}_2 = \widehat{\gamma_1\gamma_2}$$

and

$$\hat{\gamma}_1 + \hat{\gamma}_2 = \widehat{\gamma_1 + \gamma_2} \tag{5.4.44}$$

(5.4.42) and (5.4.43) lead to the following solutions:

$$\hat{\gamma}_1^* = \tfrac{1}{2}(b + \sqrt{b^2-4a})$$

and

$$\hat{\gamma}_2^* = \tfrac{1}{2}(b - \sqrt{b^2-4a}). \tag{5.4.45}$$

For $\underline{p > 2}$, we use the technique of Restricted Maximum Likelihood Estimation (see, for example, Silvey, 1975) to incoporate the information from the exact moments of $|A_1A_2^{-1}|$ and $\mathrm{Tr}(A_1A_2^{-1})$ as constraints into the maximum likelihood equations obtained from Chang's (1970) formula (5.3.5) for the limiting joint density function of the g_i.

Using the same reparameterisation as before to get around the problem of the "*inadmissible singularities*" in *the likelihood function* (see (5.3.22)) and reformulating the constraint (5.4.42)

$$\widehat{\prod_{i=1}^{p} \gamma_i} = a$$

for algehraic convenience by taking logarithms on both sides, we obtain the following constrained maximization problem (see (5.3.24), (5.4.42) and (5.4.43)).

Maximise:

$$L = f^*(g) - \tfrac{1}{2}(\nu_1+\nu_2-p+1)\sum_{j=1}^{p} \log(1+g_j(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}))$$

$$- \tfrac{1}{2}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \log(\sum_{k=i+1}^{j} e^{\delta_k+\varepsilon_k})$$

subject to:

(i) $$-\sum_{j=1}^{p} \log(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k}) = \log a$$

and (ii) $$\sum_{j=1}^{p} (\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k})^{-1} = b \qquad (5.4.46)$$

where $f^*(g)$ is a funtion of the g_j only. (Note that, because of the first constraint, the term: $\tfrac{1}{2}\nu_1 \sum_{j=1}^{p} \log(\sum_{k=1}^{j} e^{\delta_k+\varepsilon_k})$ in the objective function of (5.4.46) is a constant and has therefore been incorporated into $f^*(g)$).

Remark 5.4.4 Although the estimated value of $\sum_{i=1}^{p} \gamma_i^2$, obtained from the variance of $\mathrm{Tr}(A_1 A_2^{-1})$ could also have been brought in as a constraint, it was felt that it would be unrealistic to do so, particularly in view of the indirect method in which it is obtained. $\square$

The constrained maximization problem (5.4.46) is a nonlinear programming problem and is therefore most readily solved using one of the standard algorithms (see, for example Walsh, 1975) for the restricted maximum likelihood estimator $\hat{\underset{\sim}{\delta}}^*$. Finally, by transforming back via (5.3.22) we obtain the restricted maximum likelihood estimator $\hat{\underset{\sim}{\gamma}}^*$ of $\underset{\sim}{\gamma}$.

5.4.6 Large Sample Distribution of the Restricted Maximum Likelihood Estimators of the γ_i

Silvey (1975) shows that for large sample sizes the restricted maximum likelihood estimator $\hat{\underset{\sim}{\gamma}}^*$ is approximately normally distributed with mean vector $\underset{\sim}{\gamma}$ and covariance matrix Σ, where Σ is obtained by the following matrix equality:

$$\begin{pmatrix} B_{\underset{\sim}{\gamma}} & H \\ H' & 0 \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma & Q \\ Q' & R \end{pmatrix} \qquad (5.4.47)$$

where $B_{\underset{\sim}{\gamma}}$ is Fisher's Information Matrix given by (5.3.30), (5.3.33) and (5.3.38) and H is the (p×2) matrix of partial derivatives:

$$H = \begin{pmatrix} \frac{\partial}{\partial\gamma_1}(\sum_{j=1}^{p} \text{Log}\ \gamma_j - \log a) & \frac{\partial}{\partial\gamma_1}(\sum_{j=1}^{p} \gamma_j - b) \\ \vdots & \\ \frac{\partial}{\partial\gamma_p}(\sum_{j=1}^{p} \log \gamma_j - \log a) & \frac{\partial}{\partial\gamma_p}(\sum_{j=1}^{p} \gamma_j - b) \end{pmatrix}$$

$$= \begin{pmatrix} \gamma_1^{-1} & 1 \\ \vdots & \vdots \\ \gamma_p^{-1} & 1 \end{pmatrix}. \qquad (5.4.48)$$

It follows from (5.4.47) that the elements of $\underset{\sim}{\gamma}^*$ will tend to have smaller approximate variances than those of the "unrestricted" maximum likelihood estimators $\hat{\underset{\sim}{\gamma}}$, discussed in Section 5.3, for, as shown by Silvey (1975), Appendix A:

$$\Sigma = B_{\underset{\sim}{\gamma}}^{-1} - B_{\underset{\sim}{\gamma}}^{-1} H(H'B_{\underset{\sim}{\gamma}}^{-1}H)^{-1} H'B_{\underset{\sim}{\gamma}}^{-1} . \qquad (5.4.49)$$

The result now follows, since $B_{\underset{\sim}{\gamma}}^{-1}$ is the approximate covariance matrix of $\hat{\underset{\sim}{\gamma}}$ and $H(H'B_{\underset{\sim}{\gamma}}^{-1}H)^{-1}H'$ is a positive semidefinite matrix.

However, the above result could be rather misleading in our situation, since formulae (5.4.47) and (5.4.48) are based on the assumption that the two constraints

$$\sum_{j=1}^{p} \log \gamma_j = a$$

and

$$\sum_{j=1}^{p} \gamma_j = b$$

are deterministic, whereas, in fact, they are stochastic since a and b are random variables. Thus result (5.4.49) will tend to give too optimistic a picture of the large sample behaviour of the restricted maximum estimator $\hat{\underset{\sim}{\gamma}}^*$.

This point is illustrated in Table 5.4.1 below, which gives the approximate large sample standard deviations for the elements of $\hat{\underset{\sim}{\gamma}}$ and $\hat{\underset{\sim}{\gamma}}^*$ as well as the corresponding standard deviations obtained from the simulation experiments on $\hat{\underset{\sim}{\gamma}}^*$ described in the next section, for two of the sets of parameter values used earlier in Example 5.3.1.

Table 5.4.1

		Standard Deviations		
Degrees of Freedom	True $\underset{\sim}{\gamma}$	Approx for $\hat{\underset{\sim}{\gamma}}$	Approx for $\hat{\underset{\sim}{\gamma}}^*$	From simulated $\hat{\underset{\sim}{\gamma}}^*$
$\nu_1 = 60$	6	1.45	0.59	1.06
$\nu_2 = 244$	4	0.92	0.79	0.59
	2	0.42	0.20	0.33
$\nu_1 = 30$	16	5.13	0.58	4.33
$\nu_2 = 124$	4	1.32	1.01	1.21
	2	0.61	0.43	0.51

As is evident from Table 5.4.1 there is a marked reduction in the approximate standard deviations when moving from the unrestricted to the restricted maximum likelihood estimator for $\underset{\sim}{\gamma}$, the reduction being by far the greatest for the estimator of the largest eigenvalue γ_1. However, it is also clear that most of this reduction is not realised in practice.

Nevertheless, the simulation experiments described in the next section do suggest that with regard to both bias and standard deviation the restricted maximum likelihood estimator $\hat{\underset{\sim}{\gamma}}^*$ is a slight improvement over its unrestricted counterpart $\hat{\underset{\sim}{\gamma}}$.

5.5 Simulation Experiments on the Various Estimators of $\{\gamma_i\} = \text{Eigs}\{\Sigma_1\Sigma^{-1}\}$

In this section we describe some simulation experiments that were carried out to evaluate the performances of the various estimators of $\{\gamma_i\} = \text{Eigs}\{\Sigma_1\Sigma^{-1}\}$ that have been proposed in the earlier sections. In addition, because of the problems associated with some of these estimators under various circumstances, another, "hybrid" estimator, defined below, was also considered. Specifically, the following five estimators of γ_i, $i = 1,\ldots,p$ were considered:

(1) The maximum likelihood estimator $\hat{\gamma}_i^{(1)} = \ell_i$ where $\{\ell_i\} = \text{Eigs}\{S_1S_2^{-1}\}$.

(2) The approximate maximum marginal likelihood estimator $\hat{\gamma}_i^{(2)}$, given by (5.3.20) obtained as an approximate solution to the maximum marginal likelihood equations (5.3.11) derived from Chang's limiting distribution of the g_i.

(3) The "hybrid" estimator $\hat{\gamma}_i^{(3)}$, defined below.

(4) The "unrestricted" maximum marginal likelihood estimator $\hat{\gamma}_i^{(4)}$ obtained by solving equations (5.3.11) numerically, as described in Section 5.3.3.

(5) The "restricted" maximum marginal likelihood estimator $\hat{\gamma}_i^{(5)}$ obtained by solving the constrained maximization problem (5.4.46).

In a sense that shall be made clear later, and excluding for the moment the "hybrid" estimator $\hat{\gamma}^{(3)}$, the "goodness" of the estimators increase in the above order, $\underset{\sim}{\hat{\gamma}}^{(1)}$ being worst and $\underset{\sim}{\hat{\gamma}}^{(5)}$ best. However, the reliability of these estimators, defined as their ability to produce meaningful results over a wide range of parameter values, increases in the reverse order. In fact, $\hat{\gamma}^{(5)}$ and $\hat{\gamma}^{(4)}$ generally only produce meaningful results when the sample sizes are large and the eigenvalues well separated, whereas $\hat{\gamma}^{(1)}$ is completely reliable.

$\underset{\sim}{\hat{\gamma}}^{(2)}$ can produce meaningless results in the following ways:

(i) the $\hat{\gamma}_i^{(2)}$ may not be monotonically decreasing with i,

or (ii) some of the $\hat{\gamma}_i^{(2)}$ may be negative,

or (iii) both (i) and (ii) may occur.

However, in many cases when failure of any one of the above three kinds occurs, the first few $\hat{\gamma}_i^{(2)}$ are well-behaved and the failure only affects the estimates of the lower-valued parameters.

For this reason, and because:

(i) the greatest improvement occurs between estimators $\underset{\sim}{\hat{\gamma}}^{(1)}$ and $\underset{\sim}{\hat{\gamma}}^{(2)}$, the incremental improvement between $\underset{\sim}{\hat{\gamma}}^{(4)}$and $\underset{\sim}{\hat{\gamma}}^{(5)}$being relatively much smaller,

(ii) $\underset{\sim}{\hat{\gamma}}^{(2)}$ fails less frequently than $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$,

and (iii) $\underset{\sim}{\hat{\gamma}}^{(2)}$ is far simpler to evaluate than $\underset{\sim}{\hat{\gamma}}^{(4)}$ or $\underset{\sim}{\hat{\gamma}}^{(5)}$,

the "hybrid" estimator $\underset{\sim}{\hat{\gamma}}^{(3)}$ has been defined as that combination of $\underset{\sim}{\hat{\gamma}}^{(1)}$ and $\underset{\sim}{\hat{\gamma}}^{(2)}$ that makes maximal use of $\underset{\sim}{\hat{\gamma}}^{(2)}$, yet never produces meaningless results. Thus $\underset{\sim}{\hat{\gamma}}^{(3)}$ is defined to be equal to $\underset{\sim}{\hat{\gamma}}^{(2)}$ whenever the latter does not fail; otherwise it uses as much of the "meaningful" part of $\underset{\sim}{\hat{\gamma}}^{(2)}$ as possible and uses $\underset{\sim}{\hat{\gamma}}^{(1)}$ for the rest. This leads to the following formal definition of $\underset{\sim}{\hat{\gamma}}^{(3)}$:

Let s be one of the integers {0,1,...,p} such that, s = p if $\underset{\sim}{\hat{\gamma}}^{(2)}$ does not fail; otherwise s is the largest integer for which both

(i) failure of $\hat{\gamma}_i^{(2)}$ occurs for the first time when i > s

and (ii) $\hat{\gamma}_s^{(2)} > \hat{\gamma}_{s+1}^{(1)}$.

Then $\underset{\sim}{\hat{\gamma}}^{(3)} = (\hat{\gamma}_1^{(3)}, \ldots, \hat{\gamma}_p^{(3)})'$ is defined as:

$$\begin{aligned} \hat{\gamma}_i^{(3)} &= \hat{\gamma}_i^{(2)}\ , \ i = 1,\ldots,s \qquad \text{(unless } s = 0) \\ \hat{\gamma}_i^{(3)} &= \hat{\gamma}_i^{(1)}\ , \ i = s+1,\ldots,p \quad \text{(unless } s = p). \end{aligned} \tag{5.5.1}$$

5.5.1 The Experimental Setup

The experiments, performed on the Council for Scientific and Industrial Research's CDC Cyber 174 computer, consisted in:

(a) selecting the parameters p, ν_1, ν_2 and $\underset{\sim}{\gamma}$,

(b) generating two random matrices A_1 and A_2 from Wishart distributions with the selected values of the parameters,

(c) computing the eigenvalues $\{g_i\} = \text{eigs}(A_1A_2^{-1})$,.

(d) computing the five estimators $\underset{\sim}{\hat{\gamma}}^{(1)}$ to $\underset{\sim}{\hat{\gamma}}^{(5)}$ and

(e) repeating steps (b) to (d) a hundred times and computing summary statistics, separately for each selection of parameter values.

All the computer programs were written in FORTRAN IV making use of the University of the Witwatersrand's multivariate statistical library developed largely by Prof. D.M. Hawkins, as well as of the IMSL (1975) and the NAG (1975) program libraries.

(a) Selecting the Parameters

As is often the case in simulation experiments, the computer programs were developed and tested using a particular set of parameter values, and many of the conclusions could be obtained from just this one set of values. It also became apparent during the development stage that some of the estimators broke down for particular parameter values and this, to a large extent, guided the choice of parameter values (in particular the degrees of freedom ν_1 and ν_2) used in the experiments.

(i) The dimension p Four values, 2 (see comment below), 3("small"), 5("medium") and 10("large") were used. For values greater than 10 the computing time associated with estimators $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ became too large to allow enough simulation runs to be performed for meaningful conclusions to be drawn from them. The value $p = 2$ was included to test the estimator (5.4.45).

(ii) The degrees of freedom ν_1 and ν_2 Here again, four sets of values were chosen, corresponding to "small", "medium", "large" and "very large" sized samples. Clearly, the "largeness" of the samples depends very much on the dimension p, so "small" samples were considered to have $\nu_1 = 2p$ "medium" samples $\nu_1 = 5p$ and "large" samples $\nu_1 = 10p$. The "very large" category ($\nu_1 = 20p$) was included because of the tendency for the estimators $\hat{\underset{\sim}{\gamma}}^{(4)}$ and $\hat{\underset{\sim}{\gamma}}^{(5)}$ to fail for the smaller sample sizes. This was particularly so for the larger values of p and the "equal separations" choice of eigenvalues (see (iii) below). ν_2 has, by definition, to be greater than ν_1 and since the results were not very sensitive to variations in ν_2, almost all the simulation runs reported here were done assuming that there were n = 5 observations per group, so that $\nu_2 = 4(\nu_1+1)$. A few runs were also performed with n = 10 observations per group.

(iii) The Eigenvalues $\{\gamma_i\}$ Since $\Sigma_1 = \Sigma + nT$, and T is a nonnegative definite matrix, the γ_i cannot be less than 1. This is easily seen by noting that the γ_i all satisfy the relationship:

$$|\Sigma_1\Sigma^{-1} - I\gamma_i| = 0$$

and since

$$\Sigma_1\Sigma^{-1} = (\Sigma+nT)\Sigma^{-1} = I + nT\Sigma^{-1},$$

we have that

$$|\Sigma_1\Sigma^{-1} - I\gamma_i| = |nT\Sigma^{-1} - I(\gamma_i-1)| = 0.$$

Therefore, since $nT\Sigma^{-1}$ is a nonnegative definite matrix

$$\gamma_i - 1 \geq 0, \quad \text{i.e.} \quad \gamma_i \geq 1.$$

Furthermore, we may assume that all the $\gamma_i > 1$, since $\gamma_i = 1$ corresponds to $\lambda_i = 0$, and in the practical situation we would have tested for this (see Section 5.2) and if accepted we would have no further use for that eigenvalue.

Finally, bearing in mind the fact that the γ_i should all be different from each other for Chang's expression (5.3.5) for the limiting joint density of the g_i to be valid, the following two sets of γ_i were selected for the simulation experiments:

Equal separations	20	18	16	14	12	10	8	6	4	2
Increasing separations	1024	512	256	128	64	32	16	8	4	2

For $p < 10$ the lower p values were used.

(b) Generating the Random Wishart Matrices

As discussed in Section 5.4.2, there exists a nonsigular matrix V that simultaneously diagonalizes Σ to the identity matrix and Σ_1 to a diagonal matrix Δ whose diagonal elements are the eigenvalues of $\Sigma_1\Sigma^{-1}$. As the eigenvalues of $A_1A_2^{-1}$, where $A_1 \sim W_p(\nu_1,\Sigma_1)$ and $A_2 \sim W_p(\nu_2,\Sigma)$ independently, are invariant under this transformation, we may assume that, for the purpose of the simulation,

$$A_1 \sim W_p(\nu_1,\Delta)$$

and $$A_2 \sim W_p(\nu_2,I) \quad \text{independently}$$

where $$\Delta = \mathrm{diag}\{\gamma_i\} .$$

Given values for p,ν_1, ν_2 and $\{\gamma_i\}$, two random matrices from the $W_p(\nu_1,I)$ and $W_p(\nu_2,I)$ distributions, respectively, were generated as described below and then A_1 was obtained by equating its $(i,j)^{th}$ element to $\sqrt{\gamma_i\gamma_j}$ times the $(i,j)^{th}$ element of the first random matrix,$\forall i,j$, and A_2 was obtained by equating it to the second random matrix.

The most efficient procedure for generating a random $W_p(\nu,I)$ matrix is that of Odell and Feiveson (1966), a good description of which is given by Johnson and Hegemann (1974). To apply their procedure,

$p(p-1)/2$ independent standard normal random variables $\{x_{ij}, i < j = 1, 2,\ldots,p\}$ must be generated, as well as a sequence of independent chi-square random variables $\{v_j, j = 1,\ldots,p\}$ where for each j, $v_j \sim \chi^2_{\nu-j+1}$. The random $W_p(\nu,I)$ matrix $W = (w_{ij})$ is then constructed as follows:

$$\begin{aligned}
w_{11} &= v_1 \\
w_{jj} &= v_j + \sum_{i=1}^{j-1} x_{ij}^2 && j = 2,\ldots,p \\
w_{1j} &= x_{1j}\sqrt{v_1} && j = 2,\ldots,p \\
w_{ij} &= x_{ij}\sqrt{v_i} + \sum_{k=1}^{i-1} x_{ki}\, x_{kj} && i,j = 2,\ldots,p;\ i < j. \qquad (5.5.2)
\end{aligned}$$

Subroutine RANDN, from the Witwatersrand library, an exceptionally fast routine that generates random samples from the standard normal distribution by transforming a uniform (0,1) random variable by interpolation in a table of the normal inverse probability transformation (with exact evaluation in the tails), was used to generate the x_{ij}.

The v_j were generated by first generating k uniform (0,1) random variables u_i, where k is the integer part of $\frac{1}{2}(\nu-j+1)$, and letting

$$v_j = \begin{cases} -2\log_e \prod_{i=1}^{k} u_i & \text{for } \nu - j + 1 \text{ even} \\ -2\log_e \prod_{i=1}^{k} u_i + x^2 & \text{for } \nu - j + 1 \text{ odd} \end{cases} \qquad (5.5.3)$$

where x is a random variable from the standard normal distribution. The u_i were generated by the CDC built-in mixed congruential generator RANF.

Subroutine WSHRT was written to generate random Wishart matrices as described above.

(c) Computing the Eigenvalues $\{g_i\}$ of $A_1A_2^{-1}$

Subroutine CANON (Fatti and Hawkins (1976)) was used to find the eigenvalues $\{g_i\}$ of $A_1A_2^{-1}$.. This subroutine solvesthe eigen problem:

$$(B - \lambda A)Z = 0, \qquad (5.5.4)$$

where A is a p×p symmetric, positive definite matrix (generally an error covariance matrix) and B is a p×p symmetric matrix (generally an hypothesis covariance matrix) by first obtaining the Cholesky inverse square root $A^{-\frac{1}{2}}$, where $A^{-\frac{1}{2}}$ is a real, nonsingular lower triangular matrix such that

$$A^{-\frac{1}{2}}A(A^{-\frac{1}{2}})' = I.$$

$A^{-\frac{1}{2}}$ is computed efficiently in the following manner. Note that if $X = (x_1,x_2,\ldots,x_p)'$ is a random vector with observed covariance matrix A, then, for $i = 2$ to p, the residual, y_i, on its predictor based on the least-squares regression line of x_i on $x_1,x_2,\ldots,x_{i-1}$ is uncorrelated with $x_1,x_2,\ldots,x_{i-1}$. So, if we standardize y_i to have unit variance by dividing it by the square root of the residual mean square of x_i on $x_1,x_2,\ldots,x_{i-1}$ for $i = 2$ to p, and let $y_1 = x_1/\sqrt{\mathrm{var}(x_1)}$, then $y = (y_1,y_2,\ldots,y_p)'$ has covariance matrix I, the p-dimensional identity matrix.

Clearly Y is obtained from X by the transformation:

$$Y = CX,$$

where C is a lower triangular matrix whose elements may be computed from A by performing successive pivotal sweeps on A using the diagonal elements of A as pivots, as described in Beale, Kendall, and Mann (1967).

Finally, we note that the covariance matrix of Y is

$$CAC' = I,$$

so $C = A^{-\frac{1}{2}}$.

The eigen problem

$$(A^{-\frac{1}{2}}B(A^{-\frac{1}{2}})' - \lambda I)W = 0 \qquad (5.5.5)$$

then is solved using the two subroutines TDIAG and LRVT (Sparks and Todd, 1973) and finally the matrix Z of eigenvectors of the original system (5.5.4) is obtained by transforming the W matrix:

$$Z = (A^{-\frac{1}{2}})'W.$$

(d) Cr [illegible] e five Estimators

[illegible] g_i, $i = 1,\ldots,p$, estimators $\hat{\underset{\sim}{\gamma}}^{(1)}$ and $\hat{\underset{\sim}{\gamma}}^{(2)}$ were comp in a straightforward manner from their definitions,

$$\hat{\gamma}^{(1)} = \ell_i = \frac{\nu_2}{\nu_1} g_i \qquad i = 1,\ldots,p$$

$$\hat{\gamma}_i^{(2)} = \ell_i \left(\frac{1 - \frac{p-1}{\nu_2} + \frac{1}{\nu_2} \sum\limits_{j \neq i} \frac{\ell_j}{\ell_j - \ell_i}}{1 - \frac{1}{\nu_1} \sum\limits_{j \neq i} \frac{\ell_j}{\ell_j - \ell_i}} \right) \qquad i = 1,\ldots,p$$

and then $\hat{\underset{\sim}{\gamma}}^{(3)}$ was computed from $\hat{\underset{\sim}{\gamma}}^{(2)}$ and $\hat{\underset{\sim}{\gamma}}^{(1)}$ according to definition (5.5.1).

$\underset{\sim}{\hat{\gamma}}^{(4)}$ was computed by the Newton-Raphson iterative procedure as described in Section 5.3.3.

Subroutine GRAD was written to compute the vector of first derivatives of the log likelihood function (in terms of the new parameters $\underset{\sim}{\delta}$) as given in equation (5.3.25) and subroutine HESS was written to compute the Hessian matrix whose elements are given in equations (5.3.28) and (5.3.29). Finally, the Newton-Raphson iterative procedure was carried out by subroutine UNREST, using as convergence criteria both the value of the vector of first derivatives at the previous iteration and the change in value of the log likelihood function (computed by subroutine FUNCT) over the previous two iterations.

To compute the restricted maximum marginal likelihood estimator $\underset{\sim}{\hat{\gamma}}^{(5)}$ the NAG (1975) library subroutine E04HAF was used to solve the non-linear programming problem (5.4.46). This subroutine uses a penalty function technique (Lootsma, 1972) to solve constrained minimization problems. A full description of this subroutine is given in volume 1 of the NAG manual (1975). Subroutines FUNCT, GRAD and HESS were used to compute the values of the function and its first- and second derivatives, respectively, at the various trial solutions, as required by E04HAF.

(e) Repeating and Computing Summary Statistics

One hundred simulation runs were performed for each combination of parameter values given in (a). Because of the large amount of computation required for each evaluation of $\underset{\sim}{\hat{\gamma}}^{(4)}$, the even larger amount required for $\underset{\sim}{\hat{\gamma}}^{(5)}$ and the fact that in many of the simulation runs both failed to produce meaningful results, the following procedure was adopted for each selection of parameter values:

(i) First perform 100 simulation runs, computing only $\underset{\sim}{\hat{\gamma}}^{(1)}$, $\underset{\sim}{\hat{\gamma}}^{(2)}$ and $\underset{\sim}{\hat{\gamma}}^{(3)}$, and compute summary statistics on them. Because of the efficiency of subroutine WSHRT and CANON and the small amount of computation required to obtain these three estimators, the time required for this step was fairly small.

(ii) Repeat the 100 simulation runs, this time computing $\underset{\sim}{\hat{\gamma}}^{(1)}$, $\underset{\sim}{\hat{\gamma}}^{(2)}$, $\underset{\sim}{\hat{\gamma}}^{(3)}$ and $\underset{\sim}{\hat{\gamma}}^{(4)}$ on each run. If $\underset{\sim}{\hat{\gamma}}^{(4)}$ failed on any run, then none of the estimators from that run were included in the summary statistics. If $\underset{\sim}{\hat{\gamma}}^{(4)}$ produced meaningful results, then $\underset{\sim}{\hat{\gamma}}^{(5)}$ was computed and if that too produced meaningful results all five estimators were included in the summary statistics. Otherwise none of them were included.

In this way a considerable amount of computing time was saved, since $\underset{\sim}{\hat{\gamma}}^{(5)}$, which requires by far the greatest amount of computer time, was only computed in those situations where it was likely to produce meaningful results. ($\underset{\sim}{\hat{\gamma}}^{(5)}$ very rarely produces meaningful results when $\underset{\sim}{\hat{\gamma}}^{(4)}$ does not, whereas the reverse occurs more frequently.)

The reason for performing steps (i) and (ii) above separately is twofold. Firstly, step (i) gives a larger number of runs on which to evaluate the first three estimators. (For some sets of parameter values, especially for the larger values of p, $\underset{\sim}{\hat{\gamma}}^{(4)}$ or $\underset{\sim}{\hat{\gamma}}^{(5)}$ never produced meaningful results.) Secondly, $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ are far more likely to produce meaningful results when the $\{\ell_i\} = \text{Eigs}(S_1 S_2^{-1})$ are spaced widely apart than when they are closer together, with the result that the estimators in step (ii) have a built-in bias towards larger spacing between the eigenvalues. Therefore the results from step (ii) are only useful for evaluating the *relative* performances of the five estimators.

The summary statistics for each of the estimators were computed and printed using the Witwatersrand Library's COVUP (Hawkins, 1974) and PRINT subroutines, producing mean vectors, standard deviations, covariance and correlation matrices over the various sets of simulation runs.

5.5.2 Results

Summary statistics in the form of mean vectors and vectors of standard deviations for each of the five estimators are given in Tables 5.5.1 to 5.5.4, separately for each selection of parameter values. From considerations of space and because the same conclusions seem to hold in all cases, correlation matrices are only given for the case of $p = 3$ dimensions and four combinations of the other parameter values in Table 5.5.5.

As mentioned earlier, two sets of simulation runs were performed for each selection of parameter values, only the first three estimators being computed in the first set which always consisted of a hundred runs, and all five being computed in the second set, but only on those occasions when $\hat{\underset{\sim}{\gamma}}^{(4)}$ and $\hat{\underset{\sim}{\gamma}}^{(5)}$ both produced meaningful results. The only exception occurred in the case $p = 10$ when, because of convergence problems in the nonlinear programming package E04HAF, $\hat{\underset{\sim}{\gamma}}^{(5)}$ was mostly not computed at all. Because $\hat{\underset{\sim}{\gamma}}^{(2)}$ never fails when either $\hat{\underset{\sim}{\gamma}}^{(4)}$ or $\hat{\underset{\sim}{\gamma}}^{(5)}$ produce meaningful results, $\hat{\underset{\sim}{\gamma}}^{(2)}$ and $\hat{\underset{\sim}{\gamma}}^{(3)}$ were identical (see definition (5.5.1)) for all of the simulation runs in the second set. Therefore summary statistics for $\hat{\underset{\sim}{\gamma}}^{(3)}$ are not included in Tables 5.5.1 to 5.5.5 for those simulation runs.

