UNIVERSITY COLLEGE, CORK

ON CLASSICAL AND OTHER METHODS OF DISCRIMINANT ANALYSIS

AND

ESTIMATION OF LOG-ODDS

by

BRENDAN J. MURPHY

DEPARTMENT OF STATISTICS

PROFESSOR M.A. MORAN

Thesis submitted for the Degree of Doctor of Philosophy
of the National University of Ireland

MARCH 1981

Researched under the supervision of Professor M.A. Moran

CONTENTS

SUMMARY

For two multinormal populations with equal covariance matrices the likelihood ratio discriminant function, an alternative allocation rule to the sample linear discriminant function when $n_1 \neq n_2$, is studied analytically. With the assumption of a known covariance matrix its distribution is derived and the expectation of its actual and apparent error rates evaluated and compared with those of the sample linear discriminant function. This comparison indicates that the likelihood ratio allocation rule is robust to unequal sample sizes.

The quadratic discriminant function is studied, its distribution reviewed and evaluation of its probabilities of misclassification discussed. For known covariance matrices the distribution of the sample quadratic discriminant function is derived. When the known covariance matrices are proportional exact expressions for the expectation of its actual and apparent error rates are obtained and evaluated. The effectiveness of the sample linear discriminant function for this case is also considered.

Estimation of true log-odds for two multinormal populations with equal or unequal covariance matrices is studied. The estimative, Bayesian predictive and a kernel method are compared by evaluating their biases and mean square errors. Some algebraic expressions for these quantities are derived. With equal covariance matrices the predictive method is preferable. Where it derives this superiority is investigated by considering its performance for various levels of fixed true log-odds. It is also shown that the predictive method is sensitive to $n_1 \neq n_2$. For unequal but proportional covariance matrices the unbiased estimative method is preferred.

Product Normal kernel density estimates are used to give a kernel estimator of true log-odds. The effect of correlation in the variables with product kernels is considered. With equal covariance matrices the kernel and parametric estimators are compared by simulation. For moderately correlated variables and large dimension sizes the product kernel method is a good estimator of true log-odds.

ACKNOWLEDGEMENTS

I would like to express my appreciation to Aidan Moran for affording me the opportunity of undertaking this thesis and for the many stimulating discussions we have had. Thanks are also due to Michael Crowley with whom programming problems were discussed. My wife Rosaleen and children Ruth and Judith are also to be praised for their patience and understanding. Finally my thanks to Eileen Fenton for her excellent typing.

LIST OF TABLES

List of Tables (continued) Page

List of Tables(continued) Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Introduction

Discriminant analysis is concerned with the allocation of an observation known a priori to have come from one of two or more populations. Allocation of the observation is made using an allocation rule or discriminant function. The exact forms of the discriminant functions depend on the distribution of the variables in the various populations under consideration. In this thesis our interest will centre on the classic case of two multivariate Normal populations with equal or unequal covariance matrices.

When the population distributions and parameters are known, optimal allocation rules, which minimise the number of observations misclassified may be constructed. In practice however the population parameters are unknown but the type of population density is assumed, and sample allocation rules constructed from sample observations whose true allocation is known, are used. These sample allocation rules also misclassify observations and their error rates or probabilities of misclassification are often used to assess their performance.

This thesis has two parts; the first consisting of five chapters, concerning the distribution of the likelihood ratio, sample linear and sample quadratic discriminant functions; the second of four chapters concerning the estimation of true log-odds.

In the first part, the likelihood ratio discriminant function, an alternative allocation rule to the sample linear discriminant function when $n_1 \neq n_2$,is studied analytically. For the assumption of known and equal covariance matrices its conditional and unconditional distributions are derived. Expectations of its associated actual and apparent error rates are evaluated and compared with their counterparts for the linear discriminant function. With $n_1 \neq n_2$ it is found in contrast to the linear discriminant function that there is very little distortion in the expected actual probabilities of misclassification of the likelihood ratio allocation rule. The case of equal but unknown covariance matrices is also considered.

The optimal quadratic discriminant function is then studied. Its distribution and associated probabilities of misclassification are reviewed. The distribution for some special cases of the population parameters such as proportional covariance matrices is considered in detail and exact expressions are derived for the optimal error rates. For the assumption of known covariance matrices the conditional and unconditional distributions of the sample quadratic discriminant function are derived and with the additional assumption of proportional covariance matriees, exact expressions for the expected actual and apparent error rates are obtained. The effectiveness of linear discrimination in this case of known proportional covariance matrices is also examined analytically.

In the second part of the thesis, estimation of true log-odds is considered. Various methods of estimating true log-odds are compared including classical methods, the Bayesian predictive method and a kernel method. Comparison of the estimators of log-odds is based on the evaluation of their biases and mean square errors and implications for corresponding rates of misclassification are also considered. Exact expressions are derived for the bias of the parametric estimators and for mean square errors where possible. The mean square error of the predictive method and the expected actual probabilities of misclassification are estimated by simulation. With equal covariance matrices the exaggeration in true log-odds reported in the literature for the estimative method is shown to be primarily due to bias. The estimative method corrected for bias gives a distinct improvement. The predictive method remains superior however with smaller mean square error and conservative bias. The relative performance of the predictive and estimative method is examined in some detail for various levels of true log-odds, population separations and sample sizes. While the predictive methods superiority is very marked for large log-odds, it persists albeit modestly for low true log-odds. It is shown that the predictive estimator is sensitive to $n_1 + n_2$ and an appropriate adjustment is suggested. For unequal but proportional covariance matrices the superiority of the predictive method does not persist and the unbiased estimative method seems preferable.

Finally, product Normal kernel density estimates are used to give a kernel estimator of true log-odds. Quite dramatic claims for the relative allocation performance of the product kernel to the estimative allocation rule have recently been made in the literature. Such studies however have been based on the population assumption of independence of the variables. The effect of correlation in the variables on product kernel density estimates is considered. For equal covariance matrices the product kernel's allocation and estimative performance is compared by simulation with the parametric methods in the presence of correlation in the variables. It is found that the claims for the allocation performance of the product kernel method are overstated and only hold when the variables are close to independence and then only for large dimension sizes and well separated populations. For moderately correlated variables, large dimension sizes and small sample sizes, the product kernel method is however a good estimator of true log-odds.

The thesis concludes with a chapter outlining possible areas into which the work undertaken in the previous chapters might be extended.

1.2 The Linear and Quadratic Discriminant Functions

The two basic discriminant functions together with some general notation are introduced here. No specific symbol is used to denote a vector or matrix, it is hoped that such quantities will be apparent from the context in which they occur.

A (p x 1) observation vector x is assumed to have come from one of two multivariate Normal populations Π_t (t = 1,2). The p-dimensional Normal distributions will be denoted by $N_p(\mu_t, \Sigma_t)$ where μ_t is the mean vector and Σ_t the covariance matrix. If π_t(t = 1,2) are the prior probabilities of x coming from Π_t then allocation to Π_1 or Π_2 is made according as

$$\ell n \ \{f_1(x)/f_2(x)\} \gtrless C \qquad (1.2.1)$$

where $C = \ell n(\pi_2/\pi_1)$

and f_t is the probability density function of x in Π_t. The allocation rule (1.2.1) minimises the total probability of misclassification, Welch (1939). The symbol C will be referred to as the cut-off point. Anderson (1958, pp127-133) has shown that the allocation rule (1.2.1) with C now involving the costs of misclassification minimises the average cost and where prior probabilities are unknown that rules of the type (1.2.1) form a minimal complete class.

With the assumption of equal covariance matrices i.e. $\Sigma_1 = \Sigma_2 = \Sigma$, the allocation rule (1.2.1) becomes

$$L(x) = (\mu_1 - \mu_2)' \, \Sigma^{-1} \{x - \tfrac{1}{2}(\mu_1 + \mu_2)\} \gtrless C$$

the familiar linear discriminant function, Welch (1939). The distribution of L(x) for x in Π_t is obviously Normal with mean $(-1)^{t-1} \tfrac{1}{2}\Delta^2$ and variance Δ^2 where $\Delta^2 = (\mu_1 - \mu_2)' \, \Sigma^{-1} \, (\mu_1 - \mu_2)$ is Mahalonobis' squared distance between the populations, Anderson (1958, pp133-137).

With $\Sigma_1 \neq \Sigma_2$ the allocation rule (1.2.1) becomes

$$Q(x) = -\tfrac{1}{2}(x - \mu_1)' \, \Sigma_1^{-1}(x - \mu_1) + \tfrac{1}{2}(x - \mu_2)' \, \Sigma_2^{-1}(x - \mu_2) - \tfrac{1}{2}\ell n(|\Sigma_1|/|\Sigma_2|)$$
$$\gtrless C$$

the quadratic discriminant function Smith (1947). The distribution of Q(x) is considered in a later chapter.

1.3 The Sample Linear and Quadratic Discriminant Functions

When the population parameters are unknown it is customary to replace them by their sample estimates, Anderson (1958, p137). This will result in the sample linear and quadratic discriminant functions.

Random samples x_{tj} of size n_t, $1 \leq j \leq n_t$ are assumed from Π_t (t = 1,2). Denoting the sample means and covariance matrices by $\bar{x}_t$ and S_t and with

$$S = \{(n_1 - 1)S_1 + (n_2 - 1)S_2\} \, / \, (n_1 + n_2 - 2)$$

the sample linear discriminant function $\hat{L}(x)$ corresponding to L(x) is given by

$$\hat{L}(x) = (\bar{x}_1 - \bar{x}_2)' \, S^{-1}\{x - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_2)\} \gtrless K$$

With the assumption of equal priors or in the absence of information to the contrary K is set equal to zero. In some cases it is assumed that the proportions of the sample size reflect the incidence rate of the observations and K is set equal to $\ln(n_2/n_1)$. $\hat{L}(x)$ is often referred to as the W statistic from Anderson (1951).

The conditional distribution of $\hat{L}(x)$, i.e. $\bar{x}_1, \bar{x}_2$ and S are assumed fixed, is obviously Normal with mean $\hat{L}(\mu_t)$ and variance $(\bar{x}_1 - \bar{x}_2)' S^{-1} \Sigma S^{-1}(\bar{x}_1 - \bar{x}_2)$. The unconditional distribution of $\hat{L}(x)$ is however extremely complicated and has been studied by Anderson (1951), Sitgreaves (1952) and Bowker (1961). Asymptotic expansions for the cumulative distribution function of $\hat{L}(x)$ were derived by Bowker and Sitgreaves (1961) for $n_1 = n_2$ and by Okamoto (1963) for $n_1 \neq n_2$. Anderson (1973) has given an asymptotic expansion for a studentised version of $\hat{L}(x)$.

The sample quadratic discriminant function $\hat{Q}(x)$ corresponding to Q(x) is given by

$$\hat{Q}(x) = -\tfrac{1}{2}(x - \bar{x}_1)' S_1^{-1} (x - \bar{x}_1) + \tfrac{1}{2}(x - \bar{x}_2)' S_2^{-1} (x - \bar{x}_2)$$

$$-\tfrac{1}{2}\ln(|S_1|/|S_2|)$$

$$\gtrless K$$

first described by Smith (1947).

The unconditional distribution of $\hat{Q}(x)$ has proved to be more intractable than that of $\hat{L}(x)$. The results in the literature are asymptotic expansions for various special cases of the population parameters and are reviewed in Chapter 4.

1.4 Optimal Probabilities of Misclassification

These are the probabilities of misclassification when the optimal rules L(x) and Q(x) are used to allocate observations from Π_t, (t = 1,2).

With $\Sigma_1 = \Sigma_2 = \Sigma$ they are defined as

$$L_1 = \Pr\{L(x) < C \mid x \text{ in } \Pi_1\}$$

$$= \Phi\left(\frac{C - \frac{1}{2}\Delta^2}{\Delta}\right)$$

and

$$L_2 = \Pr\{L(x) > C \mid x \text{ in } \Pi_2\}$$

$$= \Phi\left(\frac{-C - \frac{1}{2}\Delta^2}{\Delta}\right)$$

where Φ denotes the standard Normal distribution function. As noted in Section 1.2 the allocation rule L(x) minimises the total probability of misclassification $\pi_1 L_1 + \pi_2 L_2$. With equal prior probabilities C = 0, and

$$L_1 = L_2 = \Phi\ (-\tfrac{1}{2}\Delta).$$

With $\Sigma_1 \neq \Sigma_2$ the optimal probabilities of misclassification will be denoted by Q_1 and Q_2 where

$$Q_1 = \Pr\{Q(x) < C \mid x \text{ in } \Pi_1\}$$

and

$$Q_2 = \Pr\{Q(x) > C \mid x \text{ in } \Pi_2\}\ .$$

The evaluation of Q_t, t = 1 and 2 is considered in Chapter 3 where the distribution of Q(x) is derived.

1.5 Actual and Apparent Probabilities of Misclassification

The actual probabilities of misclassification are the misclassification rates of the sample based rules when used to classify future observations from Π_1 and Π_2. They are also referred to as the conditional probabilities of misclassification, $\bar{x}_t$ and S_t being assumed fixed.

For $\hat{L}(x)$ the actual probabilities of misclassification are given by

$$L_1^* = \Pr\{\hat{L}(x) < K \mid x \text{ in } \Pi_1, \bar{x}_1, \bar{x}_2, S\}$$

$$= \Phi\left(\frac{K - \hat{L}(\mu_1)}{\sqrt{(\bar{x}_1 - \bar{x}_2)' S^{-1} \Sigma S^{-1} (\bar{x}_1 - \bar{x}_2)}}\right)$$

and

$$L_2^* = \Pr\{\hat{L}(x) > K \mid x \text{ in } \Pi_2, \bar{x}_1, \bar{x}_2, S\}$$

For $\hat{Q}(x)$

$$Q_1^* = \Pr\{\hat{Q}(x) < K \mid x \text{ in } \Pi_1, \bar{x}_1, \bar{x}_2, S_1, S_2\}$$

and

$$Q_2^* = \Pr\{\hat{Q}(x) > K \mid x \text{ in } \Pi_2, \bar{x}_1, \bar{x}_2, S_1, S_2\}$$

evaluation of Q_t^* is considered in Chapter 4. Conditional probabilities of misclassification will be denoted by an asterisk. Expectations of the actual probabilities of misclassification over repeated samples of sizes n_1 and n_2 from Π_1 and Π_2 are sometimes referred to as the unconditional probabilities of misclassification and will be denoted by E (L_t^*) and E (Q_t^*).

The apparent probabilities of misclassification are defined as the proportions of the sample observations misclassified by the sample allocation rules and will be denoted by a double asterisk i.e. L_t^{**} and Q_t^{**}. They are estimates of the actual probabilities of misclassification. The method of estimation sometimes referred to as the resubstitution method was proposed by Smith (1947) and is known to result in substantially biased estimates, Hills (1966). This bias is to be expected since the sample observations are used twice, once in the construction of the allocation rules and then as observations to be allocated by these rules. Expectations of the apparent probabilities of misclassification over repeated samples from Π_1 and Π_2 will enable us to see the extent of the bias in the resubstitution method. These expectations will be denoted by $E(L_t^{**})$ and $E(Q_t^{**})$.

A clear distinction between optimal, actual and apparent error rates was made by Hills (1966). Considerable work on the estimation of these error rates and their expectations when $\Sigma_1 = \Sigma_2$ has been done by Lachenbruch (1967, 1968), Lachenbruch and Mickey (1968) who suggested several alternative estimators Broffitt and Williams (1973), Sorum (1968, 1971, 1972, 1973), Sedrask and Okamoto (1971) and Mc Lachlan (1972, 1973, 1974, 1976). Inequalities between the various error rates and their expectations have been given by Hills (1966), Das Gupta (1974) and Glick (1972). In the latter reference the consistency of sample based allocation rules is also considered.

CHAPTER 2

THE LIKELIHOOD RATIO ALLOCATION RULE OR Z STATISTIC

2.1 INTRODUCTION

The sample linear discriminant function $\hat{L}(x)$ enjoys widespread use for discriminating between multinormal populations with equal covariance matrices when some or all of the populations parameters are unknown. One of the main reasons for its popularity is its multiple regression derivation, Fisher (1936). In the literature many alternative sample allocation rules have been proposed among them the likelihood ratio allocation rule, Anderson (1958, P141-142) or Z statistic, John (1963), where

$$Z(x) = -\frac{n_1}{n_1+1}(x-\bar{x}_1)'\,S^{-1}(x-\bar{x}_1) + \frac{n_2}{n_2+1}(x-\bar{x}_2)'\,S^{-1}(x-\bar{x}_2) \qquad (2.1.1)$$

$$= \frac{2n}{n+1}\,\hat{L}(x) \quad \text{when } n_1 = n_2 = n.$$

For unequal sample sizes the Z statistic has been shown to have some advantages over $\hat{L}(x)$ and these are investigated further in this chapter for Σ known.

The conditional and unconditional distribution of Z(x) are derived for Σ known, here S in (2.1.1) is replaced by Σ. Expectations of its actual and apparent probabilities of misclassification are derived and evaluated. The imbalance in the expectations of the actual and apparent probabilities of misclassification for Z(x), for unequal sample sizes, is compared with that for $\hat{L}(x)$. Various approximations to the expectations of the probabilities of misclassification for Z(x) are then considered. The chapter concludes with a discussion of possible extensions to the case where the covariance matrix Σ is unknown.

2.2 Derivation and Notation

The likelihood ratio allocation rule as derived by Anderson (1958, P141-142) is as follows.

Let $f(x|\mu,\Sigma)$ denote the probability density function of the unidentified observation x where it is believed that $(\mu,\Sigma) = (\mu_1,\Sigma)$ or (μ_2,Σ). Then the likelihood ratio statistic for testing the hypothesis $H_1: \mu = \mu_1$ versus $H_2: \mu = \mu_2$ is

$$LR = \frac{\max\limits_{H_1} f(x|\mu_1,\Sigma) \prod\limits_{t=1}^{2} \prod\limits_{j=1}^{n_t} f\,(x_{tj}|\mu_t,\Sigma)}{\max\limits_{H_2} f(x|\mu_2,\Sigma) \prod\limits_{t=1}^{2} \prod\limits_{j=1}^{n_t} f\,(x_{tj}|\mu_t,\Sigma)} .$$

The maximum likelihood estimates of the unknown parameters under H_1 are

$$\hat{\mu}_1 = \frac{n_1 \bar{x}_1 + x}{n_1+1} , \quad \hat{\mu}_2 = \bar{x}_2 ,$$

$$\hat{\Sigma} = \frac{1}{n_1+n_2+1} \left[(n_1+n_2-2)\, S + \frac{n_1}{n_1+1} (x-\bar{x}_1)(x-\bar{x}_1)'\right].$$

There are obvious changes for the corresponding estimates under H_2 where the maximum likelihood estimate of Σ will be denoted by $\tilde{\Sigma}$. Substitution of these estimates gives after some simplification

$$LR = \{ |\tilde{\Sigma}| \,/\, |\hat{\Sigma}| \}^{\frac{1}{2}(n_1+n_2+1)}$$

$$= \left[\frac{1 + \frac{n_2}{(n_2+1)(n_1+n_2-2)} (x-\bar{x}_2)'\, S^{-1} (x-\bar{x}_2)}{1 + \frac{n_1}{(n_1+1)(n_1+n_2-2)} (x-\bar{x}_1)'\, S^{-1} (x-\bar{x}_1)} \right]^{\frac{1}{2}(n_1+n_2+1)} .$$

The hypothesis H_1 is favoured according as LR $\gtrless$ C, $C > 0$, and allocation to Π_1 or Π_2 would be made accordingly.

With the cut off C = 1, which may be interpreted as our belief that the unknown prior probabilities are equal, the likelihood ratio allocation rule

$$Z(x) = -\frac{n_1}{n_1+1} (x-\bar{x}_1)' S^{-1} (x-\bar{x}_1) + \frac{n_2}{n_2+1} (x-\bar{x}_2)' S^{-1} (x-\bar{x}_2)$$

$$\gtrless 0$$

is obtained. We note that if the sample sizes are equal, $n_1 = n_2 = n$, and C = 1 that

$$Z(x) = \frac{2n}{n+1} \hat{L}(x) \qquad (2.2.1)$$

and for allocation purposes the two rules are equivalent.

An equivalent intuitive rule to Z(x) was proposed by Rao (1954) who suggested comparing tests of significance of the unidentified observation x coming from Π_1 or Π_2. If the test rejects the hypothesis that x comes from Π_1 at a level α_1 and the hypothesis that x comes from Π_2 at a level α_2 then x is assigned to Π_1 or Π_2 according as $\alpha_1 \lesseqgtr \alpha_2$. For multinormally distributed populations the conventional test criterion is $\frac{n_t}{n_t+1} (x-\bar{x}_t)' S^{-1} (x-\bar{x}_t)$ which is distributed as $\frac{n_1+n_2-p-1}{p(n_1+n_2-2)}$ times an F variate with degrees of freedom p and (n_1+n_2-p-1). Thus the procedure amounts to allocating to Π_1 or Π_2 according as

$$\frac{n_1}{n_1+1} (x-\bar{x}_1)' S^{-1}(x-\bar{x}_1) \lessgtr \frac{n2}{n_2+1} (x-\bar{x}_2)' S^{-1} (x-\bar{x}_2)$$

i.e. according as $Z(x) \gtrless 0$.

With the restriction that the covariance matrix Σ is known a similar likelihood ratio analysis gives the allocation rule

$$Z(x) = - \frac{n_1}{n_1+1} (x-\bar{x}_1)' \, \Sigma^{-1}(x-\bar{x}_1) + \frac{n_2}{n_2+1} (x-\bar{x}_2)' \, \Sigma^{-1}(x-\bar{x}_2)$$

$$= - N_1 \; (x-\bar{x}_1)' \, \Sigma^{-1}(x-\bar{x}_1) + N_2 \; (x-\bar{x}_2)' \, \Sigma^{-1}(x-\bar{x}_2) \qquad (2.2.2)$$

where $N_t = \frac{n_t}{n_t+1} \quad t = 1$ and 2.

Further reference to the Z statistic will be to (2.2.2) and Σ known unless otherwise indicated.

The actual probability of misclassification of the Z statistic for x in Π_1 is

$$Z_1^* = \Pr\{Z(x) < 0 \mid x \text{ in } \Pi_1, \bar{x}_1, \bar{x}_2\}$$

with expectation

$$E(Z_1^*) = \Pr\{Z(x) < 0 \mid x \text{ in } \Pi_1\} .$$

The corresponding Z_2^* and $E(Z_2^*)$ for x in Π_2 are similarly defined.

The apparent probability of misclassification of the Z statistic, defined as the proportion of the sample observations from Π_1 misclassified by Z(x), will be denoted by Z_1^{**}. Its expectation is

$$E(Z_1^{**}) = \Pr\{Z(x_{1j}) < 0 \mid x_{1j} \text{ a random member of the sample from } \Pi_1\}$$

with Z_2^{**} and $E(Z_2^{**})$ similarly defined for the observations from Π_2.

2.3 A Review of the Literature on the Z Statistic

As noted in Section 2.2, Anderson (1958) proposed and derived the likelihood ratio allocation rule and Rao (1954) had suggested an equivalent allocation rule to Z(x).

Ellison (1962) for Σ known and a zero-one loss function has shown that Z(x) is an admissable translation-invariant Bayes rule. For Σ known and a loss function dependent on Δ, n_1 and n_2 Das Gupta (1965) has shown that Z(x) is an unbiased, admissible minimax rule. By the term unbiased Das Gupta means that $E(Z_t^*) < \frac{1}{2}$ for all (n_1, n_2), p and Δ. He has also shown that this is true for Σ unknown. Schaafsma (1973) for Σ known, p = 1 and a zero-one loss function, has shown that Z(x) has uniform minimum risk in an invariant class, however $\hat{L}(x)$ is not a member of this class.

John (1960a) obtained a complicated expression for $E(Z_t^*)$ when Σ is known. He showed that these expectations were bounded above by $\frac{1}{2}$, anticipating Das Gupta's result on unbiasedness. John (1963) for Σ known expressed $E(Z_t^*)$ as the difference of two weighted non-central chi-squares and suggested Abdel-Aty's (1954) approximation to a non-central chi-square be used for evaluation. Schaafsma and Van Vark (1977) obtained expressions for $E(Z_t^*)$ when p = 1 and Σ known in terms of standard Normal distribution functions, similar to John's (1961) and Hills (1966) expressions for $E(L_t^*)$.

Memon (1970) for Σ known and Memon and Okamoto (1971) for Σ unknown provide asymptotic expansions for the unconditional distribution of Z(x) similar to Okamoto's (1963) expansion for

$\hat{L}(x)$. Siotani and Wang (1977) compared for Σ unknown the allocation rules $\hat{L}(x)$ and $Z(x)$ by considering the difference $\{E(L_1^*) + E(L_2^*)\} - \{E(Z_1^*) + E(Z_2^*)\}$ for various combinations of $n_1 \neq n_2$, p and Δ. Here $E(L_t^*)$ and $E(Z_t^*)$ were estimated from Okamoto's and Memon and Okamoto's asymptotic expansions with extensions to include a cubic term. They conclude that the superiority of one procedure over the other was not consistent but depended on the configuration of the parameter set $\{n_1, n_2, p, \Delta\}$.

2.4 The Conditional Distribution of Z(x)

The conditional distribution of $Z(x)$, given $\bar{x}_1$ and $\bar{x}_2$, is derived and evaluation of the actual probabilities of misclassification considered.

$$Z(x) = - N_1 \; (x-\bar{x}_1)' \; \Sigma^{-1} \; (x-\bar{x}_1) + N_2 \; (x-\bar{x}_2)' \; \Sigma^{-1} \; (x-\bar{x}_2)$$

$$= (N_2-N_1) \; \{x - \frac{(N_2\bar{x}_2-N_1\bar{x}_1)}{N_2-N_1}\}' \; \Sigma^{-1} \; \{x - \frac{(N_2\bar{x}_2-N_1\bar{x}_1)}{N_2-N_1}\}$$

$$- \frac{N_1N_2}{(N_2-N_1)} (\bar{x}_1-\bar{x}_2)' \; \Sigma^{-1} \; (\bar{x}_1-\bar{x}_2) \qquad \text{if } n_1 \neq n_2$$

$$= (N_2-N_1) \; \{(x-a)' \; \Sigma^{-1}(x-a)\} - b$$

where $a = \frac{N_2\bar{x}_2-N_1\bar{x}_1}{N_2-N_1}$ and $b = \frac{N_1N_2}{N_2-N_1} (\bar{x}_1-\bar{x}_2)' \; \Sigma^{-1}(\bar{x}_1-\bar{x}_2)$.

Conditional on $\bar{x}_1$ and $\bar{x}_2$

$$(x-a)' \; \Sigma^{-1}(x-a) \sim \chi^2 \; (p, \; (\mu_t-a)' \; \Sigma^{-1}(\mu_t-a))$$

for x in Π_t, $t = 1$ and 2, where $\chi^2(\nu,\lambda)$ denotes a non-central chi-squared variate with degrees of freedom ν and non-centrality parameter λ. The non-centrality parameters $(\mu_t\text{-}a)'\ \Sigma^{-1}(\mu_t\text{-}a)$ contain unknown population parameters. Hence the evaluation of the actual probabilities of misclassification

$$Z_1^* = \Pr\{(N_2-N_1)\ \chi^2(p,\ (\mu_1-a)'\ \Sigma^{-1}(\mu_1-a)) \ < \ b\}$$

and

$$Z_2^* = \Pr\{(N_2-N_1)\ \chi^2(p,\ (\mu_2-a)'\ \Sigma^{-1}(\mu_2-a)) \ > \ b\}$$

is impossible in practice although it may be undertaken in simulation studies. In Section 3.6 closed expressions are given for the probabilities of non-central chi-squares with odd degrees of freedom.

With $n_1=n_2$

$$Z(x) = \frac{2n}{n+1}\hat{L}(x) \ \gtrless\ 0$$

and the actual probabilities of misclassification of $Z(x)$ are the same as those of $\hat{L}(x)$ as given in Section 1.5, if $K = 0$ and S is replaced by Σ i.e.

$$Z_1^* = \Phi\left(\frac{-\hat{L}(\mu_1)}{\sqrt{(\bar{x}_1-\bar{x}_2)'\ \Sigma^{-1}(\bar{x}_1-\bar{x}_2)}}\right) .$$

2.5 The Unconditional Distribution of Z(x)

Here the method of proof is similar to that employed by John (1960b) and Moran (1974) in deriving the expectations of the actual probabilities of misclassification of $\hat{L}(x)$. The method is given in some detail as it will be used again on several occasions. John's (1963) expressions for the expectation of the actual probabilities of misclassification of the likelihood ratio allocation rule are somewhat involved and not suitable for evaluation purposes.

Let the cumulative distribution function of the Z statistic be given by

$$G_t(a) = \Pr\{Z(x) \leqslant a \mid x \text{ in } \Pi_t\} \qquad t = 1 \text{ and } 2.$$

Thus for x in Π_1

$$G_1(a) = \Pr\{-N_1(x-\bar{x}_1)'\ \Sigma^{-1}(x-\bar{x}_1) + N_2(x-\bar{x}_2)'\ \Sigma^{-1}(x-\bar{x}_2) \leqslant a\}$$

$$= \Pr\{-y'y + w'w \leqslant a\}$$

where $y = \Sigma^{-\frac{1}{2}}(x-\bar{x}_1)\sqrt{N_1}$ and $w = \Sigma^{-\frac{1}{2}}(x-\bar{x}_2)\sqrt{N_2}$

$$= \Pr\{(w-y)'\ (w+y) \leqslant a\}$$

$$= \Pr\{\ r's \leqslant a\}$$

where $r = (w-y)$ and $s = (w+y)$.

Now r and s are multinormally distributed with means $\Sigma^{-\frac{1}{2}}(\mu_1-\mu_2)\sqrt{N_2}$ and covariance matrices

$$\frac{d_1^2}{n_1n_2}\ I \qquad \text{and} \qquad \frac{d_2^2}{n_1n_2}\ I$$

respectively, where

$$d_1 = \{n_1N_2 + n_2N_1 + n_1n_2\ (\sqrt{N_1} - \sqrt{N_2}\ )^2\}^{\frac{1}{2}}$$

$$d_2 = \{n_1N_2 + n_2N_1 + n_1n_2\ (\sqrt{N_1} + \sqrt{N_2}\ )^2\}^{\frac{1}{2}}.$$

With $c_1 = \frac{(n_1n_2)^{\frac{1}{2}}}{d_1}$ and $c_2 = \frac{(n_1n_2)^{\frac{1}{2}}}{d_2}$

$u = c_1r$ and $v = c_2s$ are multinormally distributed with covariance matrices I. The correlation ρ_1 between the pairwise elements of u and v is given by

$$\rho_1 = \{n_1N_2 - n_2N_1 + n_1n_2\ (N_2-N_1)\}\ /\ d_1d_2.$$

If $n_1N_2(n_2+1) = n_2N_1(n_1+1)$, then $\rho_1 = 0$, which is always true for the Z statistic. The correlation parameter ρ_1 even when zero will be retained in the remainder of the derivation for illustrative purposes as the method of proof will be employed again in situations where the correlation is non-zero.

Now $$\begin{aligned} G_1(a) &= \Pr\{r's \leqslant a\} \\ &= \Pr\{u'v \leqslant a\ c_1c_2\} \\ &= \Pr\{(u+v)'(u+v) - (u-v)'(u-v) \leqslant 4a\ c_1c_2\} \end{aligned}$$

and (u+v) and (u-v) are independently multinormally distributed with covariance matrix $2(1+\rho_1)$ I and $2(1-\rho_1)$ I respectively.

Hence $\omega_1 = \frac{1}{2(1+\rho_1)}\ (u+v)'(u+v)$ and $\omega_2 = \frac{1}{2(1-\rho_1)}\ (u-v)'(u-v)$ are independently distributed as non-central chi-squares, $\chi^2(p,\delta_1)$ and $\chi^2(p,\delta_2)$, where the non-centralities δ_1 and δ_2 are given by

$$\begin{aligned} \delta_1 &= \{2(1+\rho_1)\}^{-1}\ (c_1+c_2)^2\ N_2\ \Delta^2 \\ \delta_2 &= \{2(1-\rho_1)\}^{-1}\ (c_1-c_2)^2\ N_2\ \Delta^2\ . \end{aligned} \qquad (2.5.1)$$

Writing

$$\begin{aligned} G_1(a) &= \Pr\{\frac{(1+\rho_1)}{2c_1c_2}\ \omega_1\ -\ \frac{(1-\rho_1)}{2c_1c_2}\ \omega_2 \leqslant a\} \\ &= \Pr\{\ \frac{1}{2c_1c_2}\ \chi^2(p,\delta_1)\ -\ \frac{1}{2c_1c_2}\ \chi^2(p,\delta_2) \leqslant a\}\ , \end{aligned}$$

as $\rho_1 = 0$, we see that the unconditional distribution of Z(x) is that of an indefinite non-central quadratic form.

A similar analysis holds for x in Π_2 with the equivalent parameters given by

$$\rho_2 = \rho_1 = 0, \quad c_3 = c_1, \quad c_4 = c_2$$

$$\delta_3 = \{2(1+\rho_2)\}^{-1} \quad (c_3-c_4)^2 \ N_1 \ \Delta^2$$

and $\delta_4 = \{2(1-\rho_2)\}^{-1} \quad (c_3+c_4)^2 \ N_1 \ \Delta^2$. (2.5.2)

The expectations of the actual probabilities of misclassification are obtained easily as

$$E(Z_1^*) = G_1(0)$$

$$= \Pr\{\chi^2(p,\delta_1) - \chi^2(p,\delta_2) < 0\} \qquad (2.5.3)$$

$$E(Z_2^*) = 1 - G_2(0)$$

$$= \Pr\{\chi^2(p,\delta_3) - \chi^2(p,\delta_4) > 0\}$$

From the expressions (2.5.1) and (2.5.2) for the non-centralities in $E(Z_t^*)$ we note that

$$E(Z_2^* \mid (n_1,n_2)) = E(Z_1^* \mid (n_2,n_1)). \qquad (2.5.4)$$

Evaluation of $E(Z_t^*)$ and the unconditional distribution of Z(x) is considered in Sections 2.9 and 2.10.

2.6 The Expectations of the Apparent Probabilities of Misclassification of Z(x)

The expectations of $E(Z_t^{**})$ may be derived in a similar manner to $E(Z_t^{*})$ the difference being that the random observation x from Π_t is now replaced by a member x_{tj} of the sample of n_t from Π_t.

Hence for t = 1

$$E(Z_1^{**}) = \Pr\{-N_1(x_{1j}-\bar{x}_1)'\ \Sigma^{-1}(x_{1j}-\bar{x}_1) + N_2(x_{1j}-\bar{x}_2)'\ \Sigma^{-1}(x_{1j}-\bar{x}_2) < 0\}.$$

Allowing for the correlation of x_{1j} and $\bar{x}_1$ it follows that

$$E(Z_1^{**}) = \Pr\{(1+\rho_3)\ \chi^2(p,\delta_5) - (1-\rho_3)\ \chi^2(p,\delta_6) < 0\}$$

where $d_5 = \{n_1N_2 - n_2N_1 + 2n_2\sqrt{N_1N_2} + n_1n_2(\sqrt{N_1} - \sqrt{N_2})^2\}^{\frac{1}{2}}$

$$d_6 = \{n_1N_2 - n_2N_1 - 2n_2\sqrt{N_1N_2} + n_1n_2(\sqrt{N_1} + \sqrt{N_2})^2\}^{\frac{1}{2}}$$

$$c_5 = (n_1n_2)^{\frac{1}{2}}/d_5 \quad , \quad c_6 = (n_1n_2)^{\frac{1}{2}}/d_6$$

$$\rho_3 = \{n_1N_2 + n_2N_1 - n_1n_2\ (N_1-N_2)\}/d_5d_6 \qquad (2.6.1)$$

$$\delta_5 = \{2(1+\rho_3)\}^{-1} \quad (c_5+c_6)^2\ N_2\ \Delta^2$$

$$\delta_6 = 2(1-\rho_3)\}^{-1} \quad (c_5-c_6)^2\ N_2\ \Delta^2 .$$

The relationship

$$E(Z_2^{**} \mid (n_1,n_2)) = E(Z_1^{**} \mid (n_2,n_1)) \qquad (2.6.2)$$

holds here also, use of which will facilitate the evaluation of $E(Z_2^{**})$. The evaluation of $E(Z_t^{**})$ is considered in Sections 2.9 and 2.10.

2.7 Alternative expressions for $E(Z_t^*)$ and $E(Z_t^{**})$ when the dimension p = 1

As may be seen in Sections 2.5 and 2.6 both expectations $E(Z_t^*)$ and $E(Z_t^{**})$ involve the probability

$$Pr(uv < 0).$$

Now for p = 1, u and v are Normally distributed with variances 1 and means a and b respectively and

$$\begin{aligned} Pr(uv < 0) &= Pr(u < 0) + Pr(v < 0) - 2Pr(u < 0, v < o) \\ &= \Phi(-a) + \Phi(-b) - 2\Phi(-a, -b, \rho) \end{aligned}$$

where $\Phi(h, k, \rho)$ denotes the standard bivariate Normal distribution function with correlation ρ.

For $E(Z_1^*)$, Section 2.5 shows

$$a = c_1\sqrt{N_2}\, \Sigma^{-\frac{1}{2}} (\mu_1-\mu_2), \quad b = c_2\sqrt{N_2}\, \Sigma^{-\frac{1}{2}} (\mu_1-\mu_2), \quad \rho = 0$$

giving $E(Z_1^*) = \Phi(-a) + \Phi(-b) - 2\Phi(-a)\,\Phi(-b).$ (2.7.1)

For $E(Z_1^{**})$, Section 2.6 shows

$$a = c_5\sqrt{N_2}\, \Sigma^{-\frac{1}{2}} (\mu_1-\mu_2), \quad b = c_6\sqrt{N_2}\, \Sigma^{-\frac{1}{2}} (\mu_1-\mu_2), \quad \rho = \rho_3$$

giving $E(Z_1^{**}) = \Phi(-a) + \Phi(-b) - 2\Phi(-a, -b, \rho_3)$.

Similar expressions to these were obtained by John (1961) and Hills (1966) for $\hat{L}(x)$. Schaafsma and Van Vark (1977) have given the expression (2.7.1) for $E(Z_1^*)$.

2.8 Bounds and Inequalities for the Expectations of the Actual and Apparent Probabilities of Misclassification of $Z(x)$ and $\hat{L}(x)$.

As noted in Section 2.3 both John (1960a) and Das Gupta (1965) have shown that $Z(x)$ is an unbiased allocation rule i.e. $E(Z_t^*) < \frac{1}{2}$ for all n_1, n_2, p and $\Delta > 0$. This is a simple consequence of the following lemma which will be required elsewhere.

Lemma:- If x and y are two random variables independently distributed as $\chi^2(p,\lambda_1)$ and $\chi^2(p,\lambda_2)$ where $\lambda_1 > \lambda_2 > 0$ then

$$\Pr\{x-y \leqslant 0\} < \tfrac{1}{2} .$$

Proof:-
$$\Pr\{x-y \leqslant 0\} = \Pr\{x \leqslant y\}$$
$$= E_y[\Pr\{x \leqslant y \mid y\}].$$

Now
$$\Pr\{x \leqslant y \mid y\} < \Pr\{w \leqslant y \mid y\}$$

where w is distributed as $\chi^2(p,\lambda_2)$ independently of y. This follows as $\lambda_1 > \lambda_2$ and $\Pr\{\chi^2(p,\lambda) > c\}$ is a monotonic increasing function of λ, Ghosh (1973).

Hence
$$\Pr\{x-y \leqslant 0\} < \Pr\{w-y \leqslant 0\}$$
$$= \tfrac{1}{2}$$

as w and y are identically distributed.

Corollary:- With x and y as before, $\lambda_1 > \lambda_2 > 0$ and $\alpha \geqslant \beta > 0$ then $\Pr\{\alpha x - \beta y \leqslant 0\} < \frac{1}{2}$.

Proof:-
$$\Pr\{\alpha x - \beta y \leqslant 0\} = \Pr\{x \leqslant \beta/\alpha\, y\}$$
$$\leqslant \Pr\{x \leqslant y\}$$
$$< \tfrac{1}{2} \text{ by lemma.}$$

From (2.5.1) we see that $\delta_1 > \delta_2 > 0$ for all n_1, n_2 and $\Delta > 0$ and since

$$E(Z_1^*) = \Pr\{\chi^2(p,\delta_1) - \chi^2(p,\delta_2) < 0\}$$
$$< \tfrac{1}{2} \quad \text{by lemma,} \qquad (2.8.1)$$

from (2.5.2) $\delta_4 > \delta_3 > 0$ and from (2.5.3) and the lemma $E(Z_2^*) < \frac{1}{2}$ also.

In Section 2.6 we expressed $E(Z_1^{**})$ as the difference of two weighted non-central chi-squares. It is easily shown that the correlation ρ_3 (2.6.1) is positive for all n_1 and n_2 and that the non-centrality $\delta_5 > \delta_6$. By the corollary to the lemma and the relationship (2.6.2) it follows that

$$E(Z_t^{**}) < \tfrac{1}{2} \quad t = 1 \text{ and } 2, \quad \text{for all } n_1,\ n_2,\ p \text{ and } \Delta > 0.$$

In the evaluations of $E(Z_t^*)$ and $E(Z_t^{**})$ Section 2.10 it will be noted that $E(Z_1^*) < E(Z_2^*)$ and $E(Z_1^{**}) < E(Z_2^{**})$ when $n_1 < n_2$, for all p and $\Delta \neq 0$. Attempts to prove these inequalities algebraically failed except for $E(Z_t^*)$ when $p = 1$.

Using the alternative expressions in Section 2.7 for $E(Z_t^*)$ when $p = 1$ we have

$$E(Z_1^*) = \Phi(-a) + \Phi(-b) - 2\Phi(-a)\,\Phi(-b)$$
$$= \tfrac{1}{2} - 2\{\Phi(-a) - \tfrac{1}{2}\}\{\Phi(-b) - \tfrac{1}{2}\}$$

where $a = c_1\sqrt{N_2}\ \Sigma^{-\frac{1}{2}}(\mu_1-\mu_2)$, $b = c_2\sqrt{N_2}\ \Sigma^{-\frac{1}{2}}(\mu_1-\mu_2)$,
c_1 and c_2 positive.

Similarly $E(Z_2^*) = \frac{1}{2} - 2\{\Phi(-f) - \frac{1}{2}\}\{\Phi(-g) - \frac{1}{2}\}$

where $f = c_1\sqrt{N_1}\ \Sigma^{-\frac{1}{2}}(\mu_1-\mu_2)$, $g = c_2\sqrt{N_1}\ \Sigma^{-\frac{1}{2}}(\mu_1-\mu_2)$.

We wish to show that $E(Z_1^*) < E(Z_2^*)$ when $n_1 < n_2$, $\Delta \neq 0$.

If $n_1 < n_2$ then $N_1 < N_2$ and the following inequalities hold between a, b, f and g.

Case (i) with $\mu_1 > \mu_2$ then $0 < f < a$

and $0 < g < b$

Case (ii) with $\mu_1 < \mu_2$ then $0 > f > a$

and $0 > g > b$ (2.8.2)

For $E(Z_1^*) < E(Z_2^*)$ we need to show that

$$\{\Phi\ (-a) - \tfrac{1}{2}\}\{\Phi\ (-b) - \tfrac{1}{2}\} > \{\Phi\ (-f) - \tfrac{1}{2}\}\{\Phi\ (-g) - \tfrac{1}{2}\}$$

that is $AB > FG$

where $A = \{\Phi\ (-a) - \tfrac{1}{2}\}$ and B, F and G are similarly defined.

Case (i):- With $\mu_1 > \mu_2$, $A < 0$, $B < 0$, $F < 0$ and $G < 0$ from (2.8.2) and the function Φ. It also follows from the function Φ that

$$A < F \quad \text{and} \quad B < G.$$

Thus $-A > -F > 0$ and $-B > -G > 0$

and' $AB > FG$ as required.

Case (ii):- With $\mu_1 < \mu_2$, $A > 0$, $B > 0$, $F > 0$ and $G > 0$ from (2.8.2) and the function Φ, also, $A > F > 0$, $B > G > 0$ and so

$AB > FG$ as required.

Thus $E(Z_1^*) < E(Z_2^*)$ when $n_1 < n_2$, $p = 1$ and $\Delta \neq 0$, from (2.5.4) it follows that

$$E(Z_1^*) > E(Z_2^*) \text{ when } n_1 > n_2,\ p = 1 \text{ and } \Delta \neq 0.$$

As $Z(x)$ is an alternative allocation rule to $\hat{L}(x)$ when $n_1 \neq n_2$ we will have occasion to evaluate and compare $E(Z_t^*)$ and $E(L_t^*)$ for various values of $n_1 \neq n_2$, p and Δ. For Σ known

$$\hat{L}(x) = (\bar{x}_1-\bar{x}_2)' \, \Sigma^{-1} \, \{x - \tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\}$$

and the expectations $E(L_t^*)$ may be obtained in a similar manner to our derivation of $E(Z_t^*)$. From Moran (1974)

$$E(L_1^*) = \Pr\{(1+\rho_4)\, \chi^2(p,\delta_7) - (1-\rho_4)\, \chi^2(p,\delta_8) < 0\} \qquad (2.8.3)$$

where

$$d_7 = (n_1+n_2)^{\frac{1}{2}} \;, \qquad d_8 = (n_1+n_2+4n_1n_2)^{\frac{1}{2}}$$

$$c_7 = (n_1n_2)^{\frac{1}{2}}/d_7 \qquad c_8 = 2(n_1n_2)^{\frac{1}{2}}/d_8$$

$$\rho_4 = (n_1-n_2)/d_7d_8$$

$$\delta_7 = \{2(1+\rho_4)\}^{-1} \quad (c_7+c_8/2)^2\Delta^2$$

$$\delta_8 = \{2(1-\rho_4)\}^{-1} \quad (c_7-c_8/2)^2\Delta^2 \; .$$

The relationship

$$E(L_2^* \mid (n_1,n_2)) = E(L_1^* \mid (n_2,n_1)) \qquad (2.8.4)$$

holds here also. For $n_1 = n_2$ $Z(x)$ and $\hat{L}(x)$ are equivalent and it follows from (2.8.1) that

$$E(L_t^*) < \tfrac{1}{2} \quad t = 1 \text{ and } 2, \text{ if } n_1 = n_2, \text{ for all } p \text{ and } \Delta > 0.$$

It also follows from the lemma and its corollary that

$$E(L_1^*) < \tfrac{1}{2} \quad \text{if } n_1 > n_2 \text{ as } \rho_4 > 0 \text{ and } \delta_7 > \delta_8, \text{ for all } p \text{ and } \Delta > 0.$$

$$E(L_2^*) < \tfrac{1}{2} \quad \text{if } n_1 < n_2 \text{ from (2.8.4), for all } p \text{ and } \Delta > 0.$$

Moran's (1974) evaluation of $E(L_t^*)$ contains a case where $E(L_1^*) > \frac{1}{2}$ when $n_1 < n_2$, thus $\hat{L}(x)$ is an unbiased allocation rule if and only if $n_1 = n_2$. Moran also derived the expectations $E(L_t^{**})$. By use of the lemma and its corollary it is easily shown that

$$E(L_t^{**}) < \tfrac{1}{2}, \ t = 1 \text{ and } 2, \text{ for all } n_1, n_2, \ p \text{ and } \Delta > 0.$$

For Σ known from a result of Das Gupta (1974) on $\hat{L}(x)$

$$E(Z_t^{**}) < L_t < E(Z_t^*) \qquad t = 1 \text{ and } 2, \ n_1 = n_2, \text{ for all } p \text{ and } \Delta > 0,$$

where L_t is the optimal probability of misclassification, Section 1.4. However with $n_1 \neq n_2$ it may be seen from our evaluation of $E(Z_t^*)$ and $E(Z_t^{**})$ Tables 2.10.1 and 2.10.2 that

$$E(Z_t^*) \nmid L_t \text{ and } L_t \nmid E(Z_t^{**}), \text{ for all } n_1 \neq n_2, \ t = 1 \text{ and } 2.$$

With $\quad w_t(x) = (x-\bar{x}_t)' \, \Sigma^{-1} \, (x-\bar{x}_t) \qquad t = 1 \text{ and } 2$

$Z(x)$ and $\hat{L}(x)$ may be written as

$$Z(x) = -N_1 \, w_1(x) + N_2 \, w_2(x)$$

and $\quad \hat{L}(x) = -\tfrac{1}{2} \, w_1(x) + \tfrac{1}{2} \, w_2(x).$

The allocation rule $Z(x) \gtrless 0$ is equivalent to

$$w_1(x) \lessgtr \frac{N_2}{N_1} \, w_2(x)$$

and $\hat{L}(x) \gtrless 0$ is equivalent to

$$w_1(x) \lessgtr w_2(x).$$

Clearly if $n_1 < n_2$ then $N_2/N_1 > 1$ and $Z(x)$ allocates more observations to population Π_1 than does $\hat{L}$, hence

$$E(Z_1^*) < E(L_1^*) \quad \text{and} \quad E(Z_2^*) > E(L_2^*), \; n_1 < n_2.$$

Similarly if $n_1 > n_2$

$$E(Z_1^*) > E(L_1^*) \quad \text{and} \quad E(Z_2^*) < E(L_2^*).$$

2.9 Method of Evaluating $E(Z_t^*)$ and $E(Z_t^{**})$.

In Sections 2.5 and 2.6 the expectations $E(Z_t^*)$ and $E(Z_t^{**})$ were expressed as the weighted difference of two independent non-central chi-squares. This prompts consideration of ways of evaluating

$$\begin{aligned} F(x) &= \Pr\{Q \leqslant x\} \\ &= \Pr\{\sum_{i=1}^{n} \alpha_i \; \chi^2(\nu_i,\lambda_i) \leqslant x\} \end{aligned}$$

where the $\chi^2(\nu_i,\lambda_i)$ are independently distributed. If $\alpha_i > 0$ for all i, Q is called positive definite. If $\lambda_i = 0$ for all i, Q is described as a central quadratic form. Otherwise Q is called an indefinite non-central quadratic form.

Exact expressions for F(x) have been given by Shah (1963) in terms of Laguerre polynominals and by Press (1966) in terms of confluent hypergeometric functions, neither of these however allow easy evaluation. Imhof (1961), by inverting the characteristic function of Q, expressed F(x) in terms of an infinite integral which allowed numerical evaluation.

Other approaches to evaluating F(x) have been to use approximations to non-central chi-squares such as Patnaik's (1949) and Pearsons' (1959). The accuracy of these approximations are considered by Imhof (1961) and Solomon and Stephens (1977). Normal approximations were considered by Jensen and Solomon (1972) and the Pearson system of curves were used by Solomon and Stephens (1978). The overall conclusion of these papers is that Imhof's (1961) method gives the most accurate results.

Imhof's infinite integral expression for F(x) is

$$F(x) = \tfrac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin f(y)}{y\, g(y)}\, dy \qquad (2.9.1)$$

where

$$f(y) = \tfrac{1}{2} \sum_{i=1}^{n} \{\nu_i \arctan(\alpha_i y) + (\lambda_i \alpha_i y)/(1+\alpha_i^2 y^2)\} - \tfrac{1}{2}xy$$

$$g(y) = \{ \prod_{i=1}^{n} (1+\alpha_i^2 y^2)^{\frac{1}{4}\nu_i} \} \exp\{\tfrac{1}{2} \sum_{i=1}^{n} (\lambda_i \alpha_i^2 y^2)/(1+\alpha_i^2 y^2)\}$$

and $$\lim_{y\to 0} \frac{\sin f(y)}{y\, g(y)} = \tfrac{1}{2} \sum_{i=1}^{n} \alpha_i (\nu_i+\lambda_i) - \tfrac{1}{2}x.$$

The function y g(y) increases monotonically with y to +∞. If the integral (2.9.1) is evaluated from 0 to u, possible errors in evaluation are (i) round off errors (ii) the error of integration inherent in the numerical method used to evaluate I_u where

$$I_u = 1/\pi \int_0^u \frac{\sin f(y)}{y\, g(y)}\, dy,$$

and (iii) the truncation error T_u where

$$T_u = 1/\pi \int_u^\infty \frac{\sin f(y)}{y\, g(y)}\, dy.$$

Imhof bounded T_u as follows

$$T_u \leqslant \left[\pi k u^k \prod_{i=1}^{n} |\alpha_i|^{\frac{1}{2}\nu_i} \exp \{ \tfrac{1}{2} \sum_{i=1}^{n} \lambda_i \alpha_i^2 u^2/(1+\alpha_i^2 u^2) \} \right]^{-1}$$

(2.9.2)

where $k = \frac{1}{2} \sum_{i=1}^{n} \nu_i$.

This bound (2.9.2) on T_u may be used to obtain the upper limit of integration u of I_u for a predetermined truncation error. Having done so, the problem of evaluating I_u must then be considered. Two numerical methods, namely a composite approach with 40 point Gaussian quadrature and a trapezoidal rule with Romberg extrapolation, Ralston (1965, pp121-129), were tried. The latter was chosen for routine use given its guaranteed convergence. When it fails to attain a prescribed tolerence level an examination of intermediate results indicates the accuracy obtained.

The accuracy of the methods was assessed by comparing the computed values with known exact values given in the literature. Those used were Imhof (1961), Jensen and Solomon (1972), Tiku's (1970) double non-central F tables and Moran's (1974) evaluation of $E(L_t^*)$ and $E(L_t^{**})$. The trapezoidal rule with Romberg extrapolation was the more accurate and convenient of the two methods. All computations were carried out in double precision. Any inaccuracy noted was in the fifth decimal place and was of the order 10^{-5}.

2.10 Evaluation of the Expectations $E(Z_t^*)$ and $E(Z_t^{**})$.

Imhof's (1961) method was used to evaluate the unconditional distributions $G_t(a)$ of Section 2.5. The expectations $E(Z_t^*)$ and $E(Z_t^{**})$ are evaluated for a zero cut-off point. This is not necessary for applying Imhof's method, but was necessary for Moran's (1974) evaluation of $E(L_t^*)$ and $E(L_t^{**})$. By (2.8.3)

$$\begin{aligned} E(L_1^*) &= \Pr\{(1+\rho_4)\,\chi^2(p,\delta_7) - (1-\rho_4)\,\chi^2(p,\delta_8) < 0\,\} \\ &= \Pr\{\, \frac{\chi^2(p,\delta_7)}{\chi^2(p,\delta_8)} < \frac{(1-\rho_4)}{(1+\rho_4)} \,\} \\ &= \Pr\{\, F(p,p,\delta_7,\delta_8) < \frac{1-\rho_4}{1+\rho_4} \,\} \end{aligned}$$

where $F(\nu_1,\nu_2,\lambda_1,\lambda_2)$ is a double non-central F variate with degrees of freedom ν_1 and ν_2 and non-centralities λ_1 and λ_2. Moran using Price's (1964) finite term expressions for $F(\nu_1,\nu_2,\lambda_1,\lambda_2)$ was able to evaluate $E(L_t^*)$ and $E(L_t^{**})$ with the dimension p even.

In Tables 2.10.1 and 2.10.2 the expectations $E(Z_1^*)$, $E(Z_1^{**})$ and $E(L_1^*)$ and the averages, $\bar{E}(Z^*) = \frac{1}{2}\{E(Z_1^*) + E(Z_2^*)\}$ with $\bar{E}(Z^{**})$ and $\bar{E}(L^*)$ similarly defined, are given for a range of values of p, Δ and (n_1,n_2). The emphasis is on cases where $n_1 \neq n_2$ since for equal sample sizes the allocation rules $\hat{Z}(x)$ and $\tilde{L}(x)$ are identical and the results of Moran (1974) may be consulted. Results for x in Π_2 may be obtained from the relationship

$$E(Z_2^* \mid (n_1,n_2)) = E(Z_1^* \mid (n_2,n_1)) \qquad (2.10.1)$$

which holds for $E(Z_t^{**})$ and $E(L_t^*)$ too. Univariate results were checked by use of the results given in Section 2.7.

The imbalance in the expectations of the actual probabilities of misclassification is measured by

$$\text{DIF } Z^* = \{E(Z_1^*) - E(Z_2^*)\} \times 10^4$$

with DIF Z^{**}, DIF L^* and DIF L^{**} similarly defined. It follows from (2.10.1) that

$$\text{DIF } Z^* (n_1, n_2) = - \text{DIF } Z^{**} (n_2, n_1).$$

These differences are given in Tables 2.10.3 and 2.10.4, with DIF L^{**} calculated from Moran's (1974) evaluation of $E(L_t^{**})$.

Table 2.10.1

Expectations of the actual and apparent probabilities of misclassification of the Z and sample linear allocation rules for moderately unequal sample sizes (Σ known).

		(n_1 = 16, n_2 = 24)			(n_1 = 24, n_2 = 16)					
	p	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$\bar{E}(Z^*)$	$\bar{E}(L^*)$	$\bar{E}(Z^{**})$
L_1 = 40% Δ = 0.5067	1	.4121	.4156	.3916	.4133	.4099	.3993	.4127	.4127	.3955
	2	.4248	.4346	.3663	.4259	.4161	.3859	.4254	.4254	.3761
	4	.4385	.4571	.3290	.4395	.4210	.3641	.4390	.4390	.3466
	8	.4518	.4824	.2767	.4527	.4223	.3306	.4523	.4524	.3037
	16	.4636	.5104	.2083	.4643	.4180	.2817	.4640	.4642	.2450
	32	.4732	.5421	.1293	.4738	.4057	.2163	.4735	.4739	.1728
L_1 = 30% Δ = 1.0488	1	.3017	.3026	.2956	.3035	.3025	.2995	.3026	.3026	.2976
	2	.3110	.3157	.2833	.3129	.3083	.2939	.3120	.3120	.2886
	4	.3256	.3367	.2616	.3276	.3166	.2833	.3266	.3267	.2725
	8	.3461	.3678	.2264	.3481	.3268	.2640	.3471	.3473	.2452
	16	.3712	.4092	.1751	.3731	.3363	.2314	.3722	.3727	.2033
	32	.3977	.4592	.1112	.3994	.3403	.1825	.3986	.3998	.1469
L_1 = 20% Δ = 1.6832	1	.2019	.2030	.1945	.2042	.2030	.1993	.2031	.2030	.1969
	2	.2064	.2094	.1885	.2088	.2050	.1965	.2076	.2076	.1925
	4	.2147	.2213	.1772	.2173	.2108	.1912	.2160	.2160	.1842
	8	.2294	.2428	.1573	.2321	.2190	.1809	.2308	.2309	.1691
	16	.2527	.2788	.1254	.2556	.2308	.1624	.2542	.2548	.1439
	32	.2853	.3329	.0822	.2883	.2440	.1320	.2868	.2885	.1071

Table 2.10.1 (continued)

		$(n_1 = 16, n_2 = 24)$			$(n_1 = 24, n_2 = 16)$					
	p	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$\bar{E}(Z^*)$	$\bar{E}(L^*)$	$\bar{E}(Z^{**})$
L_1 = 10% Δ = 2.5631	1	.1018	.1029	.0948	.1040	.1029	.0993	.1029	.1029	.0971
	2	.1036	.1055	.0924	.1059	.1040	.0982	.1048	.1048	.0953
	4	.1o72	.1106	.0878	.1096	.1062	.0960	.1084	.1084	.0919
	8	.1141	.1208	.0793	.1167	.1102	.0917	.1154	.1155	.0855
	16	.1270	.1403	.0650	.1299	.1178	.0837	.1284	.1288	.0744
	32	.1496	.1765	.0442	.1528	.1284	.0699	.1512	.1525	.0571
L_1 = 5% Δ = 3.2897	1	.0513	.0522	.0461	.0531	.0522	.0495	.0522	.0522	.0478
	2	.0522	.0534	.0450	.0539	.0527	.0490	.0531	.0531	.0470
	4	.0539	.0558	.0430	.0557	.0537	.0480	.0548	.0548	.0455
	8	.0572	.0607	.0391	.0591	.0557	.0460	.0582	.0582	.0426
	16	.0637	.0706	.0325	.0659	.0594	.0423	.0648	.0650	.0374
	32	.0763	.0908	.0226	.0789	.0659	.0358	.0776	.0784	.0292
L_1 = 1% Δ = 4.6526	1	.0105	.0108	.0086	.0111	.0108	.0098	.0108	.0108	.0092
	2	.0107	.0110	.0084	.0113	.0109	.0097	.0110	.0110	.0091
	4	.0110	.0115	.0081	.0117	.0111	.0095	.0114	.0113	.0088
	8	.0116	.0125	.0074	.0123	.0115	.0092	.0120	.0120	.0083
	16	.0129	.0145	.0063	.0137	.0123	.0085	.0133	.0134	.0074
	32	.0157	.0190	.0045	.0167	.0138	.0073	.0162	.0164	.0059

Table 2.10.2

Expectations of the actual and apparent probabilities of misclassification of the Z and sample linear allocation rules for very unequal sample sizes (Σ known).