Failure of $\hat{\underset{\sim}{\gamma}}^{(4)}$ or $\hat{\underset{\sim}{\gamma}}^{(5)}$ to produce meaningful results can be detected when any of the δ_i assumes a large negative value. This is immediately clear from the definition of the δ_i given in expression (5.3.22) since it implies that $\hat{\gamma}_{i-1}$ and $\hat{\gamma}_i$ effectively differ only by the arbitrary constant ε_i^{-1} or, for $i = 1$, that $\hat{\gamma}_1$ is effectively equal to ε_1^{-1}. As earlier experimentation had shown that the values of $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ are unaffected by the choice of values of the ε_i over a fairly wide range (for the actual simulation runs the ε_i were chosen to be ten per cent of $(1/\hat{\gamma}_i^{(3)} - 1/\hat{\gamma}_{i-1}^{(3)})$, or for $i = 1$, of $1/\hat{\gamma}_1^{(3)}$; $\underset{\sim}{\hat{\gamma}}^{(3)}$ was also used as initial value in the maximization algorithms) a large negative value of δ_i implies that the maximisation algorithm has found a "false" maximum near one of the "inadmissible singularities" in Chang's formula (5.3.5).

Since, as is clear from Tables 5.5.1 to 5.5.4, failures of $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ occur far more frequently for smaller values of ν_1 and ν_2 and for closer separations between the g_i, it would appear that under these circumstances the likelihood surface (5.3.24) may either:

(i) have no maxima within the admissible region, or

(ii) have extremely flat maxima within the admissible region, or

(iii) have very localised maxima which may be missed by the maximization algorithms.

In order to try and establish which of the above three possibilities pertain, the subroutine FUNCT was used to evaluate the likelihood function (5.3.24) over a two-dimensional grid for the case $p = 2$ dimensions. A number of cases were tried, resulting in the following conclusions: For small enough values of ν_1 and ν_2 and sufficiently closely spaced g_i, case (i) pertains, but as the degrees of freedom and/or the spacings increase a single maximum (for the case $p = 2$, at least) develops. Case (iii) never holds.

In the remainder of this sub-section some comments are made on the results of the simulations as may be gleaned from Tables 5.5.1 to 5.5.5, under the headings of bias, standard deviation and correlation.

<u>Bias</u>:

In general, the top few eigenvalues are over-estimated and the bottom few under-estimated, although this bias is different for the different estimators. This effect decreases as the degrees of freedom ν_1 and ν_2 increase, but it is more efficient to increase them by increasing the number (k) of groups than by increasing the number (n) of observations per group, where $\nu_1 = k - 1$, $\nu_2 = k(n-1)$.

More specifically:

(1) $\hat{\underset{\sim}{\gamma}}^{(1)}$ has the greatest bias, both in the upper and lower few eigenvalues. Roughly speaking, the proportional bias in the top and bottom eigenvalues are the same.

(2) $\hat{\underset{\sim}{\gamma}}^{(2)}$ has markedly less bias than $\hat{\underset{\sim}{\gamma}}^{(1)}$, both in the upper and lower eigenvalues. For low degrees of freedom and equal separations of the γ_i, there are some anomalous results in th middle values, reflecting the relatively frequent occurrence of meaningless results amongst these values.

(3) $\hat{\underset{\sim}{\gamma}}^{(3)}$ has slightly greater bias than $\hat{\underset{\sim}{\gamma}}^{(2)}$ in the upper and lower eigenvalues, but there are no anamalies on the middle values. The difference between $\hat{\underset{\sim}{\gamma}}^{(3)}$ and $\hat{\underset{\sim}{\gamma}}^{(2)}$ virtually disappears for higher degrees of freedom and increasing separations of the eigenvalues. As mentioned earlier, in the cases where either $\hat{\underset{\sim}{\gamma}}^{(4)}$ or $\hat{\underset{\sim}{\gamma}}^{(5)}$ produce meaningfuly rasults, $\hat{\underset{\sim}{\gamma}}^{(2)}$ and $\hat{\underset{\sim}{\gamma}}^{(3)}$ are identical.

(4) $\hat{\underset{\sim}{\gamma}}_4$ has slightly less bias than $\hat{\underset{\sim}{\gamma}}^{(2)}$ (or $\hat{\underset{\sim}{\gamma}}^{(3)}$) in both the upper and lower eigenvalues. (When it produces meaningful results). The Newton-Raphson procedure (with checks to prevent the $\hat{\delta}_i$ from getting two large or too small) nearly always converges, but is unlikely to produce meaningful results for equal separations of the eigenvalues and low degrees of freedom, unless the dimension is small (p=2 or 3). For p = 10 meaningful results were only produced for increasing separations of the eigenvalues.

(5) The elements of $\hat{\underset{\sim}{\gamma}}^{(5)}$ are all smaller than the corresponding elements of $\hat{\underset{\sim}{\gamma}}^{(4)}$, the proportional differences being approximately constant. As a result, $\hat{\underset{\sim}{\gamma}}^{(5)}$ has the lowest bias of all in its top element but tends to have a slightly worse bias than $\hat{\underset{\sim}{\gamma}}^{(4)}$ and $\hat{\underset{\sim}{\gamma}}^{(2)}$ (or $\hat{\underset{\sim}{\gamma}}^{(3)}$) in its bottom one. For p = 10 the non-linear programming package E04HAF had convergence problems, with the result that values of $\hat{\underset{\sim}{\gamma}}^{(5)}$ could be computed in one case only. For p = 2, where $\hat{\underset{\sim}{\gamma}}_5$ is given explicitly by (5.4 45), the same conclusions as above hold. In this case meaningless results are characterised by imaginary solutions to (5.4.45), and as before, the frequency of their occurrence decreases as the degrees of freedom increase or when the separation between γ_1 and γ_2 increases (relative to γ_2).

Standard Deviation

(1) Whereas $\hat{\underset{\sim}{\gamma}}^{(1)}$ has the greatest bias, its standard deviations, apart from that of its top element, are generally the smallest. Using Girshick's (1939) result (see, for example Press, 1972), and the comments following Remark 5.3.1, that the $\hat{\gamma}_i^{(1)} = \ell_i$ are asymptotically independent, unbiased, normally distributed estimators of the corresponding γ_i, with standard deviations $SD(\hat{\gamma}_i^{(1)}) = \sqrt{2/(\nu_1-1)}\gamma_i$ as a reference, it is clear that for very large ν_1 and ν_2 this standard deviation is approximately correct. Otherwise, the standard deviations of the top (few) $\hat{\gamma}_i^{(1)}$ tend to be larger than $\sqrt{2/(\nu_1-1)}\gamma_i$ and those of the bottom (few) smaller. This tendency is more marked in the smaller sample sizes and when the γ_i have increasing separations.

(2) The standard deviation the top element of $\underset{\sim}{\hat{\gamma}}^{(2)}$ is usually approximately the same as that of the corresponding element of $\underset{\sim}{\hat{\gamma}}^{(1)}$ but those of the other elements are always larger. For small sample sizes some of the middle elements can have extremely large standard deviations, reflecting the frequency of occurrence of meaningless results amongst them.

(3) The standard deviation of $\underset{\sim}{\hat{\gamma}_1}^{(3)}$ is sometimes slightly less than that of $\underset{\sim}{\hat{\gamma}_1}^{(1)}$ whereas those of the other elements of $\underset{\sim}{\hat{\gamma}}^{(3)}$ are always slightly larger than those of their counterparts in $\underset{\sim}{\hat{\gamma}}^{(1)}$

(4) The standard deviations of $\underset{\sim}{\hat{\gamma}}^{(4)}$ are slightly, but consistently larger than those of their counterparts in $\underset{\sim}{\hat{\gamma}}^{(2)}$ (or $\underset{\sim}{\hat{\gamma}}^{(3)}$) but that of $\hat{\gamma}_1^{(4)}$ may still sometimes be smaller than that of $\hat{\gamma}_1^{(1)}$.

(5) The standard deviations of $\underset{\sim}{\hat{\gamma}}^{(5)}$ are always smaller than the corresponding ones of $\underset{\sim}{\hat{\gamma}}^{(4)}$ and sometimes even smaller than those of $\underset{\sim}{\hat{\gamma}}^{(2)}$ (or $\underset{\sim}{\hat{\gamma}}^{(3)}$). $\hat{\gamma}_1^{(5)}$ frequently has the smallest standard deviation of all the estimators of γ_1. This confirms that the reduction in standard deviations (especially of the estimates of the top eigenvalue) suggested in Sub-section 5.4.6 by expression (5.4.49) and Table 5.4.1 for the case where the constraints are deterministic, is at least partially realised in our situation, where the constraints are stochastic. For the case p = 2, $\hat{\gamma}_1^{(5)}$ always has the smallest standard deviation, and that of $\hat{\gamma}_2^{(5)}$ is always larger than that of $\hat{\gamma}_2^{(1)}$ but smaller than those of the rest.

Correlation

The correlation coefficients in Table 5.5.5 were computed only from those simulation runs in which $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ produced meaningful results, and therefore, because of the built-in bias towards larger spacings between the eigenvalues resulting from this, these correlations have to be treated with some caution. Nevertheless certain trends are clearly evident:

(i) For any estimator $\hat{\underset{\sim}{\gamma}}^{(k)}$, the correlation coefficient between $\hat{\gamma}_i^{(k)}$ and $\hat{\gamma}_j^{(k)}$, $j \neq i$, can be quite large, especially for adjacent pairs, but it tends to decrease as the degrees of freedom are increased. Increasing the separation between γ_i and γ_j tends, however, to eliminate this correlation completely.

(ii) The correlation coefficients are appreciably smaller for $\hat{\underset{\sim}{\gamma}}^{(2)}$ (or $\hat{\underset{\sim}{\gamma}}^{(3)}$) than for $\hat{\underset{\sim}{\gamma}}^{(1)}$ and slightly smaller again for $\hat{\underset{\sim}{\gamma}}^{(4)}$, although there is generally little difference between those of $\hat{\underset{\sim}{\gamma}}^{(4)}$ and $\hat{\underset{\sim}{\gamma}}^{(5)}$.

5.5.3 Conclusions

Going back to the expression for the distribution of δ_{ij}^2 given in Theorem 3.1.1

$$\delta_{ij}^2 \sim 2 \sum_{s=1}^{r} \lambda_s \nu_s$$

where

$$\nu_s \sim \chi_1^2 \ , \quad \text{independently}, \quad s = 1,\ldots,r$$

$$\lambda_s = \tfrac{1}{n}(\gamma_s - 1) \qquad\qquad s = 1,\ldots,r$$

$$\gamma_r > \gamma_{r-1} > \ldots > \gamma_1 > 1$$

and

$$r = r(T)$$

it is clear that γ_1, being the largest, will have the greatest influence on the distribution, and γ_r the smallest.

From this point of view therefore, $\underset{\sim}{\hat{\gamma}}^{(5)}$ is the best estimator, since $\hat{\gamma}_1^{(5)}$ has the lowest bias and often has the lowest standard deviation amongst the five estimators. The drawback to this estimator is that, apart from the case p = 2, it requires a nonlinear programming algorithm for its evaluation and frequently produces meaningless results. Moreover, for large values of p it may be difficult to obtain convergence of the nonlinear program (although other algorithms may give better performance than E04HAF).

Next in line is $\underset{\sim}{\hat{\gamma}}^{(4)}$, its only advantages over $\underset{\sim}{\hat{\gamma}}^{(5)}$ being that it occasionally produces meaningful results when the latter does not, and that (for dimensions up to 10, at least) it does not have convergence problems.

$\underset{\sim}{\hat{\gamma}}^{(3)}$ is perhaps the most practical of all the estimators, being simple to compute and, by definition, never producing meaningless results. In terms of bias, it is a considerable improvement over $\underset{\sim}{\hat{\gamma}}^{(1)}$ and not much worse than $\underset{\sim}{\hat{\gamma}}^{(4)}$ or $\underset{\sim}{\hat{\gamma}}^{(5)}$. A regards spread, its standard deviations are not much larger than those of $\underset{\sim}{\hat{\gamma}}^{(1)}$ (the standard deviation for $\hat{\gamma}_1^{(3)}$ can in fact, be smaller than that of $\hat{\gamma}_1^{(1)}$) whereas they are always slightly smaller than those of $\underset{\sim}{\hat{\gamma}}^{(4)}$ and are often even smaller than those of $\underset{\sim}{\hat{\gamma}}^{(5)}$.

As $\underset{\sim}{\hat{\gamma}}^{(3)}$ retains all of the good points of $\underset{\sim}{\hat{\gamma}}^{(2)}$ and circumvents the problem of its unreliability, there is no reason for preferring the latter. Because of its large bias $\underset{\sim}{\hat{\gamma}}^{(1)}$ should not be used.

If the programs are available and computer time no object, the following practical procedure for estimating $\underset{\sim}{\gamma}$ is recommended:

(1) Compute $\{\ell_i\} = \mathrm{Eigs}(S_1S_2^{-1})$ and hence $\underset{\sim}{\hat{\gamma}}^{(2)}$ from formula (5.3.20).

(2) If $\hat{\underset{\sim}{\gamma}}^{(2)}$ does not give meaningful results use $\hat{\underset{\sim}{\gamma}}^{(3)}$ as defined by (5.5.1) as estimator of $\underset{\sim}{\gamma}$.

(3) If $\hat{\underset{\sim}{\gamma}}^{(2)}$ does give meaningful results, compute $\hat{\underset{\sim}{\gamma}}^{(5)}$ and use this as estimator if it gives meaningful results. If it does not, compute $\hat{\underset{\sim}{\gamma}}^{(4)}$ and if that also does not give meaningful results, go back to $\hat{\underset{\sim}{\gamma}}^{(2)}$.

Remark 5.5.1 It is interesting that, even when the likelihood function apparently has no maximum outside the "inadmissible" regions, the approximate solution to the maximum marginal likelihood equations, $\hat{\underset{\sim}{\gamma}}^{(2)}$, is a better estimator than $\hat{\underset{\sim}{\gamma}}^{(1)}$, and if it does not produce meaningful results then $\hat{\underset{\sim}{\gamma}}^{(3)}$ is still usually better than $\hat{\underset{\sim}{\gamma}}^{(1)}$.

Remark 5.5.2 It is clear from the results of the simulations that for reliable estimation the number of populations, k, needs to be large, preferably at least ten times the number of dimensions, p. If there is a choice, it is generally better to increase k than it is to increase n, the number of observations per group (so long as n is at least equal to 2).

Appendix 5.1 Proof of Theorem 5.1

We will consider the more general case with (possibly) different sample sizes from *each of* the k groups. i.e. our training sample is: $\{x_{ij};\ j = 1,\ldots,n_i;\ \ i = 1,\ldots,k\}$. Then, *analogously to* Table 5.1.1, we define:

$$A_1 = \sum_{i=1}^{k} n_i(x_{i.}-x_{..})(x_{i.}-x_{..})'$$

and $A_2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij}-x_{i.})(x_{ij}-x_{i.})'$

where

$$x_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$x_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{N} \sum_{j=1}^{k} n_i x_{i.}$$

and

$$N = \sum_{i=1}^{k} n_i.$$

Therefore,

$$x_{i.-(i,j)} = \frac{n_i x_{i.} - x_{ij}}{n_i - 1}$$

$$= x_{i.} - \frac{x_{ij} - x_{i.}}{n_i - 1}$$

$$= x_{i.} - e$$

Similarly,

$$\underset{\sim}{x}_{..-(i,j)} = \frac{Nx_{..}-x_{ij}}{N-1}$$

$$= x_{..} - \frac{x_{ij}-x_{..}}{N-1}$$

$$= x_{..} - f.$$

Applying the above two results, we obtain,

$$A_{i-(i,j)} = \sum_{\ell \neq i}^{k} n_\ell (x_{\ell.}-x_{..-(i,j)})(x_{\ell.}-x_{..-(i,j)})'$$

$$+ (n_i-1)(x_{i.-(i,j)} - x_{..-(i,j)})(x_{i.-(i,j)} - x_{..-(i,j)})'$$

$$= \sum_{\ell \neq i} n_\ell (x_{\ell.}-x_{..}+f)(x_{\ell.}-x_{..}+f)'$$

$$+ (n_i-1)(x_{i.}-x_{..}+f-e)(x_{i.}-x_{..}+f-e)'$$

$$= A_i + f\sum_{\ell \neq i} n_\ell (x_{\ell.}-x_{..})' + \sum_{\ell \neq i} n_\ell (x_{\ell.}-x_{..})f' + (N-n_i)ff'$$

$$- (x_{i.}-x_{..})(x_{i.}-x_{..})' + (n_i-1)(f-e)(x_{i.}-x_{..})'$$

$$+ (n_i-1)(x_{i.}-x_{..})(f-e)' + (n_i-1)(f-e)(f-e)'$$

$$= A_i - n_i f(x_i-x_{..})' - n_i(x_{i.}-x_{..})f' + (N-n_i)ff'$$

$$- (N-n_i)f(x_{i.}-x_{..})' + (n_i-1)(x_{i.}-x_{..})(f-e)' + (n_i-1)(f-e)(f-e)'$$

since

$$(n_i-1)(f-e) - (x_{i.}-x_{..}) = -(N-n_i)f$$

and

$$\sum_{\ell=1}^{k} n_\ell(x_{\ell.}-x_{..}) = 0$$

$$= A_1 - Nfg' - g(f+(n_i-1)e)' + (N-n_i)ff' + (n_i-1)(f-e)(f-e)'$$

$$= A_1 + G.$$

Furthermore

$$A_{2-(i,j)} = \sum_{\ell\neq i}^{k} \sum_{j=1}^{n_\ell} (x_{\ell j}-x_{\ell.})(x_{\ell j}-x_{\ell.})'$$

$$+ \sum_{r\neq j}^{n_i} (x_{ir}-x_{i.-(i,j)})(x_{ir}-x_{i.-(i,j)})'$$

$$= \sum_{\ell\neq i}^{k} \sum_{j=1}^{n_\ell} (x_{\ell j}-x_{\ell.})(x_{\ell j}-x_{\ell.})'$$

$$+ \sum_{r\neq j}^{n_i} (x_{ir}-x_{i.}+e)(x_{ir}-x_{i.}+e)'$$

$$= A_2 - (x_{ij}-x_{i.})(x_{ij}-x_{i.})' - e(x_{ij}-x_{i.})' - (x_{ij}-x_{i.})e'$$

$$+ (n_i-1)ee'$$

since

$$\sum_{r=1}^{n_i} (x_{ir}-x_{i.}) = 0$$

$$= A_2 - (n_i-1)^2ee' - (n_i-1)ee' - (n_i-1)ee' + (n_i-1)ee'$$

$$= A_2 - n_i(n_i-1)ee'$$

which agrees with Lachenbruch's (1967) result.

Now, applying the Binomial inverse theorem (Press, 1972):

$$(A+UBV)^{-1} = A^{-1} - A^{-1}UB(B+BVA^{-1}UB)^{-1}BVA^{-1}$$

which reduces to the following, for u and v column vectors and B = I:

$$(A+uv')^{-1} = A^{-1} - A^{-1}uv'A^{-1}/(1+v'A^{-1}u) ,$$

to the above expression for $A^{-1}_{2-(i,j)}$, we get:

$$A^{-1}_{2-(i,j)} = (A_2-n_i(n_i-1)ee')^{-1}$$

$$= A_2^{-1} + n_i(n_i-1)A_2^{-1}ee'A_2^{-1}/(1-n_i(n_i-1)e'A_2^{-1}e)$$

$$= A_2^{-1} + A_2^{-1}E.$$

So

$$A_{1-(i,j)}A^{-1}_{2-(i,j)} = A_1A_2^{-1} + A_1A_2^{-1}E + GA_2^{-1}(I+E)$$

Whence

$$\mathrm{Tr}(A_1A_2^{-1})_{-(i,j)} = \mathrm{Tr}(A_1A_2^{-1}) + \mathrm{Tr}(A_1A_2^{-1}E) + \mathrm{Tr}(GA_2^{-1}F).$$

<u>Remark</u> When $n_i = n$, V_i, we just remove the subscripts from all the n_i's appearing in the above formulae.

Table 5.5.1

Means and Standard Deviations of the five estimators of the $\{\gamma_i\} = \mathrm{Eigs}(\Sigma_1\Sigma^{-1})$ from the simulation experiments for the case $p = 2$

A. Degrees of Freedom $\nu_1 = 10$, $\nu_2 = 44$.

A.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
4	6.76	5.90	6.07	4.21	4.10	4.05
2	1.20	-3.46	1.56	1.12	189.41	1.29
8	11.89	10.78	10.78	7.99	7.67	7.67
2	1.41	2.36	1.85	1.26	2.74	1.40

A.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
53	4	8.53	7.90	7.73	6.97	4.80	4.52	4.53	4.04
	2	0.65	0.91	0.96	0.83	0.44	0.62	0.66	0.57
66	8	14.44	13.44	13.23	11.91	8.38	7.94	7.97	7.11
	2	0.89	1.24	1.30	1.12	0.57	0.81	0.88	0.74

B. Degrees of Freedom $\nu_1 = 10$, $\nu_2 = 44$

B.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
4	5.24	4.83	4.85	2.24	2.21	2.19
2	1.46	1.86	1.73	0.82	1.28	1.00
8	9.65	9.13	9.18	4.63	4.60	4.55
2	1.67	2.06	1.92	1.05	1.57	1.20

B.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
65	4	5.81	5.15	5.44	5.22	2.36	2.29	2.29	2.18
	2	1.08	1.26	1.29	1.21	0.51	0.60	0.64	0.58
83	8	10.38	9.95	9.89	9.45	4.67	4.56	4.56	4.34
	2	1.36	1.56	1.58	1.49	0.74	0.87	0.89	0.83

C. Degrees of Freedom $\nu_1 = 20$, $\nu_2 = 84$

C.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, and $\hat{\gamma}^{(3)}$ from all 100 simulations

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
4	4.50	4.21	4.26	1.41	1.46	1.42
2	1.71	2.09	1.87	0.70	1.47	0.75
8	8.57	8.31	8.31	2.93	2.93	2.92
2	1.82	1.99	1.96	0.78	0.92	0.86

C.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
60	4	4.91	4.72	4.68	4.59	1.42	1.42	1.41	1.37
	2	1.42	1.55	1.57	1.51	0.48	0.53	0.55	0.53
95	8	8.71	8.47	8.44	8.24	2.91	2.88	2.90	2.82
	2	1.73	1.86	1.87	1.82	0.64	0.70	0.72	0.69

D. Degrees of Freedom $\nu_1 = 40$, $\nu_2 = 164$

D.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations .

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
4	4.15	4.00	4.00	0.96	0.98	0.98
2	1.88	2.01	2.00	0.53	0.60	0.59
8	8.06	7.92	7.92	2.02	2.01	2.01
2	1.95	2.03	2.03	0.56	0.60	0.60

D.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
81	4	4.33	4.20	4.17	4.14	0.93	0.93	0.94	0.92
	2	1.78	1.87	1.89	1.86	0.47	0.51	0.53	0.51
100	8	8.06	7.92	7.91	7.82	2.02	2.01	2.02	1.99
	2	1.95	2.03	2.03	2.01	0.57	0.60	0.60	0.59

Table 5.5.2

Means and Standard Deviations of the five estimators of the $\{\gamma_i\} = \text{Eigs}\{\Sigma_1\Sigma^{-1}\}$ from the simulation experiments for p = 3 dimensions

A. Degrees of Freedom $\nu_1 = 6$, $\nu_2 = 28$

A.1. Estimators $\underset{\sim}{\hat{\gamma}}^{(1)}$, $\underset{\sim}{\hat{\gamma}}^{(2)}$ and $\underset{\sim}{\hat{\gamma}}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
6	9.61	7.78	8.09	4.99	4.62	4.49
4	3.19	4.68	3.59	1.67	5.76	1.97
2	0.78	2.04	1.07	0.57	3.80	0.76
16	20.63	17.94	18.15	13.14	12.33	12.11
4	3.87	4.63	4.47	1.95	3.29	2.44
2	0.85	1.31	1.29	0.63	7.72	0.94

A.2. Estimators $\underset{\sim}{\hat{\gamma}}^{(1)}$, $\underset{\sim}{\hat{\gamma}}^{(2)}$, $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ from n simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
9	6	18.06	16.01	15.60	14.50	2.91	2.55	2.49	2.31
	4	3.07	3.62	3.66	3.36	0.91	1.11	1.17	1.07
	2	0.40	0.63	0.65	0.59	0.13	0.21	0.23	0.20
34	16	28.33	25.36	24.81	23.09	13.45	12.38	12.60	11.45
	4	3.81	4.45	4.51	4.13	1.56	1.86	1.94	1.74
	2	0.53	0.84	0.89	0.79	0.33	0.55	0.63	0.53

B. Degrees [illegible] $\nu_1 = 15, \nu_2 = 64$

B.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
6	7.68	6.70	6.83	2.98	2.94	2.89
4	3.53	4.12	3.71	1.29	2.44	1.51
2	1.37	1.85	1.62	0.53	0.86	0.70
16	17.89	16.71	16.74	8.28	8.13	8.09
4	4.04	4.30	4.21	1.58	2.21	1.84
2	1.44	1.87	1.76	0.56	0.83	0.75

B.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
11	6	9.47	8.70	8.51	8.32	2.53	2.39	2.39	2.29
	4	3.28	3.50	3.52	3.39	0.89	1.01	1.09	1.05
	2	0.93	1.15	1.19	1.13	0.32	0.41	0.44	0.41
47	16	19.07	17.89	17.74	17.24	7.74	7.45	7.48	7.22
	4	4.50	4.74	4.72	4.56	1.41	1.53	1.56	1.50
	2	1.22	1.48	1.53	1.46	0.52	0.66	0.69	0.65

C. Degrees of Freedom $\nu_1 = 30, \nu_2 = 124$

C.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviati		
True γ	1	2	3	1	2	3
6	6.75	6.09	6.20	1.65	1.71	1.63
4	3.84	4.31	3.99	1.01	2.13	1.12
2	1.64	1.90	1.82	0.52	0.68	0.62
16	16.15	15.56	15.56	4.68	4.63	4.63
4	4.28	4.33	4.34	1.22	1.36	1.34
2	1.67	1.94	1.87	0.53	0.81	0.62

C.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
25	6	8.00	7.52	7.42	7.34	1.71	1.64	1.63	1.60
	4	3.70	3.83	3.85	3.77	0.86	0.92	0.95	0.93
	2	1.47	1.65	1.68	1.63	0.38	0.44	0.46	0.44
74	16	16.74	16.13	16.08	15.84	4.43	4.39	4.41	4.33
	4	4.60	4.71	4.70	4.62	1.12	1.20	1.23	1.21
	2	1.56	1.73	1.75	1.72	0.45	0.51	0.52	0.51

D. <u>Degrees of Freedom $\nu_1 = 60$, $\nu_2 = 244$</u>

D.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
6	6.52	6.10	6.19	1.31	1.42	1.33
4	4.07	4.32	4.18	0.76	1.05	0.83
2	1.86	2.00	1.96	0.39	0.45	0.40
16	16.56	16.26	16.26	3.65	3.63	3.63
4	4.25	4.26	4.26	[illegible]	0.89	0.88
2	1.88	2.01	1.99	[illegible]	0.46	0.43

D.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
44	6	7.42	7.14	7.10	7.06	1.07	1.06	1.07	1.06
	4	3.83	3.90	3.91	3.87	0.55	0.58	0.60	0.59
	2	1.77	1.89	1.89	1.87	0.30	0.33	0.33	0.33
91	16	16.75	16.44	16.43	16.30	3.61	3.59	3.59	3.66
	4	4.34	4.37	4.36	4.32	0.81	0.84	0.85	0.84
	2	1.82	1.93	1.94	1.92	0.36	0.39	0.40	0.40

Table 5.5.3

Means and Standard Deviations of the five estimators of the $\{\gamma_i\} = \mathrm{Eigs}\{\Sigma_1\Sigma^{-1}\}$ from the simulation experiments for $p = 5$ dimensions.

A. Degrees of Freedom $\nu_1 = 10$, $\nu_2 = 44$

A:1: Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
10	17.86	13.11	14.83	9.38	8.72	8.10
8	8.73	7.91	8.52	3.13	15.65	3.32
6	4.27	13.58	4.60	1.71	69.15	2.12
4	1.98	8.09	2.29	0.81	49.28	1.12
2	0.72	1.79	0.90	0.42	3.50	0.64
32	42.94	33.84	36.40	25.86	24.33	22.70
16	16.15	11.00	16.06	7.22	53.48	7.69
8	6.36	8.76	6.93	3.27	17.91	4.10
4	2.51	1.86	2.90	1.07	26.59	1.47
2	0.83	1.81	0.99	0.51	4.77	0.64

A.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

Failure of either $\hat{\gamma}^{(4)}$ or $\hat{\gamma}^{(5)}$ in all simulations.