		$(n_1 = 8,\ n_2 = 32)$			$(n_1 = 32,\ n_2 = 8)$					
	p	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$\bar{E}(Z^*)$	$\bar{E}(L^*)$	$\bar{E}(Z^{**})$
L_1 = 40% Δ = 0.5067	1	.4185	.4369	.3712	.4238	.4054	.4106	.4212	.4212	.3909
	2	.4317	.4739	.3244	.4364	.3949	.4065	.4341	.4344	.3655
	4	.4453	.5196	.2619	.4493	.3768	.3970	.4473	.4482	.3295
	8	.4580	.5747	.1848	.4612	.3477	.3789	.4596	.4612	.2819
	16	.4687	.6411	.1028	.4712	.3042	.3490	.4700	.4726	.2259
	32	.4772	.7202	.0369	.4790	.2440	.3042	.4781	.4821	.1706
L_1 = 30% Δ = 1.0488	1	.3008	.3059	.2866	.3090	.3040	.3054	.3049	.3050	.2960
	2	.3142	.3363	.2591	.3230	.3015	.3075	.3186	.3189	.2833
	4	.3332	.3831	.2164	.3423	.2949	.3080	.3378	.3390	.2622
	8	.3572	.4499	.1575	.3661	.2804	.3036	.3617	.3652	.2306
	16	.3840	.5382	.0898	.3920	.2531	.2897	.3880	.3957	.1898
	32	.4101	.6452	.0328	.4168	.2090	.2610	.4135	.4271	.1469
L_1 = 20% Δ = 1.6832	1	.1993	.2045	.1844	.2095	.2045	.2059	.2044	.2045	.1952
	2	.2060	.2197	.1712	.2170	.2033	.2070	.2115	.2115	.1891
	4	.2178	.2482	.1482	.2293	.2006	.2084	.2236	.2244	.1783
	8	.2374	.2995	.1125	.2498	.1942	.2091	.2436	.2468	.1608
	16	.2664	.3844	.0669	.2793	.1805	.2059	.2729	.2825	.1364
	32	.3032	.5093	.0254	.3157	.1545	.1934	.3095	.3319	.1094

Table 2.10.2 (continued)

		$(n_1 = 8, n_2 = 32)$			$(n_1 = 32, n_2 = 8)$					
	p	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$E(Z_1^*)$	$E(L_1^*)$	$E(Z_1^{**})$	$\bar{E}(Z^*)$	$\bar{E}(L^*)$	$\bar{E}(Z^{**})$
L_1 = 10% Δ = 2.5631	1	.0994	.1043	.0856	.1093	.1043	.1057	.1044	.1043	.0957
	2	.1020	.1105	.0806	.1122	.1038	.1061	.1071	.1072	.0934
	4	.1072	.1231	.0714	.1180	.1027	.1068	.1126	.1129	.0891
	8	.1169	.1487	.0563	.1287	.1004	.1078	.1228	.1245	.0821
	16	.1345	.2005	.0353	.1479	.0952	.1084	.1412	.1479	.0719
	32	.1637	.3007	.0142	.1788	.0845	.1062	.1713	.1926	.0602
L_1 = 5% Δ = 3.2897	1	.0495	.0533	.0395	.0571	.0533	.0543	.0533	.0533	.0469
	2	.0507	.0562	.0374	.0585	.0531	.0545	.0546	.0546	.0460
	4	.0531	.0621	.0335	.0613	.0526	.0549	.0572	.0573	.0442
	8	.0578	.0747	.0269	.0668	.0515	.0555	.0623	.0631	.0412
	16	.0670	.1025	.0173	.0774	.0493	.0562	.0722	.0759	.0368
	32	.0844	1052	.0073	.0969	.0445	.0562	.0907	.1049	.0318
L_1 = 1% Δ = 4.6526	1	.0098	.0112	.0065	.0128	.0112	.0116	.0113	.0112	.0091
	2	.0100	.0118	.0062	.0130	.0112	.0117	.0115	.0115	.0090
	4	.0105	.0129	.0056	.0136	.0111	.0118	.0121	.0120	.0080
	8	.0114	.0155	.0046	.0148	.0109	.0119	.0131	.0132	.0083
	16	.0132	.0215	.0030	.0172	.0105	.0121	.0152	.0160	.0076
	32	.0172	.0374	.0013	.0223	.0097	.0124	.0198	.0236	.0069

Table 2.10.3

The difference in the expectations of the actual and apparent probabilities of misclassification of the Z and sample linear allocation rules for moderately unequal sample sizes (Σ known).

$$\text{DIF } Z^* = \{E(Z_1^*) - E(Z_2^*)\} \times 10^4$$

$(n_1 = 16, n_2 = 24)$

	p	DIF Z^*	DIF L^*	DIF Z^{**}	DIF L^{**}		p	DIF Z^*	DIF L^*	DIF Z^{**}	DIF L^{**}
$L_1 = 40\%$ $\Delta = 0.5067$	1	- 12	57	- 77	- 11	$L_1 = 10\%$ $\Delta = 2.5631$	1	- 32	0	- 45	- 24
	2	- 11	185	- 196	- 13		2	- 23	15	- 58	- 23
	4	- 10	361	- 351	- 16		4	- 24	44	- 82	- 24
	8	- 9	601	- 539	- 19		8	- 26	106	- 124	- 23
	16	- 7	924	- 734	- 24		16	- 30	230	- 187	- 21
	32	- 6	1364	- 870	- 26		32	- 32	481	- 257	- 19
$L_1 = 30\%$ $\Delta = 1.0488$	1	- 18	21	- 39	- 19	$L_1 = 5\%$ $\Delta = 3.2897$	1	- 18	0	- 34	- 18
	2	- 19	74	- 106	- 21		2	- 17	7	- 40	- 18
	4	- 20	201	- 217	- 21		4	- 18	21	- 50	- 17
	8	- 20	410	- 376	- 23		8	- 19	50	- 69	- 17
	16	- 19	729	- 563	- 25		16	- 22	112	- 98	- 15
	32	- 17	1189	- 713	- 26		32	- 26	249	- 132	- 14
$L_1 = 20\%$ $\Delta = 1.6832$	1	- 23	0	- 48	- 25	$L_1 = 1\%$ $\Delta = 4.6526$	1	- 6	0	- 12	- 6
	2	- 24	36	- 80	- 25		2	- 6	1	- 13	- 6
	4	- 26	105	- 140	- 25		4	- 7	4	- 14	- 6
	8	- 27	238	- 236	- 26		8	- 7	10	- 18	- 6
	16	- 29	480	- 370	- 26		16	- 8	22	- 22	- 6
	32	- 30	889	- 498	- 25		32	- 10	52	- 28	- 4

Table 2.10.4

The difference of the expectations of the actual and apparent probabilities of misclassification of the Z and sample linear allocation rules for very unequal sample sizes (Σ known).

$$\text{DIF } Z^* = \{E(Z_1^*) - E(Z_2^*)\} \times 10^4$$

$(n_1 = 8,\ n_2 = 32)$

	p	DIF Z^*	DIF L^*	DIF Z^{**}	DIF L^{**}		p	DIF Z^*	DIF L^*	DIF Z^{**}	DIF L^{**}
L_1 = 40% Δ = 0.5067	1	-53	315	-394	-52	L_1 = 10% Δ = 2.5631	1	-99	0	-201	-108
	2	-47	790	-821	-65		2	-102	67	-255	-108
	4	-40	1428	-1351	-81		4	-108	204	-354	-106
	8	-32	2270	-1941	-100		8	-118	483	-515	-101
	16	-25	3369	-2472	-118		16	-134	1053	-731	-93
	32	-18	4761	-2673	-118		32	-151	2162	-920	-76
L_1 = 30% Δ = 1.0488	1	-82	19	-118	-90	L_1 = 5% Δ = 3.2897	1	-76	0	-148	-80
	2	-88	348	-484	-95		2	-78	31	-171	-79
	4	-91	882	-961	-103		4	-82	95	-214	-77
	8	-89	1695	-1461	-112		8	-90	232	-286	-73
	16	-80	2851	-1991	-120		16	-104	532	-389	-64
	32	-67	4362	-2282	-113		32	-125	1207	-489	-51
L_1 = 20% Δ = 1.6832	1	-102	0	-215	-25	L_1 = 1% Δ = 4.6526	1	-30	0	-51	-27
	2	-110	164	-358	-117		2	-30	4	-55	-27
	4	-115	476	-602	-118		4	-31	18	-62	-27
	8	-124	1053	-966	-119		8	-34	46	-73	-24
	16	-129	2039	-1390	-116		16	-40	110	-91	-21
	32	-125	3548	-1680	-102		32	-51	277	-111	-16

2.11 Discussion of the Results of Tables 2.10.1, 2.10.2, 2.10.3 and 2.10.4.

Considering first the individual expectations of the actual probabilities of misclassification. It is noted that $E(Z_1^*)$ increases with increasing dimension p for all (n_1,n_2) and Δ. The same is not true of $E(L_1^*)$ where for very unequal sample sizes, Table 2.10.2, it decreases with increasing p when the larger sample is from population one. A consequence of this is that in some cases $E(L_1^*)$ is less than the optimal probability of misclassification. When the larger sample is from population two, $E(L_1^*)$ in some cases exceeds $\frac{1}{2}$ and $\hat{L}(x)$ is not an unbiased allocation rule. The expectation $E(Z_1^*)$ in general exceeds the optimal probability of misclassification and is as shown in Section 2.8 always less than $\frac{1}{2}$. The difference between individual $E(Z_1^*)$ and $E(L_1^*)$ becomes more emphatic with increasing unequal sample sizes; being in some cases of the order 1 : 2.

The average expectations $\bar{E}(Z^*)$ and $\bar{E}(L^*)$ are close for all (n_1,n_2) and Δ, with increasing p, $\bar{E}(Z^*) < \bar{E}(L^*)$. Naturally both exceed the optimal error rate L_1. Our results here for Σ known agree with those of Siotani and Wang (1977) who compared the allocation rules $\hat{L}(x)$ and $Z(x)$ for Σ unknown by considering the difference $\bar{E}(L^*) - \bar{E}(Z^*)$, derived from asymptotic approximations, for various values of $n_1 \neq n_2$, p and Δ.

The imbalance between actual expectations for the Z statistic and the sample linear discriminant function may be judged from the differences DIF Z^* and DIF L^*, Tables 2.10.3 and 2.10.4. Both imbalances become more pronounced with increasingly unequal sample sizes. For fixed Δ, DIF L^* increases with increasing p while DIF Z^* tends to remain constant. With fixed p and increasing Δ, DIF L^* decreases but DIF Z^* increases and then decreases. The robustness of the Z statistic to unequal sample sizes as measured by $E(Z_t^*)$ is confirmed by these results given that the range of differences for Z(x) is .001 to .02 whereas for $\hat{L}(x)$ it is 0.0 to .5. It is also noted for $p > 1$ that $E(Z_1^*) < E(Z_2^*)$ if $n_1 < n_2$ and $\Delta > 0$, the opposite being true of $E(L_1^*)$ and $E(L_2^*)$. This inequality in $E(Z_t^*)$ was proved analytically in Section 2.7 for $p = 1$.

For the expectations of the apparent probabilities of misclassification of the Z statistic, Tables 2.10.1 and 2.10.2, we note that the individual $E(Z_1^{**})$ decrease with increasing p for fixed Δ. With very unequal samples $E(Z_1^{**})$ exceeds the optimal error rate L_1 when the larger sample is from population one. This is not so for $E(L_1^{**})$ as may be seen in Moran (1974). However the average expectation $\bar{E}(Z^{**})$ is less than L_1 for all (n_1,n_2), p and Δ. It is also noted that $E(Z_1^{**}) < E(Z_1^*)$ for all (n_1,n_2), p and Δ.

The imbalance in $E(Z_t^{**})$ is considerable as may be seen from DIF Z^{**}, Tables 2.10.3 and 2.10.4. It is in the same direction as DIF Z^* but at least twice its size. This implies that use of the apparent error rates will indicate an imbalance in the actual probabilities of misclassification of Z(x) which does not exist. Conversely DIF L^{**} is quite small regardless of inequalities in n_1 and n_2, while as previously noted $E(L_t^*)$ displays considerable inequality when $n_1 \neq n_2$.

Overall the Z statistic, while equating the expectations of the actual probabilities of misclassification for unequal sample sizes, imbalances the expectations of the apparent probabilities, the converse being true of the sample linear discriminant function. This behaviour of $\hat{L}(x)$ when $n_1 \neq n_2$ does not appear to be well known. Siotani and Wang (1977) who compared $\hat{L}(x)$ and $Z(x)$ when $n_1 \neq n_2$ failed to note the effect of unequal sample sizes on the individual expectations. This may be due to the fact that their comparison was based on the difference of the average expectations $\bar{E}(L^*)$ and $\bar{E}(Z^*)$ rather than the individual expectations. Our results indicate that for Σ known, $n_1 \neq n_2$ and cut off zero, $\hat{L}(x)$ may be an inappropriate discriminant function for multinormally distributed populations. On average the actual probabilities of misclassification L_t^* unlike the optimal probabilities L_t will be unequal and the resubstitution method will fail to indicate this. As the actual probabilities of misclassification of the Z statistic are nearly equal even when $n_1 \neq n_2$, the Z statistic is recommended as the better allocation rule. Here however the resubstitution method will indicate an imbalance in the actual probabilities of Z(x) which does not exist.

2.12 Approximations to the Expectations of the Actual Probabilities of Misclassification of $E(Z_t^*)$

Three approximations to $E(Z_t^*)$ suggested in the literature namely, Memon and Okamoto's (1971) asymptotic expansion, a Normal approximation similar to Lachenbruch's (1968) approximation of $E(L_t^*)$ and an approximate method of evaluating non-central chi-squares suggested by John (1963), are considered here. The latter approximation requires that Σ be known the others do not.

Memon and Okamoto (1971) for Σ unknown gave an asymptotic expansion of $E(Z_t^*)$ as a function of n_1, n_2, p and Δ. This was obtained by taking a Taylor series expansion of the characteristic function of Z(x) conditional on $\bar{x}_1$, $\bar{x}_2$ and S, obtaining its unconditional expectation and inverting it. The resulting expansion for Σ known is

$$\tilde{E}_{mo}(Z_1^*) = \Phi(-\tfrac{1}{2}\Delta) + a_1/n_1 + a_2/n_2 + a_3/(n_1+n_2-2)$$
$$+ \text{ terms of order } n_1^{-2},\ n_2^{-2},\ (n_1n_2)^{-1} \text{ and } (n_1+n_2-2)^{-2},$$

where
$$a_1 = (2\Delta^2)^{-1}\{-d_o^4 + (p-4)\,d_o^2\}$$
$$a_2 = (2\Delta^2)^{-1}\{3d_o^4 + (p+8)\,d_o^2\} \qquad (2.12.1)$$
$$a_3 = \tfrac{1}{4}(p-1)\,d_o^2$$

and
$$d_o^i = (d^i/dy^i)\,\Phi(y)\,\Big|_{y=-\frac{1}{2}\Delta} \qquad i = 2,4.$$

After some simplification the expansion to order n_1^{-1} and n_2^{-1} is

$$\tilde{E}_{mo}(Z_1^*) = \Phi(-\tfrac{1}{2}\Delta) + \phi(-\tfrac{1}{2}\Delta)\ \{\tfrac{1}{4}(p-1)/\Delta - \tfrac{1}{16}\Delta\}\ /\ n_1$$

$$+ \phi(-\tfrac{1}{2}\Delta)\ \{\tfrac{1}{4}(p-1)/\Delta + \tfrac{3}{16}\Delta\}\ /\ n_2 \qquad (2.12.2)$$

where $\phi(y) = (d/dy)\ \Phi(y)$.

The usual relationship

$$\tilde{E}_{mo}(Z_2^* \mid (n_1,n_2)) = \tilde{E}_{mo}(Z_1^* \mid (n_2,n_1))$$

holds here. It is also interesting to note that the inequality $\tilde{E}_{mo}(Z_1^*) < \tilde{E}_{mo}(Z_2^*)$ if $n_1 < n_2$ holds for all p and $\Delta > 0$. This property was noted to hold for $E(Z_t^*)$ in the evaluations of Section 2.10.

The second approximation is similar in spirit to Lachenbruch's (1968) approximation of $E(L_t^*)$. Here we approximate the unconditional distribution of Z(x) by a Normal variable with mean and variance that of Z(x) when $\bar{x}_1$ and $\bar{x}_2$ vary and Σ is assumed known. The unconditional mean and variance of Z(x) may be obtained from its distribution in Section 2.5. With δ_1 and δ_2 as defined in (2.5.1), $E(Z_1^*)$ is approximated by

$$\tilde{E}_{nor}(Z_1^*) = \Phi\left(\frac{-\tfrac{1}{2}\ (\delta_1-\delta_2)}{\sqrt{p+\delta_1+\delta_2}}\right) .$$

The relationship

$$\tilde{E}_{nor}(Z_2^* \mid (n_1,n_2)) = \tilde{E}_{nor}(Z_1^* \mid (n_2,n_1))$$

holds here also. In his approximation of $E(L_1^*)$ Lachenbruch (1968) did not use the exact variance of $\hat{L}(x)$ but the expectation of its conditional variance, use of the exact variance results in a better approximation. Moran's (1974) evaluations of $\tilde{E}_{nor}(L_1^*)$ do not as a result correspond to $\tilde{E}_{nor}(Z_1^*)$ when $n_1 = n_2$.

The final approximation is that suggested by John (1963) who proposed use of Abdel-Aty's (1954) approximation to a non-central chi-square in evaluating $E(Z_t^*)$. Abdel-Aty's result is that $\{\chi^2(p,\lambda) / (p+\lambda)\}^{\frac{1}{3}}$ is approximately Normally distributed with mean $1-\{2(p+2\lambda) / 9(p+\lambda)^2\}$ and variance $= 1 -$ mean. Hence

$$\tilde{E}_j(Z_1^*) = \Phi\left(\frac{-am_1+bm_2}{\sqrt{a^2\sigma_1^2+b^2\sigma_2^2}}\right)$$

where $a = (p+\delta_1)^{\frac{1}{3}}$, $m_1 = 1-\{2(p+2\delta_1) / 9(p+\delta_1)^2\}$, $\sigma_1^2 = 1-m_1$

$b = (p+\delta_2)^{\frac{1}{3}}$, $m_2 = 1-\{2(p+2\delta_2) / 9(p+\delta_2)^2\}$, $\sigma_2^2 = 1-m_2$.

The relationship

$$\tilde{E}_j(Z_2^* \mid (n_1,n_2)) = \tilde{E}_j(Z_1^* \mid (n_2,n_1))$$

holds again. As noted this approximation specifically requires that Σ be known, since the unconditional distribution of $Z(x)$ is for this assumption a difference of two independent non-central chi-squares. A discussion of the accuracy of Abdel-Aty's approximation may be found in Johnson and Kotz (1970, vol 2, pp141-142). Although John proposed the above approximation and commented that it would be interesting to compare it and the exact values, he did not pursue the matter.

In Table 2.12.1 the difference between the exact and approximate values of $E(Z_t^*)$ are given for the three approximations listed above. The range of parameters, (n_1,n_2), p and Δ considered,is similar to that of Section 2.10. Additional results for equal sample sizes may be found in Moran (1974) where the above approximations to $E(L_t^*)$ were also considered.

Table 2.12.1

Errors in approximations for the expectation of the actual probabilities of misclassification of Z(x) (Σ known).

M & O = Memon and Okamoto, Nor = Normal, J = John

(True-approx.) x 10^4

		$(n_1 = 16, n_2 = 24)$				$(n_1 = 24, n_2 = 16)$			
	p	True	M & O	Nor	J	True	M & O	Nor	J
$L_1 = 30\%$ $\Delta = 1.0488$	4	.3256	-17	-45	38	.3276	-16	-45	37
	8	.3461	-158	-20	12	.3481	-157	-20	11
	16	.3712	-597	-5	2	.3731	-597	-9	2
	32	.3977	-1714	0	0	.3994	-1716	0	0
$L_1 = 20\%$ $\Delta = 1.6832$	4	.2147	-1	-55	49	.2173	0	-54	59
	8	.2294	-28	-40	29	.2321	-25	-40	29
	16	.2527	-141	-24	11	.2556	-137	-23	11
	32	.2853	-508	-10	3	.2883	-503	-9	3
$L_1 = 10\%$ $\Delta = 2.5631$	4	.1072	1	-42	27	.1096	1	-43	28
	8	.1141	1	-39	22	.1167	1	-38	23
	16	.1270	-15	-32	15	.1299	-8	-30	16
	32	.1496	-74	-22	9	.1528	-66	-22	8
$L_1 = 5\%$ $\Delta = 3.2897$	4	.0539	1	-29	12	.0557	3	-29	13
	8	.0572	2	-28	11	.0591	3	-29	12
	16	.0637	1	-26	10	.0659	6	-26	10
	32	.0763	-3	-22	7	.0789	5	-22	8

Table 2.12.1 (continued)

$(n_1 = 8, n_2 = 32)$ (True-approx.) x 10^4 $(n_1 = 32, n_2 = 8)$

	p	True	M & O	Nor	J	True	M & O	Nor	J
$L_1 = 30\%$ $\Delta = 1.0488$	4	.3332	-49	-43	34	.3423	-44	-40	30
	8	.3572	-327	-15	8	.3661	-324	-13	7
	16	.3840	-1095	-1	1	.3920	-1101	-1	0
	32	.4101	-2906	2	0	.4168	-2925	2	0
$L_1 = 20\%$ $\Delta = 1.6832$	4	.2178	-8	-70	54	.2293	-3	-69	54
	8	.2374	-72	-46	27	.2498	-58	-43	25
	16	.2664	-301	-22	9	.2793	-283	-19	8
	32	.3032	-973	-7	1	.3157	-958	-6	1
$L_1 = 10\%$ $\Delta = 2.5631$	4	.1072	1	-58	36	.1180	3	-61	39
	8	.1169	-9	-51	27	.1287	3	-52	28
	16	.1346	-47	-39	16	.1479	-19	-38	17
	32	.1637	-183	-24	7	.1788	-138	-22	7
$L_1 = 5\%$ $\Delta = 3.2897$	4	.0531	1	-41	16	.0613	3	-45	19
	8	.0578	-1	-39	15	.0668	9	-41	17
	16	.0670	-7	-35	12	.0774	17	-36	14
	32	.0844	-29	-27	9	.0969	16	-27	9

As was to be expected from formula (2.12.2) Memon and Okamoto's approximation deteriorates with increasing size p and decreasing Δ. It tends in general to be an overestimate of the true value. It should be noted that the inclusion of extra terms in the expansion may alter this overestimation to underestimation and improve the performance of the approximation. The Normal approximation is fairly insensitive to increasing Δ and improves with increasing p. It is also an overestimate of the true value. John's approximation is on the other hand an underestimate of the true value. It is fairly insensitive to increasing Δ and improves with increasing p. For p = 32 the error of approximation is within .001 of the true value.

All approximations reflect the robustness of the Z statistic to unequal sample sizes in that the magnitude of the errors for both populations are the same. The consistency of the Normal approximation has much to recommend it. Eventhough, the best approximation is John's, it is as noted specific to Σ known, whereas the Normal approximation may be applied in general.

Use of the asymptotic approximation of Memon and Okamoto (1971) for $E(Z_t^*)$ and the corresponding approximation of Okamoto (1963) for $E(L_t^*)$ gives a simple guide as to when the difference between the actual probabilities of misclassification of $Z(x)$ will exceed that of $\hat{L}(x)$ i.e.

$$| \text{DIF } Z^* | > | \text{DIF } L^* |$$

where DIF $Z^* = E(Z_1^*) - E(Z_2^*)$, with DIF L^* similarly defined.

From Okamoto (1963)

$$\tilde{E}_o(L_1^*) = \Phi(-\tfrac{1}{2}\Delta) + \phi(-\tfrac{1}{2}\Delta) \{\tfrac{3}{4}(p-1)/\Delta + \tfrac{1}{16}\Delta \}/n_1$$

$$+ \phi(-\tfrac{1}{2}\Delta) \{-\tfrac{1}{4}(p-1)/\Delta + \tfrac{1}{16}\Delta \}/n_2$$

$$+ a_3 / (n_1+n_2-2)$$

with a_3 as given in the corresponding expansion of $E(Z_1^*)$, (2.12.1). Also $\tilde{E}_o(L_2^* \mid (n_1,n_2)) = \tilde{E}_o(L_1^* \mid (n_2,n_1))$.

Hence $\text{DIF } L^* \simeq \phi(-\tfrac{1}{2}\Delta)(\frac{1}{n_1} - \frac{1}{n_2})\ (p-1)/\Delta.$ (2.12.3)

and from (2.12.1) and (2.12.2)

$$\text{DIF } Z^* \simeq \phi(-\tfrac{1}{2}\Delta)(\frac{1}{n_1} - \frac{1}{n_2})\ (-\tfrac{1}{4}\Delta). \tag{2.12.4}$$

We note immediately from (2.12.3) that if $n_1 < n_2, \tilde{E}_o(L_1^*) > \tilde{E}_o(L_2^*)$ as is the case in Tables 2.10.1 and 2.10.2 for $E(L_1^*)$ and $E(L_2^*)$. Also, that the approximated DIF L^* increases as p increases and the approximated DIF Z^* is independent of p, similar behaviour was noted in the exact differences, Tables 2.10.3 and 2.10.4. From (2.12.3) and (2.12.4)

$$|\text{ DIF } Z^* | > | \text{ DIF } L^* | \text{ when } \Delta > 2\sqrt{p-1} .$$

The results of Tables 2.10.3 and 2.10.4 indicate that for p > 1 this is a reasonable guide. More refined and complicated guides may be derived from the other approximations considered here.

2.13 $E(Z_t^*)$ when the Covariance Matrix Σ is Unknown

With Σ unknown the Z statistic is given by

$$Z(x) = -N_1 \ (x-\bar{x}_1)' \ S^{-1}(x-\bar{x}_1) + N_2 \ (x-\bar{x}_2) \ S^{-1} \ (x-\bar{x}_2) \gtrless 0.$$

Retaining the notation of Section 2.5, the expectation $E(Z_1^*)$ may be expressed as

$$E(Z_1^*) = \Pr\{u' \ A^{-1}v < 0\}$$

$$= \Pr\{(u+v)' \ A^{-1}(u+v) \ - \ (u-v)' \ A^{-1} \ (u-v) < 0\}$$

where u and v are multinormally distributed with covariance matrices I, and (n_1+n_2-2) A, where $A = \Sigma^{-\frac{1}{2}} \ S \ \Sigma^{-\frac{1}{2}}$, has a Wishart distribution with n_1+n_2-2 degrees of freedom and dispersion matrix I, independent of u and v.

Use of Hotelling's T^2 distribution gives

$$\frac{n_1+n_2-p-1}{2(n_1+n_2-2)p} \ (u+v)' \ A^{-1} \ (u+v) \ \sim \ F(p, \ n_1+n_2-p-1, \ \delta_1)$$

$$\frac{n_1+n_2-p-1}{2(n_1+n_2-2)p} \ (u-v)' \ A^{-1} \ (u-v) \ \sim \ F(p, \ n_1+n_2-p-1, \ \delta_2)$$

with δ_1 and δ_2 as in Section 2.5 and $F(m,n,\lambda)$ denoting a non-central F distribution with degrees of freedom m and n and non-centrality λ. Hence

$$E(Z_1^*) = \Pr\left(\frac{F(p,n_1+n_2-p-1 \ , \ \delta_1)}{F(p,n_1+n_2-p-1 \ , \ \delta_2)} < 1\right) \qquad (2.13.1)$$

unfortunately the numerator and denominator of (2.13.1) are not independent since they both contain S.

Asymptotic expansions of the distribution of the ratio of non-independent T^2 variables have been given by Siotani (1956) and Chou and Siotani (1974). However since Siotani's original expansion requires both non-centralities to be zero and his later expansion requires one of them to be zero, neither is applicable here where both non-centralities are non-zero. It is to be hoped that the required expansion will be forthcoming in the literature at some future date.

For dimension $p = 1$ the allocation rule $Z(x)$ with Σ unknown is equivalent to the allocation rule $Z(x)$ with Σ known. Hence the expectations,inequalities and bounds given in Sections 2.7 and 2.8 for $E(Z_t^*)$with $p = 1$ and Σ known,hold here when Σ is unknown. Given the reappearance of the non-centrality parameters δ_1 and δ_2 in the unconditional distribution of $Z(x)$ when Σ is unknown, it is anticipated that, for $p > 1$ and Σ unknown, the Z statistic is still robust to unequal sample sizes. A Monte Carlo study in Section 8.7 supports this view.

CHAPTER 3

THE DISTRIBUTION OF THE QUADRATIC DISCRIMINANT FUNCTION

3.1 Introduction

The linear discriminant function L(x) is the optimal rule for discriminating between two multinormally distributed populations with equal covariance matrices. However when the covariance matrices are unequal the optimal allocation rule is the quadratic discriminant function Q(x) where

$$Q(x) = -\tfrac{1}{2}(x - \mu_1)'\,\Sigma_1^{-1}(x - \mu_1) + \tfrac{1}{2}(x - \mu_2)'\,\Sigma_2^{-1}(x - \mu_2) - \tfrac{1}{2}\ln(|\Sigma_1|/|\Sigma_2|) \gtrless C \qquad (3.1.1)$$

with $C = \ln(\pi_2/\pi_1)$, where $\pi_t (t = 1,2)$ are the prior probabilities of x in Π_t.

Much of the earlier work on discriminant analysis concentrated on the linear allocation rule L(x) and its sample counterpart $\hat{L}(x)$, deriving their distributions and various estimates of their error rates. Work on the quadratic discriminant function has been sparse due in part to the difficulty in evaluating the distribution of Q(x).

In this chapter we review the results on the distribution of Q(x), which for various assumptions on the population parameters are scattered throughout the literature. This review is necessary to establish the notation and basis of subsequent chapters where the distribution of the sample quadratic discriminant function $\hat{Q}(x)$ is derived and evaluated. As well as reviewing the distribution of Q(x) we consider the evaluation of the optimal probabilities of misclassification Q_t and obtain exact expressions for Q_t when the covariance matrices are proportional with and without the additional assumption of zero mean.

3.2 Review of the Literature on Q(x)

Smith (1947) is one of the earliest references to the quadratic discriminant function for Normally distributed populations. He considered Q(x) for the bivariate case and the multivariate case when all the covariances are zero. Smith also proposed the sample rule $\hat{Q}(x)$ where the parameters of Q(x) are replaced by their sample estimates. Okamoto (1961) derived the distribution of Q(x) for the assumption of zero mean i.e. $\mu_1 = \mu_2$, as a weighted sum of central chi-squares and proposed an approximation to this sum in evaluating the optimal error rates. Okamoto also considered the choice of cut-off point C that equates the error rates. Bartlett and Please (1963) considered the allocation rule Q(x) with the assumptions $\mu_1 = \mu_2$ and $\alpha\Sigma_1 = \Sigma_2$ where $\alpha > 0$ and $\alpha \neq 1$ i.e. zero mean and proportional covariance matrices. They further restricted the covariance matrix to the inter-class correlation matrix where all covariances are assumed equal. With these assumptions they were able to show that Q(x) depends on the "size component" $\sum_{i=1}^{p} x_i$, the concept of size and shape components being introduced by Penrose (1947). Han (1968) extended Bartlett and Please's work to the case where $\mu_1 \neq \mu_2$ and showed that Q(x) now involves the shape component $(\mu_1 - \mu_2)' x$. Han (1969) for the assumption of proportional covariance matrices derived the distribution of Q(x) as a non-central chi-squared and showed that with Bartlett and Please's additional assumption of zero mean, that this reduces to a central chi-squared. Gilbert (1969), in a comparison of the performances of the linear and quadratic discriminant functions for population parameters known, also derived the distribution of Q(x) for the assumption of proportional covariance matrices and used Patnaik's (1949) approximation to a non-central chi-square to evaluate the optimal probabilities of misclassification Q_t. Han (1970), derived the distribution of Q(x) when the covariance matrices Σ_t are circular, Press (1972, p14). The distribution is in this instance a sum of weighted non-central chi-squares. Use of

Patnaik's (1949) approximation is suggested in evaluating Q_t. Hildebrandt, Michaelis and Koller (1973 in German), derived the distribution of Q(x) in general and suggest some possible approximations for evaluating Q_t when the distribution is that of a positive definite non-central quadratic form.