B. Degrees of Freedom $\nu_1 = 25$, $\nu_2 = 104$

B.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
10	13.97	1.35	12.21	3.08	3.27	2.77
8	9.07	11.50	8.91	1.92	21.52	2.13
6	5.46	6.15	5.64	1.22	2.08	1.46
4	3.18	4.38	3.44	0.80	2.81	0.98
2	1.43	2.02	1.62	0.43	1.15	0.59
32	37.34	33.19	33.60	9.89	9.93	9.64
16	17.81	18.22	17.76	5.01	7.20	5.62
8	7.96	8.85	8.47	2.32	6.46	2.87
4	3.59	4.32	4.02	1.00	1.62	1.16
2	1.48	1.96	1.78	0.46	0.72	0.60

B.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.
For $\{\gamma_i\} = \{10,8,6,4,2\}$ either $\hat{\gamma}^{(4)}$ or $\hat{\gamma}^{(5)}$ failed in all simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
	32	44.76	41.12	40.76	40.12	3.87	3.53	3.54	3.50
	16	16.97	16.76	16.63	16.28	3.07	3.26	3.40	3.32
6	8	7.59	8.14	8.11	7.93	1.18	1.30	1.28	1.25
	4	3.17	3.70	3.77	3.67	1.01	1.27	1.40	1.35
	2	1.06	1.34	1.36	1.32	0.41	0.54	0.56	0.54

C. Degrees of Freedom $\nu_1 = 50$, $\nu_2 = 204$

C.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
10	12.38	10.88	11.38	2.66	2.94	2.59
8	8.28	8.44	8.11	1.50	2.97	1.63
6	5.66	6.26	5.78	0.98	4.83	1.08
4	3.61	4.21	3.81	0.68	1.24	0.77
2	1.75	2.09	1.87	0.35	0.82	0.41
32	35.19	33.21	33.24	9.25	9.29	9.23
16	16.44	16.39	16.42	3.41	3.85	3.74
8	7.99	8.34	8.25	1.73	2.25	2.04
4	3.94	4.30	4.22	0.86	1.18	1.01
2	1.78	2.04	2.00	0.36	0.50	0.43

C.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations. For $\{\gamma_i\} = \{10,8,6,4,2\}$ either $\hat{\gamma}^{(4)}$ or $\hat{\gamma}^{(5)}$ failed in all simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
	32	38.04	36.11	35.89	35.60	9.87	9.79	9.88	9.77
	16	16.81	16.76	16.79	16.62	2.99	3.18	3.29	3.26
36	8	7.89	8.13	8.12	8.03	1.45	1.61	1.67	1.65
	4	3.62	4.11	4.13	4.08	0.68	0.78	0.82	0.81
	2	1.69	1.91	1.93	1.90	0.37	0.43	0.44	0.43

D. Degrees of Freedom $\nu_1 = 100$, $\nu_2 = 404$

D.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
10	11.26	10.43	10.64	1.63	1.81	1.64
8	8.04	8.00	7.93	0.95	1.44	1.05
6	5.88	6.19	6.01	0.78	1.12	0.86
4	3.74	4.02	3.88	0.65	0.86	0.69
2	1.90	2.05	2.00	0.27	0.30	0.30
32	33.64	32.56	32.71	5.83	6.21	5.82
16	15.92	16.54	15.84	2.40	7.81	2.56
8	8.29	8.50	8.46	1.34	1.55	1.45
4	3.85	3.98	3.98	0.68	0.75	0.75
2	1.92	2.05	2.04	0.27	0.30	0.30

D.2. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
	10	12.65	12.07	12.01	11.98	1.41	1.42	1.44	1.43
	8	7.95	7.84	7.81	7.78	0.56	0.56	0.55	0.55
8	6	5.37	5.50	5.52	5.49	0.37	0.39	0.40	0.40
	4	3.25	3.40	3.40	3.37	0.41	0.47	0.50	0.49
	2	1.86	2.00	2.01	2.00	0.26	0.30	0.31	0.30
	32	33.87	32.92	32.87	32.73	5.32	5.34	5.36	5.33
	16	15.97	15.91	15.90	15.82	1.96	2.06	2.08	2.07
85	8	8.14	8.28	8.28	8.24	1.28	1.38	1.41	1.40
	4	3.91	4.05	4.05	4.03	0.69	0.76	0.78	0.77
	2	1.90	2.02	2.03	2.02	0.26	0.29	0.30	0.29

E. Degrees of Freedom $\nu_1 = 60$, $\nu_2 = 459$

Estimators of $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
	32	36.66	35.27	35.10	34.99	7.15	7.18	7.25	7.21
	16	16.70	16.79	16.80	16.72	3.10	3.37	3.50	3.49
53	8	7.91	[illegible]	8.23	8.18	1.35	1.50	1.56	1.55
	4	3.89	4.19	4.21	4.18	0.62	0.71	0.73	0.72
	2	1.72	1.94	1.95	1.94	0.39	0.46	0.47	0.46

Table 5.5.4

Means and Standard Deviations of the five estimators of the $\{\gamma_i\} = Eigs(\Sigma_1\Sigma^{-1})$ from the simulation experiments for p = 10 dimensions

A. Degrees of Freedom $\nu_1 = 20$, $\nu_2 = 84$

A.1. Estimators $\hat{\underset{\sim}{\gamma}}^{(1)}$, $\hat{\underset{\sim}{\gamma}}^{(2)}$ and $\hat{\underset{\sim}{\gamma}}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
20	41.19	28.01	33.83	9.25	9.09	7.49
18	27.22	23.75	25.25	5.52	10.05	6.08
16	18.68	18.26	18.14	3.88	10.07	4.01
14	13.50	10.82	13.49	2.39	46.23	2.45
12	9.60	12.59	9.66	2.17	11.57	2.25
10	6.70	10.08	6.74	1.47	20.27	1.49
8	4.50	6.84	4.52	1.00	16.42	1.01
6	2.79	4.46	2.81	0.82	9.13	0.83
4	1.63	3.83	1.65	0.57	4.56	0.60
2	0.75	1.66	0.76	0.34	2.24	0.41
1024	1352.	1108.	1141.	427.5	398.2	376.0
512	564.2	368.8	527.7	188.5	1545.	199.3
256	246.3	260.3	247.0	73.97	143.1	84.20
128	120.8	127.9	127.2	36.43	144.8	41.82
64	54.50	67.03	58.30	17.38	29.58	19.16
32	24.28	30.25	26.86	8.51	37.38	10.58
16	11.43	17.29	12.99	3.52	8.44	5.06
8	4.94	6.88	5.63	1.77	11.10	2.34
4	2.25	5.37	2.54	0.81	9.63	1.09
2	0.88	2.34	1.00	0.38	2.99	0.49

A.2. Estimators $\hat{\underset{\sim}{\gamma}}^{(1)}$, $\hat{\underset{\sim}{\gamma}}^{(2)}$, $\hat{\underset{\sim}{\gamma}}^{(4)}$ and $\hat{\underset{\sim}{\gamma}}^{(5)}$ from n simulations.

Failure of both $\hat{\underset{\sim}{\gamma}}^{(4)}$ or $\hat{\underset{\sim}{\gamma}}^{(5)}$ in all simulations.

B. Degrees of Freedom $\nu_1 = 50$, $\nu_2 = 204$

B.1. Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True γ	1	2	3	1	2	3
20	29.90	23.84	26.37	4.77	5.31	4.57
18	22.10	19.89	21.00	3.00	4.97	3.37
16	16.88	15.40	16.65	2.00	3.77	2.17
14	13.62	18.42	13.61	1.89	29.14	1.97
12	10.59	10.[illegible]	10.60	1.35	8.75	1.36
10	8.25	9.72	8.27	1.18	7.75	1.18
8	6.19	8.04	6.21	0.87	3.53	0.89
6	4.45	6.27	4.46	0.75	3.27	0.75
4	2.90	3.81	2.90	0.59	1.81	0.59
2	1.57	2.19	1.58	0.38	0.71	0.38
1024	1139.	1044.	1048.	261.6	257.7	255.8
512	547.0	534.0	534.1	124.0	144.1	133.1
256	256.7	260.1	257.1	52.45	65.23	57.01
128	122.9	127.5	126.5	25.14	31.28	28.68
64	59.26	63.70	62.21	12.90	18.48	14.96
32	28.83	31.92	30.66	6.15	8.65	6.78
16	14.30	16.24	15.58	3.20	4.47	3.97
8	6.95	8.18	7.73	1.34	2.01	1.71
4	3.30	4.02	3.73	0.78	1.27	0.9[illegible]
2	1.64	2.16	1.96	0.39	0.61	0.59

B.2 Estimators $\underset{\sim}{\hat{\gamma}}^{(1)}$, $\underset{\sim}{\hat{\gamma}}^{(2)}$, $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ from n simulations.
For $\{\gamma_i\} = \{20,18,16,14,12,10,8,6,4,2\}$, $\underset{\sim}{\hat{\gamma}}^{(4)}$ and $\underset{\sim}{\hat{\gamma}}^{(5)}$ failed in all simulations.

		Means				Standard Deviations			
n	True $\underset{\sim}{\gamma}$	1	2	4	5	1	2	4	5
9	1024	1132.	1042.	1036.	Algorithm failed to converge	202.5	193.7	195.6	Algorithm failed to converge
	512	553.6	537.3	537.7		78.07	80.61	83.19	
	256	260.0	260.2	260.5		27.48	29.47	30.62	
	128	125.9	130.0	130.1		14.13	15.83	16.51	
	64	58.37	61.87	61.89		9.68	11.13	11.58	
	32	27.43	29.71	29.61		3.87	4.50	4.66	
	16	14.36	16.31	16.43		1.53	1.84	1.92	
	8	6.50	7.51	7.48		1.25	1.60	1.67	
	4	3.18	3.83	3.84		0.31	0.39	0.40	
	2	1.57	2.02	2.04		0.34	0.46	0.48	

C. Degrees of Freedom $\nu_1 = 100$, $\nu_2 = 404$

C.1 Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True $\underset{\sim}{\gamma}$	1	2	3	1	2	3
20	25.62	22.10	23.10	3.02	3.56	2.93
18	20.23	18.69	19.17	2.13	3.83	2.50
16	16.59	16.24	16.27	1.59	3.32	1.87
14	13.69	14.32	13.69	1.33	3.20	1.56
12	11.08	11.79	11.13	1.18	2.46	1.26
10	8.83	9.62	8.88	0.93	2.39	0.93
8	7.10	8.46	7.15	0.82	2.04	0.83
6	5.06	5.77	6.12	0.67	1.12	0.72
4	3.47	4.06	3.51	0.52	0.83	0.54
2	1.75	2.01	1.77	0.28	0.36	0.29
1024	1093.	1048.	1048.	175.3	175.2	175.2
512	536.0	528.9	528.5	84.49	90.02	80.32
256	254.4	254.4	254.4	39.45	42.03	42.02
128	126.8	128.9	128.9	18.13	19.75	19.75
64	60.97	62.72	62.71	9.20	10.21	10.22
32	30.43	31.81	31.80	5.26	5.99	5.99
16	15.27	16.19	16.19	2.24	2.55	2.55
8	7.28	7.81	7.81	1.23	1.44	1.44
4	3.68	4.05	4.02	0.61	0.74	0.71
2	1.79	2.01	2.00	0.29	0.35	0.34

C.2 Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

For $\{\gamma_i\}$ = (20,18,16,14,12,10,8,6,4,2) $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ failed in all simulations.

		Means				Standard Deviations			
n	True $\underset{\sim}{\gamma}$	1	2	4	5	1	2	4	5
75	1024	1109.	1064.	1062.	Algorithm failed to converge	174.4	172.2	173.4	Algorithm failed to converge
	512	526.8	518.4	518.7		80.72	83.94	85.61	
	256	250.3	250.1	250.2		37.79	36.91	37.64	
	128	125.6	127.6	127.7		17.91	19.32	19.64	
	64	61.46	63.27	63.29		8.90	9.77	9.99	
	32	30.24	31.56	31.59		4.84	5.45	5.60	
	16	15.37	16.31	16.33		2.25	2.56	2.60	
	8	7.28	7.82	7.82		1.16	1.33	1.36	
	4	3.66	4.01	4.01		0.50	0.58	0.59	
	2	1.77	1.99	1.99		0.28	0.32	0.33	

D. Degrees of Freedom ν_1 = 200, ν_2 = 804

D.1 Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and $\hat{\gamma}^{(3)}$ from all 100 simulations.

	Means			Standard Deviations		
True $\underset{\sim}{\gamma}$	1	2	3	1	2	3
20	23.29	21.23	21.78	2.05	2.57	2.09
18	19.02	18.08	18.42	1.44	2.06	1.67
16	16.21	16.21	16.02	1.08	2.96	1.23
14	13.62	14.53	13.64	1.00	2.93	1.11
12	11.38	11.73	11.45	0.87	1.64	0.96
10	9.36	9.95	9.44	0.77	1.96	0.86
8	7.47	8.05	7.56	0.70	1.19	0.74
6	5.56	5.97	5.64	0.51	0.68	0.57
4	3.74	4.03	3.80	0.44	0.55	0.49
2	1.85	1.97	1.87	0.21	0.23	0.22
1024	1070.	1048.	1048.	124.3	124.3	124.3
512	519.3	515.1	515.1	57.61	59.44	59.44
256	256.1	256.1	256.1	29.31	30.49	30.49
128	125.7	126.7	126.7	13.98	14.82	14.82
64	62.27	63.14	63.14	7.18	7.58	7.58
32	31.21	31.88	31.88	3.33	3.54	3.54
16	15.45	15.89	15.89	1.65	1.76	1.76
8	7.71	7.98	7.98	0.79	0.85	0.85
4	3.84	4.01	4.01	0.47	0.51	0.51
2	1.87	1.98	1.98	0.21	0.23	0.23

D.2 Estimators $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

For $\{\gamma_i\} = \{20,18,16,14,12,10,8,6,4,2\}$ $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ failed in all simulations.

		Means				Standard Deviations			
n	True γ	1	2	4	5	1	2	4	5
5	1024	1121.	1097.	1096.	1094.	143.4	143.2	143.5	143.1
(out	512	567.2	563.3	563.3	561.9	47.88	50.53	50.87	50.74
of 5)	256	267.0	267.3	267.3	266.6	38.55	39.83	39.96	39.85
	128	121.1	121.8	121.8	121.6	11.95	12.45	12.49	12.46
	64	60.18	60.97	60.97	60.82	8.04	8.45	8.49	8.47
	32	31.66	32.37	32.38	32.29	3.92	4.15	4.17	4.16
	16	15.11	15.53	15.53	15.49	1.12	1.20	1.20	1.20
	8	7.30	7.53	7.53	7.51	0.58	0.62	0.63	0.63
	4	4.17	4.38	4.38	4.37	0.47	0.52	0.53	0.52
	2	1.89	2.00	2.00	2.00	0.31	0.33	0.34	0.33
98	1024	1074.	1052.	1051.	Algorithm took too long to converge	122.3	122.1	122.5	Algorithm took too long to converge
	512	518.8	514.4	514.5		56.04	57.60	58.13	
	256	256.7	256.7	256.7		29.27	30.40	30.54	
	128	125.2	126.1	126.1		13.09	13.75	13.86	
	64	62.37	63.26	63.26		7.17	7.58	7.62	
	32	31.27	31.95	31.95		3.33	3.55	3.57	
	16	15.43	15.87	15.87		1.63	1.75	1.75	
	8	7.70	7.97	7.97		0.79	0.86	0.86	
	4	3.85	4.02	4.02		0.48	0.52	0.52	
	2	1.86	1.97	1.97		0.21	0.23	0.23	

E. Degrees of Freedom $\nu_1 = 100, \nu_2 = 909$

Estimators of $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$, $\hat{\gamma}^{(4)}$ and $\hat{\gamma}^{(5)}$ from n simulations.

		Means				Standard Deviations			
n	True $\underset{\sim}{\gamma}$	1	2	4	5	1	2	4	5
83	1024	1106.	1077.	1075.	Algorithm failed to converge	155.6	155.4	156.2	Algorithm failed to converge
	512	523.4	521.0	521.0		83.32	87.34	88.51	
	256	249.7	252.3	252.5		37.78	40.53	41.50	
	128	122.5	125.3	125.3		16.67	18.05	18.34	
	64	60.40	62.60	62.65		9.14	10.10	10.32	
	32	30.24	31.73	31.73		4.41	4.92	5.00	
	16	15.50	16.53	16.54		2.31	2.63	2.67	
	8	7.21	7.75	7.75		1.05	1.21	1.22	
	4	3.68	4.04	4.04		0.54	0.63	0.64	
	2	1.77	1.99	2.00		0.29	0.33	0.34	

Table 5.5.5

Correlation matrices for the five estimators of the $\{\gamma_i\} = \text{Eigs}(\Sigma_1\Sigma^{-1})$ from the simulation experiments for p = 3 dimensions.

C. Degrees of Freedom $\nu_1 = 30$, $\nu_2 = 124$

C.1. True $\{\gamma_i\} = \{6,4,2\}$

Number of		Correlation Coefficients			
Simulations	Pair	1	2	4	5
25	(1,2)	.802	.758	.714	.723
	(1,3)	.322	.265	.241	.260
	(2,3)	.410	.328	.270	.273

C.2. True $\{\gamma_i\} = \{16,4,2\}$

Number of		Correlation Coefficients			
Simulations	Pair	1	2	4	5
74	(1,2)	.093	.043	.021	.025
	(1,3)	-.023	-0.41	-.046	-.040
	(2,3)	.308	.248	.210	.210

D. Degrees of Freedom $\nu_1 = 60$, $\nu_2 = 244$

D.1. True $\{\gamma_i\} = \{6,4,2\}$

Number of		Correlation Coefficients			
Simulations	Pair	1	2	4	5
44	(1,2)	.425	.356	.319	.324
	(1,3)	.114	.071	.061	.070
	(2,3)	.381	.314	.283	.281

D,2. True $\{\gamma_i\} = \{6,4,2\}$

Number of		Correlation Coefficients			
Simulations	Pair	1	2	4	5
91	(1,2)	-.034	-.057	-.060	-.060
	(1,3)	.042	.033	.032	.033
	(2,3)	.289	.235	.210	.211

There is no page 199

Chapter 6 The Predictive Bayesian and other Approaches

Our chief concern in this chapter is the Predictive Bayesian approach to discriminant analysis under the random effects model.

As described in Section 2.2 this approach consists in evaluating the posterior probabilities, given the training sample and underlying model together with any known parameters, that the new observation x comes from each of the k_1 populations in question, and assigning it to that population for which this probability is the largest. Therefore, in contrast to Chapters 3 and 4 where we are concerned with the expected behaviour of the standard classification rules of classical discriminant analysis under the random effects model, this chapter is concerned with the development of new classification formulae applicable to this model.

In conformity with the rest of this thesis, we will assume that the prior probabilities q_i of the k populations π_i, $i = 1,\ldots,k$ are all equal, so that the posterior probability that x comes from π_r is proportional to the predictive density of x, given the training sample and the assumption that x comes from π_r. See expression (2.2.4). (It is, however, a trivial matter to adjust the theory for the case where the q_r are unequal.)

Therefore, in the next two sections we will derive the predictive density of x under the random effects model given the training sample and the assumption that x comes from π_r, using a noninformative prior distribution for the unknown parameters, firstly for the univariate case (Section 6.1) and then for the multivariate case (Section 6.2). In Section 6.3 the predictive density of x will be investigated under two alternative prior distributions of the unknown parameters, namely, (i) Box and Tiao's noninformative prior distribution for the random effects model, and (ii) the natural conjugate prior distribution. Finally, in Section 6.4 two other Baysian approaches to discriminant analysis, the Empirical Bayes and "Semi-Bayes" approaches, respectively,

Chapter 6 The Predictive Bayesian and other Approaches

Our chief concern in this chapter is the Predictive Bayesian approach to discriminant analysis under the random effects model.

As described in Section 2.2 this approach consists in evaluating the posterior probabilities, given the training sample and underlying model together with any known parameters, that the new observation x comes from each of the k_1 populations in question, and assigning it to that population for which this probability is the largest. Therefore, in contrast to Chapters 3 and 4 where we are concerned with the expected behaviour of the standard classification rules of classical discriminant analysis under the random effects model, this chapter is concerned with the development of new classification formulae applicable to this model.

In conformity with the rest of this thesis, we will assume that the prior probabilities q_i of the k populations π_i, $i = 1,\ldots,k$ are all equal, so that the posterior probability that x comes from π_r is proportional to the predictive density of x, given the training sample and the assumption that x comes from π_r. See expression (2.2.4). (It is, however, a trivial matter to adjust the theory for the case where the q_r are unequal.)

Therefore, in the next two sections we will derive the predictive density of x under the random effects model given the training sample and the assumption that x comes from π_r, using a noninformative prior distribution for the unknown parameters, firstly for the univariate case (Section 6.1) and then for the multivariate case (Section 6.2). In Section 6.3 the predictive density of x will be investigated under two alternative prior distributions of the unknown parameters, namely, (i) Box and Tiao's noninformative prior distribution for the random effects model, and (ii) the natural conjugate prior distribution. Finally, in Section 6.4 two other Baysian approaches to discriminant analysis, the Empirical Bayes and "Semi-Bayes" approaches, respectively, will be given brief consideration.

Remark 6.1 In this chapter we have to make a distinction between the k populations used in the training sample and the $k_1(\le k)$ populations from which it is known that the new observation x derives. Clearly these k_1 populations must be represented in the training sample, but they may well have been sampled at a later stage than the rest of the training sample, possibly only at the time when the particular classification problem in question arises.

6.1 The Univariate Case

For dimension p = 1 the discriminant analysis problem under the random effects model becomes:

Given a training sample,

$$TS = \{x_{ij} : i = 1,\dots,k; \;\; j = 1,\dots,n_i\}$$

where,

$$x_{ij} \sim N(\mu_i, \sigma^2) \quad \text{independently}, \quad \forall_{i,j}$$

and

$$\mu_i \sim N(\xi, \tau^2) \quad \text{independently}, \quad \forall_i \quad ,$$

classify a new observation x of unknown origin into one of the k_1 populations π_r, $r = 1,\dots,k_1$, where π_r is characterised by a $N(\mu_r, \sigma^2)$ distribution.

For the Predictive Baysian approach we need to make an assumption about the prior distribution of the unknown parameters σ^2, ξ and τ^2, and in this section we assume that they have the following general type of noninformative joint prior density:

$$g(\sigma^2,\ \xi,\ \tau^2)d\sigma^2\ d\xi\ d\tau^2 \propto \sigma^{-v_1}\ \tau^{-v_2}\ d\sigma^2\ d\xi\ d\tau^2 \qquad (6.1.1)$$

Remark 6.1.1 *For reasons that will become clear later, we are considering* a more general form of prior distribution than the usual diffuse or invariant (Jeffreys 1961) prior distribution which has $v_1 = v_2 = 2$. The prior density (6.1.1) is also used in Geisser and Cornfield (1963) and in Geisser (1964).

Given the above assumptions, the predictive density of x, assuming that $x \in \pi_r$, is:

$$f(x|TS,\ v_1,\ v_2,\ \pi_r) = \int_{\sigma^2}\int_{\underset{\sim}{\mu}} f(x|\underset{\sim}{\mu},\ \sigma^2,\ \pi_r)P(\underset{\sim}{\mu},\ \sigma^2|TS)d\underset{\sim}{\mu}\,d\sigma^2 \qquad (6.1.2)$$

where,

$$\underset{\sim}{\mu} = (\mu_1,\ \mu_2,\ldots,\ \mu_k)'$$

$$f(x|\underset{\sim}{\mu},\ \sigma^2,\ \pi_r) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\{-\tfrac{1}{2}(\frac{x-\mu_r}{\sigma})^2\}$$

$$P(\underset{\sim}{\mu},\ \sigma^2|TS) \propto P(TS|\underset{\sim}{\mu},\ \sigma^2)P(\underset{\sim}{\mu},\ \sigma^2)$$

$$P(TS|\underset{\sim}{\mu},\ \sigma^2) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}\ \sigma}\exp\{-\tfrac{1}{2}(\frac{x_{ij}-\mu_i}{\sigma})^2\}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}N}\ \sigma^N}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2\}$$

$$N = \sum_{i=1}^{k} n_i$$

$$P(\underset{\sim}{\mu},\ \sigma^2) = P(\sigma^2)\int_{\tau^2}\int_{\xi} P(\underset{\sim}{\mu}|\xi,\tau^2)P(\xi,\tau^2)d\xi d\tau^2$$

$$P(\underset{\sim}{\mu}|\xi,\tau^2) = \prod_{i=1}^{k}\frac{1}{\sqrt{2\pi}\ \tau}\exp\{-\tfrac{1}{2}(\frac{\mu_i-\xi}{2})^2\} = \frac{1}{(2\pi)^{\frac{1}{2}k}\ \tau^k}\exp\{-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i-\xi)^2\}$$

and

$$P(\sigma^2)P(\xi, \tau^2) = g(\sigma^2, \xi, \tau^2) \propto \sigma^{-v_1} \tau^{-v_2} \qquad (6.1.3)$$

Substituting all this into equation (6.1.2) and using the notation:

$$\begin{aligned} n_i^* &= n_i \qquad \forall\, i \neq r \\ n_r^* &= n_r + 1 \\ x_{r\, n_r^*} &= x \end{aligned} \qquad (6.1.4)$$

yields, ignoring all constants of proportionality.

$$f(x|TS, v_1, v_2, \pi_r) \propto \int_{\sigma^2}\int_{\underset{\sim}{\mu}}\int_{\tau^2}\int_{\xi} \sigma^{-(N+1)} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i^*}(x_{ij}-\mu_i)^2\}$$

$$\times\, \tau^{-k} \exp\{-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i - \xi)^2\}\, \sigma^{-v_1}\, \tau^{-v_2}\, d\xi\, d\tau^2\, d\underset{\sim}{\mu}\, d\sigma^2$$

$$= \int_{\sigma^2}\int_{\underset{\sim}{\mu}} \sigma^{-(N+v_1+1)} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i^*}(x_{ij}-\mu_i)^2\}$$

$$\times \int_{\tau^2}\int_{\xi} \tau^{-(k+v_2)} \exp\{-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i - \xi)^2\}\, d\xi\, d\tau^2\, d\underset{\sim}{\mu}\, d\sigma^2$$

Considering the inner pair of integrals:

$$\int_{\tau^2}\int_{\xi} \tau^{-(k+v_2)} \exp\{-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i - \xi)^2\}\, d\xi\, d\tau^2$$

$$= \int_{\tau^2} \tau^{-(k+v_2)} \exp\{-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i - \mu_.)^2\} \int_{\xi} \exp\{-\frac{k}{2\tau^2}(\xi - \mu_.)^2\} d\xi\, d\tau^2$$

(where $\mu_. = \frac{1}{k}\sum_{i=1}^{k} \mu_i$)

$$\propto \int_{\tau^2} (\tau^2)^{-\frac{1}{2}(k+v_2-1)} \exp\{-\frac{1}{2\tau^2} S_{\mu}^2\}\, d\tau^2$$

where $S_\mu^2 = \sum_{i=1}^{k} (\mu_i - \mu.)^2$.