3.3 Canonical Forms of the Populations and Various Cases of Q(x)

It was assumed that the distribution of x in Π_t is $N_p(\mu_t, \Sigma_t)$ (t = 1,2), resulting in the optimal allocation rule (3.1.1). As Q(x) is invariant to linear transformations it may, Okamoto (1961), without loss of generality be assumed that the distribution of x in Π_1 is $N_p(0,I)$ and in Π_2 is $N_p(\nu,\Lambda)$, where I is the identity matrix, Λ the diagonal matrix of eigen values λ_i, $1 < i < p$ where $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ and $\nu = B'(\mu_2 - \mu_1)$ where B is a non-singular matrix such that $B'\Sigma_1 B = I$ and $B'\Sigma_2 B = \Lambda$.

With these canonical forms for Π_1 and Π_2, Q(x) (3.1.1) becomes

$$Q(x) = -\tfrac{1}{2}x'x + \tfrac{1}{2}(x - \nu)' \Lambda^{-1} (x - \nu) + \tfrac{1}{2}\ln(| \Lambda |) - C \gtrless 0$$

$$= \tfrac{1}{2} \sum_{i=1}^{p} \{(\frac{1}{\lambda_i} - 1)x_i^2 - 2 \frac{\nu_i}{\lambda_i} x_i + \frac{\nu_i^2}{\lambda_i} + \ln \lambda_i\} - C \gtrless 0.$$

If $\lambda_i \neq 1$ for all i then

$$Q(x) = \sum_{i=1}^{p} \{\frac{1 - \lambda_i}{\lambda_i} (x_i - \frac{\nu_i}{1-\lambda_i})^2 \} - \left[2C + \sum_{i=1}^{p} \{\frac{\nu_i^2}{1-\lambda_i} - \ln\lambda_i\}\right] \gtrless 0 \tag{3.3.1}$$

Various possible values of λ_i, where $\lambda_i > 0$ for all i, result in the following cases of Q(x).

Case (i) $\lambda_i < 1, 1 \leqslant i \leqslant p$. The case of $\lambda_i > 1$ for all i is not considered a separate case since the distribution of Q(x) may be obtained for this case by merely altering the direction of the inequality.

Case (ii) $\lambda_i > 1$ for $1 \leqslant i \leqslant r$ and $\lambda_i < 1$ for $r + 1 \leqslant i \leqslant p$. The converse of this is excluded by the assumption that the λ_i's are in descending order of magnitude.

Case (iii) $\lambda_i > 1$ for $1 \leqslant i \leqslant r$, $\lambda_i = 1$ for $r + 1 \leqslant i \leqslant s$ and $\lambda_i < 1$ for $s + 1 \leqslant i \leqslant p$. This proves to be the most awkward case, combining as it does both linear and quadratic terms. Here

$$Q(x) = \sum_{i=1}^{r} \{\frac{1-\lambda_i}{\lambda_i}(x_i - \frac{\nu_i}{1-\lambda_i})^2 - 2 \sum_{i=r+1}^{s} x_i\nu_i + \sum_{i=s+1}^{p} \{\frac{1-\lambda_i}{\lambda_i}(x_i - \frac{\nu_i}{1-\lambda_i})^2\}$$

$$- \left[2C + \sum_{i=1}^{r} \{\frac{\nu_i^2}{1-\lambda_i} - \ln\lambda_i\} - \sum_{i=r+1}^{s} \nu_i^2 + \sum_{i=s+1}^{p} \{\frac{\nu_i^2}{1-\lambda_i} - \ln\lambda_i\}\right]$$

$$\gtrless 0 \qquad (3.3.2)$$

For the assumption of zero'mean i.e. $\mu_1 = \mu_2$, $\nu_i = 0$ for all i and Q(x) (3.3.1) becomes

$$Q(x) = \sum_{i=1}^{p} \{\frac{1-\lambda_i}{\lambda_i} x_i^2\} - \left[2C - \sum_{i=1}^{p} \ln\lambda_i\right] \gtrless 0 \qquad (3.3.3)$$

Here Case (iii) does not apply as for each $\lambda_i = 1$, Q(x) is reduced by one i.e. the dimensionality of our problem drops by one and we revert to Case (i) or (ii).

For the assumption of proportional covariance matrices i.e. $\alpha\Sigma_1 = \Sigma_2$, $\alpha > 0$ and $\alpha \neq 1$, $\lambda_i = \alpha$ for all i and so Case (i) need only be considered, thus without loss of generality it is assumed that $0 < \alpha < 1$. With the additional assumption of zero mean Q(x) (3.3.3) now becomes

$$Q(x) = \frac{1-\alpha}{\alpha}\ x'x - \left[2C - p\ell n\ \alpha\right] \gtrless 0 .$$

3.4 The Distribution of the Quadratic Discriminant Function

Here the general distribution of Q(x) is given and the effect of the assumptions of zero-mean and proportional covariances matrices considered.

Let

$$y_i^2 = \frac{|1-\lambda_i|}{\lambda_i}\ (x_i - \frac{\nu_i}{1-\lambda_i})^2$$

and

$$H = 2C + \sum_{i=1}^{p} \{\frac{\nu_i^2}{1-\lambda_i} - \ell n\lambda_i\} .$$

Then for Case (i) where $\lambda_i < 1$ for all i Q(x) (3.3.1) is given by

$$\sum_{i=1}^{p} y_i^2 - H \gtrless 0. \qquad (3.4.1)$$

For Case (ii) where $\lambda_i > 1$, $1 \leqslant j \leqslant r$ and $\lambda_i < 1$, $r + 1 \leqslant i \leqslant p$ Q(x) is given by

$$-\sum_{i=1}^{r} y_i^2 + \sum_{i=r+1}^{p} y_i^2 - H \gtrless 0. \qquad (3.4.2)$$

From (3.4.1) and (3.4.2) we see that the distribution of Q(x) requires the distribution of

$$Y = \sum_{i=1}^{r} y_i^2 = y'y$$

where $1 \leqslant r \leqslant p$ and $\lambda_i \neq 1$ for all i.

Two subcases must be considered

(a) x in Π_1 , $x_i \sim N_1$ (0,1)

(b) x in Π_2 , $x_i \sim N_1$ (ν_i,λ_i)

Subcase (a) : x in Π_1, as $x_i \sim N_1$ (0,1)

$$y_i \sim N_1\left(-\sqrt{\frac{|1-\lambda_i|}{\lambda_i}}\ \frac{\nu_i}{1-\lambda_i} , \frac{|1-\lambda_i|}{\lambda_i} \right)$$

$$= N_1\ (m_{1i}, \sigma^2_{1i}).$$

The x_i's and hence the y_i's are independently distributed and so the (rxl) vector y has a multinormal distribution $N_r(m_1, D_1)$ where $m_1 = (m_{11}, m_{12}, \ldots., m_{1r})$ a vector of means and D_1 is a diagonal matrix with diagonal $(\sigma^2_{11}, \sigma^2_{12}, ----, \sigma^2_{1r})$. Hence $y' D_1^{-1} y$ has a non-central chi-squared distribution with r degrees of freedom and non-centrality $m_1' D_1^{-1} m_1$. So

$$y' D_1^{-1} y \sim \chi^2\ (r, m_1' D_1^{-1} m_1)$$

where $$y' D_1^{-1} y = \sum_{i=1}^{r} \frac{\lambda_i}{|1-\lambda_i|}\ y_i^2$$

and $$m_1 D_1^{-1} m_1 = \sum_{i=1}^{r} \left(\frac{\nu_i}{1-\lambda_i}\right)^2 .$$

Subcase (b), x in Π_2, by a similar argument to subcase (a) and retaining similar notation it follows that

$$y_i \sim N_1 \left(-\sqrt{\frac{|1-\lambda_i|}{\lambda_i}}\ \frac{\nu_i\lambda_i}{1-\lambda_i} , |1-\lambda_i| \right)$$

$$= N_1\ (m_{2i}, \sigma^2_{2i})$$

and that $y'D_2^{-1}y \sim \chi^2\ (r, m_2' D_2^{-1} m_2)$

where $$y'D_2^{-1}y = \sum_{i=1}^{r} \frac{1}{|1-\lambda_i|}\ y_i^2$$

and $$m_2'D_2^{-1}m_2 = \sum_{i=1}^{r} \left(\frac{\nu_i}{1-\lambda_i} \right)^2 \lambda_i .$$

Thus for x in Π_1

$$Y = \sum_{i=1}^{r} y_i^2 \sim \sum_{i=1}^{r} \sigma_{1i}^2 \chi^2(1, (\frac{\nu i}{1-\lambda_i})^2)$$

and for x in Π_2

$$Y \sim \sum_{i=1}^{r} \sigma_{2i}^2 \chi^2(1, (\frac{\nu i}{1-\lambda_i})^2 \lambda_i), \qquad (3.4.3)$$

a positive definite non-central quadratic form in both subcases.

From the results (3.4.1), (3.4.2) and (3.4.3) it follows that in

Case (i) the distribution of Q(x) is a positive definite non-central quadratic form.

Case (ii) the distribution of Q(x) is that of an indefinite non-central quadratic form.

In Case (iii) from (3.3.2) and Case (ii) the distribution of Q(x) is that of an indefinite non-central quadratic form plus an additional Normal variate contributed by the (s-r) linear terms where $\lambda_i = 1$.

With the assumption of zero mean, $\nu_i = 0$ for all i, and so the non-central chi-squares of (3.4.3) become central giving the distribution of Q(x) as a positive definite quadratic form Case (i) and an indefinite quadratic form Case (ii). Case (iii) does not occur here.

With the assumption of proportional covariance matrices, $\lambda_i = \alpha$ for all i where $0 < \alpha < 1$. From (3.3.1) Q(x) is given by

$$Q(x) = \frac{1-\alpha}{\alpha} (x - \frac{\nu}{1-\alpha})' (x - \frac{\nu}{1-\alpha}) - H \gtrless 0$$

$$= Y - H \gtrless 0$$

where $H = 2C - p\ell n\alpha + \frac{\nu'\nu}{1-\alpha}$.

Now from (3.4.3)

$$\text{if } x \text{ in } \Pi_1 \qquad Y \sim \frac{1-\alpha}{\alpha}\ \chi^2(p, \frac{\nu'\nu}{(1-\alpha)^2}),$$

and (3.4.4)

$$\text{if } x \text{ in } \Pi_2 \qquad Y \sim 1-\alpha\ \chi^2(p, \frac{\alpha\nu'\nu}{(1-\alpha)^2})$$

thus with proportional covariance matrices the distribution of Q(x) is that of a non-central chi-squared. For the additional assumption of zero mean the non-centralities of (3.4.4) are zero and the distribution of Y and so of Q(x) is that of a central chi-square.

3.5 The Optimal Probabilities of Misclassification and their Evaluation.

The distribution of Q(x) is in almost all cases that of a quadratic form be it central or non-central, positive definite or indefinite. Finite expressions for the probability density functions of quadratic forms are only available for some special cases of the parameters, two of which are considered in the following sections.

The optimal probabilities of misclassification Q_t were defined in section 1.4 as

$$Q_1 = \Pr\{Q(x) < C \mid x \text{ in } \Pi_1\}$$

and (3.5.1)

$$Q_2 = \Pr\{Q(x) > C \mid x \text{ in } \Pi_2\}$$

From Cases (i) and (ii) of the previous section it follows that evaluation of Q_t is equivalent to evaluating probabilities of the type

$$\Pr\ \{ \sum_{i=1}^{r} \alpha_i\ \chi^2(1,\delta_i) < z \}$$

where $1 \leqslant r \leqslant p$ and α_i in R , $1 \leqslant i \leqslant r$, i.e. the evaluation of a cumulative distribution function of a quadratic form. A comprehensive review of quadratic forms, their distributions and evaluation is given by Johnson and Kotz (1970, Vol.2, pp149-188). Finite expressions for the probability (3.5.1) are only possible in special cases, some of which are considered in following sections. In the absence of such expressions the method of Imhof (1961) as described in Section 2.9 may be used to evaluate Q_t. Evaluation of Q_t for Case (iii) does not appear to be possible except where finite expressions for the probability density functions of the quadratic forms are available, the complication here being the inclusion of a Normal variate in the indefinite quadratic form of Q(x).

Hildebrandt et al (1973), who derived the general distribution of Q(x), gave a complicated infinite expansion in confluent hypergeometric functions for Q_t in Case (i) i.e. a positive definite non-central quadratic form. Evaluation by use of this expansion was not attempted in their paper. Many of the other references cited in Section 3.2 suggest various approximations such as Patnaik's (1949) to non-central chi-squares in evaluating Q_t for the particular population situation considered.

3.6 Evaluation of Q_t when the Covariance Matrices are Proportional

The special case of quadratic discrimination with proportional covariance matrices was considered by Bartlett and Please (1963) for zero mean and by Han (1968,1969) and Gilbert (1969) for non-zero mean. Both Han (1969) and Gilbert obtained the optimal error rates Q_t and Gilbert approximated Q_t by use of Patnaik's (1949) approximation of a non-central chi-square. It will be shown here that with the dimension size p odd exact expressions may be given for Q_t when $\mu_1 \neq \mu_2$.

With $\alpha\Sigma_1 = \Sigma_2$, $0 < \alpha < 1$ and $\mu_1 \neq \mu_2$ the distribution of Q(x) is, from Section 3.4, that of a non-central chi-square. Hence the optimal probabilities of misclassification are

$$Q_1 = \Pr\{\chi^2 \;(p, \frac{1}{(1-\alpha)^2} \;\nu'\nu) < \frac{\alpha}{1-\alpha} \;H\}$$

$$Q_2 = \Pr\{\chi^2 \;(p, \frac{\alpha}{(1-\alpha)^2} \;\nu'\nu) > \frac{1}{1-\alpha} \;H\} \qquad (3.6.1)$$

where $H = 2C - p\ell n\alpha + \frac{\nu'\nu}{1-\alpha}$.

For the additional assumption of zero mean $\mu_1 = \mu_2$, $\nu = 0$ and so

$$Q_1 = \Pr\{\chi^2_p < \frac{\alpha}{1-\alpha} \;H\}$$

$$Q_2 = \Pr\{\chi^2_p > \frac{1}{1-\alpha} \;H\} \qquad (3.6.2)$$

where now $H = 2C - p\ell n\alpha$.

The symbol χ^2_η will denote a central chi-squared variate with η degrees of freedom. By standard results, Johnson and Kotz (1970, Vol.1, p173)

$$\Pr\{\chi^2_{2k} > z\} = e^{-z/2} \sum_{j=o}^{k-1} \{(\tfrac{1}{2}z)^j/j!\} \qquad (3.6.3)$$

$$\Pr\{\chi^2_{2k-1} > z\} = e^{-z/2} \sum_{j=o}^{k-2} \{(\tfrac{1}{2}z)^{j+\frac{1}{2}}/\Gamma(j+\tfrac{3}{2})\} + 2\{1 - \Phi(\sqrt{z})\}$$

and so exact values of Q_t for the assumption of zero mean are easily obtained. If the prior probabilities are such that the cut-off $H < 0$ then $Q_1 = 0$ and $Q_2 = 1$, a similar comment applies to the case where $\mu_1 \neq \mu_2$.

Exact expressions for Q_t in the non-central case (3.6.1) are not so easily obtained. When the degrees of freedom are odd, Imhof (1961) gave a finite expression for the probability

$$\Pr\{\chi^2(\eta,\delta) < z\} \qquad (3.6.4)$$

by use of a recursive property of Bessel functions. Seber (1963) noted a similar result. Han (1975) also for odd degrees of freedom was able to express the probability (3.6.4) as a finite sum of standard Normal distributions and their derivatives, his result is

$$\begin{aligned}\Pr\{\chi^2(\eta,\delta) < z\} &= \Phi(a) - \Phi(b) \qquad (3.6.5)\\ &+ \sum_{i=1}^{k} \sum_{j=1}^{i} \binom{i-1}{j-1} 2^j \{\Phi^j(a) - \Phi^j(b)\}\end{aligned}$$

where $\eta = 2k+1$, $a = \sqrt{\delta} + \sqrt{z}$, $b = \sqrt{\delta} - \sqrt{z}$ and Φ^j is the jth derivative of Φ with respect to δ. Han (1978), using the recursive property of the derivatives of Φ developed a computational formula based on his earlier result which facilitates computer evaluation of an non-central chi-square with odd degrees of freedom.

Explicit expressions for (3.6.4) are given in both Imhof (1961) for $\eta = 3$ to 9 and Han (1975) for $\eta = 1$ to 7, Imhof's expressions being the more compact are recorded here.

$$\begin{aligned}\Pr\{\chi^2(\eta,\delta) < z\} &= \Phi(a) + \Phi(b) - 1\\ &- \frac{1}{\delta^{\frac{\eta-2}{2}}} (2\pi)^{-\frac{1}{2}} [T_\eta(\delta,a) - T_\eta(\delta,-b)]\end{aligned} \qquad (3.6.6)$$

where $a = \sqrt{z} - \sqrt{\delta}$, $b = \sqrt{z} + \sqrt{\delta}$ and $T_\eta(\delta,x) = \exp(-\frac{1}{2}x^2)\, P_\eta(\delta,x)$, with $P_\eta(\delta,x)$ a polynomial of degree $\frac{1}{2}(\eta-3)$.

For $\eta = 3, 5, 7$ and 9 these polynominals are given by

$$P_3(\delta,x) \equiv 1, \qquad P_5(\delta,x) = 2\delta + x\delta^{\frac{1}{2}} - 1,$$

$$P_7(\delta,x) = 3\delta^2 + 3x\delta^{\frac{3}{2}} + (x^2-4)\delta - 3x\delta^{\frac{1}{2}} + 3,$$

and
$$P_9(\delta,x) = 4\delta^3 + 6x\delta^{\frac{5}{2}} + (4x^2-10)\delta^2 + (x^3-15x)\delta^{\frac{3}{2}} - (6x^2-18)\delta + 15x\delta^{\frac{1}{2}} - 15.$$

For $\eta = 1$ if we let $P_1(\delta,x) \equiv 0$ then the desired result is obtained,coinciding with Han's result of (3.6.5).

Gilbert (1969) having derived the error rates Q_t (3.6.1) used Patnaik's (1949) two moment central chi-square approximation of a non-central chi-square for evaluation purposes. Patnaik's approximation is

$$\Pr\{\chi^2(\eta,\delta) < z\} \approx \Pr\{\chi^2_k < z/g\} \qquad (3.6.7)$$

where g and k are chosen so that the first two moments of the central and non-central chi-squares agree i.e.

$$g = \frac{\eta+2\delta}{\eta+\delta}, \qquad k = \frac{(\eta+\delta)^2}{\eta+2\delta}.$$

The accuracy of this approximation and other approximations to non-central chi-squares are discussed in Johnson and Kotz (1970, vol.2, pp139-143). In Section5.3,where for proportional covariance matrices the optimum error rates Q_t are evaluated using the exact expressions (3.6.6), the error in using Patnaik's approximation (3.6.7) is considered.

3.7 Evaluation of Q_t for the assumptions of zero-mean and half the eigen values the reciprocals of the other half.

This final special case of quadratic discrimination is considered for three reasons: (i) The resulting density of Q(x) is one of the few indefinite quadratic forms capable of being expressed in closed form, allowing exact expressions for Q_t; (ii) A similar result will be required in the following chapter where the distribution of the sample quadratic discriminant function is considered; (iii) A cut-off point C which equates the error rates Q_t is easily obtained, a problem considered by Okamoto (1961).

Here it is assumed that $\mu_1 = \mu_2$ giving $\nu = 0$ and that when the dimension size p is even, $\lambda_i = \alpha$ for $1 \leqslant i \leqslant p/2$ and $\lambda_i = 1/\alpha$ for $p/2+1 \leqslant i \leqslant p$ where $0 < \alpha < 1$. If p is odd then the central eigen value is put equal to one and we revert to the p even case. Unfortunately these assumptions on the eigen values do not appear to have any practical interpretation in terms of the original covariance matrices Σ_t.

With these assumptions the quadratic discriminant function is by (3.3.3)

$$Q(x) = \frac{1-\alpha}{\alpha} \sum_{i=1}^{p/2} x_i^2 - (1-\alpha) \sum_{i=p/2+1}^{p} x_i^2 - 2C \gtrless 0$$

$$= \frac{1-\alpha}{\alpha} N - (1-\alpha) M - 2C \gtrless 0$$

where $N = \sum_{i=1}^{p/2} x_i^2$ and $M = \sum_{i=p/2+1}^{p} x_i^2$.

For x in Π_1, $x_i \sim N_1(0,1)$ and so N and M are independently distributed as $\chi^2_{p/2}$. For x in Π_2, $x_i \sim N_1(0,\lambda_i)$ and so $\frac{1}{\alpha}$ N and αM are independently distributed as $\chi^2_{p/2}$, giving the distribution of Q(x) as that of the weighted difference of two independent central chi-squares.

Before considering the distribution of Q(x) in detail we note that with equal prior-probabilities C = 0 and that

$$\begin{aligned} Q_1 &= \Pr\{\tfrac{1}{\alpha} N - M < 0 \mid x \text{ in } \Pi_1\} \\ &= \Pr\{\tfrac{N}{M} < \alpha\} \\ &= \Pr\{F(p/2,\ p/2) < \alpha\} \qquad (3.7.1) \end{aligned}$$

where F(n,m) denotes an F variate with degrees of freedom n and m.

Similarly

$$\begin{aligned} Q_2 &= \Pr\{\tfrac{1}{\alpha} N - M > 0 \mid x \text{ in } \Pi_2\} \\ &= \Pr\{\frac{N}{\alpha^2 M} > \frac{1}{\alpha}\} \\ &= \Pr\{F(p/2,\ p/2) > \frac{1}{\alpha}\} \\ &= \Pr\{F(p/2,\ p/2) < \alpha\} = Q_1 \qquad (3.7.2) \end{aligned}$$

Hence with C = 0 the probabilities of misclassification are equal. This result is not surprising as the assumptions on the λ_i's give the required symmetry. By a standard result for an F distribution, Johnson and Kotz (1970, vol.2, p78) and from (3.7.1) and (3.7.2)

$$Q_1 = Q_2 = I_{\alpha/_{1+\alpha}}\ (p/4,\ p/4)$$

where $I_c(a,b)$ is an incomplete Beta function as defined and tabulated in Pearson (1934). If p/2 is even then p/4 is an integer and the following finite expression for $I_c(a,b)$ and so Q_t holds, Abramowitz and Stegun (1965,p944).

$$Q_1 = Q_2 = 1 - (\frac{1}{1+\alpha})^{p/2-1} \sum_{j=o}^{p/4-1} \binom{p/2-1}{j} \alpha^j .$$

For the distribution of Q(x) we must now consider the distribution of

$$X = \beta \chi^2_{2n} - \gamma \chi^2_{2m}$$

where β and γ are positive and the χ^2's are independently distributed. Finite expressions for the probability density function f(x) of X are given by Wang (1967) and may be applied here with the assumption that 2n = 2m = p/2 i.e. p/2 even.

From Wang (1967) it follows that with n = m

$$f(x) = \begin{cases} \sum_{j=o}^{n-1} [g(x)_{2\beta,2n-2j} \{(\frac{\beta}{\beta+\gamma})^n (\frac{\gamma}{\beta+\gamma})^j d_j\}] & x \geqslant 0 \\ \sum_{j=o}^{n-1} [g(-x)_{2\gamma,2n-2j} \{(\frac{\gamma}{\beta+\gamma})^n (\frac{\beta}{\beta+\gamma})^j d_j\}] & x \leqslant 0 \end{cases} \qquad (3.7.3)$$

where $d_j = \frac{(n+j-1)!}{j!\ (n-1)!} = \binom{n-1+j}{j}$

and $g(x)_{2\beta,2n-2j} = \frac{1}{(2\beta)^{n-j}\ \Gamma(n-j)} x^{n-j-1} e^{-\frac{x}{2\beta}} \qquad x \geqslant 0$

the probability density function of a Gamma variate with parameters 2β and degrees of freedom 2n-2j. Note when β=1 it becomes the probability density function of χ^2_{2n-2j}.

For $\bar{x}$ in Π_1, letting $\beta = 1/\alpha, \gamma = 1$ and $2n=p/2$ the probability density function of $Q(x)$ apart from the additive constant $H = \frac{2C}{1-\alpha}$ is from (3.7.3)

$$f(x)= \begin{cases} \sum\limits_{j=o}^{n-1} [g^{(x)}_{2/\alpha,2n-2j} \{(\frac{1}{\alpha+1})^n (\frac{\alpha}{\alpha+1})^j d_j\}] & x \geqslant 0 \\ \\ \sum\limits_{j=o}^{n-1} [g^{(\frac{-x}{2})}_{\chi 2n-2j} \{(\frac{\alpha}{\alpha+1})^n (\frac{1}{\alpha+1})^j d_j\}] & x \leqslant 0 \end{cases} \qquad (3.7.4)$$

The optimal error rate Q_1 is now given by

$$Q_1 = \int_{-\infty}^{H} f(x)\, dx$$

and with equal priors C and so H = 0 and

$$Q_1 = \int_{-\infty}^{o} f(x)\, dx$$

$$= \sum_{j=o}^{n-1} \{(\frac{\alpha}{\alpha+1})^n (\frac{1}{\alpha+1})^j \; dj\}$$

$$= \Pr\{F(p/2,\ p/2) < \alpha\}$$

If $H < 0$ then from (3.7.4) it follows that

$$Q_1 = \sum_{j=o}^{n-1} [\Pr\{\chi^2_{2n-2j} > -H\} \{(\frac{\alpha}{\alpha+1})^n (\frac{1}{\alpha+1})^j d_j\}]$$

$$= \sum_{j=o}^{n-1} [\{e^{H/2} \sum_{i=o}^{n-j-1} (-H/2)^i/i!\} \{(\frac{\alpha}{\alpha+1})^n (\frac{1}{\alpha+1})^j d_j\}]$$

by use of the finite expansions for χ^2 probabilities (3.6.3).

Since similar expansions apply to Gamma probabilities it follows from (3.7 .4) that if H > 0

$$Q = 1 - \int_H^\infty f(x)\,dx \tag{3.7.5}$$

$$= 1 - \sum_{j=o}^{n-1} [\{e^{-H\alpha/2} \sum_{i=o}^{n-j-1} (H\alpha/2)^i/i!\}\{(\frac{1}{\alpha+1})^n (\frac{\alpha}{\alpha+1})^j d_j\}] .$$

The probability density function of Q(x) for x in Π_2 may be obtained in a similar manner and the probability of misclassification Q_2 expressed in finite form.

CHAPTER 4

THE DISTRIBUTION OF THE SAMPLE QUADRATIC DISCRIMINANT FUNCTION

4.1 Introduction

In the absence of information about the populations, the unknown population parameters of an optimal allocation rule are replaced by their sample estimates so as to obtain a sample based allocation rule. For multinormally distributed populations with unequal covariance matrices this substitution of sample estimates results in the sample quadratic discriminant function $\hat{Q}(x)$ where

$$\hat{Q}(x) = -\tfrac{1}{2}\,(x-\bar{x}_1)'\,S_1^{-1}(x-\bar{x}_1) + \tfrac{1}{2}\,(x-\bar{x}_2)'\,S_2^{-1}(x-\bar{x}_2) - \tfrac{1}{2}\,\ln(|S_1|/|S_2|) - K$$

$$\gtrless 0$$

The allocation rule $\hat{Q}(x)$ was proposed by Smith (1947), who compared its allocation performance and that of the sample linear allocation rule $\hat{L}(x)$ on two examples. As is the case for $\hat{L}(x)$, the distribution of $\hat{Q}(x)$ has proved extremely difficult to obtain and the results in the literature give asymptotic expansions for various special cases of the population parameters. Okamoto (1961) for $\mu_1 = \mu_2$ and $p = 1$ outlined such an expansion and commented that the expansion for the general case appeared to be "rather difficult". Many of the other expansions found in the literature concentrated on the proportional covariance matrices case where $\alpha\Sigma_1 = \Sigma_2 = \alpha\Sigma$, $\alpha \neq 1$, Han (1969) giving the expansion for Σ and α known, Han (1974) for either Σ or α known and Mc Lachlan (1975) for Σ and α unknown, where α was estimated by the ratio of the determinants of the sample covariance matrices i.e. $\hat{\alpha} = (|S_2|/|S_1|)^{1/p}$. Han (1970) has also given the asymptotic expansion for the distribution of $\hat{Q}(x)$ when Σ_t, $(t = 1,2)$ are circular and unknown.

In this chapter it will be assumed that the covariance matrices Σ_t $(t = 1,2)$ are known, a restrictive assumption but one which allows the exact derivation of the distribution of $\hat{Q}(x)$. The case of proportional covariance matrices is considered in detail and the exact expectations of the actual and apparent probabilities of misclassification derived. With the additional assumption of zero mean, closed expressions for these expectations are obtained and the chapter concludes with their evaluation.

4.2 The Distribution of $\hat{Q}(x)$, Conditional on $\bar{x}_1$ and $\bar{x}_2$, for Known Proportional Covariance Matrices

As the sample quadratic discriminant function is invariant to linear transformations, the canonical forms for the populations, Section 3.3, may be assumed without loss of generality. With proportional covariance matrices these canonical forms allow the following values of the population parameters

$$\Sigma_1 = I, \quad \Sigma_2 = \alpha I, \quad 0 < \alpha < 1, \quad \mu_1 = 0 \text{ and } \mu_2 = \nu$$

Thus for known proportional covariance matrices

$$\hat{Q}(x) = -\tfrac{1}{2}(x-\bar{x}_1)'(x-\bar{x}_1) + \tfrac{1}{2}\frac{1}{\alpha}(x-\bar{x}_2)'(x-\bar{x}_2) + \tfrac{1}{2}p\ell n\alpha - K \gtrless 0$$

$$= \frac{1-\alpha}{\alpha}\left\{x - \frac{(\bar{x}_2-\alpha\bar{x}_1)}{1-\alpha}\right\}'\left\{x - \frac{(\bar{x}_2-\alpha\bar{x}_1)}{1-\alpha}\right\} - \frac{(\bar{x}_2-\bar{x}_1)'(\bar{x}_2-\bar{x}_1)}{1-\alpha}$$

$$+ p\ \ell n\ \alpha - 2K \gtrless 0$$

$$= \frac{1-\alpha}{\alpha}(x-z)'(x-z) - J \gtrless 0$$

where $z = \dfrac{(\bar{x}_2-\alpha\bar{x}_1)}{1-\alpha}$ and $w = \dfrac{(\bar{x}_2-\bar{x}_1)}{1-\alpha}$

and $J = (1-\alpha)w'w - p\ \ell n\ \alpha + 2K.$

Conditional on $\bar{x}_1$ and $\bar{x}_2$

$$(x-z)'\ (x-z) \sim \chi^2(p,z'z) \quad \text{for } x \text{ in } \Pi_1$$

and $$(x-z)'\ (x-z) \sim \alpha\ \chi^2(\ p,\frac{(\nu-z)'\ (\nu-z)}{\alpha}) \text{ for } x \text{ in } \Pi_2\ ,$$

giving the conditional distribution of $\hat{Q}(x)$ as a non-central chi-squared variate. With the assumption of zero-mean i.e. ν=0 the distribution is still that of a non-central chi-squared unlike the optimal distribution Section 3.4. The general conditional distribution may be derived in a similar manner taking one dimension at a time. Depending on the various cases of λ_i, as given in Section 3.3, the distribution is that of a positive definite or indefinite non-central quadratic form. The problem of the Normal variates corresponding to the λ_i=1 still remains.