Transforming to:

$$u = \tfrac{1}{2}\tau^{-2}\, S_\mu^2$$

so that

$$d\tau^2 = -\tfrac{1}{2}\, u^{-2}\, S_\mu^2\, du$$

the integral becomes (ignoring constants of proportionality):

$$(S_\mu^2)^{-\frac{1}{2}(k + v_2 - 3)} \int_0^\infty u^{-\frac{1}{2}(k + v_2 - 3)-1} \exp(-u)\, du$$

$$= (S_\mu^2)^{-\frac{1}{2}(k + v_2 - 3)}\, \Gamma(\tfrac{1}{2}(k + v_2 - 3))$$

So:

$$f(x|TS,\, v_1,\, v_2,\, \pi_r) \propto \int_{\sigma^2} \int_{\underset{\sim}{\mu}} (\sigma^2)^{-\frac{1}{2}(N+v_1+1)} (S_\mu^2)^{-\frac{1}{2}(k+v_2-3)}$$

$$\exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i^*} (x_{ij} - \mu_i)^2\}\, d\sigma^2\, d\underset{\sim}{\mu} \quad . \qquad (6.1.5)$$

Now,

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i^*} (x_{ij} - \mu_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i^*} (x_{ij} - x_{i.}^*)^2 + \sum_{i=1}^{k} n_i^*(\mu_i - x_{i.}^*)^2$$

where $x_{i.}^* = \frac{1}{n_i^*} \sum_{j=1}^{n_i^*} x_{ij}$

$$= \begin{cases} x_{i.} & \forall\, i \neq r \\ x_{r.} + \dfrac{x - x_{r.}}{n_r + 1}\,, & i = r \end{cases} \qquad (6.1.6)$$

So (6.1.5) becomes

$$f(x|TS, v_1, v_2, \pi_r) \propto \int_{\sigma^2} (\sigma^2)^{-\frac{1}{2}(N+v_1-k+1)} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i^*} (x_{ij} - x_{i.}^*)^2\}$$

$$\times \int_{\underset{\sim}{\mu}} (S_\mu^2)^{-\frac{1}{2}(k+v_2-3)} (\sigma^2)^{-\frac{1}{2}k} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{k} n_i^* (\mu_i - x_{i.}^*)^2\} \, d\underset{\sim}{\mu} \, d\sigma^2$$

(6.1.7)

Now, apart from a constant of proportionality, the *inner integral in (6.1.7) can* be thought of as the $-\frac{1}{2}(k+v_2-3)^{th}$ moment about zero of the unnormed sample variance: $S_\mu^2 = \sum_{i=1}^{k} (\mu_i - \mu_.)^2$ *where* the individual μ_i are *independently* distributed according to the $N(x_{i.}^*, \sigma^2/n_i^*)$ *distribution.*

In order to be able to evaluate this expected value, we have to make the assumption that the n_i^* are all equal, say,

$$n_i^* = n^* \qquad i = 1,\ldots,k \qquad (6.1.8)$$

under *this* assumption S_μ^2 has σ^2/n^* times a noncentral $\chi_{k-1}^2(\lambda^*)$ distribution, with noncentrality parameter,

$$\lambda^* = \frac{n^*}{\sigma^2} \sum_{i=1}^{k} (x_{i.}^* - x_{..}^*)^2 = (\sigma^2)^{-1} A_1^* \qquad (6.1.9)$$

where,

$$x_{..}^* = \frac{1}{k} \sum_{i=1}^{k} x_{i.}^*$$

and

$$A_1^* = n^* \sum_{i=1}^{k} (x_{i.}^* - x_{..}^*)^2$$

Now, although the cumulants *of* the *noncentral* chi-squared distribution (and hence the first moments about zero) may be expressed extremely simply, general expressions *for* the moments about zero are usually in terms of infinite sums. (See, for example, Johnson and Kotz, 1970b.) The following expression for the r^{th} moment about zero of the $\chi^2_\nu(\lambda)$ distribution, derived in Appendix 6.1, is convenient for the present purpose:

$$\mu'_r = 2^r \exp(-\tfrac{1}{2}\lambda) \sum_{j=0}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} \frac{\Gamma(\frac{1}{2}\nu + r + j)}{\Gamma(\frac{1}{2}\nu + j)} \quad \text{for } r > -\tfrac{1}{2}\nu \qquad (6.1.10)$$

The inner integral in(6.1.7) is therefore proportional to:

$$\left(\frac{2\sigma^2}{n^*}\right)^{-\frac{1}{2}(k+\nu_2-3)} \exp\{-\tfrac{1}{2}\lambda^*\} \sum_{j=0}^{\infty} \frac{(\frac{1}{2}\lambda^*)^j}{j!} \frac{\Gamma(\frac{1}{2}(2-\nu_2)+j)}{\Gamma(\frac{1}{2}(k-1)+j)} \quad \text{for } \nu_2 < 2 \qquad (6.1.11)$$

The infinite series in (6.1.11) is proportional to the confluent hypergeometric function $M(\frac{1}{2}(2-\nu_2);\ \frac{1}{2}(k-1);\ \frac{1}{2}\lambda^*)$ (see, for example, Abramowitz and Stegun, 1965) and therefore it converges for all values of the parameter $\frac{1}{2}\lambda^*$. Substituting (6.1.11) into (6.1.7) and interchanging the order of integration and summation yields, ignoring the constants of proportionality:

$$f(x \mid TS, \nu_1, \nu_2, \pi_r) = \sum_{j=0}^{\infty} \frac{\Gamma(\frac{1}{2}(2-\nu_2)+j)}{\Gamma(\frac{1}{2}(k-1)+j)} \frac{(\frac{1}{2}A_1^*)^j}{j!} \int (\sigma^2)^{-\frac{1}{2}(N+\nu_1+\nu_2+2j-2)} \times \exp\{-\frac{A_3^*}{2\sigma^2}\}\, d\sigma^2$$

where,

$$A_3^* = \sum_{i=1}^{k} \sum_{j=1}^{n^*} (x_{ij} - x^*_{i.})^2 + A_1^* = \sum_{i=1}^{k} \sum_{j=1}^{n^*} (x_{ij} - x^*_{..})^2 \qquad (6.1.12)$$

Making the transformation $y = A_3^*/2\sigma^2$, the above integral may be evaluated as a gamma function, yielding eventually:

$$f(x|TS, v_1, v_2, \pi_r) \propto \sum_{j=0}^{\infty} \frac{\Gamma(\frac{1}{2}(2-v_2)+j)}{\Gamma(\frac{1}{2}(k-1)+j)} \Gamma(\tfrac{1}{2}(N+v_1+v_2+2j-4))$$

$$\times(\tfrac{1}{2}A_1^*)^j (\tfrac{1}{2}A_3^*)^{-\frac{1}{2}(N+v_1+v_2+2j-4)}$$

$$= (A_3^*)^{-\frac{1}{2}(N+v_1+v_2-4)} F(\tfrac{1}{2}(2-v_2), \tfrac{1}{2}(N+v_1+v_2-4);$$

$$\tfrac{1}{2}(k-1);\ (A_1^*/A_3^*)) \quad \text{for} \quad v_2 < 2 \qquad (6.1.13)$$

where,

$F(\alpha, \beta; \gamma; x) = \sum_{j=0}^{\infty} \frac{\alpha^{[j]} \beta^{[j]}}{\gamma^{[j]}} \frac{x^j}{j!}$ is the hypergeometric function,

and $\alpha^{[j]} = \alpha(\alpha+1) \ldots (\alpha+j-1)$.

Since, by definition, $|A_1^*/A_3^*| < 1$, the hypergeometric function in (6.1.13) converges. (See, for example, Abramowitz and Stegun (1965) or Johnson and Kotz (1969).)

Remark 6.1.1 Assumption (6.1.8) effectively implies that

$$n_r = n^* = n \qquad \forall\ r = 1,\ldots,k$$

and that when evaluating the predictive density (6.1.13) assuming that $x \in \pi_r$, for each $r = 1,\ldots,k$ in turn, one of the observations x_{rj} is chosen from $\{x_{rj},\ j = 1,\ldots,n\}$ and is replaced by x in the sample. Under these circumstances, therefore, the effective size of the training Sample becomes N - 1.

The two terms in (6.1.13)affected by the above are A_1^* and A_3^*, and it is shown in Appendix 6.2 that, for $x \in \pi_r$:

$$A_1^* = A_1 + 2(x - x_{rj})(x_{r.} - x_{..}) + \frac{k-1}{kn}(x - x_{rj})^2 \qquad (6.1.14)$$

and

$$A_2^* = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - x_{i.}^*)^2 = A_2 - \frac{1}{n}(x - x_{rj})^2 - (x_{rj} - x_{r.})^2 + (x - x_{r.})^2 \qquad (6.1.15)$$

where,

$$A_1 = n\sum_{i=1}^{k}(x_{i.} - x_{..})^2$$

and

$$A_2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - x_{i.})^2$$

are the between groups and within groups sums of squares, respectively, as defined in Table 5.1.1 for the case p = 1. Finally, A_3^* is obtained by summing A_1^* and A_2^*,

i.e.

$$A_3^* = A_1^* + A_2^* \qquad (6.1.16)$$

Formulae (6.1.14) and (6.1.16) will be useful when evaluating the predictive density (6.1.13) successively for all $r = 1,\ldots,k$.

Note also that under these circumstances N should be replaced by N - 1 in (6.1.13).

Remark 6.1.2 The fact that v_2 must be less than 2 in (6.1.13) implies that, for the predictive density to exist, τ^2 cannot have the usual diffuse prior distribution with $v_2 = 2$.

It is interesting to compare this with problems encountered by other authors studying related problems through the Bayesian approach. Lindley and Smith (1972) and Smith (1973) studying the problem of estimation under a Bayesian General Linear Model, both start off with their analysis by assuming all variances and covariances known. When passing to the situation where the variances and covariances are unknown and have prior distributions, they come up against intractable mathematical problems in evaluating the posterior distributions and means for the parameters of interest. To overcome this problem they use instead the mode of the joint posterior distribution of the parameters of interest and the nuisance parameters (the variances and covariances) and use these modal values as Bayesian estimates of the parameters. In practice, the modal values usually have to be obtained by iterative procedures. In their examples they use natural conjugate prior distributions for the variances and covariances; in Section 6.3 we will investigate this class of prior distributions for our problem.

Box and Tiao (1973) use a different type of diffuse prior distribution when considering the random effects model, in order to get around their analytical problems. This prior distribution will also be considered in Section 6.3.

It is rather remarkable that it is the prior distribution of the second stage "hyperparameter" τ^2 in our random effects model that gives the problem, while that of the corresponding first stage parameter σ^2 presents no problem at all, at least within the framework of the diffuse prior distributions (6.1.1).

Therefore, in (6.1.12) we may assign the value $v_1 = 2$, giving σ^2 a noninformative prior distribution relative to the likelihood function of the normal distribution, both in the sense that it produces a posterior distribution that is "data translated" as defined by Box and Tiao (1973) and in the sense that probability statements on σ^2 based on its posterior distribution are invariant under parameter transformations.

For v_2 we may assign the value $v_2 = 1$ so that τ^2 has a prior distribution that, while it is not noninformative, is as close as it may be to one without jeopardising the existence of the predictive density (6.1.13). Under these parameter values (6.1.13) becomes (remembering that N is replaced by $N - 1$):

$$f(x|TS, \pi_r) \propto (A_3^*)^{-\frac{1}{2}(N-2)} \sum_{j=0}^{\infty} \frac{(\frac{1}{2})^{[j]}(\frac{1}{2}(N-2))^{[j]}}{(\frac{1}{2}(k-1))^{[j]}} \frac{(A_1^*/A_3^*)^j}{j!}$$

$$= (A_3^*)^{-\frac{1}{2}(N-2)} F(\tfrac{1}{2}, \tfrac{1}{2}(N-2);\; \tfrac{1}{2}(k-1); \frac{A_1^*}{A_3^*}) \qquad (6.1.17)$$

<u>Remark 6.1.3</u> An alternative, asymptotic expression for the predictive density of x may be obtained by interchanging the order of integration in (6.1.5). This yields,

$$f(x|TS, v_1, v_2, \pi_r) \propto (A_2^*)^{-\frac{1}{2}(N+v_1-1)} \int_{\underset{\sim}{\mu}} (S_{\underset{\sim}{\mu}}^2)^{-\frac{1}{2}(k+v_2-3)}$$

$$\times \left(1 + \sum_{i=1}^{k} \left(\frac{\mu_i - x_{i.}^*}{\sqrt{A_2^*/n_i^*}}\right)^2\right)^{-\frac{1}{2}(N+v_1-1)} d\underset{\sim}{\mu} \; .$$

This integral is proportional to the $-\frac{1}{2}(k+v_2-3)^{\text{th}}$ moment of the (unnormed) sample variance $S_\mu^2 = \sum_{i=1}^{k} (\mu_i - \mu_.)^2$ where the μ_i, $i = 1,\ldots,k$ jointly have a multivariate t-distribution with common denominator (see, for example, Johnson and Kotz (1972)). Assuming that $n_i^* = n$, $\forall_i$ and that the total sample size N is large enough for the multivariate t-distribution to be approximated by that of k independent normal random variables with different means but common variance, the integral may be evaluated approximately using the $-\frac{1}{2}(k+v_2-3)^{\text{th}}$ moment of the noncentral χ_{k-1}^2 distribution. This yields, after some algebra:

$$f(x|TS, v_1, v_2, \pi_r) \mathrel{\dot{\propto}} (A_2^*)^{-\frac{1}{2}(N+k+v_1+v_2-4)} \exp\{-\tfrac{1}{2}\lambda^*\}$$

$$\times\, M(\tfrac{1}{2}(2-v_2);\ \tfrac{1}{2}(k-1);\ \tfrac{1}{2}\lambda^*) \text{ for } v_2 < 2 \tag{6.1.18}$$

where,

$$\lambda^* = (N-k)\, A_1^* / A_2^*$$

and

$$M(\alpha;\ \beta;\ x) = \sum_{j=0}^{\infty} \frac{\alpha^{[j]}}{\beta^{[j]}} \frac{x^j}{j!}$$ is the confluent hypergeometric function.

It is interesting to note that again the parameter v_2 in the prior density of τ^2 can not take the value 2 corresponding to the usual noninformative prior distribution. Assigning the values $v_1 = 2$ and $v_2 = 1$ as before, and replacing N by N - 1 (see Remark 6.1.1}, (6.1.18) becomes

$$f(x|TS, \pi_r) \mathrel{\dot{\propto}} (A_2^*)^{-\frac{1}{2}(N+k-2)} \exp\{-\tfrac{1}{2}\lambda^*\}\, M(\tfrac{1}{2};\ \tfrac{1}{2}(k-1)\};\tfrac{1}{2}\lambda^*) \tag{6.1.19}$$

Example C.1.1 To illustrate the use of the above formulae, the following hypothetical example was considered. Given the training samples of size n = 3 from each of k = 5 populations in Table 6.1.1 and an observation x = 7 of unknown origin, classify x into one of these 5 populations, assuming that they are generated by the random effects model.

Table 6.1.2 gives the quantities A_1^*, A_2^* and A_3^* for each of the five populations, as well as the ratios (A_1^*/A_3^*) and $(5A_1^*/A_2^*)$ required in formulae (6.1.17) and (6.1.19) for the exact and approximate predictive densities, assuming that v_2=1 and v_1=2. FORTRAN subroutine HYPGFN, given in Appendix 6.5, was written to compute the hypergeometric and confluent hypergeometric functions required in the above formulae. The posterior probabilities for the five populations, computed using both the *exact and approximate formulae* and assuming equal prior probabilities, are also given in Table 6.1.2. As recommended in Sub-section 6.3.3 below, the observation closest to the mean of the training sample from π_r was replaced by x when computing the predictive density given $x \in \pi_r$.

Table 6.1.1

The Hypothetical Training Sample

Populations	1	2	3	4	5
	1	3	6	7	9
Observations	2	4	7	8	10
	3	5	8	9	11

Observation x of unknown origin: 7

Table 6.1.2

Computing the Posterior Probabilities

Populations:	1.	2	3	4	5
A_1^*	87.07	111.60	122.40	119.07	102.00
A_2^*	26.67	16.00	10.00	10.67	16.00
A_3^*	113.73	127.60	132.40	129.73	118.00
A_1^* / A_3^*	0.7655	0.8746	0.9245	0.9178	0.8644
$\frac{1}{2}(N-k)A_1^* / A_2^*$	16.33	34.88	61.20	55.81	31.88
Exact Probs.	0.0065	0.0553	0.4981	0.3766	0.0635
Approximate Probs.	*0.0006*	*0.0199*	*0.5822*	0.3744	0.0228
Fixed Effect Probs.	0.0017	0.0327	0.5580	0.3749	0.0327

The last row of Table 6.1.2 gives the posterior probabilities for each of the five populations computed from formula (2.2.6) for the case where the population means μ_i are given a diffuse prior distribution - roughly speaking, this corresponds to a fixed effects model (See Box and Tiao (1973) pages 379-80 for a discussion of this point). Comparing these probabilities with their counterparts under the random effects model, computed from the exact formula (6.1.17), it is clear that in the latter case the posterior probabilities are slightly more conservative, in the sense that the highest probability (that of population 3) is somewhat lower, and those of the other populations correspondingly higher, than their counterparts under the fixed effects model. Intuitively speaking this is reasonable, as one would expect classification to be better in the situation where, a priori, the populations tend to be further apart, as is the case with the diffuse pri·· relative to the normal prior. (See Cox and Hinkley (1974) page 379 for a related discussion.)

Finally, the probabilities given by the approximate formula (6.1.19) are clearly too optimistic (in a sense complementary to conservative) giving values that differ even more from the exact probabilities than do the corresponding probabilities under the fixed effects model.

6.2 The multivariate case

Analogously to the univariate case discussed in the previous section, our discriminant analysis problem becomes:

Given a training sample from k populations,

$$TS = \{x_{ij};\ j = 1,\ldots,n_i,\ i = 1,\ldots,k\}$$

where,

$$x_{ij} \sim N_p(\mu_i, \Sigma) \quad \text{independently } \forall\ i, j$$

and

$$\mu_i \sim N_p(\xi, T) \quad \text{independently } \forall_i \ ,$$

classify a new observation x of unknown origin into one of the k_1 populations:

$$\pi_r : N_p(\mu_r, \Sigma) \qquad r = 1,\ldots,k_1$$

where $\qquad k_1 \leq k$

We assume that the unknown parameters Σ, ξ and T have the diffuse prior distribution with joint density:

$$g(\Sigma,\ \xi,\ T)d\Sigma\ d\xi\ dT \propto |\Sigma|^{-\frac{1}{2}v_1}\ |T|^{-\frac{1}{2}v_2}\ d\Sigma\ d\xi\ dT \qquad (6.2.1)$$

Remark 6.2.1 As in the univariate case, and for the same reason, we are considering the more general form of diffuse prior distribution, used by Geisser and Cornfield (1963) and Geisser(1964), than the usual one for which $v_1 = v_2 = p + 1$.

Given the above assumptions, the predictive density of x, given the hypothesis $x \in \pi_r$, becomes:

$$f(x|TS, v_1, v_2, \pi_r) = \int_{\underset{\sim}{\mu}}\int_{\Sigma} f(x|\underset{\sim}{\mu}, \Sigma, \pi_r)P(\underset{\sim}{\mu},\Sigma|TS)d\Sigma\ d\underset{\sim}{\mu} \qquad (6.2.2)$$

where,

$\underset{\sim}{\mu}$ is the p × k matrix $(\mu_1, \mu_2,\ldots,\mu_k)$

$$f(x|\underset{\sim}{\mu},\ \Sigma,\ \pi_r) = (2\pi)^{-\frac{1}{2}p}\ |\Sigma|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x - \mu_r)'\ \Sigma^{-1}(x - \mu_r)\}$$

$$P(\underset{\sim}{\mu},\ \Sigma|\ TS) \propto P(TS|\underset{\sim}{\mu},\ \Sigma)P(\underset{\sim}{\mu},\ \Sigma)$$

$$P(TS|\mu, \Sigma) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} (2\pi)^{-\frac{1}{2}p} |\Sigma|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x_{ij} - \mu_i)' \Sigma^{-1}(x_{ij} - \mu_i))$$

$$= (2\pi)^{-\frac{1}{2}Np} |\Sigma|^{-\frac{1}{2}N} \exp\{-\tfrac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)' \Sigma^{-1}(x_{ij} - \mu_i))$$

where $\quad N = \sum_{i=1}^{k} n_i$

$$P(\mu, \Sigma) = P(\Sigma) \int_T \int_\xi P(\mu|\xi, T)\, P(\xi, T) d\xi\, dT$$

$$P(\mu|\xi, T) = \prod_{i=1}^{k} (2\pi)^{-\frac{1}{2}p} |T|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\mu_i - \xi)' T^{-1}(\mu_i - \xi)\}$$

$$= (2\pi)^{-\frac{1}{2}kp} |T|^{-\frac{1}{2}k} \exp\{-\tfrac{1}{2} \sum_{i=1}^{k} (\mu_i - \xi)' T^{-1}(\mu_i - \xi)\}$$

and

$$P(\Sigma)\, P(\xi, T) = g(\Sigma, \xi, T) \propto |\Sigma|^{-\frac{1}{2}\nu_1} |T|^{-\frac{1}{2}\nu_2} \qquad (6.2.3)$$

Substituting(6.3.3) into(6.3.2) and using the notation:

$$n_i^* = n_i \qquad \forall\, i \neq r$$

$$n_r^* = n_r + 1$$

$$x_{r,n_r^*} = x \qquad (6.2.4)$$

gives:

$$f(x|TS,\nu_1,\nu_2,\pi_r) = \int_\Sigma \int_\mu |\Sigma|^{-\frac{1}{2}(N+\nu_1+1)} \exp\{-\tfrac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{n_i^*} (x_{ij}-\mu_i)' \Sigma^{-1}(x_{ij}-\mu_i))$$

$$\times \int_T \int_\xi |T|^{-\frac{1}{2}(k+\nu_2)} \exp\{-\tfrac{1}{2} \sum_{i=1}^{k} (\mu_i - \xi)' T^{-1}(\mu_i - \xi)\} d\xi\, dT \;\; d\mu\, d\Sigma$$

$$(6.2.5)$$

The inner two integrals in (6.2.5) are evaluated using the multivariate analogues of the techniques used in the univariate case, the details of which are given in Appendix 6.3, yielding:

$$\int_T \int_\xi |T|^{-\frac{1}{2}(k+v_2)} \exp\{-\tfrac{1}{2}\sum_{i=1}^{k}(\mu_i-\xi)'\,T^{-1}(\mu_i-\xi)\}d\xi\, dT \propto |A_{\underline{\mu}}|^{-\frac{1}{2}(k+v_2-p-2)} \tag{6.2.C}$$

where $\quad A_{\underline{\mu}} = \sum_{i=1}^{k}(\mu_i-\mu_.)(\mu_i-\mu_.)'$

and $\quad \mu_. = \frac{1}{k}\sum_{i=1}^{k}\mu_i$

Substituting (6.2.6) into (6.2.5) gives:

$$f(x|TS, v_1, v_2, \pi_r) \propto \int_\Sigma \int_{\underline{\mu}} |\Sigma|^{-\frac{1}{2}(N+v_1+1)} |A_{\underline{\mu}}|^{-\frac{1}{2}(k+v_2-p-2)}$$

$$\times \exp\{-\tfrac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i^*}(x_{ij}-\mu_i)'\Sigma^{-1}(x_{ij}-\mu_i)\}d\underline{\mu}\, d\Sigma$$

Now

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i^*}(x_{ij}-\mu_i)'\,\Sigma^{-1}(x_{ij}-\mu_i) &= \sum_{i=1}^{k}\sum_{j=1}^{n_i^*}(x_{ij}-x_{i.}^*)'\,\Sigma^{-1}(x_{ij}-x_{i.}^*) \\ &\quad + \sum_{i=1}^{k} n_i^*(\mu_i-x_{i.}^*)'\,\Sigma^{-1}(\mu_i-x_{i.}^*) \\ &= \text{tr}\,\Sigma^{-1}A_2^* + \sum_{i=1}^{k} n_i^*(\mu_i-x_{i.}^*)'\,\Sigma^{-1}(\mu_i-x_{i.}^*)\end{aligned} \tag{6.2.7}$$

where,

$$A_2^* = \sum_{i=1}^{k}\sum_{j=1}^{n_i^*}(x_{ij}-x_{i.}^*)(x_{ij}-x_{i.}^*)'$$

corresponds to the Within Groups Sum of Squares A_2 in Table 5.1.1, with x included in the sample from the r^{th} population, and $x^*_{i.} = \frac{1}{n^*_i} \sum_{j=1}^{n^*_i} x_{ij} = \begin{cases} x_{i.} & \forall\, i \neq r \\ x_{r.} + \frac{x - x_{r.}}{n_r + 1}, & i = r \end{cases}$

Therefore,

$$f(x|TS, v_1, v_2, \pi_r) \propto \int_{\Sigma} |\Sigma|^{-\frac{1}{2}(N+v_1-k+1)} \exp\{-\tfrac{1}{2} \operatorname{tr} \Sigma^{-1} A^*_2\} \int_{\mu} |A_{\mu}|^{-\frac{1}{2}(k+v_2-p-2)} |\Sigma|^{-\frac{1}{2}k}$$

$$\times \exp\{-\tfrac{1}{2} \sum_{i=1}^{k} n^*_i (\mu_i - x^*_{i.})' \Sigma^{-1} (\mu_i - x^*_{i.})\} d\mu \, d\Sigma \tag{6.2.8}$$

Now the inner integral in (6.2.8) is proportional to the $-\frac{1}{2}(k + v_2 - p - 2)^{th}$ moment of the generalized variance of a random sample $\mu_1, \mu_2, \ldots, \mu_k$ where the μ_i are independently distributed as $N(x_{i.}, \frac{1}{n^*_i} \Sigma)$.

In order to be able to evaluate this expected value, we have, as in the univariate case, to make the assumption that the n^*_i are all equal, say,

$$n^*_i = n^* \qquad \forall\, i \tag{6.2.9}$$

Under this assumption, $A_{\mu} = \sum_{i=1}^{k} (\mu_i - \mu_.)(\mu_i - \mu_.)'$ can be considered to have a p-dimensional noncentral Wishart distribution with (k-1) degrees of freedom, parameter matrix $\frac{1}{n^*} \Sigma$ and noncentrality matrix:

$$\Omega^* = \tfrac{1}{2}\, n^* \Sigma^{-1} \sum_{i=1}^{k} (x^*_{i.} - x^*_{..})(x^*_{i.} - x^*_{..})' \tag{6.2.10}$$

where $\qquad x^*_{..} = \frac{1}{k} \sum_{i=1}^{k} x^*_{i.}$

See, for example, Constantine (1963).
So the inner integral in (6.2.8) is proportional to the $-\frac{1}{2}(k+v_2-p-2)^{\text{th}}$ moment of the generalized variance corresponding to the $W_p(k-1;\ \frac{1}{n^*}\Sigma,\Omega^*)$ distribution.

Constantine (1963) studies the moments of the generalized variance corresponding to the $W_p(\nu;\ \Sigma;\ \Omega)$ distribution, giving the following as *one of the expressions for the* t^{th} moment:

$$\mu_t' = \frac{\Gamma_p(\frac{1}{2}\nu+t)}{\Gamma_p(\frac{1}{2}\nu)}\,|2\Sigma|^t \exp\{-\operatorname{tr}\Omega\}\,{}_1F_1(\tfrac{1}{2}\nu+t\ ;\ \tfrac{1}{2}\nu;\ \Omega)$$

$$\text{for } t > -\tfrac{1}{2}(\nu-p+1) \quad\text{and}\quad \nu > p-1 \qquad (6.2.11)$$

where $\Gamma_p(\frac{1}{2}\nu)$ is the multivariate gamma function defined in (5.3.5) and ${}_1F_1(a;\ b;\ \Omega)$ is the confluent hypergeometric function with matrix argument defined by James (1954). Thus the inner integral in (6.2.8) is equal to (6.2.11) with t replaced by $-\frac{1}{2}(k+v_2-p-2)$, ν by $k-1$, Σ by $\frac{1}{n^*}\Sigma$, and Ω by Ω^*. Substituting this into (6.2.8) and simplifying, yields:

$$f(x|TS, v_1, v_2, \pi_r) \propto (\tfrac{1}{2}n^*)^{\frac{1}{2}p(k+v_2-p-2)}\frac{\Gamma_p(\frac{1}{2}(p+1-v_2))}{\Gamma_p(\frac{1}{2}(k-1))}\int_{\Sigma}|\Sigma|^{-\frac{1}{2}(N+v_1+v_2-p-1)}$$

$$\times \exp\{-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}A_3^*\}\,{}_1F_1(\tfrac{1}{2}(p+1-v_2);\ \tfrac{1}{2}(k-1);\ \tfrac{1}{2}\Sigma^{-1}A_1^*)\,d\Sigma$$

$$\text{for } v_2 < 2 \quad\text{and}\quad k > p \qquad (6.2.12)$$

where $A_1^* = n^* \sum_{i=1}^{k}(x_{i.}^* - x_{..}^*)(x_{i.}^* - x_{..})'$

corresponds to the Between Groups Sum of Squares A_1 in Table 5.1.1, with x included in the sample from the r^{th} population, and $A_3^* = A_1^* + A_2^*$ is the corresponding Total Sum of Squares.

In order to evaluate the integral in (6.2.12) note that, by definition,

$$ {}_1F_1(\nu_1; \nu_2; \Omega) = \sum_{j=0}^{\infty} \sum_{\chi(j)} \frac{\nu_1^{\{\chi(j)\}}}{\nu_2^{\{\chi(j)\}}} C_{\chi(j)}(\Omega)/j! \qquad (6.2.13) $$

where,

$\chi(j)$ is a partition of the integer j of weight p, of the form $\{j_1, j_2, \ldots, j_p\}$ where $j_i \geq 0$ and $\sum_{i=1}^{p} j_i = j$,

$C_{\chi(j)}(\Omega)$ is the zonal polynomial in the eigenvaluesof Ω corresponding to partition $\chi(j)$,

$a^{\{\chi(j)\}} = \prod_{i=1}^{p} (a - \frac{1}{2}(i-1))^{[j_i]}$,

$b^{[j]} = b(b+1)\ldots(b+j-1)$

and $\sum_{\chi(j)}$ denotes the sum over all possible partitions $\chi(j)$ *of j* .

See, for example, Constantine (1963) or Johnson and Kotz (1972).

Substituting *(6.2.13)* into *(6.2.12)* and *interchanging* the order of *summation and integration* (For justification, see Constantine (1963)) yields:

$$ f(x|TS, v_1, v_2, \pi_r) \propto (\tfrac{1}{2}n^*)^{\frac{1}{2}p(k+v_2-p-2)} \frac{\Gamma_p(\frac{1}{2}(p+1-v_2))}{\Gamma_p(\frac{1}{2}(k-1))} \sum_{j=0}^{\infty} \sum_{\chi(j)} \frac{(\frac{1}{2}(p+1-v_2))^{\{\chi(j)\}}}{(\frac{1}{2}(k-1))^{\{\chi(j)\}} j!} $$

$$ \times \int_{\Sigma} |\Sigma|^{-\frac{1}{2}(N+v_1+v_2-p-1)} \exp(-\tfrac{1}{2} \operatorname{tr} \Sigma^{-1} A_3^*) \, C_{\chi(j)}(\tfrac{1}{2}\Sigma^{-1} A_1^*) d\Sigma \qquad (6.2.14) $$

for $v_2 < 2$ and $k > p$

The *integral in* (6.2.14) may now be evaluated using Constantine's (1963) fundamental integral identity:

$$\int_S \exp\{-\operatorname{tr} R\,S\}\,|S|^{t-\frac{1}{2}(p+1)}\,C_{\chi(j)}(ST)\,dS$$

$$=\Gamma_p(t,\chi(j))\,C_{\chi(j)}(R^{-1}T)\,|R|^{-t} \qquad (6.2.15)$$

where $\Gamma_p(t,\chi(j)) = \Gamma_p(t)\,t^{\{\chi(j)\}}$

In order to use (6.2.15) to evaluate the integral in (6.2.14) we need to make the transformation:

$$S = \Sigma^{-1}$$

with corresponding Jacobian (See, for example Press (1972)):

$$J(\Sigma \to S) = |S|^{-(p+1)}$$

This yields after some simplification, and ignoring all constants of proportionality:

$$f(x|TS,v_1,v_2,\pi_r) \propto |A_3^*|^{-\frac{1}{2}(N+v_1+v_2-2p-2)}\sum_{j=0}^{\infty}\sum_{\chi(j)}\frac{(\frac{1}{2}(p+1-v_2))^{\{\chi(j)\}}}{(\frac{1}{2}(k-1))^{\{\chi(j)\}}}$$

$$\times(\tfrac{1}{2}(N+v_1+v_2-2p-2))^{\{\chi(j)\}}\,C_{\chi(j)}(A_3^{*-1}A_1^*)/j!$$

$$= |A_3^*|^{-\frac{1}{2}(N+v_1+v_2-2p-2)}\,{}_2F_1(\tfrac{1}{2}(p+1-v_2),\tfrac{1}{2}(N+v_1+v_2-2p-2);\tfrac{1}{2}(k-1);A_3^{*-1}A_1^*) \qquad (6.2.16)$$

for $v_2 > 2$ and $k > p$

where ${}_2F_1(a_1, a_2; b_1; \Omega)$ is the hypergeometric function with matrix argument defined by James (1964).

Remark 6.2.1 Constantine (1963) states that the hypergeometric function of matrix argument ${}_2F_1(a_1, a_2; b_1; \Omega)$ converges for $\|\Omega\| < 1$, where $\|\Omega\|$ denotes the maximum of the absolute values of the eigenvalues of Ω. That $\|A_3^{*-1} A_1^*\| < 1$ is easily shown by the following argument:

For $k > p$, A_1^* is positive definite with probability 1 (See, for example Giri (1977) pages 74-6), so that under this condition A_1^{*-1} exists. Hence:

$$\begin{aligned} A_3^{*-1} A_1^* &= (A_1^* + A_2^*)^{-1} A_1^* \\ &= (I + A_1^{*-1} A_2^*)^{-1} \end{aligned}$$

Now, the eigenvalues of $(I + A_1^{*-1} A_2^*)^{-1}$ are the reciprocals of the eigenvalues of $(I+A_1^{*-1} A_2^*)$ and the eigenvalues $\{\lambda_i\}$ of $(I+A_1^{*-1} A_2^*)$ are the roots of the determinantal equation:

$$|I + A_1^{*-1} A_2^* - \lambda I| = 0$$

i.e. $$|A_1^{*-1} A_2^* - (\lambda-1)I| = 0$$

For $k > p$ and $n^* > 1$, $A_1^{*-1} A_2^*$ is positive definite, so that

$$(\lambda_i - 1) > 0 \quad \forall i$$

i.e. $$\lambda_i > 1 \quad \forall i$$

The result now follows, since $\text{eigs}(A_3^{*-1} A_1^*) = (\frac{1}{\lambda_i})$. Hence, expression (6.2.16) for the predictive density of x converges as long as $k > p$, $n^* > 1$.

Remark 6.2.2 To confirm that (6.2.16) corresponds to (6.1.13) for the case $p = 1$, note that in this case:

$$\chi(j) = j$$
$$C_{\chi(j)}(\omega) = \omega^j$$
$$a^{\{\chi(j)\}} = a^{[j]}$$

and $${}_2F_1(a_1, a_2; b; \omega) = \sum_{j=0}^{\infty} \frac{a_1^{[j]}\, a_2^{[j]}}{b^{[j]}} \frac{\omega^j}{j!}$$

So (6.2.16) becomes:

$$f(x|TS, v_1, v_2, \pi_r) \propto A_3^{*\,-\frac{1}{2}(N+v_1+v_2-4)} \sum_{j=0}^{\infty} \frac{(\frac{1}{2}(2-v_2))^{[j]}(\frac{1}{2}(N+v_1+v_2-4))^{[j]}}{(\frac{1}{2}(k-1))^{[j]}\, j!} (A_1^*/A_3^*)^j$$

for $v_2 < 2$

which is exactly expression (6.1.13).

Remark 6.2.3 As in the univariate case, assumption (6.2.9) effectively implies that:

$$n_i = n^* = n, \quad i = 1,\ldots,k$$

and that when evaluating the posterior probability that x belongs to π_r, one of the x_{rj} chosen from $\{x_{rj}, j = 1,\ldots,n\}$ is replaced by x in the sample. Under these circumstances therefore, the effective size of the training sample becomes N - 1.

Analogously to results (6.1.14) and (6.1.15) for the univariate case we have that:

$$A_1^* = A_1 + (x-x_{rj})(x_{r.}-x_{..})' + (x_{r.}-x_{..})(x-x_{rj})' + \frac{k-1}{kn}(x-x_{rj})(x-x_{rj})' \qquad (6.2.17)$$

and

$$A_2^* = A_2 - \frac{1}{n}(x-x_{rj})(x-x_{rj})' - (x_{rj}-x_{r.})(x_{rj}-x_{r.})' + (x-x_{r.})(x-x_{r.})' \tag{6.2.18}$$

where,

$$A_1 = n \sum_{i=1}^{k} (x_{i.} - x_{..})(x_{i.} - x_{..})'$$

and $$A_2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - x_{i.})(x_{ij} - x_{i.})'$$

are the between group and within groups sums of squares, respectively, as defined in Table 5.1.1 . *Finally,* A_3^* *is obtained from:*

$$A_3^* = A_1^* + A_2^* \tag{6.2.19}$$

Formulae (6.2.17) and (6.2.19) will be useful when evaluating the predictive density (6.2.16) for all groups π_r, $r = 1,\ldots,k_1$. Their proofs are the exact multivariate analogues of *those given in Appendix 6.2 for the case* $p = 1$ *and will therefore* be omitted.

It should also be noted that under these circumstances N should be replaced by N - 1 in (6.2.16)

<u>Remark 6.2.4</u> As in the univariate case, the parameter v_1 may assume the value p + 1, giving Σ the usual noninformative *prior distribution*, whereas v_2 has to assume a value less than 2 to ensure that the predictive density is properly defined. If therefore, analogously to the univariate case, we assign the values:

$$v_1 = p + 1$$

$$v_2 = 1$$

giving Σ a noninformative prior distribution relative to the likelihood function of the multivariate normal distribution, and T a *prior distribution* that is only very approximately so, then the predictive density becomes (remembering that N is replaced by N-1):

$$f(x|TS, \pi_r) \propto |A_3^*|^{-\frac{1}{2}(N-p-1)}\,{}_2F_1(\tfrac{1}{2}p, \tfrac{1}{2}(N-p-1);\ \tfrac{1}{2}(k-1);\ A_3^{*-1} A_1^*) \qquad (6.2.20)$$

for $k > p$

<u>Remark 6.2.5</u> The alternate, asymptotic expression for $f(x|TS, v_1, v_2, \pi_r)$ corresponding to that in the univariate case is obtained by reversing the *order of integration* in (6.2.8). This yields:

$$f(x|TS, v_1, v_2, \pi_r) \propto \int_{\underset{\sim}{\mu}} |A_{\underset{\sim}{\mu}}|^{-\frac{1}{2}(k+v_2-p-2)} |A_2^* + (\underset{\sim}{\mu} - \underset{\sim}{X}^*)\Lambda(\underset{\sim}{\mu} - \underset{\sim}{X}^*)'|^{-\frac{1}{2}(N+v_1-p)}\, d\underset{\sim}{\mu}$$

where $\underset{\sim}{X}^*$ is the $(p \times k)$ matrix: $(x_{1.}^*, x_{2.}^*, \ldots, x_{k.}^*)$

and $\Lambda = \mathrm{diag}\{n_i^*;\ i = 1, \ldots, k\}$

The secon [illegible] in the integrand is proportional to the density function of a $(p \times k)$ matrix T-distribution centered at $\underset{\sim}{X}^*$ (See, for example, *Dickey*, 1967) so the integral is proportional to the $-\frac{1}{2}(k + v_2 - p - 2)^{th}$ moment of the (unnormed) *sample covariance matrix*

$$A_{\underset{\sim}{\mu}} = \sum_{i=1}^{k} (\mu_i - \mu_.)(\mu_i - \mu_.)'$$

where the μ_i, $i = 1, \ldots, k$ jointly have the abovementioned distribution.

Assuming that $n_i^* = N^*$, ∀ i and that N is large enough for the matrix T-distribution to be *approximated by* the joint distribution of k independent (since Λ is a diagonal matrix), p-variate *normal random* variables with different mean vectors $x_{i.}^*$, $i = 1, \ldots, k$ but common covariance matrix, the above integral may be evaluated approximately using the $-\frac{1}{2}(k+v_2-p-2)^{th}$ moment of the generalized variance of the noncentral Wishart distribution $W_p(k-1; \frac{1}{n^*} S_2^*; \Omega^*)$

where:

$$S_2^* = \frac{1}{N-k} A_2^* \qquad \text{(assuming } N = kn^* = kn\text{)}$$

and $\Omega^* = \frac{1}{2} S_2^{*-1} A_1^*$

Theis yields, after some algebra:

$$f(x|TS, v_1, v_2, \pi_r) \dot{\propto} |A_2^*|^{-\frac{1}{2}(N+k+v_1+v_2-2p-2)} \exp\{-\mathrm{tr}\,\Omega^*\}\ {}_1F_1(\tfrac{1}{2}(p-v_2+1); \tfrac{1}{2}(k-1); \Omega^*) \tag{6.2.21}$$

$$\text{for } v_2 < 2 \quad \text{and} \quad k > P$$

Once again, the parameter v_2 has to assume a value less than 2 so that T cannot have the usual noninformative prior distribution. Assigning the values $v_1 = p+1$ and $v_2 = 1$ as before and replacing N by N-1 (see Remark 6.2.3), (6.2.21) *becomes:*

$$f(x|TS, \pi_r) \dot{\propto} |A_2^*|^{-\frac{1}{2}(N+k-p-1)} \exp\{-\ \mathrm{tr}\,\Omega^*\}\ {}_1F_1(\tfrac{1}{2}\ p; \tfrac{1}{2}(k-1); \Omega^*) \tag{6.2.22}$$

6.2.1 On Evaluating the Predictive Densities in the Multivariate case

The exact and approximate formulae (6.2.20) and (6.2.22) for the *predictive density* of x given that it comes from π_r are expressed in terms of the hypergeometric function of matrix argument ${}_2F_1(\frac{1}{2}\ p, \frac{1}{2}(N-p-1); \frac{1}{2}(k-1); A_3^{*-1} A_1^*)$ and the comfulent hypergeometric function of matrix argument ${}_1F_1(\frac{1}{2}\ p; \frac{1}{2}(k-1); \Omega^*)$ respectively. In order to try and evaluate these functions, the suite of FORTRAN programs of van der Westhuizen and Nagel (1979) for computing the zonal polynomials in the eigenvalues of a matrix Ω, corresponding to all the partitions of an integer j, were used. This suite consists of a number of programs that generate tables of all the partition vectors, symmetric functions, elementary symmetric function weights and Chi-coefficients (James, 1961, 1968) corresponding to all the partitions of the integers of interest, and then store them on files in the computer. The zonal polynomials corresponding to these integers are then computed by the last program in the suite, using these tables and the eigenvalues of the matrix in question.

Although the actual computation of the zonal polynomials is quite rapid once the files containing the abovementioned tables exist, the generation of these tables is very heavy on computer time, particularly for large integers, where the number of possible partitions becomes very large. As an indication of this, it took about 20 hours on the University of South Africa's Burroughs B6800 computer to generate the tables corresponding to all the partitions of all the integers up to 18.

Unfortunately in all the examples considered, the number of terms required for either of the two abovementioned hypergeometric functions to converge was far in excess of what could reasonably be computed without incurring prohibitive computing costs. An attempt was made to get an indication of the values, or relative values, of the hypergeometric functions in the predictive densities corresponding to different populations by studying the successive sums of the individual terms in the hypergeometric series for integers $j = 1$ to 18. However, the graphs of neither the values of these successive sums against j nor of the ratios of these sums corresponding to different populations against j, provided any insight, except that the values and relative values of the hypergeometric functions would be very different from the values and relati values of the sums of the first eighteen terms in the corresponding hypergeometric series.

Therefore, the unhappy conclusion is that although the programs of van der Westhuizen and Nagel (1979) are very useful for computing the

the partition vectors, symmetric functions, elementary symmetric function weights and Chi-coefficients (James, 1961, 1968) corresponding to all the partitions of the integers of interest, and then store them on files in the computer. The zonal polynomials corresponding to these integers are then computed by the last program in the suite, using these tables and the eigenvalues of the matrix in question.

Although the actual computation of the zonal polynomials is quite rapid once the files containing the abovementioned tables exist, the generation of these tables is very heavy on computer time, particularly for large integers, where the number of possible partitions becomes very large. As an indication of this, it took about 20 hours on the University of South Africa's Burroughs B6800 computer to generate the tables corresponding to all the partitions of all the integers up to 18.

Unfortunately in all the examples considered, the number of terms required for either of the two abovementioned hypergeometric functions to converge was far in excess of what could reasonably be computed without incurring prohibitive computing costs. An attempt was made to get an indication of the values, or relative values, of the hypergeometric functions in the predictive densities corresponding to different populations by studying the successive sums of the individual terms in the hypergeometric series for integers j = 1 to 18. However, the graphs of neither the values of these successive sums against j nor of the ratios of these sums corresponding to different populations against j, provided any insight, except that the values and relative values of the hypergeometric functions would be very different from the values and relative values of the sums of the first eighteen terms in the corresponding hypergeometric series.

Therefore, the unhappy conclusion is that although the programs of van der Westhuizen and Nagel (1979) are very useful for computing the values of individual zonal polynomials, they are unfortunately not of much practical use, given the computers presently available, for evaluating the hypergeometric functions of matrix argument appearing in the predictive densities under the random effects model.

6.3 The Predictive Bayesian Approach using different prior Distributions

In this section we investigate the use of two different prior distributions in the evaluation of the predictive density of a new observation x of unknown origin, given the training sample TS = $\{x_{ij}, j=1,\ldots,n_i; \; i=1,\ldots,k\}$ and the hypothesis that $x \in \pi_r$, one of the k populations in the training sample.

The reason for doing this is twofold:

Firstly, other authors have considered different prior distributions for the parameters in Bayesian analyses associated with the normal distribution, and it is interesting to investigate their use in the present context.

Secondly, in the light of the problems encountered with the parameter v_2 (the exponent of τ^{-1} and $|T|^{-\frac{1}{2}}$) when using the noninformative prior distribution in evaluating the *predictive density of* x, it is interesting to see whether similar problems occur when different prior assumptions are used.

The following two cases will therefore be investigated in Subsections 6.3.1 and 6.3.2, respectively:

(1) using the distribution that Box and Tiao (1973) use as reference prior when considering the random effects model in the context of one-way analysis of variance, and

(2) using the natural conjugate prior distribution for the parameters σ^2(or Σ), ξ and τ^2 (or T).

Because of the fact that the results for the univariate and multivariate situations are, apart from algebraic complexity, essentially the same, the above two cases will be investigated only for the univariate situation. In the first case the result obtained will, however, also be given for the corresponding multivariate situation.

Finally, some general comments about the Predictive Bayesian approach under the random effect model will be made in Sub-section 6.3.3.

6.3.1 Box and Tiao's Prior Distribution

Box and Tiao (1973), *Chapter 5*, make the point that under the random effects model with equal sample sizes from each group, the sampling theory estimator $\hat{\tau}^2$ for the variance τ^2 of the population means μ_i, given by:

$$\hat{\tau}^2 = \frac{S_1-S_2}{n}$$

where S_1 and S_2 are the between groups and within groups mean squares, respectively, as defined in Table 5.1.1 for p = 1 dimension, may be negative.

In order to avoid this possibility within the Bayesian framework, they propose the following noninformative joint prior density for the parrmeters σ^2, ξ and τ^2:

$$g(\sigma^2,\xi,\tau^2)d\sigma^2 d\xi d\tau^2 \propto (\sigma^2)^{-1}(\sigma^2+n\tau^2)^{-1}d\sigma^2 d\xi d\tau^2 \ . \qquad (6.3.1)$$

<u>Remark 6.3.1</u> This prior distribution can be criticised because of the fact that the within-groups sample size n appears in expression (6.3.1) for its density. Thus the prior distribution is in this sense dependent on the actual likelihood function of the sample itself, and not only on the form of the likelihood function, as is usually the case.

As before, we will generalise expression (6.3.1) for the prior density slightly by using the following form:

$$g(\sigma^2,\xi,\tau^2)d\sigma^2 d\xi d\tau^2 \propto (\sigma^2)^{-\frac{1}{2}\nu_1}(\sigma^2+n\tau^2)^{-\frac{1}{2}\nu_2}d\sigma^2 d\xi d\tau^2 \ . \qquad (6.3.2)$$

The form used by Box and Tiao is therefore given by (6.3.2) with $\nu_1 = \nu_2 = 2$.

Substituting (6.2.2) into (6.1.2) and (6.1.3) of Section 6.1 and using the same notation as in (6.1.4) gives:

$$f(x|TS,\nu_1,\nu_2,\pi_r) \propto \int_{\sigma^2}\int_{\underline{\mu}}\int_{\tau^2}\int_{\xi} (\sigma^2)^{-\frac{1}{2}N} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2\}$$

$$\times \tau^{-k}\exp\{-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i-\xi)^2\}(\sigma^2)^{-\frac{1}{2}\nu_1}(\sigma^2+n\tau^2)^{-\frac{1}{2}\nu_2}d\xi d\tau^2 d\underline{\mu} d\sigma^2$$

$$(6.3.3)$$

where, as before, it has been assumed that $n_i^* = n$, $\forall i$, so that the j^{th} observation x_{rj} from π_r has been replaced by x, x has been re-labelled x_{rj} and N has been replaced by N-1.

As shown in appendix 6.4, this eventually yields:

$$f(x|TS, v_1, v_2, \pi_r) \propto (A_1^*)^{-\frac{1}{2}(v_2+k-3)} (A_2^*)^{-\frac{1}{2}(N+v_1-k-2)} E_z\left[\Gamma_{\frac{A_1^* z}{A_2^*}}(\tfrac{1}{2}(v_2+k-3))\right] \qquad (6.3.4)$$

where,

$$A_1^* = n \sum_{i=1}^{k} (x_{i.}^* - x_{..}^*)^2$$

and

$$A_2^* = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - x_{i.}^*)^2$$

are the between group and within group sums of squares, respectively, with x replacing one of the observations, x_{rj}, from the r^{th} group, and $x_{i.}^*$ and $x_{..}^*$ are the corresponding adjusted i^{th} group and overall means,

$$\Gamma_y(m) = \int_0^y w^{m-1} e^{-w} \, dw$$

is the incomplete gamma function, and the expectation is taken over the distribution of z, where z has a gamma distribution with parameter $\frac{1}{2}(N+v_1-k-2)$. Therefore, for $v_1 = v_2 = 2$, the predictive density of x, given the training sample, Box and Tiao's prior distribution and the hypothesis that $x \in \pi_r$, is, from (6.3.4)

$$f(x|TS, \pi_r) \propto (A_1^*)^{-\frac{1}{2}(k-1)} (A_2^*)^{-\frac{1}{2}(N-k)} E_z\left[\Gamma_{\frac{A_1^* z}{A^*_2}}(\tfrac{1}{2}(k-1))\right] \qquad (6.3.5)$$

where z has a gamma distribution with parameter $\frac{1}{2}(N-k)$.

To evaluate expression (6.3.5), the easiest approach is to use Pearson's (1922) formula for the incomplete gamma function (the formula given by Pearson is for the incomplete gamma function ratio $\Gamma_y(m)/\Gamma(m)$):

$$\Gamma_y(m) = m^{-1}\exp(-y) \sum_{j=0}^{\infty} y^{m+j}/(1+m)^{[j]} . \qquad (6.3.6)$$

Applying (6.3.6) to (6.3.5) and interchanging the order of integration and summation (justified by the uniform convergence of (6.3.6) for all y), yields:

$$f(x|TS,\pi_r) = (A_1^*)^{-\frac{1}{2}(k-1)}(A_2^*)^{-\frac{1}{2}(N-k)} \sum_{j=0}^{\infty} \frac{(A_1^*/A_2^*)^{\frac{1}{2}(k-1)+j}}{(\frac{1}{2}(k+1))^{[j]}}$$

$$\times \int_{\varepsilon}^{\infty} \exp\{-(A_1^*/A_2^*)z\}z^{\frac{1}{2}(k-1)+j}z^{\frac{1}{2}(N-k)-1}\exp\{-z\}dz$$

$$= (A_2^*)^{-\frac{1}{2}(N-1)} \sum_{j=0}^{\infty} \frac{(A_1^*/A_2^*)^{j}}{(\frac{1}{2}(k+1))^{[j]}} \int_0^{\infty} z^{\frac{1}{2}(N-1)+j-1} \exp\{-(A_3^*/A_2^*)z\}dz$$

where $A_3^* = A_1^* + A_2^*$.

The integral may be evaluated as a gamma function, and after some simplification this eventually yields the following expression for the predictive density:

$$f(x|TS,\pi_r) = (A_3^*)^{-\frac{1}{2}(N-1)} \sum_{j=0}^{\infty} \frac{(\frac{1}{2}(N-1))^{[j]}}{(\frac{1}{2}(k+1))^{[j]}} (A_1^*/A_3^*)^{j}$$

$$= (A_3^*)^{-\frac{1}{2}(N-1)}F(1,\tfrac{1}{2}(N-1);\ \tfrac{1}{2}(k+1);\ A_1^*/A_3^*) \qquad (6.3.7)$$

where $F(\alpha,\beta;\gamma;x)$ is the hypergeometric function defined in (6.1.13).

Remark 6.3.2 It is interesting to note the close similarity between expressions (6.3.7) and (6.1.17), the former based on Box and Tiao's noninformative prior distribution (6.3.1) for the random effects model, and the latter on the noninformative prior distribution (6.1.1), with $v_1 = 2$ and $v_2 = 1$. In order to establish just how similar these two expressions are, (6.3.7) was applied to the data of Example 6.1.1, yielding the following posterior probabilities for each of the five populations, assuming equal prior probabilities:

Population	1	2	3	4	5
Posterior prob.	.0072	.0565	.4921	.3766	.0676

These probabilities agree, to two decimal places, with those obtained using (6.1.17), confirming that the choice of noninformative prior distribution has little effect on the predictive densities.

Finally, it is interesting to note that we do not experience any problems with the parameters v_1 and v_2 in the Box and Tiao prior distribution, in contrast to the case with the more usual noninformative prior.

Remark 6.3.3 In the multivariate case, Box and Tiao's prior distribution for the Random Effects model is:

$$P(\xi,T,\Sigma) \propto |\Sigma|^{-\frac{1}{2}(p+1)}|\Sigma+nT|^{-\frac{1}{2}(p+1)} \qquad (6.3.8)$$

and the predictive density of x becomes, in an analogous manner to (6.3.7):

$$f(x|TS,\pi_\mu) \propto |A_3^*|^{-\frac{1}{2}(N-p)} {}_2F_1(\tfrac{1}{2}(p+1),\tfrac{1}{2}(N-p);\ \tfrac{1}{2}(k+1);\ A_3^{*-1}A_1^*) \qquad (6.3.9)$$

where A_1^* and A_3^* are defined in Section 6.2 and ${}_2F_1(a_1, a_2; b_1; \Omega)$ is the hypergeometric function of matrix argument defined in (6.2.16).

6.3.2 Natural Conjugate Prior Distributions

The joint natural conjugate prior distribution for the case $p = 1$ for the mean and variance of the normal distribution is the Normal-inverted χ^2 distribution (see, for example, Press (1972)) with density function:

$$f(\xi,\tau^2) \propto (\tau^2)^{-\frac{1}{2}(\nu_2+1)}\exp\{-\tfrac{1}{2}((\tfrac{\xi-b}{c\tau})^2 + \tfrac{d^2}{\tau^2})\} \qquad (6.3.10)$$

where ν_2, b,c and d are constants and $\nu_2 > 2$. The natural conjugate prior distribution for σ^2 is the inverted χ^2 distribution, with density function (see, for example, Box and Tiao (1973)):

$$g(\sigma^2) \propto (\sigma^2)^{-\frac{1}{2}\nu_1} \exp(-\tfrac{1}{2} \frac{a^2}{\sigma^2}) \qquad (6.3.11)$$

where ν_1 and a are constants and $\nu_1 > 2$, and it would seem reasonable to assume that σ^2 is independent of (ξ,τ^2).

Substituting (6.3.10) and (6.3.11) into (6.1.2) and (6.1.3) yields the following expression for the predictive density of x, where we have assumed that $n_i = n$, $\forall_i$ and that x has replaced some x_{rj} in the training sample from π_r:

$$f(\underline{x}|TS,a,b,c,d,\nu_1,\nu_2,\pi_r) \propto \int_{\sigma^2}\int_{\underline{\mu}}\int_{\tau^2}\int_{\xi} \sigma^{-N} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2\}$$

$$\times\, \tau^{-k} \exp\{-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i-\xi)^2\}\sigma^{-\nu_1} \exp\{-\tfrac{1}{2}\frac{a^2}{\sigma^2}\}$$

$$\times\, \tau^{-(\nu_2+1)} \exp\{-\tfrac{1}{2}((\tfrac{\xi-b}{c\tau})^2 + \tfrac{d^2}{\tau^2})\}d\xi d\tau^2 d\underline{\mu} d\sigma^2$$

$$\propto \int_{\sigma^2}\int_{\underline{\mu}} \sigma^{-(N+v_1)} \exp\{-\frac{1}{2\sigma^2}(a^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2)\}$$

$$\int_{\tau^2}\int_{\xi} \tau^{-(k+v_2+1)} \exp\{-\frac{1}{2\tau^2}(\sum_{i=1}^{k}(\mu_i-\xi)^2+(\frac{\xi-b}{c})^2+d^2)\}d\xi d\tau^2 d\underline{\mu} d\sigma^2 \quad (6.3.12)$$

The inner pair of integrals I_1 in (6.3.12) are evaluated in a manner analogous to that used to evaluate the corresponding integrals in Section 6.1 , yielding

$$I_1 \propto (g(\underline{\mu}))^{-\frac{1}{2}(k+v_2-2)} \quad (6.3.13)$$

where $\quad g(\underline{\mu}) = \sum_{i=1}^{k}(\mu_i-\mu_.)^2 + \frac{k}{1+c^2k}(\mu_.-b)^2 + d^2$

Therefore,

$$f(x|TS,a,b,c,d,v_1,v_2) \propto \int_{\sigma^2}(\sigma^2)^{-\frac{1}{2}v_1} \exp\{-\frac{1}{2\sigma^2}(a^2 + A_2^*)\}$$

$$\int_{\underline{\mu}}(g(\underline{\mu}))^{-\frac{1}{2}(k+v_2-2)} \sigma^{-N} \exp\{-\frac{n}{2\sigma^2}\sum_{i=1}^{k}(\mu_i-x_{i.}^*)^2\}d\underline{\mu} d\sigma^2 \quad (6.3.14)$$

where $\quad A_2^* = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-x_{i.}^*)^2$

The inner integral in (6.3.14) can be considered as the expected value of

$(g(\underline{\mu}))^{-\frac{1}{2}(k+v_2-2)}$ where the μ_i, $i = 1,\ldots,k$, are independently distributed $N(x_{i.}^*, \frac{\sigma^2}{n})$ random variables.

A way of evaluating this expected value is to assume that $c^2 \ll k$, so that

$$\frac{k}{c^2k+1} \doteq k$$

and

$$g(\underline{\mu}) \doteq \sum_{i=1}^{k}(\mu_i - b)^2 + d^2$$

Remark 6.3.4 The assumption $c^2 \ll k$ implies that, a priori, ξ has a distribution that is narrowly concentrated around the value $\xi = b$ and that the information from this prior distribution far outweighs the information contained in the training sample.

Under this assumption it is clear that $g(\underline{\mu})$ is distributed as:

$$g(\underline{\mu}) \sim \frac{\sigma^2}{n}\ \chi_k^2(\lambda^*) + d^2$$

where $\chi_k^2(\lambda^*)$ represents a noncentral chi-squared random variable with k degrees of freedom and noncentrality parameter:

$$\lambda^* = \frac{n}{\sigma^2}\sum_{i=1}^{k}(x_{i.}^* - b)^2$$

So the inner integral in (6.3.14) can be considered to be proportional to the $-\frac{1}{2}(k+v_2-2)^{th}$ moment of $\frac{n}{\sigma^2}$ times a $\chi_k^2(\lambda^*)$ distribution that has been shifted an amount $\frac{nd^2}{\sigma^2}$ to the right. From Appendix 6.1 we know that this moment will exist only if

$$-\tfrac{1}{2}(k + v_2 - 2) > -\frac{k}{2}$$

i.e. only if $\quad v_2 < 2$

However, this condition violates the condition $v_2 > 2$ that is necessary for the natural conjugate prior to be a proper distribution.

On the surface it would therefore appear that when the parameters σ^2, ξ and τ^2 follow their natural conjugate prior distributions, then the predictive density of x does not exist. However, this contradicts the fact that since the joint distribution of x, $\underline{\mu}$, σ^2, ξ and τ^2 is proper, the marginal distribution, and therefore the predictive density, of x must exist. The reason for this contradiction clearly lies in the approximating assumption on $g(\underline{\mu})$ which has apparently been so powerful as to have rendered improper the predictive distribution of x.

As it does not appear to be possible to evaluate the integral (6.3.14) analytically without this approximating assumption on $g(\underline{\mu})$, we will not pursue the matter any further. It is nevertheless interesting to compare the situation found here with that when σ^2, ξ and τ^2 follow diffuse prior distributions. Under those circumstances the predictive density of x does not exist when the parameter v_2 in the prior density of τ^2 is given the value 2, required for it to be noninformative in the usual sense.