The actual probabilities of misclassification Q^*_t as defined in Section 1.5 may, for proportional covariance matrices, be written as

$$Q_1^* = \Pr\{\chi^2(p,z'z) < \frac{\alpha}{1-\alpha}\quad J\}$$

$$Q_2^* = \Pr\{\chi^2(\ p,\frac{(\nu-z)'(\nu-z)}{1-\alpha}) > \frac{1}{1-\alpha}\ J\}$$

Imhof (1961) or Han's (1975) closed form of a non-central chi-square Section 3.6 may be used to evaluate Q_t^*. However since the non-centralities require a knowledge of the population parameter ν this evaluation would be impossible in practice but may be undertaken in simulation studies.

It is noted that with $\nu \neq 0$ and proportional covariance matrices, if α is set equal to one in $\hat{Q}(x)$, then the sample linear allocation rule $\hat{L}(x)$ is obtained. This relationship may be used to check expressions derived for $\hat{Q}(x)$.

4.3 The Unconditional Distribution of $\hat{Q}(x)$ for Known Proportional Covariance Matrices

The exact unconditional distribution of $\hat{Q}(x)$ is derived initially for the case of known proportional covariance matrices as it illustrates the method of proof to be followed in the general case. The method of proof is similar to that used in Section 2.5 to derive the unconditional distribution of the Z statistic. Han (1969) has given a complicated asymptotic expansion for the cumulative distribution function of $\hat{Q}(x)$ in this case of known proportional covariance matrices.

Expectations of the actual probabilities of misclassification are derived and their evaluation considered. With the additional assumption of zero-mean closed expressions for these expectations are given.

Now $$\hat{Q}(x) = -(x-\bar{x}_1)'(x-\bar{x}_1) + \frac{1}{\alpha}(x-\bar{x}_2)'(x-\bar{x}_2) - J \gtrless 0$$

where $$J = 2K - p\,\ln\alpha .$$

This may be re-written as

$$\begin{aligned} \hat{Q}(x) &= \{\tfrac{1}{\sqrt{\alpha}}(x-\bar{x}_2) - (x-\bar{x}_1)\}'\{\tfrac{1}{\sqrt{\alpha}}(x-\bar{x}_2) + (x-\bar{x}_1)\} - J \gtrless 0 \\ &= \{(\tfrac{1}{\sqrt{\alpha}} -1)x - (\tfrac{1}{\sqrt{\alpha}}\bar{x}_2-\bar{x}_1)\}'\{(\tfrac{1}{\sqrt{\alpha}} +1)x - (\tfrac{1}{\sqrt{\alpha}}\bar{x}_2+\bar{x}_1)\} - J \gtrless \\ &= r's - J \gtrless 0 \end{aligned}$$

where $$r = \{(\tfrac{1}{\sqrt{\alpha}} - 1)x - (\tfrac{1}{\sqrt{\alpha}}\bar{x}_2 - \bar{x}_1)\}$$

and $$s = \{(\tfrac{1}{\sqrt{\alpha}} + 1)x - (\tfrac{1}{\sqrt{\alpha}}\bar{x}_2 + \bar{x}_1)\} .$$

Let the cumulative distribution function of $\hat{Q}(\bar{x})$ be given by

$$F_t(a) = \Pr\{\hat{Q}(x) \leqslant a \mid x \text{ in } \Pi_t\} \qquad (t = 1,2)$$
$$= \Pr\{r's \leqslant a + J \mid x \text{ in } \Pi_t\} .$$

For x in Π_1

$$F_1(a) = \Pr\{r's \leqslant a + J \mid x \text{ in } \Pi_1\}$$

where r and s are p-dimensional Normally distributed with means $-\nu/\sqrt{\alpha}$ and covariance matrices

$$\frac{d_1^2}{\alpha n_1 n_2} I \quad \text{and} \quad \frac{d_2^2}{\alpha n_1 n_2} I$$

respectively, where

$$d_1 = \{\alpha n_1 + \alpha n_2 + (1-\sqrt{\alpha})^2 n_1 n_2\}^{\frac{1}{2}}$$
$$d_2 = \{\alpha n_1 + \alpha n_2 + (1+\sqrt{\alpha})^2 n_1 n_2\}^{\frac{1}{2}} .$$

With $c_1 = (\alpha n_1 n_2)^{\frac{1}{2}}/ d_1$ and $c_2 = (\alpha n_1 n_2)^{\frac{1}{2}}/ d_2$

then $u = c_1 r$ and $v = c_2 s$

are multinormally distributed with covariance matrices I. The correlation ρ_1 between the pairwise elements of u and v is given by $\rho_1 = \{\alpha n_1 - \alpha n_2 + (1-\alpha) n_1 n_2\} / d_1 d_2$.

Now $F_1(a) = \Pr\{u'v \leqslant c_1 c_2 (a + J)\}$

$$= \Pr\{(u+v)'(u+v) - (u-v)'(u-v) \leqslant 4 c_1 c_2 (a + J)\}$$

where (u+v) and (u-v) are independently multinormally distributed with covariance matrices $2(1+\rho_1)I$ and $2(1-\rho_1)I$ respectively.

Hence

$$\omega_1 = (u+v)'(u+v) \;/\; 2(1+\rho_1) \quad \text{and} \quad \omega_2 = (u-v)'(u-v) \;/\; 2(1-\rho_1)$$

are independently distributed as non-central chi-squares

$$\chi^2(p, \delta_1) \quad \text{and} \quad \chi^2(p, \delta_2),$$

where the non-centralities δ_1 and δ_2 are given by

$$\delta_1 = \{2(1+\rho_1)\}^{-1} n_1 n_2 (d_1^{-1} + d_2^{-1})^2 \nu'\nu$$

$$\delta_2 = \{2(1-\rho_1)\}^{-1} n_1 n_2 (d_1^{-1} - d_2^{-1})^2 \nu'\nu.$$

Thus

$$F_1(a) = \Pr\{\frac{(1+\rho_1)}{2c_1c_2} \chi^2(p, \delta_1) - \frac{(1-\rho_1)}{2c_1c_2} \chi^2(p, \delta_2) \leq a + J\}$$

and the unconditional distribution of $\hat{Q}(x)$ is that of an indefinite non-central quadratic form.

A similar argument for x in Π_2 gives

$$F_2(a) = \Pr\{\frac{(1+\rho_2)}{2c_3c_4} \chi^2(p, \delta_3) - \frac{(1-\rho_2)}{2c_3c_4} \chi^2(p, \delta_4) \leq a + J\}$$

where

$$d_3 = \{n_1 + n_2 + (1-\sqrt{\alpha})^2 n_1 n_2\}^{\frac{1}{2}}$$

$$d_4 = \{n_1 + n_2 + (1+\sqrt{\alpha})^2 n_1 n_2\}^{\frac{1}{2}}$$

$$c_3 = (n_1 n_2)^{\frac{1}{2}} \;/\; d_3 \quad \text{and} \quad c_4 = (n_1 n_2)^{\frac{1}{2}} \;/\; d_4$$

$$\rho_2 = \{n_1 - n_2 + (1-\alpha)\, n_1 n_2\} \;/\; d_3 d_4$$

$$\delta_3 = \{2(1+\rho_2)\}^{-1} n_1 n_2 (d_3^{-1} - d_4^{-1})^2 \nu'\nu$$

$$\delta_4 = \{2(1-\rho_2)\}^{-1} n_1 n_2 (d_3^{-1} + d_4^{-1})^2 \nu'\nu.$$

As noted in the previous section, by letting α=1 the above parameters of the unconditional distribution of $\hat{Q}(x)$ for known proportional covariance matrices may be checked against those for $\hat{L}(x)$ in Section 2.8.

The expectations of the actual probabilities of misclassification are given by

$$E(Q_1^*) \;=\; \Pr\{\hat{Q}(x) < 0 \mid x \text{ in } \Pi_1\}$$

$$= F_1(0)$$

and $$E(Q_2^*) \;=\; 1 - F_2(0).$$

These expectations are evaluated in Chapter 5, Section 5.5, the method of evaluation is that of Imhof (1961) as discussed in Section 2.9.

With the additional assumption of zero-mean $\nu=0$, the non-central chi-squares in the cumulative distribution functions F_1 and F_2 become central and the unconditional distribution of $\hat{Q}(x)$ is now an indefinite central quadratic form. If the dimension p is even, the results of Section 3.7 may be used to obtain the probability density function of $\hat{Q}(x)$ and to give exact expressions for the expectations of the actual error rates.

For x in Π_1 and $\nu = 0$

$$F_1(a) = \Pr\{\frac{(1+\rho_1)}{2c_1c_2}\chi_p^2 - \frac{(1-\rho_1)}{2c_1c_2}\chi_p^2 \leqslant a + J\}.$$

Retaining the notation of Section 3.7, then if p is even and with

$$\beta_1 = \frac{1+\rho_1}{2c_1c_2}\ , \quad \gamma_1 = \frac{1-\rho_1}{2c_1c_2} \quad \text{and} \quad n = p/2$$

the probability density function of $\hat{Q}(x)$ is

$$f(x) = \begin{cases} \sum_{j=o}^{n-1} [g_{2\beta_1}^{(x)},_{2n-2j} \{ \left(\frac{1+\rho_1}{2}\right)^n \left(\frac{1-\rho_1}{2}\right)^j d_j\}] & x > 0 \\ & (4.3.1) \\ \sum_{j=o}^{n-1} [g_{2\gamma_1,2n-2j}^{(-x)} \{ \left(\frac{1-\rho_1}{2}\right)^n \left(\frac{1+\rho_1}{2}\right)^j d_j\}] & x < 0 \end{cases}$$

where g(x) and d_j are as defined in (3.7.3).

Hence $$E(Q_1^*) = \int_{-\infty}^{J} f(x)\, dx$$

and as $J = 2K - p \ln \alpha$ if we assume equal prior-probabilities $K = o$, and since $o < \alpha < 1$, $J > o$, then by (3.7.5) and (4.3.1)

$$E(Q_1^*) = 1 - \sum_{j=o}^{n-1} \{e^{-J/2\beta_1} \sum_{i=o}^{n-j-1} \frac{1}{i!} \left(\frac{J}{2\beta_1}\right)^i \}\{ \left(\frac{1+\rho_1}{2}\right)^n \left(\frac{1-\rho_1}{2}\right)^j d_j\} \tag{4.3.2}$$

The probability density function of $\hat{Q}(x)$ and $E(Q_2^*)$ for x in Π_2 may similarly be expressed in closed form. With

$$\beta_2 = \frac{1+\rho_2}{2c_3c_4} \quad \text{and} \quad n = p/2$$

p even and equal prior-probabilities, it follows that

$$E(Q_2^*) = \sum_{j=o}^{n-1} \{e^{-J/2\beta_2} \sum_{i=o}^{n-j-1} \frac{1}{i!} \left(\frac{J}{2\beta_2}\right)^i \}\{ \left(\frac{1+\rho_2}{2}\right)^n \left(\frac{1-\rho_2}{2}\right)^j d_j\}. \tag{4.3.3}$$

The expectations $E(Q_t^*)$ (4.3.2) and (4.3.3) are evaluated in Section 4.6 for a range of values of p, α and (n_1,n_2).

4.4 The Expectations of the Apparent Probabilities of Misclassification of the Sample Quadratic Discriminant Function for Known Proportional Covariance Matrices.

The apparent probabilities of misclassification were defined in Section 1.5 as the proportions of the sample observations misclassified by the sample allocation rule. For quadratic discrimination these apparent error rates are denoted by Q_1^{**} and Q_2^{**}.

Here for known proportional covariance matrices the expectations of the apparent error rates will be derived by the method used in the previous section, the difference being that the random observation x from Π_t is replaced by a member x_{tj} of the sample of n_t from Π_t. Hence

$$E(Q_1^{**}) = \Pr\{-(x_{1j}-\bar{x}_1)'(x_{1j}-\bar{x}_1) + \frac{1}{\alpha}(x_{1j}-\bar{x}_2)(x_{1j}-\bar{x}_2) - J < 0\}$$

where $J = 2K - p \ln \alpha$.

Allowing for the correlation of x_{1j} and $\bar{x}_1$ it follows that

$$E(Q_1^{**}) = \Pr\{\frac{(1+\rho_3)}{2c_5c_6}\chi^2(p, \delta_5) - \frac{(1-\rho_3)}{2c_5c_6}\chi^2(p, \delta_6) < J\}$$

where $d_5 = \{\alpha n_1 + (2\sqrt{\alpha}-\alpha)n_2 + (1-\sqrt{\alpha})^2 n_1n_2\}^{\frac{1}{2}}$

$d_6 = \{\alpha n_1 - (2\sqrt{\alpha}+\alpha)n_2 + (1+\sqrt{\alpha})^2 n_1n_2\}^{\frac{1}{2}}$

$c_5 = (\alpha n_1n_2)^{\frac{1}{2}} / d_5$ and $c_6 = (\alpha n_1n_2)^{\frac{1}{2}} / d_6$

$\rho_3 = \{\alpha n_1 + \alpha n_2 + (1-\alpha) n_1n_2) / d_5d_6$

$\delta_5 = \{2(1+\rho_3)\}^{-1} n_1n_2 (d_5^{-1} + d_6^{-1})^2 \nu'\nu$

$\delta_6 = \{2(1-\rho_3)\}^{-1} n_1n_2 (d_5^{-1} - d_6^{-1})^2 \nu'\nu$.

Similarly

$$E(Q_2^{**}) = \Pr\{\frac{(1+\rho_4)}{2c_7c_8}\chi^2(p, \delta_7) - \frac{(1-\rho_4)}{2c_7c_8}\chi^2(p,\delta_8) > J\}$$

where $d_7 = \{(-1 + 2\sqrt{\alpha}) n_1 + n_2 + (1-\sqrt{\alpha})^2 n_1n_2\}^{\frac{1}{2}}$

$d_8 = \{(-1 - 2\sqrt{\alpha}) n_1 + n_2 + (1+\sqrt{\alpha})^2 n_1n_2\}^{\frac{1}{2}}$

$c_7 = (n_1n_2)^{\frac{1}{2}} / d_7$ and $c_8 = (n_1n_2)^{\frac{1}{2}} / d_8$

$\rho_4 = \{-n_1 - n_2 + (1-\alpha) n_1n_2\} / d_7d_8$

$\delta_7 = \{2(1+\rho_4)\}^{-1} n_1n_2 (d_7^{-1} - d_8^{-1})^2 \nu'\nu$

$\delta_8 = \{2(1-\rho_4)\}^{-1} n_1n_2 (d_7^{-1} + d_8^{-1})^2 \nu'\nu$.

With $\alpha = 1$ the results given coincide with Moran's (1974) expressions for the expectations of the apparent error rates of $\hat{L}(x)$, the sample linear discriminant function.

As is the case for the expectations of the actual error rates with the additional assumption of zero-mean, the non-centralities δ_5, δ_6, δ_7 and δ_8 are zero and exact expressions similar to (4.3.2) and (4.3.3) may be obtained for the expectations of the apparent error rates in this case also.

If $K = 0$ then $J > 0$ and with p even, $n = p/2$ and $\beta_3 = \frac{1+\rho_3}{2c_5c_6}$

$$E(Q_1^{**}) = 1 - \sum_{j=o}^{n-1} \{e^{-J/2\beta_3} \sum_{i=o}^{n-j-1} \frac{1}{i!} \left(\frac{J}{2\beta_3}\right)^i \}\{\left(\frac{1+\rho_3}{2}\right)^n \left(\frac{1-\rho_3}{2}\right)^j \quad d_j\} \tag{4.4.1}$$

and with $\beta_4 = \frac{1+\rho_4}{2c_7c_8}$

$$E(Q_2^{**}) = \sum_{j=o}^{n-1} \{e^{-J/2\beta_4} \sum_{i=o}^{n-j-1} \frac{1}{i!} \left(\frac{J}{2\beta_4}\right)^i \}\{ \left(\frac{1+\rho_4}{2}\right)^n \left(\frac{1-\rho_4}{2}\right)^j \quad d_j\} \tag{4.4.2}$$

The expectations $E(Q_t^{**})$ (4.4.1) and (4.4.2) are evaluated in Section 4.6 for a range of values of p, α and (n_1,n_2).

4.5 The Unconditional Distribution of $\hat{Q}(x)$ for Known Covariance Matrices

The general derivation for Cases (i),(ii) and (iii) of the eigen values, Section 3.3, is presented here. The canonical forms of the populations being $N_p(0,I)$ and $N_p(\nu, \Lambda)$, the sample quadratic discriminant function for known covariance matrices is

$$\hat{Q}(x) = -(x-\bar{x}_1)'(x-\bar{x}_1) + (x-\bar{x}_2)'\,\Lambda^{-1}\,(x-\bar{x}_2) + \ln|\Lambda| - 2K \gtrless 0$$

$$= \sum_{i=1}^{p} \{-(x_i-\bar{x}_{1i})^2 + \frac{1}{\lambda_i}\,(x_i-\bar{x}_{2i})^2\} - J \gtrless 0 \qquad (4.5.1)$$

where $J = 2K - \sum_{i=1}^{p} \ln \lambda_i$.

Case (i) and (ii) of the eigen values; $\lambda_i \neq 1$ for all i. Here

$$\hat{Q}(x) = \sum_{i=1}^{p}\left[\frac{1-\lambda_i}{\lambda_i}\{x_i - \frac{(\bar{x}_{2i}-\lambda_i\bar{x}_{1i})}{1-\lambda_i}\}^2 - \frac{(\bar{x}_{2i}-\bar{x}_{1i})^2}{1-\lambda_i}\right] - J \gtrless 0$$

$$= \sum_{i=1}^{p}\left[\frac{1-\lambda_i}{\lambda_i}\,(r_i^2 - s_i^2)\right] - J \gtrless 0$$

where $r_i = \left\{x_i - \frac{(\bar{x}_{2i}-\lambda_i\bar{x}_{1i})}{1-\lambda_i}\right\}$ and $s_i = \frac{\sqrt{\lambda_i}}{(1-\lambda_i)}\,(\bar{x}_{2i}-\bar{x}_{1i})$.

Without loss of generality we let x in Π_1, as a similar proof holds for x in Π_2.

If $c_{1i} = (1-\lambda_i)\,(n_1n_2)^{\frac{1}{2}} \,/\, \{(1-\sqrt{\lambda_i})^2\,\lambda_i\,(n_1+n_2) + (1-\lambda_i)^2\,n_1n_2\}^{\frac{1}{2}}$

and $c_{2i} = (1-\lambda_i)\,(n_1n_2)^{\frac{1}{2}} \,/\, \{(1+\sqrt{\lambda_i})^2\,\lambda_i\,(n_1+n_2) + (1-\lambda_i)^2\,n_1n_2\}^{\frac{1}{2}}$

then $u_i = c_{1i}\,(r_i+s_i)$ and $v_i = c_{2i}\,(r_i-s_i)$

are Normally distributed with unit variances.

Hence

$$\hat{Q}(x) = \sum_{i=1}^{p} \left[\frac{1-\lambda_i}{\lambda_i c_{1i} c_{2i}} \quad u_i v_i \right] - J \gtrless 0$$

$$= \sum_{i=1}^{p} \left[\frac{1-\lambda_i}{4\lambda_i c_{1i} c_{2i}} \{(u_i+v_i)^2 - (u_i-v_i)^2\} \right] - J \gtrless 0 \ .$$

Now (u_i+v_i) and (u_i-v_i) are independently Normally distributed with variances $2(1+\rho_i)$ and $2(1-\rho_i)$ where

$$\rho_i = -c_{1i} c_{2i} \lambda_i^{3/2} (n_1+n_2) \ / \ (1-\lambda_i)^2 \ n_1 n_2$$

is the correlation between u_i and v_i. Let

$$\omega_{1i} = \frac{(u_i+v_i)^2}{2(1+\rho_i)} \qquad \text{and} \qquad \omega_{2i} = \frac{(u_i-v_i)^2}{2(1-\rho_i)}$$

which are independently distributed as non-central chi-squares with non-centralities δ_{1i} and δ_{2i} and degrees of freedom 1, where

$$\delta_{1i} = \{2(1+\rho_i)\}^{-1} \{c_{1i}(1-\sqrt{\lambda_i}) + c_{2i}(1+\sqrt{\lambda_i})\}^2 \ \frac{\nu_i^2}{(1-\lambda_i)^2}$$

$$\delta_{2i} = \{2(1-\rho_i)\}^{-1} \{c_{1i}(1-\sqrt{\lambda_i}) - c_{2i}(1+\sqrt{\lambda_i})\}^2 \ \frac{\nu_i^2}{(1-\lambda_i)^2} \ .$$

If follows that

$$\Pr\{\hat{Q}(x) \leq a\} = \Pr\left[\sum_{i=1}^{p} \eta_i \ \{2(1+\rho_i) \ \omega_{1i} - 2(1-\rho_i) \ \omega_{2i}\} \leq a + J \right]$$

where $$\eta_i = \frac{1-\lambda_i}{4\lambda_i c_{1i} c_{2i}}$$

$$= \Pr\left[\sum_{i=1}^{p} \eta_i \{2(1+\rho_i) \ \chi^2(1,\delta_{1i}) - 2(1-\rho_i) \ \chi^2(1,\delta_{2i})\} \leq a+J \right] .$$

Hence for x in Π_t and for Cases (i) and (ii) of the eigen values the unconditional distribution of $\hat{Q}(x)$ is that of an indefinite non-central quadratic form.

Case (iii) $\lambda_i = 1$ for some i $1 \leqslant i \leqslant p$:- The extension of the proof to Case (iii) where $\lambda_i = 1$ for some i is straight forward in contrast to the optimal and conditional distributions.

Let x in Π_1 and $\lambda_k = 1$ where $1 \leqslant k < p$. Then a term of the type

$$-(x_k-\bar{x}_{1k})^2 + (x_k-\bar{x}_{2k})^2$$

is contributed to the sum (4.5.1) for each $\lambda_i = 1$. Rewriting the above term as

$$(\bar{x}_{1k}-\bar{x}_{2k})\ \{2x_k-(\bar{x}_{1k}+\bar{x}_{2k})\}$$

i.e. a univariate sample linear discriminant function, with

$$u_k = c_{1k}\ (\bar{x}_{1k}-\bar{x}_{2k}) \quad \text{and} \quad v_k = c_{2k}\ \{2x_k-(\bar{x}_{1k}+\bar{x}_{2k})\}$$

where $c_{1k} = \{(n_1n_2)\ /\ (n_1+n_2)\}^{\frac{1}{2}}$

$$c_{2k} = \{(n_1n_2)\ /\ (n_1+n_2+\ 4n_1n_2)\}^{\frac{1}{2}}$$

it may be shown as previously that this term is distributed as the weighted difference of two independent non-central chi-squares with degrees of freedom 1. Retaining the previous notation the appropriate terms are

$$\rho_k = (n_1-n_2)\ /\ \{(n_1+n_2)\ (n_1+n_2+4n_1n_2)\}^{\frac{1}{2}}$$

$$\delta_{1k} = \{2(1+\rho_k)\}^{-1}\ (c_{1k}+c_{2k})^2\ \ \nu_k^2$$

$$\delta_{2k} = \{2(1-\rho_k)\}^{-1}\ (c_{1k}-c_{2k})^2\ \ \nu_k^2$$

and $\eta_k = 1\ /\ 4\ c_{1k}c_{2k}$.

In general the degrees of freedom will be the number of λ_i which equal one. Since the remaining terms corresponding to $\lambda_i \neq 1$ fall into Case (i) or (ii), it follows that the unconditional distribution of $\hat{Q}(x)$ is in all cases that of an indefinite non-central quadratic form.

The expectations of the actual probabilities of misclassification may now be derived and evaluated by Imhof's (1961) method as given in Section 2.9.

4.6 Zero-Mean Discrimination with Known Proportional Covariance Matrices. An evaluation of the optimal and expectations of the actual and apparent error rates.

The optimal probabilities of misclassification Q_t and the expectations $E(Q_t^*)$ and $E(Q_t^{**})$ are evaluated here for the assumptions of zero-mean and known proportional covariance matrices. While admitting that these assumptions are restrictive, they allow a reasonable range of parameters i.e. p,α,n_1, and n_2, to be considered, which would not be the case in general. The present assumptions also allow exact expressions for the expectations of the actual and apparent error rates which are easily evaluated.' Linear discrimination is not applicable here since the assumption of zero-mean and common covariance matrix would make the two populations indistinguishable. Applications of zero-mean discrimination with proportional covariance may be found in Bartlett and Please (1963) and Desu and Geisser (1973).

It is assumed that the distribution of x in Π_1 is $N_p(0,I)$ and in Π_2 is $N_p(\nu,\alpha I)$ where $o < \alpha < 1$, that the mean $\nu = o$ and that the prior-probabilities in both the optimal and sample allocation rules are equal. It is also assumed that the dimension p is even, the latter assumption is required in obtaining the closed expressions for $E(Q_t^*)$ and $E(Q_t^{**})$.

The individual and total error rates and expectations;

$$Q_t, \quad E(Q_t^*), \quad E(Q_t^{**}) \qquad\qquad t = 1,2$$

$$\bar{Q} = \tfrac{1}{2}\{Q_1+Q_2\}, \quad \bar{E}(Q^*) = \tfrac{1}{2}\{E(Q_1^*)+E(Q_2^*)\} \text{ and } \bar{E}(Q^{**}) = \tfrac{1}{2}\{E(Q_1^{**})+E(Q_2^{**})\}$$

are compared for the range of parameters given in Table 4.6.1. The necessary formulae for evaluating the probabilities and expectations are given in Sections 3.6, 4.3 and 4.4. Optimal probabilities Q_t are given in Table 4.6.2, the expectations for equal sample sizes in Table 4.6.3 and for unequal sizes in Table 4.6.4.

In Table 4.6.2 it is noted that for fixed p the optimal probabilities Q_1 and Q_2 increase as α approaches one. This is to be expected as the two populations become less distinguishable with increasing α. However, for p = 2 it is easily shown that $\lim_{\alpha\to 1} Q_1 = 1-e^{-1} = .632$ and $\lim_{\alpha\to 1} Q_2 = e^{-1} = .368$. With fixed proportionality α in Table 4.6.2, Q_1 decreased with increasing p as does Q_2 except for $\alpha = .9$ where Q_2 increases for p = 8 and 16 and then decreases. By standardising the chi-square probabilities of Q_1 and Q_2 (3.6.2) it may be shown that for fixed α, Q_1 and Q_2 go to zero as p approaches infinity. The total probability of misclassification $\bar{Q}$ in Table 4.6.2 increases and approaches $\frac{1}{2}$ as α approaches one for fixed p, and decreases with increasing p for fixed α.

Table 4.6.1

Range of parameters used in evaluating Q_t, $E(Q_t^*)$ and $E(Q_t^{**})$ for zero-mean and known proportional covariance matrices.

Dimension	p	2 , 8 , 16 , 32
Proportionality	α	.1 , .5 , .9
Sample Sizes	(n_1,n_2)	(20,20) , (40,40) , (8,32) , (32,8)

Table 4.6.2

The Optimal Probabilities of Misclassification Q_t for Zero-mean and Proportional Covariance Matrices.

	α = .1			α = .5			α = .9		
p	Q_1	Q_2	$\bar{Q}$	Q_1	Q_2	$\bar{Q}$	Q_1	Q_2	$\bar{Q}$
2	.226	.077	.151	.500	.250	.375	.613	.349	.481
8	.021	.009	.015	.302	.197	.299	.525	.393	.459
16	.001	.001	.001	.196	.138	.167	.488	.395	.442
32	.000	.000	.000	.098	.072	.085	.450	.385	.417

Table 4.6.3

Expectations of the actual and apparent probabilities of misclassification of $\hat{Q}(x)$ for equal sample sizes, zero-mean and known proportional covariance matrices.

		$n_1 = n_2 = 20$						$n_1 = n_2 = 40$					
	p	$E(Q_1^*)$	$E(Q_2^*)$	$E(Q_1^{**})$	$E(Q_2^{**})$	$\bar{E}(Q^*)$	$\bar{E}(Q^{**})$	$E(Q_1^*)$	$E(Q_2^*)$	$E(Q_1^{**})$	$E(Q_2^{**})$	$\bar{E}(Q_1^*)$	$\bar{E}(Q_2^{**})$
$\alpha = .1$	2	.230	.086	.227	.064	.158	.145	.228	.082	.227	.071	.155	.149
	8	.022	.012	.021	.005	.017	.013	.021	.010	.021	.007	.015	.014
	16	.001	.001	.001	.000	.001	.001	.001	.001	.001	.000	.001	.001
	32	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
$\alpha = .5$	2	.515	.275	.484	.218	.395	.351	.509	.263	.493	.233	.386	.363
	8	.334	.237	.278	.139	.285	.208	.320	.219	.291	.165	.269	.228
	16	.235	.184	.171	.077	.209	.124	.217	.163	.184	.103	.190	.143
	32	.135	.117	.077	.027	.126	.052	.117	.096	.087	.044	.106	.065
$\alpha = .9$	2	.555	.434	.482	.352	.494	.417	.569	.417	.519	.361	.493	.440
	8	.514	.466	.356	.294	.490	.325	.519	.454	.409	.335	.486	.372
	16	.503	.470	.282	.233	.486	.257	.503	.459	.348	.291	.481	.319
	32	.492	.470	.197	.157	.481	.177	.489	.458	.295	.230	.473	.262

Table 4.6.4

Expectations of the actual and apparent probabilities of misclassification of $\hat{Q}(x)$ for unequal sample sizes, zero-mean and known proportional covariance matrices.

		$n_1 = 8, \; n_2 = 32$						$n_1 = 32, \; n_2 = 8$					
	p	$E(Q_1^*)$	$E(Q_2^*)$	$E(Q_1^{**})$	$E(Q_2^{**})$	$\bar{E}(Q^*)$	$\bar{E}(Q^{**})$	$E(Q_1^*)$	$E(Q_2^*)$	$E(Q_1^{**})$	$E(Q_2^{**})$	$\bar{E}(Q^*)$	$\bar{E}(Q^{**})$
$\alpha = .1$	2	.729	.076	.230	.064	.156	.147	.227	.105	.225	.052	.166	.138
	8	.024	.008	.022	.005	.016	.013	.028	.021	.020	.003	.024	.011
	16	.002	.001	.002	.000	.001	.001	.001	.003	.001	.000	.002	.001
	32	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
$\alpha = .5$	2	.549	.247	.476	.215	.398	.345	.489	.326	.471	.192	.407	.331
	8	.409	.180	.269	.129	.294	.199	.290	.338	.258	.101	.314	.179
	16	.332	.117	.162	.068	.225	.115	.184	.317	.150	.045	.250	.097
	32	.245	.054	.070	.021	.149	.045	.088	.276	.061	.010	.192	.036
$\alpha = .9$	2	.601	.387	.454	.345	.494	.399	.492	.503	.455	.334	.497	.394
	8	.631	.348	.311	.266	.489	.288	.391	.600	.314	.248	.495	.281
	16	.672	.300	.228	.195	.486	.211	.332	.556	.231	.175	.494	.203
	32	.729	.235	.138	.115	.482	.126	.260	.725	.142	.096	.492	.119

In Table 4.6.3 and 4.6.4 it is noted that in all cases the expectations of the actual probabilities of misclassification exceed the expectations of the apparent probabilities of misclassification i.e.

$$E(Q_t^*) > E(Q_t^{**}) \qquad t = 1 \text{ and } 2.$$

This is not surprising given the inherent bias in the apparent method of estimation. A similar result holds for linear discrimination and the Z statistic, Section 2.10. In all cases the total probabilities and expectations are ordered as

$$\bar{E}(Q^*) > \bar{Q} > \bar{E}(Q^{**})$$

for a similar linear result see Hills (1966). However even for $n_1 = n_2$

$$E(Q_t^*) \nmid Q_t \nmid E(Q_t^{**})$$

unlike the linear case or Z statistic, Section 2.8.

The effect of unequal sample sizes on the individual expectations may be seen in Table 4.6.4, the expectations of the actual rather than the apparent probabilities of misclassification being effected most. Comparing the results of Table 4.6.4 with these for $n_1 = n_2 = 20$ (Table 4.6.3) we see that $E(Q_1^*)$ has increased and $E(Q_2^*)$ decreased with some slight increase in $\bar{E}(Q^*)$, while $E(Q_1^{**})$ and $E(Q_2^{**})$ have both decreased. A similar pattern was noted by Moran (1974) for the sample linear discriminant function with Σ known. But is not the case for the Z statistic as shown in Section 2.11.

CHAPTER 5

An Analytical Comparison of Linear and Quadratic Discrimination with Proportional Covariance Matrices.

5.1 Introduction

A linear discriminant function is commonly used for discriminating between two populations mainly because of its simplicity of form and concept. For multinormally distributed populations with equal covariance matrices the optimal discriminant function is linear. Lack of normality of the populations is not considered here, but it is noted that Fisher's (1936) derivation of a linear discriminant function for populations with equal covariance matrices is distribution free, and that the optimal discriminant function in other cases may also be linear. The inequality of the covariance matrices in Normally distributed populations would suggest use of the optimal quadratic discriminant function but this is not the case in practice where the simplicity of the linear allocation rule and its "distribution free" derivation have given it wide usage even in non-optimal situations.

Some of the resistance to quadratic discrimination is due to the lack of results on its distribution and the behaviour of its error rates, the main work on these problems being concentrated on the linear discriminant function. However the main source of resistance is due to its poor performance in practice when the dimension is large and the sample sizes are small, given the need of estimating two covariance matrices. As early as 1947 Smith had considered the behaviour of linear and quadratic discriminant functions when applied to an example where the covariance matrices are unequal. One of the earliest analytical comparisons of linear and quadratic discrimination was undertaken by Gilbert (1969), who derived and evaluated the optimal error rates when the populations are multinormally distributed with proportional covariance matrices. By considering the difference between the quadratic and linear error rates for various selected values of the population parameters such as dimension, proportionality and separation, Gilbert was able to indicate the ranges of these parameters where an application of linear discrimination would give misleading results.

Gilbert's study required a knowledge of all the population parameters and Marks and Dunn (1974) extended her comparison to sample allocation rules; by simulation they considered the effect of various population parameters and sample sizes on applying linear discrimination to the quadratic situation of unequal convariance matrices. Wahl and Kronmal (1977) further elaborated on Marks and Dunn's study, the main extension being their use of larger sample sizes. Michaelis (1973) considered the effects of applying linear discrimination to more than two Normally distributed populations with unequal covariance matrices. Using actual data to give a range of population parameters for simulation purposes, Michaelis noted that certain behaviour of the error rates and their estimates indicate when an application of quadratic discrimination is appropriate.

In this chapter we compare linear and quadratic discriminant functions when applied to multinormally distributed populations with proportional covariance matrices. We do this firstly for all population parameters known using exact results derived in previous chapters. Then the sample discriminant functions are compared for known proportional covariance matrices. Thus our study is midway between that of Gilbert who assumed all population parameters known and the simulation studies of Marks and Dunn, Wahl and Kronmal and Michaelis where all the population parameters are assumed unknown. The expectations of the actual and apparent probabilities of misclassification of the linear discriminant function are derived for the assumption of known proportional covariance matrices and compared to the corresponding quadratic expectations. Some comments on when an application of quadratic discrimination is appropriate conclude the chapter.

5.2 Population Assumptions and the Optimal Probabilities of Misclassification

It is assumed that the distribution of x in Π_t is $N_p(\mu_t, \Sigma_t)$ (t=1,2) and that the covariance matrices are proportional i.e. $\alpha\Sigma_1 = \Sigma_2$, $0 < \alpha < 1$. Equal prior probabilities, $\pi_1 = \pi_2 = \frac{1}{2}$ are also assumed and the joint covariance matrix Σ is given by

$$\begin{aligned}\Sigma &= \pi_1\Sigma_1 + \pi_2\Sigma_2 \\ &= \tfrac{1}{2}(\Sigma_1 + \Sigma_2) \\ &= \frac{1+\alpha}{2}\,\Sigma_1.\end{aligned}$$

With these assumptions the quadratic Q(x) and linear L(x) discriminant functions are given by

$$Q(x) = -\tfrac{1}{2}(x-\mu_1)'\,\Sigma_1^{-1}\,(x-\mu_1) + \tfrac{1}{2}\,\frac{1}{\alpha}\,(x-\mu_2)'\,\Sigma_1^{-1}\,(x-\mu_2) + \tfrac{1}{2}\,p\ln\alpha \gtrless 0$$

and

$$L(x) = \frac{2}{1+\alpha}\,(\mu_1-\mu_2)'\,\Sigma_1^{-1}\,\{x - \tfrac{1}{2}(\mu_1+\mu_2)\} \gtrless 0.$$

As both Q(x) and L(x) are invariant to linear transformations the following canonical forms of the populations may be assumed without loss of generality, Section 3.3.

x in Π_1 is distributed as $N_p(o,I)$ and in Π_2 as $N_p(\nu,\alpha I)$.

Where the mean vector $\nu = (\Delta_1,o,o,\ldots,o)'$ with

$$\Delta_t^2 = (\mu_1-\mu_2)'\ \Sigma_t^{-1}(\mu_1-\mu_2) \qquad t = 1 \text{ and } 2.$$

Hence $\Sigma=\frac{1+\alpha}{2}$ I and the separation Δ where $\Delta^2 = (\mu_1-\mu_2)'\ \Sigma^{-1}(\mu_1-\mu_2)$ becomes

$$\Delta^2 = \frac{2}{1+\alpha}\ \nu'\nu = \frac{2}{1+\alpha}\ \Delta_1^2 .$$

Q(x) and L(x) may now be written as

$$Q(x) = \tfrac{1}{2}\{\frac{1-\alpha}{\alpha}\ (x - \frac{\nu}{1-\alpha})'\ (x - \frac{\nu}{1-\alpha}) - \frac{\nu'\nu}{1-\alpha} + p\ \ell n\ \alpha\} \gtrless 0$$

and

$$L(x) = -(\frac{2}{1+\alpha})\ \nu'\{x - \tfrac{1}{2}\nu\} \gtrless 0$$

with optimal probabilities of misclassification for Q(x) given in Section 3.6 by

$$Q_1 = \Pr\{\chi^2(p, \frac{1+\alpha}{(1-\alpha)^2}\ \Delta^2/2) < \frac{\alpha}{1-\alpha}\ H\}$$

$$Q_2 = \Pr\{\chi^2(p, \frac{\alpha(1+\alpha)}{(1-\alpha)^2}\ \Delta^2/2) > \frac{1}{1-\alpha}\ H\}$$

where $H = -p\ell n\ \alpha + \frac{1+\alpha}{1-\alpha}\ \Delta^2/2$,

and for L(x) by

$$L_1 = \Phi(-\sqrt{\frac{1+\alpha}{2}}\ \Delta/2)$$

$$L_2 = \Phi(-\sqrt{\frac{1+\alpha}{2\alpha}}\ \Delta/2).$$

It is interesting to note that in the present case of proportional covariance matrices, of all linear allocation rules, L(x) minimises the total probability of misclassification, Anderson and Bahadur (1962).

5.3 A Comparison of Linear and Quadratic Discrimination for Proportional Covariance Matrices with All Population Parameters Known.

For various separations Δ, dimension p and proportionality α we compare the error rates Q_t and L_t of Section 5.2, where the populations are multinormally distributed with proportional covariance matrices. Imhof's (1961) closed expressions for a non-central chi-square, Section 3.6, were used to evaluate Q_t. Such a comparison was undertaken by Gilbert (1969) who however approximated Q_t by Patnaik's (1949) approximation to a non-central chi-square, Section 3.6. The criteria of comparison used by Gilbert was the difference in the total probabilities of misclassification $\bar{Q}$ and $\bar{L}$ where $\bar{Q} = \frac{1}{2}(Q_1+Q_2)$ and $\bar{L} = \frac{1}{2}(L_1+L_2)$. In Marks and Dunn's (1974) simulation comparison of $\hat{L}(x)$ and $\hat{Q}(x)$ the case of proportionate covariance matrices was included as was a comparison of Q_t and L_t, the former being estimated from the simulation runs. Their criterion of comparison was the ratio of $\bar{Q}$ to $\bar{L}$. Wahl and Kronmal's (1977) elaboration of Marks and Dunn's study also included a comparison of Q_t and L_t, where Q_t was, as in Gilbert (1969) approximated by Patnaik's (1949) approximation.

For the range of parameters listed in Table 5.3.1 the optimal probabilities of misclassification Q_t and L_t as given in Section 5.2 were evaluated. Patnaik's (1949) approximation to Q_t, Section 3.6, given by

$$Q_1 \approx \Pr\{\chi^2_{k_1} < \frac{\alpha H}{(1-\alpha)g_1}\} \text{ and } Q_2 \approx \Pr\{\chi^2_{k_2} > \frac{H}{(1-\alpha)g_2}\}$$

Table 5.3.1

Range of Parameters used in Comparing Q_t and L_t for Proportional Covariance Matrices.

Dimension	p	3 , 5 , 7 , 9
Proportionality	α	.1 , .3 , .5 , .7 , .9
Separation	Δ	1.0488 , 1.6832 , 2.5631*

* The separations Δ in Table 5.3.1 were chosen to give (when α=1) optimal linear error rates of 30%, 20% and 10% respectively.

Table 5.3.2

Optimal Probabilities of Misclassification of the Linear Discriminant Function for the case of Proportional Covariance Matrices.

Δ	1.0488			1.6832			2.5631		
α	L_1	L_2	$\bar{L}$	L_1	L_2	$\bar{L}$	L_1	L_2	$\bar{L}$
.1	.349	.109	.229	.266	.024	.145	.171	.001	.086
.3	.336	.220	.278	.249	.108	.178	.151	.030	.090
.5	.325	.260	.293	.233	.151	.192	.134	.058	.096
.7	.314	.282	.298	.219	.177	.198	.119	.079	.099
.9	.305	.295	.300	.206	.194	.200	.106	.094	.100

where

$$k_1 = \frac{1}{(1-\alpha)^2}\left[\frac{\{p(1-\alpha)^2 + (1+\alpha)\ \Delta^2/2\}^2}{p(1-\alpha)^2 + (1+2)\ \Delta^2}\right] \qquad g_1 = \frac{(1-\alpha)^2 p + (1+\alpha)\ \Delta^2}{(1-\alpha)^2 p + (1+\alpha)\ \Delta^2/2}$$

$$k_2 = \frac{1}{(1-\alpha)^2}\left[\frac{\{p(1-\alpha)^2 + \alpha(1+\alpha)\ \Delta^2/2\}^2}{p(1-\alpha)^2 + \alpha(1+\alpha)\ \Delta^2}\right] \qquad g_2 = \frac{(1-\alpha)^2 p + \alpha(1+\alpha)\ \Delta^2}{(1-\alpha)^2 p + \alpha(1+\alpha)\ \Delta^2/2}$$

was also calculated.

In Table 5.3.2 the linear probabilities of misclassification L_t, (t=1 and 2), and their total $\bar{L}$ are given. In Table 5.3.3 the quadratic probabilities of misclassification Q_t, (t=1 and 2) are listed as are

DIF = Difference in the total probabilities of misclassification i.e. $\bar{L} - \bar{Q}$.

R = $\bar{Q}/\bar{L}$ x 100

† = case where percentage error in Patnaik's approximation to Q_t is in excess of ten percent i.e.

$$(1 - \frac{\text{approx}}{Q_t}) \times 100 > 10.$$

In Table 5.3.2 we note as may be seen from the formulae Section 5.2, that $L_2 < L_1$ for all $\Delta > 0$ and $\alpha < 1$ and that as α approaches one, L_1 and L_2 converge with L_1 decreasing and L_2 increasing.

As R is always $\leqslant$ 100 it is the size of the ratio in Table 5.3.3 that is of interest. With fixed p and Δ the ratio approaches 100 as α approaches one, since the quadratic model is tending to the linear. With fixed Δ and α, the ratio decreases with increasing dimension p, as the quadratic probabilities allow for increasing dimension the linear do not. With p and α fixed the ratio is fairly insensitive to increasing separation Δ.

Table 5.3.3

Optimal Probabilities of Misclassification of the Quadratic Discriminant Function
for the case of Proportional Covariance Matrices

	Δ	1.0488				1.6832				2.5631			
	α	Q_1	Q_2	DIF	R	Q_1	Q_2	DIF	R	Q_1	Q_2	DIF	R
p = 3	.1	.117	.042	.150	35	.086	.029	.088	39	.043†	.014	.058	33
	.3	.262	.122	.086	69	.185	.086	.043	76	.092†	.044	.022	75
	.5	.328	.183	.037	87	.218	.130	.018	91	.107	.068	.008	91
	.7	.338	.236	.011	96	.220	.165	.005	97	.108	.085	.002	98
	.9	.316	.282	.001	100	.208	.191	.000	100	.103	.096	.000	100
p = 5	.1	.052	.020	.193	16	.039	.015	.119	18	.020†	.007	.073	16
	.3	.192	.099	.133	52	.139	.071	.073	59	.072†	.036	.036	60
	.5	.286	.171	.064	78	.196	.121	.033	83	.098	.063	.015	84
	.7	.324	.231	.021	93	.214	.162	.010	95	.106	.083	.004	96
	.9	.315	.281	.002	99	.208	.190	.001	100	.103	.096	.000	100
p = 7	.1	.025	.010	.212	8	.018	.007	.132	9	.010†	.004	.079	8
	.3	.146	.079	.166	40	.107	.057	.096	46	.057†	.030	.047	48
	.5	.253	.159	.087	70	.178	.113	.047	76	.090	.059	.021	78
	.7	.312	.226	.029	90	.208	.158	.015	92	.103	.082	.006	94
	.9	.314	.280	.003	99	.207	.190	.001	99	.103	.096	.001	99
p = 9	.1	.019	.005	.221	4	.009	.004	.139	4	.005†	.002	.083	4
	.3	.113	.064	.190	32	.084	.046	.113	37	.045†	.024	.055	39
	.5	.226	.146	.106	64	.161	.105	.057	69	.083	.055	.027	72
	.7	.300	.221	.038	87	.202	.155	.019	90	.101	.080	.008	92
	.9	.313	.279	.004	99	.207	.189	.002	99	.103	.096	.001	99

One concludes from the difference columns of Table 5.3.3 that an application of the linear discriminant function to the quadratic case of proportional covariance matrices, will when $\Delta = 1.0488$ and $\alpha < .5$ result in an exaggeration of the true total probability of misclassification of at least .05. This exaggeration will increase as α gets smaller and or p increases. With increasing Δ to maintain this error of .05, α must be smaller than .5.

Similar trends may be seen in Gilbert's (1969) figures. However Gilbert, Marks and Dunn (1974) and Wahl and Kronmal (1977) concentrated on the total probabilities of misclassification and not on the behaviour of the individual probabilities Q_t and L_t. We note that in general Q_1 increases with increasing α unlike L_1 and is larger than L_1 when $\alpha = .9$. Use by Gilbert of Patnaik's (1949) approximation in evaluating Q_t leads to a positive percentage error in excess of 10% in nine cases indicated by † in Table 5.3.3. The error in these cases was to be expected since it is known that Patnaik's approximation is poor when estimating the lower tail probabilities of a non-central chi-square distribution, Johnson and Kotz (1970, vol 2, P142). Use of Patnaik's approximation does not alter the above comments.

5.4 Sample Linear and Quadratic Discriminant Functions for Known Proportional Covariance Matrices. The Expectations of the Actual and Apparent Probabilities of Misclassification of $\hat{L}(x)$.

The comparison of Section 5.3 will be extended in the following section to the case where the unknown population parameters of the discriminant functions Q(x) and L(x) are replaced by their sample estimates. Assuming that the populations have known proportional covariance matrices and retaining the canonical forms $N_p(0,I)$ and $N_p(\nu,\alpha I)$ of Section 5.2 with random samples of size n_t from Π_t and equal prior-probabilities here also, the sample quadratic and linear discriminant functions $\hat{Q}(x)$ and $\hat{L}(x)$ are given by

$$\hat{Q}(x) \ = \ -\tfrac{1}{2}(x-\bar{x}_1)'\ (x-\bar{x}_1) + \tfrac{1}{2}\,\frac{1}{\alpha}\,(x-\bar{x}_2)'\ (x-\bar{x}_2) + \tfrac{1}{2}\ p\ln\alpha \ \gtrless 0$$

and

$$\hat{L}(x) \ = \ \frac{2}{1+\alpha}\ (\bar{x}_1-\bar{x}_2)'\ \{x - \tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\} \ \gtrless \ 0.$$

In comparing $\hat{Q}(x)$ and $\hat{L}(x)$ the expectations of the actual and apparent probabilities of misclassification i.e. $E(Q_t^*)$, $E(Q_t^{**})$, $E(L_t^*)$ and $E(L_t^{**})$ will be required. Expressions for the quadratic expectations were derived in Sections 4.3 and 4.4. Similar expressions for the linear expectations are now derived. The method of derivation being similar to that used in previous chapters only an outline derivation is presented here.

Expectations of the Actual Probabilities of Misclassification of $\hat{L}(x)$ for Known Proportional Covariance Matrices.

$$E(L_1^*) = \Pr\{\hat{L}(x) < 0 \mid x \text{ in } \Pi_1\}$$
$$= \Pr\{u'v < 0\}$$

where $u = c_1\sqrt{\frac{2}{1+\alpha}}\,(\bar{x}_1-\bar{x}_2)$ and $v = c_2\sqrt{\frac{2}{1+\alpha}}\,\{x - \tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\}$ with

$$c_1 = \{(1+\alpha)\, n_1n_2/\ 2(\alpha n_1+n_2)\}^{\frac{1}{2}}$$

and

$$c_2 = \{(1+\alpha)\, 4n_1n_2/\ 2(\alpha n_1+n_2+4n_1n_2)\}^{\frac{1}{2}}$$

chosen to give u and v covariance matrix I.

Now $$E(L_1^*) = \Pr\{(u+v)'(u+v) - (u-v)'(u-v) < 0\}$$
$$= \Pr\{(1+\rho_1)\,\omega_1 - (1-\rho_1)\,\omega_2 < 0\}$$

where $\omega_1 = (u+v)'(u+v)/2(1+\rho_1)$ and $\omega_2 = (u-v)'(u-v)/2(1-\rho_1)$

with $\rho_1 = (\alpha n_1-n_2)\ /\ \{(\alpha n_1+n_2)(\alpha n_1+n_2+4n_1n_2)\}^{\frac{1}{2}}$.

As ω_1 and ω_2 are independently distributed as non-central chi-squares $\chi^2(p,\delta_1)$ and $\chi^2(p,\delta_2)$ where

$$\delta_1 = \{2(1+\rho_1)\}^{-1}\,(c_1+c_2/2)^2\ \Delta^2$$

$$\delta_2 = \{2(1-\rho_1)\}^{-1}\,(c_1-c_2/2)^2\ \Delta^2$$

then $$E(L_1^*) = \Pr\ \{(1+\rho_1)\,\chi^2(p,\delta_1) - (1-\rho_1)\,\chi^2(p,\delta_2) < 0\}. \qquad (5.4.1)$$

Similarly with x in Π_2

$$E(L_2^*) = \Pr\{(1+\rho_2)\ \chi^2(p,\delta_3) - (1-\rho_2)\ \chi^2(p,\delta_4) > 0\} \tag{5.4.2}$$

where $c_3 = c_1 = \{(1+\alpha)\ n_1 n_2 \ /\ 2\ (\alpha n_1+n_2)\}^{\frac{1}{2}}$

$c_4 = \{(1+\alpha)\ 4n_1 n_2 \ /\ 2\ (\alpha n_1+n_2+4\alpha n_1 n_2)\}^{\frac{1}{2}}$

$\rho_2 = (\alpha n_1-n_2)\ /\ \{(\alpha n_1+n_2)\ (\alpha n_1+n_2+4\alpha n_1 n_2)\}^{\frac{1}{2}}$

$\delta_3 = \{2(1+\rho_2)\}^{-1}\ (c_3-c_4/2)^2\ \Delta^2$

$\delta_4 = \{2(1-\rho_2)\}^{-1}\ (c_3+c_4/2)^2\ \Delta^2$.

It is noted that Moran's (1974) expressions for $E(L_t^*)$ (Section 2.8) when the covariance matrices are equal may be obtained from (5.4.1) and (5.4.2) by letting $\alpha = 1$.

Expectations of the Apparent Probabilities of Misclassification of $\hat{L}(x)$ for Known Proportional Covariance Matrices.

The expectations of the apparent probabilities of misclassification $E(L_t^{**})$ of $\hat{L}(x)$ as defined in Section 1.5 may be obtained in a manner similar to the expectations of the actual probabilities of misclassification. Moran's (1974) expression for $E(L_t^{**})$ when $\Sigma_1 = \Sigma_2$ may once again be obtained by letting $\alpha=1$.

Thus $E(L_1^{**}) = \Pr\{(1+\rho_3)\ \chi^2(p,\delta_5) - (1-\rho_3)\ \chi^2(p,\delta_6) < 0\}$

where $c_5 = c_1 = \{(1+\alpha)\ n_1 n_2 \ /\ 2\ (\alpha n_1+n_2)\}^{\frac{1}{2}}$

$c_6 = \{(1+\alpha)\ 4n_1 n_2 \ /\ 2\ (\alpha n_1-3n_2+4n_1 n_2)\}^{\frac{1}{2}}$

$\rho_3 = \{(\alpha n_1+n_2)\ /\ (\alpha n_1-3n_2+4n_1 n_2)\}^{\frac{1}{2}}$

$$\delta_5 = \{2(1+\rho_3)\}^{-1} \; (c_5+c_6/2)^2 \;\; \Delta^2$$

$$\delta_6 = \{2(1-\rho_3)\}^{-1} \; (c_5-c_6/2)^2 \;\; \Delta^2$$

and $$E(L_2^{**}) = \Pr\{(1+\rho_4)\; \chi^2(p,\delta_7) - (1-\rho_4)\; \chi^2(p,\delta_8) > 0\}$$

where $$c_7 = c_5 = \{(1+\alpha)\; n_1 n_2/2(\alpha n_1+n_2)\}^{\frac{1}{2}}$$

$$c_8 = \{(1+\alpha)\; 4n_1 n_2/2(-3\alpha n_1+n_2+4\alpha n_1 n_2)\}^{\frac{1}{2}}$$

$$\rho_4 = \{-(\alpha n_1+n_2)/(-3\alpha n_1+n_2+4\alpha n_1 n_2)\}^{\frac{1}{2}}$$

$$\delta_7 = \{2(1+\rho_4)\}^{-1} \; (c_7-c_8/2)^2 \;\; \Delta^2$$

$$\delta_8 = \{2(1-\rho_4)\}^{-1} \; (c_7+c_8/2)^2 \;\; \Delta^2 .$$

5.5 A Comparison of Sample Linear and Quadratic Discrimination for Known Proportional Covariance Matrices.

Our comparison here will involve the probabilities, expectations and totals

$$Q_t, L_t, \quad E(Q_t^*), \quad E(L_t^*), \quad E(Q_t^{**}), \quad E(L_t^{**}) \qquad t = 1 \text{ and } 2$$

$$\bar{Q}, \bar{L}, \quad \bar{E}(Q^*), \quad \bar{E}(L^*), \quad \bar{E}(Q^{**}), \quad \bar{E}(L^{**}) = \tfrac{1}{2}\{E(L_1^{**}) + E(L_2^{**})\} \qquad (5.5.1)$$

for various values of the parameters (n_1,n_2), p, α and Δ. The number of probabilities and parameters involved limits the range of results that may be presented. Table 5.5.1 gives the range of parameters used. These were chosen in the light of the simulation of Marks and Dunn (1974) and Wahl and Kronmal (1977) and the range of parameters considered in Section 5.3.

Table 5.5.1

Range of Parameters used in Comparing Sample Linear and Quadratic Discrimination for Known Proportional Covariance Matrices.

Dimension	p	3 , 9		
Proportionality	α	.1 , .5 , .9		
Sample Size	(n_1,n_2)	(25,25)	(100,100)	
Separation	$\tilde{\Delta}$	1.0488 ,	1.6832 ,	2.5631

The various expectations and totals (5.5.1) are given in Table 5.5.2, the exact formulae for the expectations may be found in Sections 4.3, 4.4 and 5.4. The differences D^* and D^{**} where

$$D^* = \bar{E}(L^*) - \bar{E}(Q^*) \quad \text{and} \quad D^{**} = \bar{E}(L^{**}) - \bar{E}(Q^{**})$$

are given in Table 5.5.3.

Table 5.5.2

Totals and Expectations of the Actual and Apparent Probabilities of Misclassification of the Sample Quadratic and Linear Discriminant functions for known Proportional Covariance Matrices.

Sample Sizes (n_1,n_2) = (25,25)

	α	p	$E(Q_1^*)$	$E(L_1^*)$	$E(Q_1^{**})$	$E(L_1^{**})$	$E(Q_2^*)$	$E(L_2^*)$	$E(Q_2^{**})$	$E(L_2^{**})$	$\bar{Q}$	$\bar{E}(Q^*)$	$\bar{E}(L^*)$	$\bar{E}(Q^{**})$	$\bar{E}(L^{**})$
Δ = 1.0488	.1	3	.120	.377	.118	.336	.047	.111	.034	.100	.080	.088	.244	.076	.218
		9	.013	.440	.012	.309	.007	.084	.003	.065	.009	.010	.262	.008	.187
	.5	3	.336	.346	.319	.311	.201	.269	.166	.244	.256	.269	.308	.243	.278
		9	.246	.389	.211	.280	.176	.275	.112	.207	.186	.211	.332	.162	.244
	.9	3	.331	.321	.302	.290	.297	.309	.265	.280	.299	.314	.315	.284	.285
		9	.361	.351	.266	.257	.325	.333	.231	.246	.291	.338	.342	.249	.252
Δ = 1.6832	.1	3	.087	.283	.085	.251	.033	.028	.024	.025	.058	.060	.156	.055	.141
		9	.009	.325	.009	.241	.005	.022	.002	.017	.007	.007	.174	.006	.129
	.5	3	.224	.245	.213	.224	.141	.157	.119	.143	.174	.183	.201	.166	.184
		9	.175	.274	.151	.206	.126	.163	.082	.125	.133	.151	.219	.117	.166
	.9	3	.217	.216	.200	.197	.200	.202	.181	.185	.200	.209	.209	.191	.191
		9	.232	.235	.181	.179	.217	.218	.161	.166	.198	.225	.227	.171	.173
Δ = 2.5631	.1	3	.044	.180	.042	.164	.016	.002	.011	.002	.029	.030	.091	.027	.083
		9	.005	.204	.005	.156	.003	.002	.001	.001	.004	.004	.103	.003	.079
	.5	3	.111	.140	.104	.128	.074	.062	.062	.055	.088	.093	.101	.083	.092
		9	.090	.155	.078	.119	.066	.064	.044	.049	.069	.018	.110	.061	.084
	.9	3	.108	.111	.099	.101	.101	.099	.091	.089	.100	.105	.105	.095	.095
		9	.115	.120	.091	.093	.109	.106	.082	.082	.100	.112	.113	.081	.088

Table 5.5.2 (continued)

Sample Sizes (n_1,n_2) = (100,100)

	α	p	$E(Q_1^*)$	$E(L_1^*)$	$E(Q_1^{**})$	$E(L_1^{**})$	$E(Q_2^*)$	$E(L_2^*)$	$E(Q_2^{**})$	$E(L_2^{**})$	$\bar{Q}$	$\bar{E}(Q^*)$	$\bar{E}(L^*)$	$E(Q^{**})$	$E(L^{**})$
Δ = 1.0488	.1	3	.118	.356	.117	.346	.043	.110	.040	.107	.080	.081	.233	.078	.177
		9	.012	.376	.012	.338	.005	.102	.004	.096	.009	.009	.289	.008	.217
	.5	3	.330	.330	.325	.321	.188	.262	.179	.256	.256	.259	.296	.252	.289
		9	.232	.344	.223	.313	.154	.266	.139	.245	.186	.193	.305	.176	.279
	.9	3	.320	.304	.313	.301	.285	.299	.277	.291	.299	.303	.304	.295	.296
		9	.324	.318	.300	.292	.293	.307	.266	.282	.291	.309	.313	.283	.287
Δ = 1.6832	.1	3	.086	.270	.086	.264	.030	.035	.028	.024	.058	.058	.148	.072	.144
		9	.009	.282	.009	.260	.004	.024	.003	.022	.007	.007	.153	.006	.141
	.5	3	.220	.236	.217	.231	.133	.153	.128	.149	.174	.177	.195	.173	.190
		9	.165	.244	.159	.226	.110	.154	.098	.144	.133	.138	.199	.129	.185
	.9	3	.210	.208	.206	.204	.193	.196	.188	.191	.200	.202	.202	.197	.198
		9	.213	.214	.200	.199	.197	.200	.182	.186	.198	.205	.207	.191	.194
Δ = 2.5631	.1	3	.043	.173	.043	.169	.014	.001	.013	.001	.029	.028	.087	.028	.085
		9	.005	.179	.005	.167	.002	.001	.002	.001	.004	.004	.090	.004	.084
	.5	3	.108	.135	.106	.132	.069	.059	.067	.057	.088	.089	.097	.081	.095
		9	.085	.139	.082	.130	.058	.060	.052	.056	.069	.072	.100	.067	.093
	.9	3	.104	.107	.102	.105	.097	.095	.095	.093	.100	.101	.101	.098	.094
		9	.106	.109	.100	.102	.099	.097	.092	.091	.100	.103	.103	.096	.092

From Tables 5.3.2, 5.3.3 and 5.5.2 we note that the following inequalities held in general

$$\begin{aligned} &E(Q_t^*) > Q_t \geq E(Q_t^{**}) \\ &\text{and} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.5.2) \\ &E(L_t^*) > L_t > E(L_t^{**}) \qquad t = 1 \text{ and } 2 \end{aligned}$$

the bounds becoming tighter with increasing sample size. The linear expectations $E(L_t^*)$ and $E(L_t^{**})$ do not in general contain the optimal probabilities of misclassification Q_t except when α is close to one. This is to be expected since as α approaches one the quadratic and linear models converge and for $n_1 = n_2$ and $\Sigma_1 = \Sigma_2$ the inequality $E(L_t^*) > L_t > E(L_t^{**})$ holds, Section 2.8.