6.3.3 Final Remarks

From the results of the previous two sub-sections we therefore know that:

1) the posterior probabilities of the k_1 populations from which the observation x could have come are not materially affected by the form of noninformative prior distribution used for the parameters σ^2 (or Σ), ξ and τ^2 (or T), be it the more usual (Jeffreys, 1961) invariant prior distribution (with modification to the parameter v_2) or Box and Tiao's (1973) prior distribution for the random effects model;

2) if the abovementioned parameters follow their natural conjugate prior distributions then the corresponding predictive densities cannot be evaluated.

The formulae for the predictive densities derived in this section and in the previous two are all expressed compactly in terms of hypergeometric functions, which are readily evaluated on a computer or even a modern programmable pocket calculator for the case p=1. For higher dimensions however, in spite of the existence of the programs of van der Westhuizen and Nagel (1979) for computing zonal polynomials, described in sub-section 6.2.1, the computation of the hypergeometric functions of matrix argument, and hence the predictive densities and posterior probabilities, is not yet a practical proposition.

The only ambiguity in all the abovementioned formulae derives from the fact that x can replace any one of the n observations x_{rj}, $j = 1,\ldots,n$ in the training sample from π_r when computing the quantities A_1^*, A_2^* and A_3^* appearing in them.

A sensible rule for getting around this ambiguity would be to replace that observation x_{rj_*} that is closest to the sample mean from the r^{th} population, as measured by the Mahalanobis distance. i.e. Choose x_{rj_*} such that

$$d_r^2(x_{rj}) = (x_{rj} - x_{r.})' \, S^{-1}(x_{rj} - x_{r.})$$

is minimised when $j = j_*$.
This rule would avoid the possibility of anomalous results due to, for example, an extreme observation from π_r being replaced by x.

6.4 Other Bayesian Type Approaches

In this section two further approaches to discriminant analysis, the Empirical Bayes and Semi-Bayes approaches, are discussed in the context of the random effects model. In each case the discussion is confined to a brief description of the approach, its application to the present problem, the derivation of preliminary results and recommendations for further research.

6.4.1 The Empirical Bayes Approach

Good descriptions of the Empirical Bayes approach to statistical inference may be found in many texts (see, for example, Maritz (1970), Cox and Hinkley (1974) and van Niekerk (1978)) and therefore a brief sketch here will suffice.

Suppose we have an observation x made on a random variable X whose distribution function F(X|λ) depends on an unknown (vector) parameter λ. In both the "pure" Bayes and Empirical Bayes methods the parameter λ is assumed to have a prior distribution, the point of departure between the two being the way in which this prior distribution is treated. As we have seen, the "pure" Bayes approach assumes that the prior distribution of λ is either completely specified or that any unknown parameters in it themselves have prior distributions that are completely specified. In contrast, the Empirical Bayes (EB) approach gives the prior distribution of λ a frequency interpretation whose parameters may be estimated from previous data by classical techniques. Therefore the E.B. approach uses the mathematical techniques and results of the "pure" Bayes approach, but avoids the problem in this approach of having to specify the prior distribution completely.

For example, it is well known (see, for example, Maritz (1970)) that the Bayes point estimator of λ given x is, using a quadratic loss function:

$$\hat{\lambda}(x) = \frac{\int \lambda \, dF(x|\lambda) \, dG(\lambda)}{\int dF(x|\lambda) \, dG(\lambda)} \tag{6.4.1}$$

where,

G(λ) is the prior distribution function of λ and the integration is performed with respect to G(λ).

The E.B. estimator of λ is now obtained from (6.4.1) by replacing G(λ) by $\hat{G}(\lambda)$, the sample-based estimator of the prior distribution function of λ .

We may apply formula (6.4.1) to our random effects model as follows. Assume that

$$X|\mu \sim N_p(\mu, \Sigma) \tag{6.4.2}$$

where, a priori,

$$\mu \sim N_p(\xi, T) \tag{6.4.3}$$

Given an observation x of X, our Bayesian point estimator of the corresponding μ is:

$$\hat{\mu}(x) = \frac{\int \mu\, f(x|\mu)\, g(\mu)\, d\mu}{\int f(x|\mu)\, g(\mu)\, d\mu} \tag{6.4.4}$$

where $f(x|\mu)$ and $g(\mu)$ are the density functions of the distributions (6.4.2) and (6.4.3) respectively.

This yields, after some algebra (see, for example, Maritz (1970) for the univariate case):

$$\hat{\mu}(x) = x - \Sigma(\Sigma + T)^{-1}\ (x - \xi) \tag{6.4.5}$$

The E.B. estimator of μ is now obtained by replacing the unknown parameters Σ, ξ and T in (6.4.5) by their sample-based estimators $\hat{\Sigma}$, $\hat{\xi}$ and $\hat{T}$, respectively.

In practice, particularly in discriminant analysis, we will generally have more than one observation x on which to base our estimator of μ . In the situation considered in this thesis, where we have a training sample $\{x_{ij},\ j = 1,\ldots,n\}$ of size n from each of k populations π_i, $i = 1,\ldots,k$, as described in Section 5.1, then our E.B. estimator of the mean μ_i of π_i will be based on the sample mean $x_{i.}$. Remembering that

$$x_{i.}|\mu_i \sim N_p(\mu_i, \frac{1}{n}\Sigma)$$

and using the notation of Table 5.1.1, the E.B. estimator of μ_i is, from (6.4.5):

$$\begin{aligned}\hat{\mu}_i(EB) &= x_{i.} - \frac{1}{n}\hat{\Sigma}(\frac{1}{n}\hat{\Sigma} + \hat{T})^{-1}(x_{i.} - \hat{\xi}) \\ &= x_{i.} - S_2 S_1^{-1}(x_{i.} - x_{..}) \qquad (6.4.6)\end{aligned}$$

where S_1 and S_2 are the between group and within group mean square matrices, respectively.

Coming now to our discriminant analysis problem, the Bayesian classification rule that minimises the expected loss from misclassification (assuming equal costs of misclassification) is to classify the observation x of unknown origin into that population π_i for which:

$$(x - \tfrac{1}{2}(\mu_i + \mu_j))'\Sigma^{-1}(\mu_i - \mu_j) > \log\frac{q_j}{q_i} \qquad \forall j = 1,\ldots,k;\ j \neq i \quad (6.4.7)$$

where q_j is the prior probability that x comes from π_j. (See (2.1.3) in Chapter 2).

As mentioned in Sub-section 2.1.1, Anderson's (1951) "plug-in" rule (2.1.19) obtained by replacing the unknown parameters μ_i, μ_j and Σ in (6.4.7) by their maximum likelihood estimators $x_{i.}$, $x_{j.}$ and S_2, respectively, is an E.B. procedure under the fixed effects model. Under the random effects model the E.B. procedure is to replace μ_i and μ_j by $\hat{\mu}_i$(EB) and $\hat{\mu}_j(EB)$ respectively, given in (6.4.6), and Σ by S_2. This yields the following E.B. classification rule:

Classify x into that population π_i for which

$$(x-\tfrac{1}{2}(I-S_2 S_1^{-1})(x_{i.}+x_{j.})+S_2 S_1^{-1} x_{..})' S_2^{-1}(I-S_2 S_1^{-1})(x_{i.}-x_{j.}) > \log \frac{q_j}{q_i}$$

$$\forall j = 1,\ldots,k; \quad j \neq i \qquad (6.4.8)$$

Therefore under the random effects model, the classification rule corresponding to Anderson's (1951) rule for the fixed effects case is given by (6.4.8).

The properties and behaviour of classification rule (6.4.8) have not yet been studied, and this indicates a promising area for future research.

It is interesting to note that the E.B. estimator (6.4.6) for μ_i, which may also be written as:

$$\hat{\mu}_i(EB) = (I-A)x_{i.} + A(x_{..}) \qquad (6.4.9)$$

where $A = S_2 S_1^{-1}$

is the multivariate analogue of the James - Stein (1961) "shrinkage" estimator (slightly modified) of μ_i. See, for example, Cox and Hinkley (1974). It also corresponds to the approximate large sample posterior mean of μ_i under the random effects model, given by Box and Tiao (1973) when their prior distribution, discussed in Sub-section 6.3.1, is used.

6.4.2 The Semi-Bayes Approach

Geisser (1967) coins the term "Semi-Bayes" to describe the Bayesian analysis of the properties of the classical approach to discriminant analysis based on the Linear Discriminant Function (or the Quadratic Discriminant Function in the case of unequal within-group covariance matrices). Considering the two population problem, he investigates both situations where the parameters are known and the classification rule (given in (2.1.6)) is based on the population discriminant function:

$$U_{12}(x) = (x - \tfrac{1}{2}(\mu_1 + \mu_2))' \, \Sigma^{-1}(\mu_1 - \mu_2) \qquad (6.4.10)$$

and where they are unknown, and the classification rule (given in (2.1.19)) is based on the sample discriminant function:

$$V_{12}(x) = (x - \tfrac{1}{2}(x_{1.} + x_{2.}))' \, S^{-1}(x_{1.} - x_{2.}) \qquad (6.4.11)$$

Given training samples of size n_1 and n_2 (denoted collectively by TS) from the two populations π_1 and π_2, respectively, and assuming a diffuse prior distribution for the parameters μ_1, μ_2 and Σ^{-1}, the joint posterior density of these parameters becomes:

$$f(\mu_1, \mu_2, \Sigma^{-1}|TS) \propto |\Sigma^{-1}|^{\frac{1}{2}(\nu-p+1)} \exp\{-\tfrac{1}{2} \operatorname{Tr} \Sigma^{-1}[\nu S + n_1(x_{1.} - \mu_1)(x_{1.} - \mu_1)' + n_2(x_{2.} - \mu_2)(x_{2.} - \mu_2)']\} \qquad (6.4.12)$$

where the notation is the same as that used in earlier sections. Using (6.4.12) as his starting point, Geisser (1967) first investigates the posterior distribution of $U_{12}(x)$ and hence obtains expressions for the posterior limits on the "true" probabilities of misclassification when classification rule (2.1.6), based on $U_{12}(x)$, is used. It turns out that these [illegible] may be obtained directly from the posterior distribution of $\delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)$, for which the following expression for its density function is derived:

$$f_{\delta^2}(x) = \sum_{j=0}^{\infty} w_j \, g_{p+2j}(x) \tag{6.4.13}$$

where,

the w_j are the individual terms of a negative binomial density

and $g_{p+2j}(\cdot)$ is the density function of the χ^2_{p+2j} distribution.

Remark 6.4.1 It is interesting to note the similarity between (6.4.13) and expressions (3.1.11) and (3.1.12) for the density function of δ^2 under the random effects model.

Secondly, *Geisser* (1967) obtains posterior limits on the conditional or "index" probabilities of misclassification when using classification rule (2.1.19) based on the sample discriminant function $V_{12}(x)$. *Because* of the complicated distribution theory involved, asymptotic theory is used to obtain approximate limits in terms of the standard normal integral which he shows should be reasonably accurate even for moderate sample sizes. Finally, he obtains expressions, in terms of the t-distribution function, for the unconditional (or posterior predictive) probabilities of misclassification when the sample-based classification rule is used.

To apply this Semi-Bayesian approach to our random effects model, we need first to obtain the joint posterior distribution of the parameters in this model corresponding to expression (6.4.12) in the fixed effects case. In what follows, therefore, we will derive this distribution using a diffuse prior on the parameters Σ^{-1}, ξ and T^{-1}. As shall be seen, however, applying this distribution to the discriminant analysis problem in a manner analogous to Geisser (1967) does not promise to be a straightforward matter.

Considering first the two-group case, the joint *posterior density* of the parameters μ_1, μ_2, Σ^{-1}, ξ and T, given the training sample $TS = \{ x_{ij};\ j = 1,\dots,n_i;\ i = 1,2\}$, may be written:

$$P(\mu_1,\mu_2,\Sigma^{-1},\xi,T^{-1}) \propto f(TS|\mu_1,\mu_2,\Sigma^{-1})P(\mu_1,\mu_2|\xi,T^{-1})\ \ P(\Sigma^{-1},\xi,T) \qquad (6.4.14)$$

where,

$$f(TS|\mu_1,\mu_2,\Sigma^{-1}) = \prod_{i=1}^{2}\prod_{j=1}^{n_i}(2\pi)^{-\frac{1}{2}p}|\Sigma^{-1}|^{\frac{1}{2}}\exp\{-\tfrac{1}{2}(x_{ij}-\mu_i)'\Sigma^{-1}(x_{ij}-\mu_i)\}$$

$$P(\mu_1,\mu_2|\xi,T^{-1}) = \prod_{i=1}^{2}(2\pi)^{-\frac{1}{2}p}|T^{-1}|^{\frac{1}{2}}\exp\{-\tfrac{1}{2}(\mu_i-\xi)'\ T^{-1}(\mu_i-\xi)\}$$

and

$$P(\Sigma^{-1},\xi,T^{-1}) \propto |\Sigma|^{\frac{1}{2}(p+1)}|T|^{\frac{1}{2}(p+1)}$$

After some simplification, and assuming that $n_1 = n_2 = n$, this becomes:

$$P(\mu_1,\mu_2,\Sigma^{-1},\xi,T^{-1}|TS) \propto |\Sigma|^{-\frac{1}{2}(N-p-1)}|T|^{\frac{1}{2}(p-1)}\exp\{-\tfrac{1}{2}\,Tr\,\Sigma^{-1}A_2\}$$

$\times \exp\{-\tfrac{1}{2}\sum_{i=1}^{2}[n(x_{i.}-\mu_i)'\Sigma^{-1}(x_{i.}-\mu_i)+($ [illegible] $^{-1}(\mu_i-\xi)]\}$ (6.4.15)

where,

$$N = 2n$$

and $$A_2 = \sum_{i=1}^{2}\sum_{j=1}^{n}(x_{ij} - x_{i.})(x_{ij} - x_{i.})'\ .$$

We *may simplify the exponent* in (6.4.15) by using the following identity given by Box and Tiao (1973) in their appendix A7.1:

$$(x-a)'\ A(x-a)+(x-b)'\ B(x-b)=(x-c)'(A+B)(x-c)+(a-b)'(A^{-1}+B^{-1})^{-1}(a-b)$$

where,

x, a and b are p-dimensional vectors, A and B are $(p \times p)$ symmetric nonsingular matrices

and $c = (A+B)^{-1}(Aa+Bb)$

This finally yields the following expression for the joint posterior density of μ_1, μ_2, Σ^{-1}, ξ and T^{-1}:

$$P(\mu_1, \mu_2, \Sigma^{-1}, \xi, T^{-1}) \propto |\Sigma|^{-\frac{1}{2}(N-p-1)}|T|^{\frac{1}{2}(p-1)}$$

$$\times \exp\{-\tfrac{1}{2}\operatorname{Tr}[\Sigma^{-1}A_2+(n\Sigma^{-1}+T^{-1})\sum_{i=1}^{2}(\mu_i-c_i)(\mu_i-c_i)'$$

$$+(n^{-1}\Sigma+T)^{-1}(\sum_{i=1}^{2}(x_{i.}-x_{..})(x_{i.}-x_{..})'+2(x_{..}-\xi)(x_{..}-\xi)')]\} \tag{6.4.16}$$

where $c_i = (n\Sigma^{-1}+T^{-1})^{-1}(n\Sigma^{-1}x_{i.}+T^{-1}\xi) \quad i = 1,2$

For the general k-group case (6.4.16) becomes, assuming $n_i = n, \ \forall i = 1,\ldots,k$:

$$P(\mu_1,\ldots,\mu_k, \Sigma^{-1},\xi, T^{-1}) \propto |\Sigma|^{-\frac{1}{2}(N-p-1)}|T|^{-\frac{1}{2}(k-p-1)}$$

$$\times \exp(-\tfrac{1}{2}\operatorname{Tr}[\Sigma^{-1}A_2+(n\Sigma^{-1}+T^{-1})\sum_{i=1}^{k}(\mu_i-c_i)(\mu_i-c_i)'$$

$$+(n^{-1}\Sigma+T)^{-1}(A_1+k(x_{..}-\xi)(x_{..}-\xi)')]\} \tag{6.4.17}$$

wher≥,

c_i is the same as in (6.4.16)

$A_1 = \sum_{i=1}^{k}(x_{i.}-x_{..})(x_{i.}-x_{..})'$

and $A_2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-x_{i.})(x_{ij}-x_{i.})'$

Expressions (6.4.16) or (6.4.17) should therefore be used instead of (6.4.12) as starting point for the Semi-Bayesian analysis under the random effects model.

Comparing these expressions, it is apparent that the Semi-Bayesian analysis under the random effects model will be considerably more difficult than under the fixed effects model, and we will therefore not proceed *any further* with it is *this thesis*.

Nevertheless, this promises to be an interesting direction for research, especially if it is applied to the classification rule based on the modified discriminant function (6.4. 8) derived in Sub-section 6.4.1 using the Empirical Bayes approach.

Finally it is interesting to note that the approach of Chapter 3 and 4 is the classical analogue, under the random effects model, of Geisser's Semi-Bayesian approach to the analysis of the properties of the classical rules of discriminant analysis.

Appendix 6.1 Derivation of the r^th moment of the $\chi^2_\nu(\lambda)$ distribution

The density function of $X \sim \chi^2_\nu(\lambda)$ can be written in the following form (see, for example C.R. Rao, 1965):

$$f_X(x) = \exp(-\tfrac{1}{2}\lambda) \sum_{j=0}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} g_{\nu+2j}(x) \qquad \text{(A6.1.1)}$$

where $g_{\nu+2j}(x)$ is the density of the central $\chi^2_{\nu+2j}$ distribution. Therefore,

$$E[X^r] = \exp(-\tfrac{1}{2}\lambda) \sum_{j=0}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} E[(\chi^2_{\nu+2j})^r] \qquad \text{(A6.1.2)}$$

by the uniform convergence of the infinite series in (A6.1.1). Now, it is well known that

$$E[(\chi^2_{\nu+2j})^r] = 2^r \frac{\Gamma(\frac{1}{2}\nu+j+r)}{\Gamma(\frac{1}{2}\nu+j)} \qquad \text{(A6.1.3)}$$

for $r > -\frac{1}{2}(\nu+2j)$ and is not defined otherwise. Substituting (A6.1.3) into (A6.1.2) yields:

$$E[X^r] = 2^r \exp(-\tfrac{1}{2}\lambda) \sum_{j=0}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} \frac{\Gamma(\frac{1}{2}\nu+j+r)}{\Gamma(\frac{1}{2}\nu+j)} \qquad \text{(A6.1.4)}$$

where $X \sim \chi^2_\nu(\lambda)$.

Appendix 6.2

Derivation of the computational formulae for

$$A^*_{2(r,\ell)} = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij}-x_{i.})^2_{(r,\ell)} \qquad \text{(A6.2.1)}$$

and

$$A^*_{1(r,\ell)} = n \sum_{i=1}^{k} (x_{i.}-x_{..})^2_{(r,\ell)} \qquad \text{(A6.2.2)}$$

where the subscript (r,ℓ) denotes that observation $x_{r\ell}$ from the r^{th} population has been replaced by x. The computational formulae derive immediately from the following two general results:

Let x_j be replaced by x in the sample $\{x_i,\ i = 1,\ldots,n\}$. Then:

$$\text{(i)}\quad x_{.(j)} = x_. + \frac{x-x_j}{n} \qquad \text{(A6.2.3)}$$

$$\text{(ii)}\quad SS_{(j)} = SS - (x_j-x_.)^2 + (x-x_.)^2 - \frac{(x-x_j)^2}{n}. \qquad \text{(A6.2.4)}$$

where ,

$$x_. = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$SS = \sum_{i=1}^{n} (x_i-x_.)^2$$

and the subscript (j) has the same meaning as above.

Proof:

$$\text{(i)}\quad x_{.(j)} = \frac{1}{n}\{\sum_{i=1}^{n} x_i-x_j+x\} = x_. + \frac{x-x_j}{n}$$

(ii) $$SS_{(j)} = \sum_{i=1}^{n} (x_i - x_{.(j)})^2 - (x_j - x_{.(j)})^2 + (x - x_{.(j)})^2$$

$$= \sum_{i=1}^{n} \left(x_i - x_{.} - \frac{x-x_j}{n}\right)^2 - \left(x_j - x_{.} - \frac{x-x_j}{n}\right)^2 + \left(x - x_{.} - \frac{x-x_j}{n}\right)^2$$

$$= \sum_{i=1}^{n} \quad x_{.})^2 + \frac{(x-x_j)^2}{n} - (x_j - x_{.})^2 + \frac{2}{n}(x_j - x_{.})(x - x_j) - \frac{(x-x_j)^2}{n^2}$$

$$+ (x - x_{.})^2 - \frac{2}{n}(x - x_{.})(x - x_j) + \frac{(x-x_j)^2}{n^2}$$

$$= SS - (x_j - x_{.})^2 + (x - x_{.})^2 + \frac{(x-x_j)^2}{n} - \frac{2}{n}(x - x_j)(x - x_{.} - x_j + x_{.})$$

$$= SS - (x_j - x_{.})^2 + (x - x_{.})^2 - \frac{(x-x_j)}{n} . \qquad \text{Q.E.D.}$$

Applying (A6.23) and (A6.24) to $A^*_2(r,\ell)$ yields:

$$A^*_{2(r,\ell)} = \sum_{\substack{i=1 \\ i \neq r}}^{k} \sum_{j=1}^{n} (x_{ij} - x_{i.})^2 + \sum_{j=1}^{n} (x_{rj} - x_{r.})^2_{(r,\ell)}$$

$$= \sum_{\substack{i=1 \\ i \neq r}}^{k} \sum_{j=1}^{n} (x_{ij} - x_{i.})^2 + \sum_{j=1}^{n} (x_{rj} - x_{r.})^2 - (x_{r\ell} - x_{r.})^2 + (x - x_r)^2$$

$$- \frac{(x - x_{r\ell})^2}{n}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - x_{i.})^2 - (x_{r\ell} - x_{r.})^2 + (x - x_{r.})^2 - \frac{1}{n}(x - x_{r\ell})^2 . \tag{A6.2.5}$$

Considering $A^*_{1(r,\ell)}$, note that, from (A6.2.3),

$$x_{r.(r,\ell)} = x_{r.} + \frac{x - x_{r\ell}}{n} .$$

Therefore, applying(A6.2.4) to (A6.2.2), with n replaced by k, x_i by $x_{i.}$, $x_.$ by $x_{..}$, x_j by $x_{r.}$ and x by $x_{r.} + \frac{x-x_{r\ell}}{n}$ gives:

$$\begin{aligned}\frac{1}{n}A^*_{1(r,\ell)} &= \sum_{i=1}^{k}(x_{i.}-x_{..})^2 - (x_{r.}-x_{..})^2 + (x_{r.}+\frac{x-x_{r\ell}}{n} - x_{..})^2 \\ &\qquad - \frac{1}{k}\left(\frac{x-x_{r\ell}}{n}\right)^2 \\ &= \sum_{i=1}^{k}(x_{i.}-x_{..})^2 - (x_{r.}-x_{..})^2 + (x_{r.}-x_{..})^2 + \frac{2}{n}(x_{r.}-x_{..})(x-x_{r\ell}) \\ &\qquad + \frac{1}{n^2}(x-x_{r\ell})^2 - \frac{1}{kn^2}(x-x_{r\ell})^2 \\ &= \sum_{i=1}^{k}(x_{i.}-x_{..}) + \frac{2}{n}(x-x_{r\ell})(x_{r.}-x_{..}) + \frac{k-1}{kn^2}(x-x_{r\ell})^2\end{aligned} \qquad \text{(A6.2.6)}$$

<u>Appendix 6.3</u>

Evaluating:

$$I = \int_T \int_\xi |T|^{-\frac{1}{2}(k+v_2)} \exp(-\tfrac{1}{2}\sum_{i=1}^{k}(\mu_i-\xi)'T^{-1}(\mu_i-\xi))d\xi dT \qquad \text{(A6.3.1)}$$

where,

T is a (p×p) symmetric matrix

and ξ is a p×1 vector.

Note that:

$$\sum_{i=1}^{k}(\mu_i-\xi)'T^{-1}(\mu_i-\xi) = \sum_{i=1}^{k}(\mu_i-\mu_{.})'T^{-1}(\mu_i-\mu_{.}) + k(\xi-\mu_{.})'T^{-1}(\xi-\mu_{.})$$

$$= \mathrm{Tr}(T^{-1}A_{\underline{\mu}}) + k(\xi-\mu_{.})'T^{-1}(\xi-\mu_{.}) \qquad (A6.3.2)$$

where

$$\mu_{.} = \frac{1}{k}\sum_{i=1}^{k}\mu_i$$

and

$$A_{\underline{\mu}} = \sum_{i=1}^{k}(\mu_i-\mu_{.})(\mu_i-\mu_{.})'.$$

Substituting (A6.3.2) into (A6.3.1) *yields:*

$$I = \int_T |T|^{-\frac{1}{2}(k+\nu_2-1)}\exp\{-\tfrac{1}{2}\mathrm{Tr}(T^{-1}A_{\underline{\mu}})\}\int_\xi |T|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}k(\xi-\mu_{.})'T^{-1}(\xi-\mu)\}d\xi dT$$

$$(A6.3.3)$$

Since the integrand of the inner integral in (A6.3.3) is proportional to the multivariate normal density function, we have that:

$$I = \int |T|^{-\frac{1}{2}(k+\nu_2-1)}\exp\{-\tfrac{1}{2}\mathrm{Tr}(T^{-1}A_{\underline{\mu}})\}dT. \qquad (A6.3.4)$$

The integrand in (A6.3.4) is proportional to the density function of the Inverted Wishart Distribution (see, for example Press, 1972), the constant of proportionality being

$$C = |A_{\underline{\mu}}|^{\frac{1}{2}(k+\nu_2-p-2)}/2^{\frac{1}{2}p(k+\nu_2-p-2)}\Gamma_p(\tfrac{1}{2}(k+\nu_2-p-2) \qquad (A6.3.6)$$

where $\Gamma_p(\frac{1}{2}\nu)$ is the multivariate gamma function defined in (5.3.5). Hence,

$$I \propto C^{-1} \propto |A_{\underset{\sim}{\mu}}|^{-\frac{1}{2}(k+\nu_2-p-2)} \qquad (A6.3.6)$$

Appendix 6.4

Evaluating

$$I = \int_{\sigma^2}\int_{\underset{\sim}{\mu}}\int_{\tau^2}\int_{\xi} \sigma^{-(N+\nu_1)}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2\}$$

$$\times\, \tau^{-k}(\sigma^2+n\tau^2)^{-\frac{1}{2}\nu_2}\exp(-\frac{1}{2\tau^2}\sum_{i=1}^{k}(\mu_i-\xi)^2\}d\xi d\tau^2 d\underset{\sim}{\mu} d\sigma^2 \qquad (A6.4.1)$$

The exponent in the integrand in (A6.4.1)can be written:

$$-\tfrac{1}{2}\{\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-x^*_{i.})^2 + \frac{n}{\sigma^2}\sum_{i=1}^{k}(\mu_i - x^*_{i.})^2 + \frac{1}{\tau^2}\sum_{i=1}^{k}(\mu_i-\xi)^2\}$$

where x^*_i is defined in (6.1.6)

$$= -\tfrac{1}{2}(A^*_2/\sigma^2 + \sum_{i=1}^{k}(\frac{n}{\sigma^2}(\mu_i-x^*_{i.})^2 + \frac{1}{\tau^2}(\mu_i-\xi)^2)\}$$

where

$$A^*_2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-x^*_{i.})^2.$$

Using the result given by Box and Tiao (1973) in their equation (A1.1.5), viz:

$$A(z-a)^2 + B(z-b)^2 = (A+B)(z-c)^2 + \frac{AB}{A+B}(a-b)^2$$

where

$$c = \frac{1}{A+B}(Aa+Bb),$$

the exponent becomes:-

$$-\tfrac{1}{2}\{A_2^*/\sigma^2 + \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\sum_{i=1}^{k}(\mu_i - c_i)^2 + \left(\frac{1}{\sigma^2+n\tau^2}\right)(A_1 + nk(\xi - x_{..})^2)\} \quad (A6.4.2)$$

where

$$c_i = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{x_{i.}^*}{\sigma^2} + \frac{\xi}{\tau^2}\right),$$

$$A_1 = n\sum_{i=1}^{k}(x_{i.}^* - x_{..}^*)^2,$$

and $x_{..}$ is defined in (6.1.9).