When $\Sigma_1 = \Sigma_2$ and $n_1 = n_2$, $E(L_1^*) = E(L_2^*)$ and $E(L_1^{**}) = E(L_2^{**})$. Hence the considerable inequality displayed by $E(L_t^*)$ and $E(L_t^{**})$ in Table 5.5.2, when α is small, might be used in practice when $n_1 = n_2$ to indicate that an application of quadratic discrimination is appropriate.

For the total probabilities Table 5.5.2 inequalities corresponding to the individual ones (5.5.2) hold

$$\bar{E}(Q^*) \geq \bar{Q} \geq \bar{E}(Q^{**})$$

and

$$\bar{E}(L^*) \geq \bar{L} \geq \bar{E}(L^{**})$$

as do comments on when the optimal total quadratic probabilities will be bounded by the linear expectations. The additional inequalities

$$\bar{E}(L^*) > \bar{E}(Q^*) \quad \text{and} \quad \bar{E}(L^{**}) > \bar{E}(Q^{**})$$

also hold.

Considerable difference in the latter may indicate when $n_1 = n_2$ that an application of quadratic discrimination is appropriate. Michaelis(1973) in his simulation of $\hat{Q}(x)$ and $\hat{L}(x)$, Σ_t unknown has noted similar inequalities to those presented here.

Table 5.5.3

Differences in the Expectations of the Actual and Apparent Probabilities of Misclassification of the Sample linear and Quadratic Allocation Rules for Known Proportional Covariance Matrices.

			$(n_1,n_2) = (25,25)$		$(n_1,n_2) = (100,100)$	
	α	p	D*	D**	D*	D**
Δ = 1.0488	.1	3	.161	.142	.153	.078
		9	.252	.180	.230	.209
	.5	3	.039	.035	.037	.037
		9	.121	.082	.062	.103
	.9	3	.001	.002	.002	.001
		9	.004	.003	.004	.004
Δ = 1.6832	.1	3	.096	.087	.090	.092
		9	.167	.124	.147	.110
	.5	3	.019	.018	.018	.018
		9	.068	.049	.062	.057
	.9	3	.001	.001	.001	.001
		9	.002	.002	.002	.003
Δ = 2.5631	.1	3	.061	.057	.059	.057
		9	.099	.076	.087	.081
	.5	3	.009	.009	.009	.008
		9	.032	.023	.028	.026
	.9	3	.001	.000	.001	.001
		9	.001	.001	.001	.001

For the differences in the total actual and apparent expectations, Table 5.5.3, we note that for (n_1,n_2), Δ and p fixed the differences decrease as α approaches one since the two models converge. With (n_1,n_2), p and α fixed, the differences decrease with increasing separation Δ of the populations. For p, α and Δ fixed the decrease is slight with increasing sample sizes; this is in contrast to the simulation studies of Marks and Dunn (1974) and Wahl and Kronmal (1977) with Σ_t unknown. Here however as Σ_t are known the linear and quadratic models need only estimate the same number of parameters which is not the case in practice. With α, Δ and (n_1,n_2) fixed the differences increase with increasing p due possibly to the quadratic models utilisation of the p differences in the covariances.

In conclusion we note that with known proportional covariance matrices, if α is small the sample quadratic discriminant function is superior to the linear, becoming more so with increasing p. For moderate α the sample quadratic rule is better than the linear especially for large p, and for α close to one the sample quadratic rule is only slightly better than the linear, improving with increasing p. With known covariance matrices, increasing sample size has a small but similar effect on the sample linear and quadratic allocation rules. This will not be the case when the covariance matrices are unknown, where sample size relative to dimension size has a greater effect on the quadratic allocation rule.

CHAPTER 6

ESTIMATORS OF LOG-ODDS

6.1 INTRODUCTION

Allocation in discriminant analysis can be effected without assessing the odds of an observation coming from one rather than another of the populations, as for instance a method based on ranks described by Kendall & Stuart (1976, Vol. 3, pp 346-349). However observations with widely differing odds may be allocated to the same population. An assessment of the size of the odds would enable one to gauge the strength of the allocation.

The estimation of true log-odds, in the classical case of two multivariate Normal populations, with equal or unequal covariance matrices, is the problem considered in the latter half of this thesis. Our reasons for estimating log-odds rather than odds are; the distribution of true odds is skew whereas the distribution of true log-odds is, for equal covariance matrices, univariate Normal. The natural bias of odds is a multiplicative factor, with log-odds the bias is additive. We will also see that the log-transformation renders the estimation problems amenable to standard statistical techniques. Our concentration on point estimation rather than interval estimation reflects the intractability of the distributions of the estimators, with only some asymptotic results available in the literature.

Point estimation leads to a consideration of bias and mean square error. Bias is of interest here as it is directly related to the misclassification of observations, its sign and relative size to true log-odds being important.

In this chapter we establish our notation and review some of the methods of estimating true log-odds proposed in the literature.

6.2 True Log-Odds

Let an unidentified observation x come from one of two populations Π_t (t=1,2) with probability density functions f_t. If the prior probabilities of x from Π_t are π_t, the posterior probability $Pr(\Pi_t|x)$ of coming from Π_t given x is by Bayes theorem,

$$Pr(\Pi_t|x) = \frac{\pi_t \ f_t(x)}{\pi_1 \ f_1(x) + \pi_2 \ f_2(x)} \qquad t = 1,2 \qquad (6.2.1)$$

and thus the log-odds in favour of x from Π_1 is

$$\ell n \ \{\frac{Pr(\Pi_1|x)}{Pr(\Pi_2|x)}\} = \ell n(\frac{\pi_1}{\pi_2}) + \ell n\{\frac{f_1(x)}{f_2(x)}\}. \qquad (6.2.2)$$

The observation x is allocated to Π_1 or Π_2 according as the log-odds (6.2.2) are $\gtrless 0$.

The posterior probabilities $Pr(\Pi_t|x)$ were described by Cornfield (1962) as the risk of belonging to Π_t given x. The logit of risk where

$$\text{logit } \{Pr(\Pi_1|x)\} = \ell n\{\frac{Pr(\Pi_1|x)}{1 - Pr(\Pi_1|x)}\}$$

$$= \ell n\{\frac{\pi_1 \ f_1(x)}{\pi_2 \ f_2(x)}\} \qquad (6.2.3)$$

is in fact the log-odds in favour of x from Π_1, Cox (1966).

If $f_t(x|\mu_t,\Sigma_t) = (2\pi)^{-\frac{1}{2}p}\ |\Sigma_t|^{-\frac{1}{2}}\ \exp\{-\frac{1}{2}\omega_t(x)\}$ (6.2.4)

where $\omega_t(x) = (x-\mu_t)'\ \Sigma_t^{-1}(x-\mu_t)$

the log-odds in favour of x from Π_1 (6.2.2), assuming equal prior-probabilities is

$$LO(T) = \ln\{\frac{f_1(x)}{f_2(x)}\}$$

where LO(T) denotes the phrase "log-odds true".

If $\Sigma_1 = \Sigma_2$ $\quad LO(T_e) = \frac{1}{2}\{-\omega_1(x) + \omega_2(x)\}$

$$= (\mu_1-\mu_2)'\Sigma^{-1}\{x-\tfrac{1}{2}(\mu_1+\mu_2)\}$$

and if $\Sigma_1 \neq \Sigma_2$ $\quad LO(T_u) = \frac{1}{2}\{-\omega_1(x) + \omega_2(x) - \ln(|\Sigma_1|/|\Sigma_2|)\}$,

where the subscripts e and u will denote whether $\Sigma_1 = \Sigma_2$ or $\Sigma_1 \neq \Sigma_2$. It is noted that $LO(T_e) = L(x)$ and $LO(T_u) = Q(x)$ the linear and quadratic discriminant functions of Section 1.2.

6.3 Estimators of True Log-Odds

Some of the methods of estimating LO(T) proposed in the literature are now considered. As LO(T) is a function of the probability densities f_t, a number of the methods reviewed are based on density estimates.

The Estimative Method:

The estimative or frequentist method corresponds to Anderson's (1958) method of replacing the unknown population parameters in the optimal allocation rules L(x) and Q(x) by their sample estimates. For $\Sigma_1 \neq \Sigma_2$ the density estimator of f_t obtained on this basis is

$$r_t(x|\bar{x}_t, S_t) = (2\pi)^{-\frac{1}{2}p} |S_t|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2} w_t(x)\}$$

where $w_t(x) = (x-\bar{x}_t)' S_t^{-1}(x-\bar{x}_t)$ is the sample counterpart of $\omega_t(x)$, (6.2.4). The resulting estimator of true log-odds is then

$$\begin{aligned} LO(E_u) &= \ln\{\frac{r_1(x)}{r_2(x)}\} \\ &= \tfrac{1}{2}\{-w_1(x) + w_2(x) - \ln(|S_1|/|S_2|)\}. \end{aligned}$$

For $\Sigma_1 = \Sigma_2$

$$r_t(x|\bar{x}_t,S) = (2\pi)^{-\frac{1}{2}p} |S|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}w_t(x)\}$$

with $\quad w_t(x) = (x-\bar{x}_t)' S^{-1} (x-\bar{x}_t),$

and $\quad LO(E_e) = \tfrac{1}{2}\{- w_1(x) + w_2(x)\}$

$$= (\bar{x}_1-\bar{x}_2)' S^{-1}\{x-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\}.$$

Again $LO(E_e) = \hat{L}(x)$ and $LO(E_u) = \hat{Q}(x)$ the sample linear and quadratic allocation rules, Section 1.3.

In the literature no specific justification apart from intuition and consistency has been given for the estimative approach. It is well known that LO(E) is a biased estimator of LO(T). This prompted consideration ofan unbiased estimator of true log-odds as well as an investigation into the size of the bias and its consequences for allocation. Mc Lachlan (1977) has derived the asymptotic bias of the estimative odds when $\Sigma_1 = \Sigma_2 = \Sigma$ is unknown.

An alternative frequentist estimator might be based on the uniform minimum variance unbiased (U.M.V.U.) estimator of f_t, viz.

$$u(x|\bar{x}_t,S_t) = \pi^{-\frac{1}{2}p}\{\frac{n_t}{(n_t-1)^2}\}^{\frac{1}{2}p} |S_t|^{-\frac{1}{2}} \{1-\frac{n_t}{(n_t-1)^2} w_t(x)\}^{\frac{1}{2}(n_t-p-3)}$$

$$\text{if } \frac{n_t}{(n_t-1)^2} w_t(x) < 1$$

$$= 0 \text{ otherwise,}$$

Ghurye and Olkin (1969). With large p and small n_t the $w_t(x)$ are often large enough to result in zero estimates for both densities and this approach seems unproductive.

The Predictive Method:

Here the true densities f_t in LO(T) are replaced by their corresponding Bayesian predictive densities. The predictive densities are obtained as follows. Assuming that the non-informative prior density for μ_t and Σ_t^{-1} is $g_t(\mu_t,\Sigma_t^{-1}) \propto |\Sigma_t|^{\frac{1}{2}(p+1)} d\mu_t \, d\Sigma_t^{-1}$, the posterior density

$h_t(\mu_t,\Sigma_t^{-1}|\bar{x}_t,S_t)$ is obtained in the usual manner. The predictive density of x given $\bar{x}_t,S_t$, $p_t(x|\bar{x}_t,S_t)$ is then derived as

$$p_t(x|\bar{x}_t,S_t) = \int\int f_t(x|\mu_t,\Sigma_t)\ h_t(\mu_t,\Sigma_t^{-1}|\bar{x}_t,S_t)\ d\mu_t\ d\Sigma_t^{-1}$$

Jeffries (1961, p139). For multinormally distributed populations the predictive density was given by Geisser (1964) as

$$p_t(x|\bar{x}_t,S_t) = \pi^{-\frac{1}{2}p}\ \frac{\Gamma(\frac{m_t+1}{2})}{\Gamma(\frac{m_t-p+1}{2})}\ (\frac{c_t}{m_t})^{\frac{1}{2}p}\ |S_t|^{-\frac{1}{2}}\{1+\frac{c_t}{m_t}\ w_t(x)\}^{-\frac{1}{2}(m_t+1)}$$

where $c_t = n_t/(n_t+1)$,

for $\Sigma_1 \neq \Sigma_2$, $m_t = n_t - 1$

and for $\Sigma_1 = \Sigma_2$, $m = m_1 = m_2 = n_1+n_2-2$ and S_t replaced by S.

The predictive density p_t belongs to the multivariate -t family of distributions.

The predictive log-odds is then

$$LO(P) = \ln\{\frac{p_1(x)}{p_2(x)}\}.$$

With $\Sigma_1 = \Sigma_2$ and $n_1 = n_2 = n$

$$LO(P_e) = -\tfrac{1}{2}(2n-1)\ \ln\{\frac{1+\frac{1}{2}\frac{n}{n^2-1}\ w_1(x)}{1+\frac{1}{2}\frac{n}{n^2-1}\ w_2(x)}\} \quad (6.3.1)$$

and for allocation purposes the predictive and estimative methods are now identical.

In the classic study of the authorship of the Federalist papers, Mosteller and Wallace (1963,1964), log-odds were extensively used. $LO(T_u)$ was estimated by the estimative and predictive approaches, the latter gave the smaller estimates, a confidence interval for $LO(T_u)$, p=1, was also given. Based on the asymptotic Normality of the standardised estimator $LO(E_e)$, Schaafsma and Van Vark (1977,1979) have obtained a confidence interval for $LO(T_e)$. Large differences in the log-odds quoted in practice by the estimative and predictive estimators were first noted by Aitchison and Kay (1973). Further examples of such differences in the estimators with small sample sizes were given by Aitchison and Dunsmore (1975, Ch 12) and Hermans and Habbema (1975). A limited simulation comparison of the estimators was undertaken by Aitchison, Habbema and Kay (1977), where the predictive estimator was distinctly superior. Mc Lachlan (1979) derived the asymptotic bias of the predictive odds when $\Sigma_1 = \Sigma_2 = \Sigma$ and compared it with the asymptotic bias of the estimative odds, Mc Lachlan (1977), he concludes that the predictive method is asymptotically less biased then the estimative.

The superiority of the predictive method must be due in part to the predictive density p_t being a better estimator of f_t than is the estimative density r_t. Adopting a measure of closeness based on the Kullback and Liebler (1951) information measure, viz

$$E_{x,\bar{x}_t,S_t}\{\ln\frac{p_t}{r_t}\} = E_{\bar{x}_t,S_t}[\int f_t(x|\mu_t,\Sigma_t)\ \ln\{\frac{p_t(x|\bar{x}_t,S_t)}{r_t(x|\bar{x}_t,S_t)}\}\ dx].$$

Aitchison (1975) showed this measure was positive and independent of the population parameters μ_t and Σ_t, this is interpreted as p_t being closer overall to f_t than is r_t. Murray (1977) showed that of all densities based on distances of the type $(x-y)' S_t^{-1}(x-y)$ and therefore invariant to non-singular linear transformations, the predictive density was closest to f_t in the sense of minimising $E_{x,\bar{x}_t S_t}\{\ell n(\frac{f_t}{p_t})\}$. If the estimated density is Normal i.e. $h_t(x) = N_p(\bar{x}_t,\alpha_t S_t)$, Murray (1979) showed that

$$E_{x,\bar{x}_t\, S_t}\{\ell n(\frac{f_t}{h_t})\}$$

is minimised

with $\qquad \alpha_t = \frac{n_t+1}{n_t}\,(\frac{n_t-1}{n_t-p-2}) \qquad$ for $\Sigma_1 \neq \Sigma_2$

and $\qquad \alpha_t = \frac{n_t+1}{n_t}(\frac{n_1+n_2-2}{n_1+n_2-p-3}) \qquad$ for $\Sigma_1 = \Sigma_2$,

with S_t replaced by S.

The bias and mean square error of the predictive estimator will be investigated in subsequent chapters and compared with alternatives.

Semi-Bayesian Method

The name Semi-Bayesian was used by Geisser (1967) to distinguish the predictive method from an alternative Bayesian approach. He viewed the problem of allocation as one of estimating LO(T) and suggested as estimator the posterior mean of LO(T) given x. For equal covariance matrices and the non-informative prior distribution $g_t(\mu_t, \Sigma_t^{-1}) \alpha |\Sigma_t|^{\frac{1}{2}(p+1)} d\mu_t \; d\Sigma_t^{-1}$ Geisser derived this posterior mean as

$$E\{LO(T_e)|x\} = LO(E_e) + \tfrac{1}{2}p(\frac{n_1-n_2}{n_1 n_2}).$$

The result was extended to the unequal covariance matrices case by Enis and Geisser (1970), to give

$$E\{LO(T_u)|x\} = LO(E_u) + \tfrac{1}{2}[p\ell n(\frac{n_2-1}{n_1-1}) + p(\frac{n_1-n_2}{n_1 n_2}) + \sum_{i=1}^{p} \{ \psi(\frac{n_1-i}{2}) - \psi(\frac{n_2-i}{2})\}]$$

where $\psi(a) = \Gamma'(a)/\Gamma(a)$ is the digamma function. They noted that with equal sample sizes, $n_1 = n_2$, the posterior means are unbiased and equivalent to the estimators given by the estimative approach.

The Likelihood Ratio Method

A likelihood ratio method of allocation was considered in Chapter 2 with $\Sigma_1 = \Sigma_2 = \Sigma$ and Σ known. This approach of choosing between the hypothesis x in Π_1 against the alternative hypothesis x in Π_2 results in a likelihood ratio estimator of posterior log-odds. If the hypothesis with the larger likelihood is favoured, then for equal prior probabilities or prior support of zero, Edwards (1972, p36), the likelihood ratio estimator or posterior support, Edwards (p36 and Ch.9) is from Chapter 2 with $\Sigma_1 = \Sigma_2$

$$LO(LR_e) = -\tfrac{1}{2}(n_1+n_2+1)\ \ell n \left[\frac{1 + \frac{n_1}{(n_1+1)(n_1+n_2-2)} w_1(x)}{1 + \frac{n_1}{(n_2+1)(n_1+n_2-2)} w_2(x)} \right] . \tag{6.3.2}$$

Similarly for $\Sigma_1 \neq \Sigma_2$

$$LO(LR_u) = \tfrac{1}{2}\ [p\ell n\{\frac{(n_1+1)(n_2-1)}{(n_2+1)(n_1-1)} - n_1\ p\ell n(\frac{n_1}{n_1+1}) + n_2\ p\ell n(\frac{n_2}{n_2+1})$$

$$- \ell n(|S_1|/|S_2|) - (n_1+1)\ \ell n\{1 - \frac{n_1}{n_1^2-1} w_1(x)\}$$

$$+ (n_2+1)\ \ell n\{1 - \frac{n_2}{n_2^2-1} w_2(x)\}] .$$

The similarity between the likelihood ratio and predictive estimators is striking. In fact for $n_1 = n_2 = n$ from (6.3.1) and (6.3.2)

$$LO(LR_e) = \frac{2n+1}{2n-1} LO(P_e)$$

and the two methods result in identical allocation. While no such simple relationship holds for $\Sigma_1 \neq \Sigma_2$, the differences in the methods when $n_1 = n_2$ are slight and the behaviour of the predictive and likelihood ratio estimators should be similar. The similarity between the methods was noted by Aitchison and Dunsmore (1975, p235), although they state incorrectly that the methods are equivalent.

The likelihood ratio estimator differs from the usual estimate approach by using such information contained in the unidentified observation x. As such differences in the methods should be apparent for n_t small and p large.

The Logistic Method

If the populations are multinormally distributed with equal covariance matrices the posterior probabilities $Pr(\Pi_t|x)$ (6.2.1) may be written as

$$Pr(\Pi_1|x) = \frac{\pi_1/\pi_2 \quad \dfrac{f_1(x)}{f_2(x)}}{1+\pi_1/\pi_2 \quad \dfrac{f_1(x)}{f_2(x)}}$$

$$= \frac{\exp\ (\beta o + \beta' x)}{1+\exp(\beta o + \beta' x)} \qquad (6.3.3)$$

and $Pr(\Pi_2|x) = 1 - Pr(\Pi_1|x)$

$$= \frac{1}{1 + \exp(\beta o + \beta' x)} \qquad (6.3.4)$$

where $\beta o = \ell n(\frac{\pi_1}{\pi_2}) - \frac{1}{2}(\mu_1-\mu_2)' \ \Sigma^{-1}(\mu_1+\mu_2)$ and β is a $p \times 1$ vector with $\beta = (\mu_1-\mu_2)' \ \Sigma^{-1}$. The functions (6.3.3) and (6.3.4) are called multivariate logistic functions, Cox (1966). The logit of risk (6.2.3) is given by

$$\begin{aligned} \text{logit}\{\Pr(\Pi_1|x)\} &= \beta_o + \beta'x \\ &= LO(T_e) \end{aligned}$$

as noted by Cox (1966). If the covariance matrices are unequal the arguments in the logistic model (6.3.3) and (6.3.4) are no longer linear but quadratic. The assumption of Normality of the populations is not however a requirement of the logistic model. The model is valid for a variety of population types including discrete and a mixture of discrete and continuous variables, as shown by Day and Kerridge (1967).

In practice if one postulates the linear logistic form of the posterior probabilities, the problem is the estimation of the p+1 unknown parameters $\beta_o, \beta_1, \ldots\ldots\ldots \beta_p$. Maximum likelihood estimates of these parameters have been given by Day and Kerridge, Anderson (1972), Prentice and Pyke (1979). The later references deal with separate samples from two populations rather than a single sample from a mixture of the populations as in Day and Kerridge. The quadratic case was considered by Anderson (1975). Anderson in Anderson and Richardson (1979) has considered correcting the maximum likelihood estimates for bias having noted in his previous 1972 paper the potential bias of all maximum likelihood procedures and in a simulation study the considerable difference between the estimates and the true values. In his 1979 paper a simulation study on univariate Normal populations with small sample sizes indicated that the corrected parameters were much closer to the true parameters than were the uncorrected. This reduction in bias was effected without an increase in the variability of the estimates and so results in a better estimate of true log-odds.

The robustness of the logistic model to population assumptions and the reduction in the parameters make it an attractice model. McLachlan and Byth (1979) compared the classification performance of the estimative and logistic allocation rules when the populations are multinormally distributed with equal covariance matrices. They derived asymptotic expressions for the expectations of the actual probabilities of misclassification of the logistic rule and compared them to the corresponding expectations, Okamoto (1963), of the estimative rule. With n large, p small and $\Delta \leqslant 3$ the classification rates of the rules were similar. This result should be related to Efron's (1975) who derived the asymptotic relative classification efficiency of the logistic to the estimative rule. Efron shows that classification efficiency deereases as Δ increases and has fallen to .641 when $\Delta = 3$. Parametric methods which assume normality of the populations are unlikely to be rivalled by the logistic method when the populations are normally distributed. As this is the case here the logistic approach is not pursued.

The Non-Parametric Method

Here the true densities f_t in the log-odds ratio are replaced by non-parametric density estimates $\hat{f}_t$. Unlike the parametric methods where the form of f_t is assumed known,the only assumption here is that f_t are continuous. Methods of non-parametric density estimation are well developed and are reviewed in Cover (1972), Wegman (1972a) and Fryer (1977).

In this thesis we have chosen the kernel approach since unlike other non-parametric density estimates the kernel estimate is itself a density. As such, kernel discrimination is undertaken by estimating true log-odds.

Kernel density estimation is described in Chapter 9 where a kernel based estimator of log-odds is compared with the estimative and predictive methods.

CHAPTER 7

ANALYTIC EXPRESSIONS FOR THE BIAS AND MEAN SQUARE ERROR OF THE ESTIMATIVE AND PREDICTIVE LOG-ODDS.

7.1 INTRODUCTION

In this chapter analytic expressions are obtained for the bias and mean square error of the estimative and predictive log-odds. Unconditional and conditional bias and mean square error, where the observation x is fixed, are derived on the assumption that the covariance matrices are known and unknown. The estimative log-odds corrected for bias is also considered.

7.2 Bias and Mean Square Error

Here a system of notation is adopted and some necessary results and conventions recorded. With estimators of true log-odds denoted by LO(M) where M denotes the method of estimation the derivation of biases and mean square errors will require consideration of the following expectations.

The expectation of LO(M), conditional on a fixed value of x, over repeated samples of size n_1 and n_2 from Π_1 and Π_2, will be denoted by

$$E\{LO(M) \mid x\}.$$

The bias of LO(M) for x fixed is then defined as

$$B\{LO(M) \mid x\} = E\{LO(M) \mid x\} - LO(T).$$

The unconditional expectation of LO(M) over repeated samples of size n_1 and n_2 from Π_1 and Π_2 and over repeated observations from Π_t (t = 1 and 2) is

$$E\{LO(M)\} = E_x[E\{LO(M) \mid x\}]$$

and the corresponding unconditional bias is

$$\begin{aligned} B\{LO(M)\} &= E_x[B\{LO(M) \mid x\}] \\ &= E\{LO(M)\} - E\{LO(T)\}. \end{aligned}$$

The mean square error (MSE) of the estimated log-odds, conditional on a fixed x in Π_t is defined as

$$\begin{aligned} MSE\{LO(M) \mid x\} &= E[\{LO(M) - LO(T)\}^2 \mid x] \\ &= V\{LO(M) \mid x\} + [E\{LO(M) \mid x\} - LO(T)]^2 \\ &= V\{LO(M) \mid x\} + [B\{LO(M) \mid x\}]^2, \end{aligned}$$

with $V\{LO(M) \mid x\}$ denoting the variance of LO(M) conditional on x. The corresponding unconditional mean square error is denoted by MSE{LO(M)}, where for x in Π_t, t = 1 and 2

$$\begin{aligned} MSE\{LO(M)\} &= E[\{LO(M) - LO(T)\}^2] \\ &= E_x[V\{LO(M) \mid x\}] + E_x[B\{LO(M) \mid x\}^2] \\ &= E_x[MSE\{LO(M) \mid x\}]. \end{aligned}$$

The conditional bias and mean square error are considered since the fixed x case is that which occurs in practice. The unconditional bias and mean square error provide an overview of the performance of the estimators. The known covariance matrices case is considered first as it gives some insight into what happens when the covariance matrices are unknown, it also allows the derivation of the mean square error of the predictive estimator.

7.3 True Log-Odds

Here for equal and unequal covariance matrices the means and variances of LO(T) are established for later use.

In Section 6.4 it was noted that $LO(T_e) = L(x)$ and $LO(T_u) = Q(x)$, the linear and quadratic discriminant functions. The distribution of $L(x)$, x in Π_t, $t = 1$ and 2, is univariate Normal with mean $\frac{1}{2}\,\Delta^2\,(-1)^{t-1}$ and variance Δ^2 where $\Delta^2 = (\mu_1 - \mu_2)'\,\Sigma^{-1}\,(\mu_1 - \mu_2)$. Hence for x in Π_t the unconditional mean and variance of $LO(T_e) = \frac{1}{2}\{-\omega_1(x) + \omega_2(x)\}$ where $\omega_t(x) = (x-\mu_t)'\,\Sigma^{-1}\,(x-\mu_t)$ are

$$E\{LO(T_e)\} = \tfrac{1}{2}\,\Delta^2\,(-1)^{t-1},\ t = 1 \text{ and } 2,\quad V\{LO(T_e)\} = \Delta^2. \tag{7.3.1}$$

The distribution of $LO(T_u) = Q(x)$ was obtained in Chapter 3. The expectation of $LO(T_u)$ may however be obtained directly as follows, noting that

$$LO(T_u) = \tfrac{1}{2}\,\{-\omega_1(x) + \omega_2(x) - \ln(|\Sigma_1|/|\Sigma_2|)\}$$

and that for x in Π_t, $\omega_t(x) = (x-\mu_t)'\,\Sigma_t^{-1}(x-\mu_t)$ is distributed as a central χ^2_p variate, hence $E\{\omega_t(x)\} = p$. Now for x in Π_1

$$\begin{aligned}E\{\omega_2(x)\} &= \text{tr}\ \Sigma_2^{-1}E\{(x-\mu_2)(x-\mu_2)'\}\\ &= \text{tr}\ \Sigma_2^{-1}\ \{\Sigma_1 + (\mu_1-\mu_2)\ (\mu_1-\mu_2)'\}\\ &= \text{tr}\ \Sigma_2^{-1}\Sigma_1 + (\mu_1-\mu_2)'\ \Sigma_2^{-1}(\mu_1-\mu_2)\\ &= \text{tr}\ \Sigma_2^{-1}\Sigma_1 + \Delta_2^2\end{aligned}$$

where $$\Delta_t^2 = (\mu_1-\mu_2)'\ \Sigma_t^{-1}(\mu_1-\mu_2)\ . \tag{7.3.2}$$

Similarly for x in Π_2

$$E\{\omega_1(x)\} = \text{tr}\ \Sigma_1^{-1}\Sigma_2 + \Delta_1^2.$$

Hence for x in Π_1 the unconditional mean of $LO(T_u)$ is

$$E\{LO(T_u)\} = \tfrac{1}{2}\{-p + \text{tr}\ \Sigma_2^{-1}\Sigma_1 + \Delta_2^2 - \ell n(|\Sigma_1|/|\Sigma_2|)\} \tag{7.3.3}$$

and for x in Π_2 is

$$E\{LO(T_u)\} = \tfrac{1}{2}\{-\text{tr}\ \Sigma_1^{-1}\Sigma_2 - \Delta_1^2 + p - \ell n(|\Sigma_1|/|\Sigma_2|)\}.$$

The variances of $LO(T_u)$ are not so easily obtained, the expectations $E\{\omega_1(x)\omega_2(x)\}$, $E\{\omega_1^2(x)\}$ and $E\{\omega_2^2(x)\}$ being required, where $\omega_1(x)$ is not independent of $\omega_2(x)$. Adoption of the standard canonical forms for Π_1 and Π_2 as in Section 3.3, gives the distribution of x in Π_1 as $N_p(0,I)$ and in Π_2 as $N_p(\nu,\Lambda)$ with Λ a diagonal matrix of eigen values λ_i, $1 < i \leqslant p$. The following results for x in Π_t, t = 1 and 2, are needed later.