Interchanging the order of integration, and using the above result, we get:

$$I = \int_{\tau^2}\int_{\sigma^2} \sigma^{-(N+\nu_1)}\tau^{-k}\exp\left\{-\frac{A_2^*}{2\sigma^2} - \frac{A_1^*}{2(\sigma^2+n\tau^2)}\right\}(\sigma^2+n\tau^2)^{-\frac{1}{2}\nu_2}$$

$$\times\int_{\xi}\exp\{-\frac{nk}{2(\sigma^2+n\tau^2)}(\xi - x_{..}^*)^2\}\int_{\underset{\sim}{\mu}}\exp(-\tfrac{1}{2}\left(\frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\right)^{-1}\sum_{i=1}^{k}(\mu_i - c_i)^2)$$

$$d\underset{\sim}{\mu}\,d\xi\,d\sigma^2 d\tau^2$$

$$\propto \int_{\tau^2}\int_{\sigma^2}\sigma^{-(N+\nu_1-k)}(\sigma^2+n\tau^2)^{-\frac{1}{2}(\nu_2+k-1)}\exp(-\tfrac{1}{2}\left(\frac{A_2^*}{\sigma^2} + \frac{A_1^*}{\sigma^2+n\tau^2}\right))d\sigma^2 d\tau^2$$

(A6.4.3)

If we now make the transformation:

$$y = \sigma^2$$
$$z = \sigma^2 + n\tau^2$$

with Jacobian $J = \frac{1}{n}$, we get:

$$I \propto \int_0^\infty y^{-\frac{1}{2}(N+\nu_1-k)} \exp(-\tfrac{1}{2}\frac{A_2^*}{y}) \int_y^\infty z^{-\frac{1}{2}(\nu_2+k-1)} \exp(-\tfrac{1}{2}\frac{A_1^*}{z}) dz dy. \quad (A6.4.4)$$

Denoting the inner integral in (A6.4.4) by I_1 and making the transformation:

$$w = \tfrac{1}{2}A_1^*/z$$

with Jacobian $J = \frac{1}{2}A_1^*/w^2$, I_1 becomes:

$$I_1 = (\tfrac{1}{2}A_1^*)^{-\frac{1}{2}(\nu_2+k-3)} \int_0^{\frac{A_1^*}{2y}} w^{\frac{1}{2}(\nu_2+k-3)-1} \exp(-w) dw$$

$$= (\tfrac{1}{2}A_1^*)^{-\frac{1}{2}(\nu_2+k-3)} \Gamma_{\frac{A_1^*}{2y}} (\tfrac{1}{2}(\nu_2+k-3)) \quad (A6.4.5)$$

where $\Gamma_x(n)$ denotes the incomplete gamma function.

Hence,

$$I \propto (A_1^*)^{-\frac{1}{2}(\nu_2+k-3)} \int_0^\infty \Gamma_{\frac{A_1^*}{2y}} (\tfrac{1}{2}(\nu_2+k-3)) y^{-\frac{1}{2}(N+\nu_1-k)} \exp\{-\tfrac{1}{2}\frac{A_2^*}{y}\} dy. \quad (A6.4.6)$$

Finally, making the transformation:

$$z = \tfrac{1}{2}\frac{A_2^*}{y}$$

with Jacobian $J = \frac{1}{2}A_{2}^{*}/z^{2}$, we get:

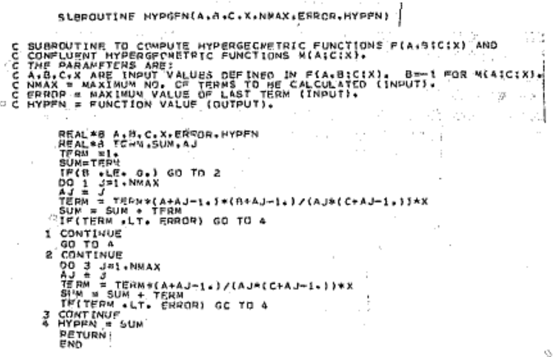
$$I \propto (A_1^*)^{-\frac{1}{2}(v_2+k-3)}(A_2^*)^{-\frac{1}{2}(N+v_1-k-2)}\int_0^\infty \sqrt{\frac{A_1^*}{A_2^*}z}^{\,(\frac{1}{2}(v_2+k-3))}z^{\frac{1}{2}(N+v_1-k-2)-1}$$

$$\times \exp(-z)dz$$

$$\propto (A_1^*)^{-\frac{1}{2}(v_2+k-3)}(A_2^*)^{-\frac{1}{2}(N+v_1-k-2)}E_z\left[\sqrt{\frac{A_1^*}{A_2^*}z}^{\,(\frac{1}{2}(v_2+k-3)}\right] \qquad (A6.4.7)$$

where z has a gamma distribution with parameter $\frac{1}{2}(N+v_1-k-2)$.

Appendix 6.5 FORTRAN Subroutine for computing the Hypergeometric and Confluent Hypergeometric Functions.



```
      SUBROUTINE HYPGEN(A,B,C,X,NMAX,ERROR,HYPFN)

C SUBROUTINE TO COMPUTE HYPERGEOMETRIC FUNCTIONS F(A,B;C;X) AND
C CONFLUENT HYPERGEOMETRIC FUNCTIONS M(A;C;X).
C THE PARAMETERS ARE:
C A,B,C,X ARE INPUT VALUES DEFINED IN F(A,B;C;X).  B=-1 FOR M(A;C;X).
C NMAX = MAXIMUM NO. OF TERMS TO BE CALCULATED (INPUT).
C ERROR = MAXIMUM VALUE OF LAST TERM (INPUT).
C HYPFN = FUNCTION VALUE (OUTPUT).

      REAL*8 A,B,C,X,ERROR,HYPFN
      REAL*8 TERM,SUM,AJ
      TERM =1.
      SUM=TERM
      IF(B .LE. 0.) GO TO 2
      DO 1 J=1,NMAX
      AJ = J
      TERM = TERM*(A+AJ-1.)*(B+AJ-1.)/(AJ*(C+AJ-1.))*X
      SUM = SUM + TERM
      IF(TERM .LT. ERROR) GO TO 4
    1 CONTINUE
      GO TO 4
    2 CONTINUE
      DO 3 J=1,NMAX
      AJ = J
      TERM = TERM*(A+AJ-1.)/(AJ*(C+AJ-1.))*X
      SUM = SUM + TERM
      IF(TERM .LT. ERROR) GO TO 4
    3 CONTINUE
    4 HYPFN = SUM
      RETURN
      END
```

with Jacobian $J = \frac{1}{2}A_2^*/z^2$, we get:

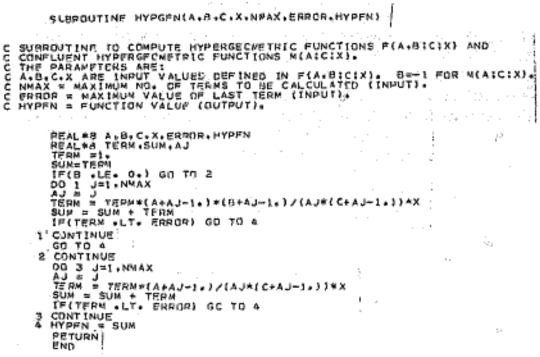
$$I = (A_1^*)^{-\frac{1}{2}(\nu_2+k-3)}(A_2^*)^{-\frac{1}{2}(N+\nu_1-k-2)}\int_0^\infty \left\lceil \frac{A_1^*}{A_2^*}\, z\right.(\tfrac{1}{2}(\nu_2+k-3))\, z^{\frac{1}{2}(N+\nu_1-k-2)-1}$$

$$\times \exp(-z)dz$$

$$= (A_1^*)^{-\frac{1}{2}(\nu_2+k-3)}(A_2^*)^{-\frac{1}{2}(N+\nu_1-k-2)}E_z\left[\left\lceil \frac{A_1^*}{A_2^*}\, z\right.(\tfrac{1}{2}(\nu_2+k-3)\right] \qquad (A6.4.7)$$

where z has a gamma distribution with parameter $\frac{1}{2}(N+\nu_1-k-2)$.

Appendix 6.5 FORTRAN Subroutine for computing the Hypergeometric and Confluent Hypergeometric Functions.



```
      SUBROUTINE HYPGFN(A,B,C,X,NMAX,ERROR,HYPFN)

C SUBROUTINE TO COMPUTE HYPERGEOMETRIC FUNCTIONS F(A,B:C:X) AND
C CONFLUENT HYPERGEOMETRIC FUNCTIONS M(A:C:X).
C THE PARAMETERS ARE:
C A,B,C,X ARE INPUT VALUES DEFINED IN F(A,B:C:X).  B=-1 FOR M(A:C:X).
C NMAX = MAXIMUM NO. OF TERMS TO BE CALCULATED (INPUT).
C ERROR = MAXIMUM VALUE OF LAST TERM (INPUT).
C HYPFN = FUNCTION VALUE (OUTPUT).

      REAL*8 A,B,C,X,ERROR,HYPFN
      REAL*8 TERM,SUM,AJ
      TERM =1.
      SUM=TERM
      IF(B .LE. 0.) GO TO 2
      DO 1 J=1,NMAX
      AJ = J
      TERM = TERM*(A+AJ-1.)*(B+AJ-1.)/(AJ*(C+AJ-1.))*X
      SUM = SUM + TERM
      IF(TERM .LT. ERROR) GO TO 4
    1 CONTINUE
      GO TO 4
    2 CONTINUE
      DO 3 J=1,NMAX
      AJ = J
      TERM = TERM*(A+AJ-1.)/(AJ*(C+AJ-1.))*X
      SUM = SUM + TERM
      IF(TERM .LT. ERROR) GO TO 4
    3 CONTINUE
    4 HYPFN = SUM
      RETURN
      END
```

Chapter 7 A Practical Application

In this chapter the theory developed in the thesis is applied to the stratigraphic problem in gold mining mentioned in Chapter 1. Given a training sample from each of fifteen strata, we will first evaluate the expected performance of classical discriminant analysis applied to this situation and then we will use the classical and Predictive Bayesian approaches to classify two observations of unknown origin into one of the strata.

After first transforming the data in the training sample to remove an unwanted dilution effect, the data is tested for multivariate normality and homoscedasticity. Using the methods described in Chapter 5, tests are performed to establish whether any of the eigenvalues $\{\lambda_i\}$ of $T\Sigma^{-1}$ are zero, and then estimates of the λ_i are obtained. These estimates are used to estimate the distribution of δ^2_{ij} and $\delta^2_j(x)$ given in Chapter 3, as well as to evaluate the expected probabilities of correct- and misclassification under classical discriminant analysis, given in Chapter 4.

Finally, using the Predictive Bayesian approach, two observations of unknown origin are each classified into one of a subset of the strata in the training sample. In this case it is possible to make direct comparisons with the results when using the Predictive Bayesian approach under the fixed effects model, as well as with those when using the classical approach. This illustrates the effect that the differences in the assumptions underlying these models have on the performance of discriminant analysis in practice.

7.1 A Problem in Stratigraphy

As mentioned in Chapter 1, this study arose out of the problem of fitting a particular band of rock encountered in a gold mine into the sedimentary succession of the area. As the trace element geochemistry of each rock band can reasonably be described by a random effects model, it seems an appropriate area for application of the theory developed in this thesis.

The concentration of trace elements in rock samples were measured by means of Instrumental Neutron Activation Analysis, a technique that allows accurate chemical analyses to be made down t/ very low concentrations. A pilot study was undertaken to assess the feasibility in general terms of using geochemical data to relocate the pay band. Five samples were taken from each of 15 bands, and 12 trace elements were measured on each sample. For the reasons given in Hawkins and Rasmussen (1973), a log transformation was applied to the data.

A complicating factor in the analysis is the presence of unknown but varying amounts of silica in the samples which tends to give a proportional decrease in the concentrations of the trace elements. This gives rise to an additive "dilution effect" or "growth affect" corresponding to each sample when using the transformed data.

The problem of statistical inference, with particular reference to canonical variate analysis, on multivariate data in the presence of additive growth effects has been studied by Gower (1976), and an interesting application to a problem in Palaeontology has been given by Reyment and Banfield (1976). Gower (1976) considers the case where a p-dimensional observation x is contaminated by m (<p) additive growth effects, each of which may be represented by a (p×1) growth vector whose elements are the relative responses of the corresponding elements of x to the unobservable growth effect. Gower (1976) uses the fact that if K is the (p×m) matrix whose columns are these growth vectors, then the symmetric idempotent matrix

$$Q = I - K(K'K)^{-1}K' \qquad (7.1.1)$$

projects x on to the space orthogonal to K so that the projected value is free from these growth effects. Therefore, if

$$y = Qx \tag{7.1.2}$$

then y is free from growth effects. Furthermore, if the sample space of x has rank r ($\leq p$) then y occupies a sample space of rank $r-r(K)$.

In the context of the present example, it is clear that the growth effect in the rock samples due to dilution by unknown quantities of silica is the same for all of the log trace element concentrations, so that it can be represented by the single p-dimensional vector

$$K = (1,1\ldots,)'. \tag{7.1.3}$$

Therefore, in the present situation

$$Q = I - \frac{1}{p} E \tag{7.1.4}$$

where E is the p×p matrix whose elements are all unity, so that the transformed variable becomes

$$y = Qx = (I - \frac{1}{p} E)x$$

i.e.

$$y_i = x_i - \frac{1}{p} \sum_{j=1}^{p} x_j$$

$$= x_i - x_{.} \quad , i = 1,\ldots,p \tag{7.1.5}$$

where x_i and y_i are the i^{th} elements of x and y, respectively. So, to remove the dilution effect from each observation we make the (intuitively reasonable) transformation of subtracting the average of all p log trace element concentration values in the sample from each these p values in turn. This will clearly reduce the dimensionality of the sample space to p-i (assuming that the original data are of full rank)and the easiest way to handle this is to drop one or more variables from the analysis.

Because of the finding in Chapter 5 that the number of populations in the training sample should be as large as possible, relative to the dimension p of the data vectors, for reliable estimation of the eigenvalues $\{\lambda_i\}$ of $T\Sigma^{-1}$, it was decided to base the discriminant analysis on a subset of four of the twelve trace elements. The following trace elements were chosen, primarily because of the fact that, out of the twelve, their marginal distributions most closely fitted the normal:

1. Cobalt (Co)
2. Iron (Fe)
3. Hafnium (Hf)
4. Gold (Au)

The data on these four elements (after log transformation and removal of dilution effect) are given in Table 7.1.1 below, and in Tables 7.1.2, 7.1.3 and 7.1.4, respectively, their mean vectors, within groups and between groups covariance matrices are given.

Table 7.1.1

The Trace Element Data (after log transformation and removal of dilution effect.)

Population	Co	Fe	Hf	Au
1	0.3858	0.0534	-0.0981	-1.1539
	0.5065	0.3371	-0.3136	-0.3335
	0.4081	0.1967	-0.7308	-0.3231
	0.3210	0.1054	-0.4605	-0.5447
	-0.2393	-0.1483	-0.2902	-0.9580
2.	0.4255	0.3744	-0.0853	0.0657
	0.4008	0.3604	-0.1572	0.0465
	0.4735	0.2852	-0.5006	0.2523
	-0.3862	0.2177	-0.4931	1.0496
	0.0569	0.2095	-0.1794	-0.0990

Table 7.1.1 continued

Population	Co	Fe	Hf	Au
3	-0.1660	0.1619	0.0849	-0.2998
	0.3160	0.3020	-0.0110	0.0073
	0.1448	0.1550	0.0443	-0.0790
	0.1572	0.1438	0.0974	-0.3205
	-0.1533	0.3362	0.0462	-0.6277
4	0.6285	0.5011	-0.1421	0.2181
	0.3091	0.3204	-0.4308	0.1654
	0.2866	0.2446	-0.5342	0.4794
	0.3784	0.1976	-0.5416	0.3688
	0.2984	0.2540	-0.2706	0.3448
5	0.5217	-0.0967	-1.0894	0.2355
	0.5099	-0.0581	-1.0972	0.0230
	0.5490	0.0535	-1.2592	0.1496
	0.2981	-0.0871	0.0159	-0.1717
	0.3222	0.1997	0.0222	-0.0271
6	0.3330	0.0663	-1.1813	0.5933
	0.6624	0.4103	-0.4863	-0.1420
	0.5272	0.0614	-0.4651	-0.3413
	0.1279	-0.0432	-0.1307	-0.9191
	0.3033	0.1018	-1.2192	-0.7925
7	0.4148	0.5829	-0.3087	-0.0125
	0.8251	0.8348	-0.5070	0.8449
	0.4799	0.4441	-0.6221	0.5807
	0.2183	0.3549	-0.2613	0.4129
	0.5873	0.7194	-0.4033	0.8994
8	-0.2589	-0.0187	0.1143	-0.3241
	-0.2214	-0.0381	0.0665	-0.2860
	0.0087	-0.0215	0.0426	-0.3333
	0.0340	-0.0688	0.1452	-0.6939
	-0.1673	-0.0867	0.1975	-0.9541
9	-0.0765	-0.0008	-0.7813	-0.4238
	-0.0739	-0.0392	-0.0604	-0.6489
	0.2947	-0.0156	-0.0310	-0.2426
	0.4301	0.3128	-0.0601	-0.2107
	-0.1776	-0.0629	-0.0250	-0.4863
10	0.5880	0.5608	-1.2298	0.4444
	0.5295	0.4700	-0.3861	0.8966
	0.4256	0.4546	-0.4285	0.3291
	-0.1759	0.1368	-0.5196	-0.0553
	0.2849	0.4989	-0.3323	0.0128

Table 7.1.1 continued

Population	Co	Fe	Hf	Au
11	0.4872	0.6116	-0.5951	0.2322
	-0.0089	0.5342	-0.4469	1.2533
	0.5219	0.6603	-0.4632	0.3363
	-0.0709	0.2994	-0.7964	0.8250
	0.0478	0.2138	-0.1079	-0.1087
12	0.1739	0.0751	0.1398	-0.3471
	0.0322	-0.1939	0.1968	-0.4244
	-0.5402	-0.5064	0.1750	-0.3985
	-0.4637	-0.4060	0.1644	-0.9229
	0.4625	0.2718	0.0311	0.0873
13	-0.3224	0.0470	0.0641	-1.0616
	-0.5506	-0.1526	0.0243	-1.0777
	-0.5330	-0.3666	0.0121	-0.1080
	-0.3700	-0.2176	0.1564	-1.0220
	-0.3114	-0.3491	0.0350	-0.4336
14	-0.1766	-0.2141	0.1451	-1.0812
	-0.4704	-0.3250	-0.1079	0.8878
	-0.4110	-0.2990	0.0718	0.6778
	-0.5465	-0.2368	0.0851	-0.9229
	-0.3710	-0.3077	0.0623	-0.3984
15	0.2866	-0.0016	0.3986	-0.2626
	0.3875	0.1105	0.2587	-0.3552
	0.3904	0.1987	0.3219	-1.0482
	0.3823	0.1131	0.3753	-1.0876
	0.2989	-0.1169	0.5886	-0.5389

Table 7.1.2

Mean Vectors

Population	Co	Fe	Hf	Au
1	0.2764	0.1089	-0.3787	-0.6625
2	0.1941	0.2895	-0.2831	0.2630
3	0.0598	0.2138	0.0523	-0.2639
4	0.3802	0.3036	-0.3839	0.3153
5	0.4402	0.0023	-0.6815	0.0419
6	0.3908	0.1193	-0.6965	-0.3203
7	0.5051	0.5872	-0.4205	0.5451
8	-0.1210	-0.0468	0.1132	-0.5183
9	0.0754	0.0389	-0.1916	-0.4026
10	0.3304	0.4242	-0.5792	0.3255
11	0.1954	0.4639	-0.4819	0.5076
12	-0.0671	-0.1519	0.1414	-0.4011
13	-0.4175	-0.2078	0.0584	-0.7406
14	-0.3961	-0.2766	0.0513	-0.1674
15	0.3491	0.0608	0.3886	-0.6585
Overall	0.1464	0.1288	-0.2194	-0.1424

Table 7.1.3

Within Groups Covariance Matrix (Degrees of Freedom 60)

	Co	Fe	Hf	Au
Co	0.0584	0.0281	-0.0061	0.0122
Fe	0.0281	0.0257	-0.0016	0.0086
Hf	-0.0061	-0.0016	0.0759	-0.0384
Au	0.0122	0.0086	-0.0384	0.1866

Table 7.1.4

Between Groups Lovariance Matrix (Degrees of Freedom 14)

	Co	Fe	Hf	Au
Co	0.4167	0.2654	-0.2990	0.3089
Fe	0.2654	0.3177	-0.2296	0.4297
Hf	-0.2990	-0.2296	0.5614	-0.4433
Au	0.3089	0.4297	-0.4433	0.9784

The data was tested for multivariate normality and homoscedasticity using the test of Hawkins (1978) based on the $N = \sum_{i=1}^{k} n_i$ sample-based Mahalanobis distances of each observation from its group mean:

$$d_i^2(x_{ij}) = (x_{ij}-x_{i.})'S^{-1}(x_{ij}-x_{i.}) \quad j = 1,\dots,n_i;\ i=1,\dots,k \qquad (7.1.6)$$

where S is the pooled covariance matrix computed from all k groups. Hawkins (1978) shows that under the null hypothesis the statistic

$$F_{ij} = \frac{(N-k-p)n_i d_i^2(x_{ij})}{p\{(n_i-1)(N-k)-n_i d_i^2(x_{ij})\}}$$

follows an F-distribution with p and N-k-p degrees of freedom, so that if

$$A_{ij} = Pr[F > F_{ij}]$$

denotes the tail area of F_{ij} under this distribution then A_{ij} is distributed exactly as a uniform variate over the range (0,1). Departures from either normality or homoscedasticity will cause departures of the A_{ij}

from the uniform distribution, and Hawkins therefore uses the Anderson-Darling test-statistic W_i computed from the n_i order statistics of the A_{ij} in group i to test for either of these types of departure in the i^{th} population, for i=1 to k. Furthermore, splitting the W_i into components allows for heteroscedasticity and non-normality to be tested separately. Finally, Hawkins uses a simulation experiment to show that, although asymptotic results are used at a few points in his theory, his test may nevertheless be used for sample sizes n_i as small as 5, as long as N is sufficiently large.

Applying the abovementioned test to the data in this example reveals moderate departures from homoscedasticity in populations 4,5,6 and 8 (5 and 6 having larger, and the other two smaller covariance matrices than the average) and also that population 4 has a slightly lighter-tailed distribution than the normal. However, because these departures are fairly minor, and so as not to reduce the number of populations in the training sample, it was decided not to remove these populations from the example.

As mentioned in Chapter 5, the first step in applying this data to the random effects model in discriminant analysis is to test the hypothesis $H_0 : T = 0$, for if it is accepted then there is no point in continuing with the analysis. Using the subroutine CANOK described earlier, the eigenvalues $\{g_i\}$ of $A_1A_2^{-1}$ were computed. These are given in Table 7.1.5, together with the two test statistics T_1 and T_2 defined in (5.2.3) and (5.2.4), respectively.

Table 7.1.5

The eigenvalues of $A_1A_2^{-1}$

g_1	g_2	g_3	g_4
74.2183	35.5174	13.0757	7.1559

$$T_1 = \sum_{i=1}^{4} \log(1+g_i) = 12.6614$$

$$T_2 = \sum_{i=1}^{4} g_i \qquad = 129.9673$$

From (5.2.5) we have that under the null hypothesis m_1T_1 has approximately a $\chi^2_{p\nu_1}$ distribution where $m_1 = \nu_2 + \frac{1}{2}(\nu_1-p-1)$ and ν_1 and ν_2 are the between groups and within groups degrees of freedom, respectively. Since $m_1T_1 = 816.7$ and $p\nu_1 = 56$, H_0 is rejected resoundingly.

In order to test whether any of the $\{\lambda_i\}$ = eigs $T\Sigma^{-1}$ are zero, we first consider the sub-hypothesis: $H_{01} : \lambda_4 = 0$. Our two test-statistics for testing H_{01} are:

$$T_{11} = \text{Log}(1+g_4) = 2.0987$$

and $$T_{21} = g_4 \qquad = 7.1559$$

(See (5.2.11) and (5.2.12)).

Using T_{11}, we have from (5.2.13) that under H_{01}, $m_{11}T_{11}$ has approximately a χ^2_f distribution

where $$f = (4-3)(14-3) = 11$$

and $$m_{11} = 60 + \frac{1}{2}(14-5) + \sum_{i=1}^{3} \lambda_i^{-1}$$

Using the estimators of the λ_i given below in the expression for m_{11} yields value $m_{11} = 65.01$, whence $m_{11}T_{11} = 136.4$ which again is highly significant. So we conclude that all the λ_i are greater than zero.

Our next step is to estimate the λ_i. Using the techniques described in Chapter 5, the five estimators $\hat{\underset{\sim}{\gamma}}^{(1)}$ to $\hat{\underset{\sim}{\gamma}}^{(5)}$ of the eigenvalues $\{\gamma_i\}$ of $\Sigma_1\Sigma^{-1} = (\Sigma+nT)\Sigma^{-1}$ were computed. Unfortunately the "unrestricted" and "restricted" maximum marginal likelihood estimators $\hat{\underset{\sim}{\gamma}}^{(4)}$ and $\hat{\underset{\sim}{\gamma}}^{(5)}$ both failed to give meaningful results, so the approximate maximum marginal likelihood estimator $\hat{\underset{\sim}{\gamma}}^{(2)}$ ($\hat{\underset{\sim}{\gamma}}^{(3)}$ was identical to $\hat{\underset{\sim}{\gamma}}^{(2)}$) was used to compute $\hat{\underset{\sim}{\lambda}}$ from the relationship

$$\hat{\lambda}_i = \frac{1}{n}(\hat{\gamma}_i - 1) \qquad i = 1,\ldots,p \ .$$

These estimates are given in Table 7.1.6. The estimation procedure was then repeated with variable 3 (Hafnium) dropped from the sample, reducing the number of variables to 3. In this case all five estimators gave meaningful results, so that $\hat{\underset{\sim}{\lambda}}$ could be obtained from $\hat{\underset{\sim}{\gamma}}^{(5)}$. These estimates are also given in Table 7.1.6.

Table 7.1.6

Estimates of $\underset{\sim}{\gamma}$ and $\underset{\sim}{\lambda}$

p = 4 variables	(Co, Fe, Hf, Au)			
	1	2	3	4
$\hat{\underset{\sim}{\gamma}}^{(1)}$	17.3176	8.2874	3.0510	1.6697
$\hat{\underset{\sim}{\gamma}}^{(2)}$(and $\hat{\underset{\sim}{\gamma}}^{(3)}$)	14.7864	8.6950	3.3602	2.5431
$\hat{\underset{\sim}{\gamma}}^{(4)}$ and $\hat{\underset{\sim}{\gamma}}^{(5)}$	failed to give meaningful results			
$\hat{\underset{\sim}{\lambda}}$ (from $\hat{\underset{\sim}{\gamma}}^{(2)}$)	2.7573	1.6390	0.4720	0.3086

<u>p = 3 variables</u> (Co, Fe, Au)

$\hat{\gamma}^{(1)}$	16.2545	6.2701	1.7935
$\hat{\gamma}^{(2)}$ (and $\hat{\gamma}^{(3)}$)	14.7184	6.7841	2.2072
$\hat{\gamma}^{(4)}$	14.2456	6.9484	2.2390
$\hat{\gamma}^{(5)}$	13.9191	6.6518	2.1261
$\hat{\lambda}$ (from $\hat{\lambda}^{(5)}$)	2.5838	1.1304	0.2252

Using the $\hat{\lambda}_i$ given in Table 7.1.6, the estimated distribution of the Mahabanobis distance

$$\delta_{ij}^2 = (\mu_i-\mu_j)'\Sigma^{-1}(\mu_i-\mu_j)$$

between two randomly selected populations, derived in Chapter 3, and that of the Mahabanobis distance

$$\delta_i^2(x) = (x-\mu_i)'\Sigma^{-1}(x-\mu_i)$$

of a random observation $x \in \pi_j$ from π_i, $i \neq j$, were computed using the subroutines given in Chapter 3. In Table 7.1.7 values of the distribution functions of δ_{ij}^2 and $\delta^2{}_i(x)$ are given at selected points, separately for the four- and three variable cases. In addition, distribution function *values for $\delta_i^2(x)$, when $x \in \pi_i$, are given at the same points for comparison.*

Table 7.1.7

Estimated Distribution Functions of δ^2_{ij}, $\delta^2_i(x)|x \in \pi_j$ and $\delta^2_i(x)|x \in \pi_i$

p = 4 variables (Co, Fe, Hf, Au)

Value of the random variable	δ^2_{ij}	Distribution Function Values $\delta^2_i(x)\|x \in \pi_j$	$\delta^2_i(x)\|x \in \pi_i$
1	.031	.012	.045
2	.098	.043	.144
3	.177	.085	.264
5	.335	.189	.496
7	.470	.298	.677
10	.626	.450	.845
15	.789	.646	.960
20	.880	.776	.990
25	.931	.858	.998
30	.960	.910	1.000
40	.986	.963	1.000
50	.995	.985	1.000

p = 3 variables (Co, Fe, Au)

Value of the random variable	δ^2_{ij}	Distribution Function Values $\delta^2_i(x)\|x \in \pi_j$	$\delta^2_i(x)\|x \in \pi_i$
1	.089	.044	.119
2	.202	.111	.279
3	.305	.184	.428
5	.474	.326	.657
7	.600	.449	.802
10	.731	.596	.917
15	.858	.760	.981
20	.923	.856	.996
25	.957	.913	.999
30	.976	.947	1.000
40	.992	.979	1.000
60	.997	.992	1.000

The expected probabilities of misclassification indicate how well classical discriminant analysis is likely to perform when applied to the problem of fitting a particular rock band into the sedimentary succession of the area, on the basis the concentrations of the four (or three) trace elements in a rock sample from that band. These were computed from the formulae derived in Chapter 4, using the subroutine PROBS for the "optimum" probabilities, where the parameters in the linear discriminant function are assumed to be known and classification rule (2.1.3) is used, and subroutine PROBS1 for the case where the sample-based classification rule (2.1.19) is used. Table 7.1.8 gives the two-population probabilities of misclassification as well as the lower and approximate upper bounds for the probabilities of correct classification for the 5- population case, for both situations where the population-based and sample-based classification rules are used.

In the situation where it is possible to make more than one observation on the unknown population (as in the case in our stratigraphic problem) it is well known that arbitrarily good classification may be achieved by increasing the number of independent observations from the unknown population and basing the classification on their mean. It is a trivial matter to show that the situation where the mean of m observations is used for classifying the unknown population is exactly equivalent, under the random effects model, to that when the eigenvalues $\{\lambda_i\}$ are all multiplied by m and a single observation is used for classification. As an illustration of this,the expected probabilities corresponding to the situation where the classification is based on $m = 2$ observations from the unknown population are also given in Table 7.1.8.

Table 7.1.8

Expected Probabilities of Correct- and Misclassification

p = 4 variables (Co, Fe, Hf, Au)

Known Parameters	Probability of misclassification with two populations	Probability of correct classification with k=5 populations	
		Lower Bound	Approx. upper Bound
One observation from unknown pop.	.1069	.5724	.8202
Two observations from unknown pop.	.0555	.7780	.8930
Unknown Parameters (degrees of freedom = 60)			
One observation from unknown pop.	.1173	.5307	.8039
Two observations from unknown pop.	.0616	.7534	.8808

p = 3 variables (Co, Fe, Au)

Known Parameters	Probability of Misclassification with two populations	Lower bound to Probability of correct classification with k=5 populations
One observation from unknown pop.	.1429	.4282
Two observations from unknown pop.	.0860	.6561
Unknown Parameters (Degrees of freedom = 60)		
One observation from unknown pop.	.1518	.3929
Two observations from unknown pop.	.0915	.6341

Note that, since p is odd, the upper bound to the probability of correct classification cannot be computed.

We now turn to the Predictive Bayesian approach. Because of our inability, at present, the compute the predictive densities under the random effects model in the multivariate case (see sub-section 6.2.1) we will consider classifying two observations of unknown origin using only the trace element Cobalt (Co). The concentration of Cobalt in each of the two unknowns, after log transformation and removal of dilution effect, are given below:

Unknown 1 :	0.2864
Unknown 2 :	-0.4075

The predictive densities under the random effects model, given by (6.1.17), were computed using the subroutine HYPGFN and are given in Table 7.1.9 for each of the fifteen populations and both unknowns. For comparison, the corresponding predictive densities under the fixed effects model, given by (2.2.6), as well as the sample-based Mahalanobis distances between each of the two unknowns and each of the fifteen populations, are also given in Table 7.1.9.

Table 7.1.9

Predictive densities of the two unknowns under the random effects and fixed effects models, as well as the corresponding Mahalanobis distances, using one variable (Co) only.

Unknown 1

	Predictive Densities		
Population	Random Effects Model	Fixed Effects Model	Mahalanobis Distances
1	.0844	.0955	0.0017
2	.0947	.0899	0.1461
3	.0598	.0660	0.8800
4	.0810	.0897	0.1506
5	.0876	.0806	0.4049
6	.0793	.0884	0.1863
7	.0681	.0677	0.8187
8	.0371	.0294	2.8426
9	.0883	.0693	0.7630
10	.0822	.0943	0.0331
11	.0978	.0900	0.1419
12	.0388	.0391	2.1401
13	.0068	.0032	8.4861
14	.0093	.0039	7.9550
15	.0849	.0929	0.0673

Unknown 2

1	.0161	.0109	8.0108
2	.0232	.0219	6.1979
3	.0649	.0580	3.7388
4	.0068	.0041	10.6259
5	.0056	.0022	12.3063
6	.0056	.0037	10.9124
7	.0026	.0011	14.2623
8	.1294	.1506	1.4059
9	.0522	.0524	3.9928
10	.0090	.0066	9.3248
11	.0232	.0217	6.2260
12	.1190	.1185	1.9847
13	.2623	.2715	0.0017
14	.2709	.2714	0.0026
15	.0094	.0055	9.8043

The posterior probabilities of each of the populations are computed from the predictive densities in Table 7.1.9 by multiplying them by their respective prior probabilities. For example, suppose that unknown 1 is equally likely to have come from one of the first five populations and from none of the others. Using the classical approach one would unhesitatingly classify it into population 1. On the other hand, although population 1 has marginally the highest posterior probability under the fixed effect model, population 2 has marginally the highest probability under the random effects model. In practice, using the Predictive Bayesian approach under either of the fixed effects or random effects models, one would consider Unknown 1 to be unclassifiable. The divergence between the classical and predictive Bayesian approaches observed here is in line with the findings of Aitchison, Habbema and Kay (1977) whose general conclusion is that the classical (or "estimative") approach tends to give too optimistic a picture of the reliability of sample-based discrimination procedures.

The picture is far clearer with Unknown 2. Assuming that it is equally likely to have come from one of the last five populations,all three classification rules come out stringly in favour of either of populations 13 or 14, the predictive approach under the random effects model giving slight preference to population 14 whereas the other two marginally favour the former.

The reason for the improved reliability of classification in the latter case is quite evident under the random effects model . Since observation 2 is much further than observation 1 from the estimated mean ξ of the individual population means μ_i, one would expect better classification with it as populations would tend to be much less clustered in its vicinity than they would be nearer to ξ.

Chapter 8 Review and Conclusions

In this, the final chapter, the theory developed in this thesis is reviewed, and the areas still requiring further work,as well as the various possible avenues for future research are pointed out. Finally, some conclusions are drawn regarding the applicability and usefulness of this theory to the solution of practical problems in discriminant analysis.

Before starting the review, some comments on the practical situation where this theory might be applicable, are in order. It is envisaged that the investigator will, in general, have two (possibly overlapping) training samples at his disposal. The first, more properly called an "estimation sample" will consist of random samples from each of a number of populations, each of them in turn being a random observation from a "super-population" under the random effects model. This sample will be used to estimate the parameters $\{\lambda_i\}$ in the manner described in Chapter 5, which will in turn be used to estimate the distributions of any of the four distance variables discussed in Chapter 3, as well as the expected probabilities of correct - and misclassification under the classical approach, derived in Chapter 4. The second training sample, which may only become available at a later date, will consist of random samples from each of k_1 populations (with possible overlap between it and the estimation sample - together they make k independent samples from the "super-population") and one or more observations x known to have come from one of these k_1 populations. The objective of the investigator is to assign x to one of these k_1 populations in the second sample.

Clearly, the information from the second training sample can be combined with that of the first to produce improved estimates of the $\{\lambda_i\}$ and of the distributions and expected probabilities of correct - and misclassification mentioned above. Under the Predictive Bayesian approach too, no distinction need be made between these two samples, except when it comes to the choice of populations into which the unknown may be classified. The device used in Chapter 7 of assigning zero prior probabilities to all those populations not involved in any particular classification problem, is a convenient way of making the abovementioned distinction without formally having to distinguish between the two samples.

8.1 Review

Starting the review at Chapter 3, it is clear that while only the distribution of δ^2_{ij} is of direct relevance to the evaluation of correct - and misclassification probabilities under the random effects model, the distributions of the other three quantities $\delta^2_i(x)$, d^2_{ij} and $d^2_i(x)$ are of interest in that they provide further insight into the likely performance of classical discriminant analysis under this model. As has been seen, the evaluation of the density and distribution functions of all four of these distance variables is a relatively straightforward matter on a computer, so that approximating them by means of, say Pearson curves, is not considered to be worth while.

Coming now to the evaluation of the probabilities of correct - and misclassification considered in Chapter 4, the two - population case where the parameters are known has clearly been solved satisfactorily and the probability of misclassification under the random effects model is readily evaluated using a computer. The k-population case is slightly less satisfactory in that only lower and (conditional and approximate) upper bounds to the probability of correct classification have been found, although it is evident from the examples considered that these two bounds can be fairly close. An exact expression for this probability will however only be found once the corresponding exact expression (4.1.24) for the conditional probability of correct classification, given δ_{ij}^2 , is available in a more tr:table form. Two further problems requiring solution are firstly, the evaluation of the upper bound on the probability of correct classification for the case where the number r of nonzero λ_i is odd, and secondly, the derivation of convenient computational formulae, when the λ_i are not all equal, for the coefficients a_j, defined in (4.1.39), appearing in formula (4.1.40) for the upper bound when r is even.

In the situation where the sample-based classification rule is used, all the results derived are based on Okamoto's(1963) asymptotic expa..ion (2.1.26) to terms of order n^{-1}. Therefore, more accurate results could be obtained, at the cost of considerable increase in complexity by including all the terms of order n^{-2} in Okamoto's expansion. [illegible] the k-population case exactly the same remarks hold as in the situation where the parameters are known.

An important piece of research that is still outstanding in Chapter 5 is to obtain unrestricted and restricted maximum marginal likelihood estimators of $\{\gamma_i\}$ = Eigs $\{\Sigma_1\Sigma^{-1}\}$ based on Khatri and Srivastava's (1978) asymptotic expansion (5.3.8) for the joint density of $\{g_i\}$ = Eigs$(A_1A_2^{-1})$ rather than on Chang's (1970) less accurate expression (5.3.5). Simulation experiments on these two estimators, corresponding to those done in Chapter 5, will give an indication of how much an improvement they are over those proposed in this chapter. A further area for research arising as a side issue out of the results of Chapter 5, is the derivation of a scaled F-approximation to the distribution of Hotelling's T_0^2 for the case where the numerator and denominator matrices have independent Wishart distributions but with different parameter matrices Σ_1 and Σ. See the comments at the end of Sub-section 5.4.2.

The treatment of the Predictive Bayesian Approach under the random effects model is fairly complete, at least for the case where the parameters Σ, ξ and T have diffuse prior distributions. A great deficiency in this approach is, however, our inability to compute the predictive densities in the multivariate case. Possible approaches towards rectifying this are, firstly, to try and evaluate the hypergeometric functions of matrix argument, appearing in the predictive densitites by using the programs of van der Westhuizen and Nagel (1979) on a very much *faster computer than the University of South Africa's* Burroughs B6800 computer. Secondly, the efficiency of these programs could possibly be improved, although a reduction in computing time by at least a few orders of magnitude would be required to ensure that a sufficient number of terms can be computed for the hypergeometric functions to converge. Two promising directions for research do, however come out of the last section in Chapter 6. Firstly there is the Empirical *Bayes approach to discriminant analysis* under the random effects model; an interesting study would be to investigate the properties of the proposed classification rule (6.4.8). Secondly, an investigation of the semi-Bayes approach under the random effects model, using the posterior density (6.4.17) as starting point would also make an interesting, if complicated, study.

8.2 Conclusions

In this thesis, discriminant analysis under the random effects model has been treated from two viewpoints. With the classical approach, the properties of the classification rules have been investigated under this model, whereas with the Predictive Bayesian approach new expressions for the predictive densities appropriate for this model have been derived.

Considering first the classical approach, the assumption of the random effects model has allowed expressions for the expected probabilities of correct - and misclassification to be derived that depend only on the eigenvalues $\{\lambda_i\}$ of $T\Sigma^{-1}$. These may be estimated with arbitrary precision as long as training samples can be drawn from a sufficient number of populations. On the other hand, under the fixed effects model, whether using Okomoto's (1963) expression (2.1.26) or Anderson's (1973a, b) expression (2.1.27) for the expected probability of misclassification with the sample-based classification rule, the value of the Mahalanobis distance δ_{12}^2 between the two populations is required. This has to be estimated using the means of the training samples from only the two populations concerned, although Σ may be estimated using training samples from other populations as well. (See Lachenbruch and Mickey (1968) for an estimator of δ_{12}^2 that partially corrects for the bias in d_{12}^2.)

Therefore it would appear that as long as there are a sufficient number of populations in the training sample (relative to the number of variables - see Section 5.5) more reliable estimates of the probabilities of correct - and misclassification will be obtained under the random effects model than under the fixed-effects model. On the other hand, the requirement that there should be a large number of populations (relative to the number of variables) in the training sample for reliable estimation under the random effects model, can also be considered to be a drawback to this model, particularly in situations where samples from many populations are hard to come by.

A topic that has not been discussed in this thesis is variable selection. Since under the random effects model the probabilities of correct - and misclassification are functions only of the eigenvalues $\{\lambda_i\}$ of $T\Sigma^{-1}$, we would want a procedure that selects variables on the basis of the values of the λ_i. Now, it is clear from (5.2.3) that the likelihood ratio statistic T_1 for testing $H_0 : T = 0$, is a monotonic increasing function of the eigenvalues g_i of $A_1A_2^{-1}$ and hence of the $\{\ell_i\} = \{\frac{\nu_2}{\nu_1} g_i\}$. Since the ℓ_i are maximum likelihood estimators of the $\gamma_i = 1 + n\lambda_i$, we would expect that variable selection based on T_1 would be appropriate for our situation. Hawkins (1976) proposes a stepwise procedure based on T_1 for selecting variables in Multivariate analysis of Variance. Although he applies the procedure to a problem in multiple discriminant analysis using the fixed effects model, it is, from the above remarks, also applicable to the random effects model.

Coming now to the Predictive Bayesian approach, an immediate conclusion that may be drawn from the examples considered is that the predictive densities (and hence posterior probabilities) are generally more conservative under the random effects model than they are under the fixed effects model. Therefore, if the predictive densities for the fixed effects model, given by (2.2.6) and (2.2.7), are computed in a situation where the random effects model holds, then they will tend to give posterior probabilities that are too optimistic. On the other hand, if the random effects model is applied to data where the fixed effects model is more appropriate, it will give results that are too conservative.

Finally, a comment on the applicability of the random effects model to discriminant analysis with unequal covariance matrices in different populations, is in order. Although it is possible, from a purely mathematical viewpoint, to perform similar analyses to those given in this thesis for the heteroscedastic situation, it is our opinion that the results would have little application in practice. The reason for this is that if different populations have different covariance matrices then it is highly unlikely, in any practical situation, that their mean vectors would come from the same distribution. A more likely situation would be that for any particular population the covariance matrix of its mean vector μ would be some function of the covariance matrix within that population.

References

1. Abramowitz, M. and Stegun, I.A. (1965). Handbook of mathematical functions. Dover, New York.

2. Aitchison, J. Habbema, J.D.F. and Kay, J.W. (1977). A critical comparison of two methods of statistical discriminant analysis. Applied Statistics, 26, 15-25.

3. Anderson, G.A. (1965). An asymptotic expansion for the distribution of the latent roots of the estimated covariance matrix. Ann. Math. Statist., 36, 1153-1173.

4. Anderson, T.W. (1951). Classification by multivariate analysis. Psychometrica,16, 31-50.

5. Anderson, T.W. (1958). An introduction to multivariate statistical analysis. Wiley, New York.

6. Anderson, T.W. (1973a). An asymptotic expansion of the distribution of the studentized classification statistic W. Ann. Statist., 1, 964-972.

7. Anderson, T.W. (1973b). Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions. Discriminant analysis and applications, T. Cacoullos, ed., Academic Press, New York, 17-35.

8. Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967). The discarding of variables in multivariate analysis. Biometrika, 54, 357-366.

8a. Bellman, R. (1970). Introduction to matrix analysis. McGraw-Hill, New York.

9. Bissell, A.F. and Ferguson, R.A. (1975). The jackknife - toy, tool or two-edged weapon? Statistician, 24, 79-100.

10. Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria. Biometrika, 36, 317-346.

11. Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, 1. Effect of inequality of variance in the one-way classification. Ann. Math. Statist., 25, 290-302.

12. Box, G.E.P. and Tiao, G.C. (1973). Bayesian inference in statistical analysis. Addison-Wesley, London.

13. Brand, L. (1955). Advanced calculus. Wiley, New York.

14. Cacoullos, T. (1973). Distance, discrimination and error. Discriminant analysis and applications, T. Cacoullos, ed. Academic Press, New York, 61-75.

15. Chang, T.C. (1970). On an asymptotic representation of the distribution of the characteristic roots of $S_1 S_2^{-1}$. Ann. Math. Statist., 41, 440-445.

16. Constantine, A.G. (1963). Some non-central distribution problems in multivariate analysis. Ann. Math. Statist., 34, 1270-1284.

17. Cox, D.R. and Hinkley, D.V. (1974). Theoretical statistics. Chapman and Hall, London.

18. Das Gupta, S. (1972). Probability inequalities and error in classification. Univ. of Minnesota, School of Statistics, Tech Rep. No. 190.

19. Davis, A.W. (1977). A differential equation approach to linear combinations of independent chi-squares. J. Amer. Statist. Assoc., 72, 212-214.

20. de Villiers, H. (1973). Human skeletal remains from Border Cave, Ingwavuma District, Kwazulu, South Africa. Annals of the Transvaal Musoum, 28, 229-256.

21. de Villiers, H. (1976). A second adult human mandible from Border Cave, Ingwavuma District, Kwazulu, South Africa. South African J. Sci., 72, 212-215.

22. de Waal, D.J. (1976). A review of tests of various hypotheses in multivariate statistical analysis. South African Statist. J., 10, 153-175.

23. Dickey, J.M. (1967). Matricvariate generalizations of the multivavariate t distributions and the inverted multivariate t

distribution. Ann. Math. Statist., 38, 511-519.

24. Downton, F. (1973). The estimation of $Pr(Y < X)$ in the normal case. Technometrics, 15, 551-558.

25. Dunn, O.J. (1971). Some expected values for probabilities of correct classification in discriminant analysis. Technometrics, 13, 345-353.

26. Dunn, O.J. and Varady, P.D. (1966). Probabilities of correct classification in discriminant analysis. Biometrics, 22, 908-924.

27. Dunsmore. I.R. (1966). A Bayesian approach to classification. J. Roy Statist. Soc. B, 28, 568-577.

28. Fatti, L.P. and Hawkins, D.M. (1976). Fortran IV program for canonical variate and principal component analysis. Computers & Geosciences; 1, 335-338

29. Fisher, R.A. (1936). The use of multiple measurement in taxonomic pr[illegible] Ann. Eugen., 7, 179-188.

30. Fujikos[illegible]). Asymptotic expansions for the distributions of [illegible]tivariate tests. Multivariate analysis IV, P.R. Krishnaiah, ed., North-Holland, Amsterdam, 55-71.

31. Geisser, S. (1964). Posterior odds for multivariate normal classifications. J. Roy. Statist. Soc. B, 26, 69-76.

32. Geisser, S. (1966). Predictive discrimination.. Multivariate analysis, P.R. Krishnaiah, ed., Academic Press, New York, 149-163.

33. Geisser, S. (1967). Estimation associated with linear discriminants. Ann. Math. Statist., 38, 807-817.

34. Geisser, S. (1973). Multiple birth discrimination. Multivariate statistical inference, D.G. Kabe and R.P. Gupta, eds., North-Holland, 49-55.

35. Geisser, S. and Cornfield, J. (1963). Posterior distributions for

multivariate normal parameters. J. Roy. Statist. Soc. B, 25, 368-376.

36. Gibbons, J.D. (1971). Nonparametric statistical inference. McGraw-Hill, London.

37. Giri, N.C. (1977). Multivariate statistical inference. Academic Press, New York.

38. Girschick, M.A. (1939). On the sampling theory of roots of determinantal equations. Ann. Math. Statist., 10, 203-224.

39. Glick, N. (1972). Sample based classification procedures derived from density estimators. J. Amer. Statist. Assoc., 67, 116-122.

40. Gower, J.C. (1976). Growth-free canonical variates and generalized inverses. Bull. Geol. Instn. Univ. Uppsala, 7, 1-10.

41. Gray, H.L. and Schucany, W.R. (1972). The generalized jackknife statistic. Marcel Dekker, New York.

42. Graybill, F.A. (1976). Theory and application of the linear model. Duxbury Press, Massachusetts.

43. Hawkins, D.M. (1974). Computing mean vectors and dispersion matrices in multivariate analysis of variance. Applied Statistics, 23, 234-238, Algorithm AS 72.

44. Hawkins, D.M. (1976). The subset problem in multivariate analysis of variance. J. Roy. Statist. Soc. B, 38, 132-139.

45. Hawkins, D.M. (1978). A new test for multivariate normality and homoscedasticity. Tech. Report. TWISK 45, 14p, CSIR, Pretoria.

46. Hawkins, D.M. and Rasmussen, S.E. (1973). Use of discriminant analysis for classification of strata in sedimentary successions. J. Math. Geology, 5, 163-177.

47. Hills, M. (1966). Allocation rules and their error rates.J. Roy. Statist. Soc. B, 28, 1-31.

48. Hughes, D.T. and Saw, J.G. (1972). Approximating the percentage points of Hotelling's generalized T_0^2 statistic. Biometrika, 59, 224-226.

49. Hutchison, R.I., Skinner, D.L. and Bowes, D.R. (1976). Discriminant trace-element analysis of strata in the Witwatersrand system. J. Math. Geology, 8, 413-427.

50. IMSL (1975). International mathematical and statistical libraries, library 1, edition 5, Houston, Texas.

51. James, A.T. (1954). Normal multivariate analysis and the orthogonal group. Ann. Math. Statist., 25, 40-75.

52. James, A.T. (1961). Zonal polynomials of the real positive definite symmetric matrices. Ann. Math., 74, 456-469.

53. James, A.T. (1966). Inference on latent roots by calculation of hypergeometric functions of matrix argument. Multivariate analysis, P.R. Krishnaiah, ed., Academic Press, New York, 209-235.

54. James, A.T. (1968). Calculation of zonal polynomial coefficients by use of the Laplace-Beltrami operator. Ann. Math. Statist., 39, 1711-1718.

55. James, W. and Stein, C. (1961). Estimation with quadratic loss. Proc. 4th Berkley Symp., 1, 361-379

56. Jeffreys, H. (1961). Theory of probability. 3rd edition. Clarendon Press, Oxford.

57. John, S. (1961). Errors in discrimination. Ann. Math. Statist., 32, 1125-1144.

57a. Johnson, D.E. and Hegemann, V. (1974). Procedures to generate random matrices with noncentral distributions. Comm.Statist., 3(7) 691-699.

58. Johnson, N.L. and Kotz, S. (1969). Distributions in statistics: discrete distributions. Houghton Mifflin, Boston.

59. Johnson, N.L. and Kotz, S. (1970a). Distributions in statistics: continuous univariate distributions - 1. Houghton Mifflin, Boston.

60. Johnson, N.L. and Kotz, S. (1970b). Distributions in statistics: continuous univariate distributions-2. Houghton Miffin, Boston.

61. Johnson, N.L. and Katz, S. (1972). Distributions in statistics: continuous multivariate distributions. Wiley, New York.

62. Johnson, N.L. and Leone, F. (1964). Statistics and experimental design, volume 2. Wiley, New York.

63. Kendall, M.G. and Stuart, A. (1969). The advanced theory of statistics, volume 1. Griffin, London.

64. Khatri, C.G. (1967). Some distribution problems associated with the characteristic roots of $S_1 S_2^{-1}$. Ann. Math. Statist., 38, 944-948.

65. Khatri, C.G. and Srivastava, M.S. (1978). Asymptotic expansions for distributions of characteristic roots of covariance matrices. South African Statist. J., 12, 161-186.

66. Kotz, S., Johnson, N.L. and Boyd, D.W. (1967a). Series representations of distributions of quadratic forms in normal variables 1. Central case. Ann. Math. Statist., 38, 823-837.

67. Kotz, S., Johnson, N.L. and Boyd, D.W. (1967b). Series representations of distributions of quadratic forms in normal variables 2. Non-central case. Ann. Math. Statist., 38, 838-848.

68. Lachenbruch, P.A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 23, 639-645.

69. Lachenbruch, P.A. (1968). On expected probabilities of misclassification in discriminant analysis,necessary sample size, and a relation with the multiple correlation coefficient. Biometrics, 24, 823-834.

70. Lachenbruch, P.A. (1973). Some results on the multiple group discriminant problem. Discriminant analysis and applications, T. Cacoullos, ed., Academic Press, New York, 193-211.

71. Lachenbruch, P.A. (1975). Discriminant Analysis. Hafner Press, New York.

72. Lachenbruch, P.A. and Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10, 1-11.

73. Lawley, D.N. (1956). Tests of significance for the latent roots of covariance and correlation matrices. Biometrika, 43, 128-136.

74. Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the Linear Model. J.Roy. Statist, Soc. B, 34, 1-41.

75. Maritz, J.S. (1970). Empirical Bayes methods. Methuen, London.

76. Mc Cracken, D.D. and Dorn, W.S. (1964). Numerical methods and fortran programming. Wiley, New York.

77. Mc Kay, R.J. (1977). Simultaneous procedures for variable selection in multiple discriminant analysis. Biometrika, 64, 283-290.

78. Mc Keon, J.J. (1974). F approximations to the distribution of Hotelling's T_0^2. Biometrika, 61, 381-383.

79. Mc Lachlan, G.J. (1974a). The asymptotic distributions of the conditional error rate and risk in discriminant analysis. Biometrika, 61, 131-135.

80. Mc Lachlan, G.J. (1974b). Estimation of the errors of misclassification on the criterion of asymptotic mean square error. Technometrics, 16, 255-260.

81. Mc Lachlan, G.J. (1974c). An asymptotic unbiased technique for estimating the error rates in discriminant analysis. Biometrics, 30, 239-249.

82. Mc Lachlan, G.J. (1975). Confidence intervals for the conditional probability of misallocation in discriminant analysis. Biometrics, 31, 161-167.

83. Mc Lachlan, G.J. (1976a). The bias of the apparent error rate in discriminant analysis. Biometrika, 63, 239-244.

84. Mc Lachlan, G.J. (1976b). A criterion for selecting variables for the linear discriminant function. Biometrics, 32, 529-534.

85. Mc Lachlan, G.J. (1977). Constrained sample discrimination with the studentized classification statistic W. Comm. Statist., A, 6, 575-583.

86. Michaelis, J. (1973). Simulation experiments with multiple group linear and quadratic discriminant analysis. Discriminant analysis and applications, T. Cacoullos, ed., Academic Press, New York, 225-238.

87. Miller, R.G. (1974). The jackknife - a review. Biometrika, 61, 1-15.

88. NAG (1975). Nottingham Algorithms Group library. Univ. of Nottingham.

88a. Odell, P.L. and Feiveson, A.H.(1966). A numerical procedure to generate a sample covariance matrix. J. Amer. Statist. Assoc., 61, 199-203.

89. Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. Ann. Math. Statist., [illegible], 1286-1301, (Correction(1968). Ann.Math.Statist., 39, 1358-1359).

90. Pearson, K. (Ed.)(1922). Tables of the Incomplete Γ-Function. H.M. Stationery Office, London, (Since 1934, Cambridge University Press).

91. Pillai, K.C.S. and Samson, P. (1959). On Hotelling's generalization of T^2. Biometrika, 46, 160-168.

92. Press, S.J. (1966). Linear combinations of non-central chi-square variates . Ann. Math. Statist., 37, 480-487.

93. Press, S.J. (1972). Applied multivariate analysis. Holt, Rinehart and Winston,New York.

94. Quenouille, M.H. (1956). Notes on bias in estimation. Biometrika, 43, 353-360.

95. Rao, C.R. (1965). Linear statistical inference and its applications. Wiley, New York.

96. Reyment, R.A. and Banfield, C.F. (1976). Growth-free canonical variates applied to fossil foraminifers. Bull. Geol. Instn. Univ. Uppsala, 7, 11-21.

97. Robbins, H.E. (1948). The distribution of a definite quadratic form. Ann. Math. Statist., 19, 266-270.

98. Robbins, H.E. and Pitman, E.J.G. (1949). Application of the method of mixtures to quadratic forms in normal variates. Ann. Math. Statist., 20, 552-560.

99. Ruben, H. (1960). Probability content of regions under spherical normal distributions, 1. Ann. Math. Statist., 31, 598-619.

100. Ruben. H. (1962). Probability content of regions under spherical normal distributions, 4. Ann. Math. Statist., 33, 542-570.

101. Silvey, S.D. (1975). Statistical inference. Chapman and Hall, London.

102. Smith, A.F.M. (1973). A general Bayesian linear model. J. Roy. Statist. Soc. B, 35, 67-75.

103. Solomon, H. and Stephens, M.A. (1977). Distribution of a sum of weighted chi-square variables. J. Amer. Statist. Assoc., 72, 881-885.

104. Sorum, M. (1972a). Three probabilities of misclassification. Technometrics, 14, 309-316.

105. Sorum, M. (1972b). Estimating the expected and the optimal probabilities of misclassification. Technometrics, 14, 935-943.

106. Sparks. D.N. and Todd, A.D. (1973). Latent roots and vectors of a symmetric matrix. Applied statistics, 22, 260-265, Algorithm AS 50.

107. Tukey, J.W. (1958). Bias and confidence in not quite large samples. (abstract), Ann. Math. Statist., 29, 614.

107a. van der Westhuizen, A.J. and Nagel, P.J.A.(1979). Programs for the evaluation of zonal polynomials. Submitted to Appl. Statistics.

108. van Niekerk, J.N. (1978). An introduction to the empirical Bayes approach to statistical decision problems. Tech. Report. TWISK 18, 14p, CSIR, Pretoria.

109. Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Statist., 15, 145-162.

110. Walsh, G.R. (1975). Methods of optimization. Wiley, London.

111. Welch, B.L. (1939). Note on discriminant functions. Biometrika 31, 218-220.

112. Wilks,S.S. (1962). Mathematical Statistics. Wiley, New York.