For x in Π_1

$$E\{\omega_1(x)\} = p$$

$$E\{\omega_2(x)\} = \sum_{i=1}^{p} \frac{1+\nu_i^2}{\lambda_i}$$

$$E\{\omega_1^2(x)\} = 2p + p^2 \tag{7.3.4}$$

$$E\{\omega_2^2(x)\} = 2\sum_{i=1}^{p} \frac{1+2\nu_i^2}{\lambda_i^2} + \left(\sum_{i=1}^{p} \frac{1+\nu_i^2}{\lambda_i}\right)^2$$

$$E\{\omega_1(x)\ \omega_2(x)\} = 2\sum_{i=1}^{p} \frac{1}{\lambda_i} + p\sum_{i=1}^{p} \frac{1+\nu_i^2}{\lambda_i} \quad .$$

For x in Π_2

$$E\{\omega_1(x)\} = \sum_{i=1}^{p} \lambda_i+\nu_i^2$$

$$E\{\omega_2(x)\} = p$$

$$E\{\omega_1^2(x)\} = 2 \sum_{i=1}^{p} \lambda_i^2 + 2\nu_i^2\lambda_i + (\sum_{i=1}^{p} \lambda_i+\nu_i^2)^2$$

$$E\{\omega_2^2(x)\} = 2p + p^2 \qquad (7.3.5)$$

$$E\{\omega_1(x)\omega_2(x)\} = 2 \sum_{i=1}^{p} \lambda_i + p \sum_{i=1}^{p} \lambda_i + \nu_i^2 \ .$$

Now $V\{LO(T_u)\} = \frac{1}{4} E[\{\omega_1(x) - \omega_2(x)\}^2] - \frac{1}{4} [(E\{\omega_1(x)\} - E\{\omega_2(x)\})^2]$ and from (7.3.4) and (7.3.5), the variance of $LO(T_u)$ for x in Π_1 is

$$V\{LO(T_u)\} = \tfrac{1}{2}p - \sum_{i=1}^{p} 1/\lambda_i + \tfrac{1}{2} \sum_{i=1}^{p} \frac{1+2\nu_i^2}{\lambda_i^2} \ ,$$

and for x in Π_2

$$V\{LO(T_u)\} = \tfrac{1}{2}p - \sum_{i=1}^{p} \lambda_i + \tfrac{1}{2} \sum_{i=1}^{p} \lambda_i^2 + 2 \nu_i^2\lambda_i \ .$$

7.4 The Estimative Method, $\Sigma_1 = \Sigma_2$ known

For Σ known

$$LO(E_e) = \tfrac{1}{2}\{-w_1(x) + w_2(x)\}$$

where $n_t\, w_t(x) = n_t(x-\bar{x}_t)'\, \Sigma^{-1}(x-\bar{x}_t) \sim \chi^2(p,\, n_t\, \omega_t(x))$ (7.4.1)

and $\omega_t(x) = (x-\mu_t)'\, \Sigma^{-1}\, (x-\mu_t)$, $t = 1$ and 2.

Now $E\{\chi^2(p,\lambda)\} = p+\lambda$ and

$$E\{LO(E_e) \mid x\} = \tfrac{1}{2}\{-p/n_1 - \omega_1(x) + p/n_2 + \omega_2(x)\}$$

$$= LO(T_e) + \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right).$$

The unconditional expectation for x in Π_t, $t = 1$ and 2, from (7.3.1) is

$$E\{LO(E_e)\} = \tfrac{1}{2}\,\Delta^2(-1)^{t-1} + \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right).$$

Thus $$B\{LO(E_e)\} = B\{LO(E_e) \mid x\} = \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right)$$

$$= 0 \text{ when } n_1 = n_2.$$

Conditional on x, $n_1 w_1(x)$ and $n_2 w_2(x)$ are independently distributed as non-central chi squares (7.4.1) with variances $2p + 4n_t\omega_t(x)$, $t = 1$ and 2. Hence

$$V\{LO(E_e) \mid x\} = \tfrac{1}{4}\{\frac{2p}{n_1^2} + \frac{4\,\omega_1(x)}{n_1} + \frac{2p}{n_2^2} + \frac{4\,\omega_2(x)}{n_2}\}$$

and as

$$MSE\{LO(E_e) \mid x\} = V\{LO(E_e) \mid x\} + [B\{LO(E_e) \mid x\}]^2$$

$$= \tfrac{1}{2}p(\frac{1}{n_1^2} + \frac{1}{n_2^2}) + \frac{\omega_1(x)}{n_1} + \frac{\omega_2(x)}{n_2} + [\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right)]^2$$

Now for x in Π_1, $\omega_1(x) \sim \chi^2_p$ and $\omega_2(x) \sim \chi^2(p,\Delta^2)$ and for x in Π_2, $\omega_1(x) \sim \chi^2(p,\Delta^2)$ and $\omega_2(x) \sim \chi^2_p$, hence the unconditional mean square error for x in Π_1 is

$$MSE\{LO(E_e)\} = \tfrac{1}{2}p\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right) + p\left(\frac{n_1+n_2}{n_1n_2}\right) + \frac{\Delta^2}{n_2} + \left[\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)\right]^2$$

and for x in Π_2

$$MSE\{LO(E_e)\} = \tfrac{1}{2}p\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right) + p\left(\frac{n_1+n_2}{n_1n_2}\right) + \frac{\Delta^2}{n_1} + \left[\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)\right]^2 .$$

With $n_1 = n_2 = n$ these reduce to

$$MSE\{LO(E_e)\} = p\left(\frac{2n+1}{n^2}\right) + \frac{\Delta^2}{n} , \quad x \text{ in } \Pi_t,\ t = 1 \text{ and } 2.$$

7.5 The Estimative Method $\Sigma_1 \neq \Sigma_2$ known

For $\Sigma_1 \neq \Sigma_2$ known

$$LO(E_u) = \tfrac{1}{2}\{-w_1(x) + w_2(x) - \ln(|\Sigma_1|/|\Sigma_2|)\}$$

where $n_t\, w_t(x) = n_t(x-\bar{x}_t)'\, \Sigma_t^{-1}(x-\bar{x}_t) \sim \chi^2(p,\, n_t\omega_t(x))$

with $\omega_t(x) = (x-\mu_t)'\, \Sigma_t^{-1}(x-\mu_t)$

and hence $E\{LO(E_u) \mid x\} = LO(T_u) + \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)$.

From (7.3.3) the unconditional mean for x in Π_1 is

$$E\{LO(E_u)\} = \tfrac{1}{2}\{-p + \text{tr}\, \Sigma_2^{-1}\Sigma_1 + \Delta_2^2 - \ln(|\Sigma_1|/|\Sigma_2|)\} + \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)$$

and for x in Π_2

$$E\{LO(E_u)\} = \tfrac{1}{2}\{-\text{tr}\, \Sigma_1^{-1}\Sigma_2 - \Delta_1^2 + p - \ln(|\Sigma_1|/|\Sigma_2|)\} + \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)$$

and the bias is

$$B\{LO(E_u)\} = B\{LO(E_u) \mid x\} = \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)$$

$$= 0 \text{ when } n_1 = n_2.$$

As in Section 7.4 the conditional mean square error of $LO(E_u)$ is

$$MSE\{LO(E_u) \mid x\} = \tfrac{1}{2}p\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right) + \frac{\omega_1(x)}{n_1} + \frac{\omega_2(x)}{n_2} + \left[\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)\right]^2 .$$

The unconditional mean square error from the results (7.3.2) is for x in Π_1

$$MSE\{LO(E_u)\} = \tfrac{1}{2}p\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right) + \frac{p}{n_1} + \frac{tr\Sigma_2^{-1}\Sigma_1}{n_2} + \frac{\Delta_2^2}{n_2} + \left[\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)\right]^2$$

and for x in Π_2 is

$$MSE\{LO(E_u)\} = \tfrac{1}{2}p\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right) + \frac{tr\Sigma_1^{-1}\Sigma_2}{n_1} + \frac{\Delta_1^2}{n_1} + \frac{p}{n_2} + \left[\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)\right]^2 .$$

7.6 The Predictive Method, $\Sigma_1 = \Sigma_2$ known

For $\Sigma_1 = \Sigma_2$ known, the predictive density for a multinormally distributed population, with the conventional non-informative prior density for μ_t as $g(\mu_t) \alpha\, d\, \mu_t$, was given by Geisser (1964) as

$$p_t(x) = (2\pi)^{-\frac{1}{2}p}\, c_t^{\frac{1}{2}p}\, \mid \Sigma \mid^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}\, c_t\, w_t(x)\}$$

where $c_t = n_t \,/\, n_t+1$ and $w_t(x) = (x-\bar{x}_t)'\,\Sigma^{-1}(x-\bar{x}_t)$, $t = 1$ and 2.

It is interesting to note that here, with Σ known, the predictive density is itself multinormal with mean $\bar{x}_t$ and covariance matrix $(1 + 1/n_t)\Sigma$ and is very similar to the estimative density $N_p(\bar{x}_t,\Sigma)$.

Thus $LO(P_e) = \ell n\, \{\frac{p_1(x)}{p_2(x)}\}$

$$= \tfrac{1}{2}\,[p\ \ell n\, \{\frac{n_1(n_2+1)}{n_2(n_1+1)}\} - \frac{n_1}{n_1+1} w_1(x) + \frac{n_2}{n_2+1} w_2(x)],$$

and for $n_1 = n_2$, $LO(P_e) = \frac{n}{n+1}\ LO(E_e)$.

From the results on LO(E_e), Section 7.4 it follows that

$$E\{LO(P_e) \mid x\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) - \frac{1}{n_1+1}\{p + n_1\omega_1(x)\}$$

$$+ \frac{1}{n_2+1}\{p + n_2\omega_2(x)\}]$$

where $\omega_t(x) = (x-\mu_t)' \, \Sigma^{-1}(x-\mu_t)$,

and for x in Π_t, t = 1 and 2

$$E\{LO(P_e)\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) + c_{3-t} \ \Delta^2(-1)^{t-1}].$$

Hence the biases are

$$B\{LO(P_e) \mid x\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) - \frac{1}{n_1+1}\{p -\omega_1(x)\}$$

$$+ \frac{1}{n_2+1}\{p - \omega_2(x)\}],$$

and for x in Π_1

$$B\{LO(P_e)\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) - \frac{1}{n_2+1} \Delta^2],$$

while for x in Π_2

$$B\{LO(P_e)\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) + \frac{1}{n_1+1} \Delta^2].$$

For $n_1 = n_2 = n$ these reduce to

$$B\{LO(P_e) \mid x\} = \frac{n}{n+1} LO(T_e) - LO(T_e) = - \frac{1}{n+1} LO(T_e)$$

$$B\{LO(P_e)\} = - \frac{1}{n+1} \ \tfrac{1}{2} \ \Delta^2(-1)^{t-1}, \ t = 1 \text{ and } 2.$$

Both biases are negative, indicating that the predictive method understates true log-odds. It is particularly noteworthy that for $n_1 = n_2$ the biases are independent of the dimension parameter p.

From Section 7.4

$$V\{LO(P_e) \mid x\} = \tfrac{1}{2}p\{\frac{1}{(n_1+1)^2} + \frac{1}{(n_2+1)^2}\} + \frac{n_1}{(n_1+1)^2}\,\omega_1(x) + \frac{n_2}{(n_2+1)^2}\,\omega_2(x)$$

and hence

$$MSE\{LO(P_e) \mid x\} = V\{LO(P_e \mid x\} + [B\{LO(P_e) \mid x\}]^2 .$$

The corresponding unconditional mean square error for x in Π_1 is

$$MSE\{LO(P_e)\} = \tfrac{1}{2}p\{\frac{1}{(n_1+1)^2} + \frac{1}{(n_2+1)^2}\} + \frac{n_1}{(n_1+1)^2}\,p + \frac{n_2}{(n_2+1)^2}\,(p+\Delta^2) + E_x[B\{LO(P_e) \mid x\}^2]$$

and for x in Π_2 is

$$MSE\{LO(P_e)\} = \tfrac{1}{2}p\{\frac{1}{(n_1+1)^2} + \frac{1}{(n_2+1)^2}\} + \frac{n_1}{(n_1+1)^2}\,(p+\Delta^2) + \frac{n_2}{n_2+1}\,p + E_x[B\{LO(P_e) \mid x\}^2],$$

where after some algebraic manipulation

$$x \text{ in } \Pi_1,\ E_x[B\{LO(P_e) \mid x\}^2] = \tfrac{1}{4}[c^2 - \frac{2c}{n_2+1}\,\Delta^2 + \frac{2p}{(n_1+1)^2} - \frac{4p}{(n_1+1)(n_2+1)} + \frac{2p+4\Delta^2+\Delta^4}{(n_2+1)^2}]$$

$$x \text{ in } \Pi_2,\ E_x[B\{LO(P_e) \mid x\}^2] = \tfrac{1}{4}[c^2 + \frac{2c}{n_1+1}\,\Delta^2 + \frac{2p+4\Delta^2+\Delta^4}{(n_1+1)^2} - \frac{4p}{(n_1+1)(n_2+1)} + \frac{2p}{(n_2+1)^2}]$$

with $c = p\,\ell n\,(c_1/c_2)$.

For $n_1 = n_2 = n$ these mean square errors simplify to

$$\text{MSE}\{\text{LO}(P_e) \mid x\} = \left(\frac{n}{n+1}\right)^2 \text{MSE}\{\text{LO}(E_e) \mid x\} + \frac{1}{(n+1)^2} \text{LO}^2(T_e)$$

$$= \frac{1}{(n+1)^2} [p + n\{\omega_1(x) + \omega_2(x)\} + \text{LO}^2(T_e)]$$

and $$\text{MSE}\{\text{LO}(P_e)\} = \frac{1}{(n+1)^2} [p(2n+1) + \Delta^2(n+1) + \tfrac{1}{4} \Delta^4],$$

x in Π_1 and Π_2.

7.7 The Predictive Method, $\Sigma_1 \neq \Sigma_2$ known

For $\Sigma_1 \neq \Sigma_2$ but known the predictive densities are $N_p(\bar{x}_t, (1+1/n_t)\Sigma_t)$, $t = 1$ and 2.

and $\text{LO}(P_u) = \tfrac{1}{2}\{p \ \ell n \ (c_1/c_2) - c_1 w_1(x) + c_2 w_2(x) - \ell n \ (|\Sigma_1|/|\Sigma_2|)\}$

where $w_t(x) = (x-\bar{x}_t)' \ \Sigma_t^{-1}(x-\bar{x}_t)$ and $c_t = n_t/n_t+1$, $t = 1$ and 2.

Again the means and mean square errors of $\text{LO}(P_u)$ may be deduced from those in Section 7.5 for $\text{LO}(E_u)$.

Thus $$\text{E}\{\text{LO}(P_u) \mid x\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) - \frac{1}{n_1+1}\{p + n_1\omega_1(x)\}$$

$$+ \frac{1}{n_2+1}\{p + n_2\omega_2(x)\} - \ell n(|\Sigma_1|/|\Sigma_2|)]$$

where $$\omega_t(x) = (x-\mu_t)' \ \Sigma_t^{-1}(x-\mu_t),$$

for x in Π_1

$$\text{E}\{\text{LO}(P_u)\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) + \frac{n_2}{n_2+1}\{-p + \text{tr}\Sigma_2^{-1}\Sigma_1 + \Delta_2^2\}$$

$$- \ell n(|\Sigma_1|/|\Sigma_2|)]$$

and for x in Π_2

$$E\{LO(P_u)\} = \tfrac{1}{2}[p \ \ell n \ (c_1/c_2) - \frac{n_1}{n_1+1}\{tr \ \Sigma_1^{-1}\Sigma_2 + \Delta_1^2 - p\}$$
$$- \ell n(|\Sigma_1|/|\Sigma_2|)],$$

where as before $\Delta_t^2 = (\mu_1-\mu_2)' \ \Sigma_t^{-1}(\mu_1-\mu_2) \qquad t = 1,2.$

The corresponding biases are

$$B\{LO(P_u) \mid x\} = \tfrac{1}{2}[p \ \ell n(c_1/c_2) - \frac{1}{n_1+1}\{p - \omega_1(x)\}$$
$$+ \frac{1}{n_2+1}\{p - \omega_2(x)\}],$$

and for x in Π_1

$$B\{LO(P_u)\} \qquad = \tfrac{1}{2}[p \ \ell n(c_1/c_2) - \frac{1}{n_2+1}\{-p + tr \ \Sigma_2^{-1}\Sigma_1 + \Delta_2^2\}]$$

while for x in Π_2

$$B\{LO(P_u)\} \qquad = \tfrac{1}{2}[p \ \ell n(c_1/c_2) + \frac{1}{n_1+1}\{tr \ \Sigma_1^{-1}\Sigma_2 + \Delta_1^2 - p\}].$$

The conditional mean square error of LO(P_u) is

$$MSE\{LO(P_u) \mid x\} = V\{LO(P_u) \mid x\} + [B\{LO(P_u) \mid x\}]^2$$

where from Section 7.5

$$V\{LO(P_u) \mid x\} = \tfrac{1}{2}p\{\frac{1}{(n_1+1)^2} + \frac{1}{(n_2+1)^2}\}$$
$$+ \frac{n_1}{(n_1+1)^2} \ \omega_1(x) + \frac{n_2}{(n_2+1)^2} \ \omega_2(x).$$

The unconditional mean square error of LO(P_u) is

$$\mathrm{MSE}\{\mathrm{LO}(P_u)\} = E_x[\mathrm{MSE}\{\mathrm{LO}(P_u) \mid x\}],$$

in terms of the canonical forms of the populations, Section 7.3,

$$E_x[V\{\mathrm{LO}(P_u) \mid x\}] = \tfrac{1}{2}p\{\frac{1}{(n_1+1)^2} + \frac{1}{(n_2+1)^2}\} + \frac{n_1}{(n_1+1)^2}p + \frac{n_2}{(n_2+1)^2}\sum_{i=1}^{p}\frac{1+\nu_i^2}{\lambda_i}, \text{for } x \text{ in } \Pi_1 \qquad (7.7.1)$$

$$E_x[V\{\mathrm{LO}(P_u) \mid x\}] = \tfrac{1}{2}p\{\frac{1}{(n_1+1)^2} + \frac{1}{(n_2+1)^2}\} + \frac{n_1}{(n_1+1)^2}\sum_{i=1}^{p}(\lambda_i+\nu_i^2) + \frac{n_2}{(n_2+1)^2}p \quad , \quad \text{for } x \text{ in } \Pi_2 .$$

As in Section 7.6 after lengthy algebraic manipulation and use of the results (7.3.4) and (7.3.5) on LO(T_u) we obtain

for x in Π_1

$$E_x[B\{\mathrm{LO}(P_u) \mid x\}^2] = \tfrac{1}{4}[c^2 + \frac{2c}{n_2+1}\{p - \sum_{i=1}^{p}\frac{1+\nu_i^2}{\lambda_i}\} + \frac{2p}{(n_1+1)^2} - \frac{4\sum_{i=1}^{p}\frac{1}{\lambda_i}}{(n_1+1)(n_2+1)} + \frac{1}{(n_2+1)^2}(\{p - \sum_{i=1}^{p}\frac{1+\nu_i^2}{\lambda_i}\}^2 + 2\sum_{i=1}^{p}\frac{1+2\nu_i^2}{\lambda_i^2})] , \qquad (7.7.2)$$

and for x in Π_2

$$E_x[B\{\mathrm{LO}(P_u) \mid x\}^2] = \tfrac{1}{4}[c^2 - \frac{2c}{n_1+1}\{p - \sum_{i=1}^{p}(\lambda_i+\nu_i^2)\} + \frac{2p}{(n_2+1)^2} - \frac{4\sum_{i=1}^{p}\lambda_i}{(n_1+1)(n_2+1)} + \frac{1}{(n_1+1)^2}(\{p - \sum_{i=1}^{p}(\lambda_i+\nu_i^2)\}^2 + 2\sum_{i=1}^{p}(\lambda_i^2+2\nu_i^2\lambda_i))]$$

where $c = p \ln (c_1/c_2)$.

Combining (7.7.1) and (7.7.2) gives the unconditional mean square errors MSE{LO(P_u)}, x in Π_t, t = 1 and 2.

7.8 The Estimative Method, $\Sigma_1 = \Sigma_2$ Unknown

For $\Sigma_1 = \Sigma_2 = \Sigma$ unknown

$$LO(E_e) = \tfrac{1}{2}\{- w_1(x) + w_2(x)\}$$

where now $w_t(x) = (x-\bar{x}_t)'\ S^{-1}\ (x-\bar{x}_t)$, $t = 1$ and 2, and $(n_1+n_2-2)S$ has a Wishart distribution, $W(p,\ n_1+n_2-2,\Sigma)$.

Conditional on x,

$$\frac{n_t}{n_1+n_2-2}\ w_t(x) \sim \frac{p}{n_1+n_2-p-1}\ F(p,\ n_1+n_2-p-1,\ n_t\omega_t(x))$$

where $\omega_t(x) = (x-\mu_t)'\ \Sigma^{-1}\ (x-\mu_t)$.

The mean of a non-central F variate, $F(\nu_1,\nu_2,\lambda)$, is

$$\frac{\nu_2(\nu_1+\lambda)}{\nu_1(\nu_2-2)},$$

and therefore

$$E\{LO(E_e)\ |\ x\} = \tfrac{1}{2}\{- \frac{n_1+n_2-2}{n_1}\ (\frac{p+n_1\omega_1(x)}{n_1+n_2-p-3}) + \frac{n_1+n_2-2}{n_2}\ (\frac{p+n_2\omega_2(x)}{n_1+n_2-p-3})\}$$

$$= \frac{n_1+n_2-2}{n_1+n_2-p-3}\ \{-\tfrac{1}{2}\ \omega_1(x) + \tfrac{1}{2}\ \omega_2(x) + \tfrac{1}{2}p(\frac{n_1-n_2}{n_1n_2})\}$$

$$= \frac{n_1+n_2-2}{n_1+n_2-p-3}\ \{LO(T_e) + \tfrac{1}{2}p\ (\frac{n_1-n_2}{n_1n_2})\}. \qquad (7.8.1)$$

The unconditional mean of $LO(E_e)$ for x in Π_t is

$$E\{LO(E_e)\} = \frac{n_1+n_2-2}{n_1+n_2-p-3}\{\tfrac{1}{2}\Delta^2(-1)^{t-1} + \tfrac{1}{2}p(\frac{n_1-n_2}{n_1n_2})\}, \quad t = 1 \text{ and } 2,$$

and corresponding biases are

$$B\{LO(E_e) \mid x\} = \frac{p+1}{n_1+n_2-p-3} LO(T_e) + \frac{n_1+n_2-2}{n_1+n_2-p-3} \tfrac{1}{2}p(\frac{n_1-n_2}{n_1n_2})$$

and

$$B\{LO(E_e)\} = \frac{p+1}{n_1+n_2-p-3} \tfrac{1}{2}\Delta^2(-1)^{t-1} + \frac{n_1+n_2-2}{n_1+n_2-p-3} \tfrac{1}{2}p(\frac{n_1-n_2}{n_1n_2})$$

$$x \in \Pi_t, \quad t = 1 \text{ and } 2.$$

For $n_1 = n_2 = n$ the biases are non-zero, unlike the case of Σ known, Section 7.4. The estimative method over-states true log-odds, this overstatement increases with increasing dimension p and separation Δ. The main source of bias is the coefficient

$$\frac{n_1+n_2-2}{n_1+n_2-p-3}$$

which arises in

$$E(S^{-1}) = \frac{n_1+n_2-2}{n_1+n_2-p-3} \Sigma^{-1}$$

Das Gupta (1968), Lachenbruch (1968).

The conditional mean square error of $LO(E_e)$ is

$$MSE\{LO(E_e) \mid x\} = E\{LO^2(E_e) \mid x\} - 2E\{LO(E_e) \mid x\} LO(T_e) + LO^2(T_e)$$

$$= E\{LO^2(E_e) \mid x\} - \left(\frac{n_1+n_2+p-1}{n_1+n_2-p-3}\right) LO^2(T_e)$$

$$- \left(\frac{n_1+n_2-2}{n_1+n_2-p-3}\right) p \left(\frac{n_1-n_2}{n_1n_2}\right) LO(T_e).$$

The conditional expectation $E\{LO^2(E_e) \mid x\}$ is in general a complicated expression and details are left to Appendix 7.A. The result for $n_1 = n_2 = n$ is

$$E\{LO^2(E_e) \mid x\} = \frac{(2n-2)^2}{d} [\{\Delta^2 + \frac{2}{n}(2n-3)\} (x-\eta)' \Sigma^{-1}(x-\eta)$$

$$+ (2n-3) \{\frac{p}{n^2} + \frac{\Delta^2}{2n}\} + (2n-p-3) LO^2(T_e)]$$

where $\eta = \frac{1}{2}(\mu_1+\mu_2)$ and $d = (2n-p-2)(2n-p-3)(2n-p-5)$.

Since for x in Π_t, $t = 1$ and 2, $(x-\eta)' \Sigma^{-1}(x-\eta) \sim \chi^2(p, \frac{1}{4}\Delta^2)$, the unconditional expectation when $n_1 = n_2 = n$ is

$$E\{LO^2(E_e)\} = \frac{(2n-2)^2}{d} [(2n-3) \{\frac{(2n+1)}{n^2} p + \frac{(n+1)}{n} \Delta^2\}$$

$$+ (2n-p-2) \Delta^4/4],$$

and the corresponding conditional and unconditional mean square errors of $LO(E_e)$ may be written as

$$MSE\{LO(E_e) \mid x\} = E\{LO^2(E_e) \mid x\} - \frac{2n+p-1}{2n-p-3} LO^2(T_e)$$

and

$$MSE\{LO(E_e)\} = E\{LO^2(E_e)\} - \frac{2n+p-1}{2n-p-3} (\Delta^2+\Delta^4/4),$$

x in Π_t, $t = 1$ and 2.

7.9 The Estimative Method, $\Sigma_1 \neq \Sigma_2$ and both unknown

For $\Sigma_1 \neq \Sigma_2$ and both unknown

$$LO(E_u) = \tfrac{1}{2}\{-w_1(x) + w_2(x) - \ell n(|S_1|/|S_2|)\}.$$

Now $\frac{n_t}{n_t-1} w_t(x) = \frac{n_t}{n_t-1} (x-\bar{x}_t)' S_t^{-1}(x-\bar{x}_t) \sim \frac{p}{n_t-p} F(p, n_t-p, n_t\omega_t(x))$

and $\omega_t(x) = (x-\mu_t)' \Sigma_t^{-1}(x-\mu_t)$ $t = 1$ and 2,

giving $E\{w_t(x) \mid x\} = \frac{n_t-1}{n_t}\left(\frac{p+n_t\omega_t(x)}{n_t-p-2}\right)$. (7.9.1)

Further $|S_t| \sim \frac{|\Sigma_t|}{(n_t-1)^p} \chi^2_{n_t-1} \chi^2_{n_t-2} \cdots \chi^2_{n_t-p}$, Anderson (1958,p171)

and $E\{\ell n |S_t|\} = \ell n |\Sigma_t| - p\ell n(n_t-1) + \sum_{i=1}^{p} E\{\ell n \chi^2_{n_t-i}\}$.

Now $E\{\ell n \chi^2_\nu\} = \Psi(\frac{\nu}{2}) + \ell n2$, Johnson and Kotz (1970, vol. 1, P196) where $\Psi(a) = \Gamma'(a) / \Gamma(a)$, the digamma function, Abramowitz and Stegun (1965, P258) and thus

$$E\{\ell n |S_t|\} = \ell n |\Sigma_t| - p\ell n(\frac{n_t-1}{2}) + \sum_{i=1}^{p} \Psi(\frac{n_t-i}{2}). \quad (7.9.2)$$

Combining the results (7.9.1) and (7.9.2) it follows that

$$E\{LO(E_u) \mid x\} = \tfrac{1}{2}[- \frac{n_1-1}{n_1-p-2} (\frac{p}{n_1} + \omega_1(x)) + \frac{n_2-1}{n_2-p-2} (\frac{p}{n_2} + \omega_2(x))$$
$$- \ell n(|\Sigma_1|/|\Sigma_2|) + p\ell n(\frac{n_1-1}{n_2-1})$$
$$- \sum_{i=1}^{p} \{\Psi(\frac{n_1-i}{2}) - \Psi(\frac{n_2-i}{2})\}]$$

and for x in Π_1

$$E\{LO(E_u)\} = \tfrac{1}{2}[- \frac{(n_1^2-1)}{n_1-p-2}\,\frac{p}{n_1} + \frac{n_2-1}{n_2-p-2}\,(\frac{p}{n_2} + tr\Sigma_2^{-1}\Sigma_1+\Delta_2^2)$$

$$- \ell n(|\ \Sigma_1\ |/|\ \Sigma_2\ |) + p\ell n(\frac{n_1-1}{n_2-1})$$

$$- \sum_{i=1}^{p} \{\Psi(\frac{n_1-i}{2}) - \Psi(\frac{n_2-i}{2})\}]$$

and for x in Π_2

$$E\{LO(E_u)\} = \tfrac{1}{2}[- \frac{n_1-1}{n_1-p-2}\,(\frac{p}{n_1} + tr\Sigma_1^{-1}\Sigma_2 + \Delta_1^2)$$

$$+ \frac{(n_2^2-1)}{n_2-p-2}\,\frac{p}{n_2} - \ell n(|\ \Sigma_1|/|\ \Sigma_2|)$$

$$+ p\ell n(\frac{n_1-1}{n_2-1}) - \sum_{i=1}^{p} \{\Psi(\frac{n_1-i}{2}) - \Psi(\frac{n_2-i}{2})\}].$$

The corresponding biases follow from the above expectations. For $n_1 = n_2 = n$ they reduce to

$$B\{LO(E_u)\ |\ x\} = \frac{p+1}{n-p-2}\,\tfrac{1}{2}\{-\omega_1(x) + \omega_2(x)\}$$

$$= \frac{p+1}{n-p-2}\,LO(T_u) + \frac{p+1}{n-p-2}\,\tfrac{1}{2}\ell n(|\ \Sigma_1|/|\Sigma_2\ |)$$

$$B\{LO(E_u)\} = \frac{p+1}{n-p-2}\,\tfrac{1}{2}[-p + tr\Sigma_2^{-1}\Sigma_1 + \Delta_2^2],\ x \text{ in } \Pi_1$$

and

$$B\{LO(E_u)\} = \frac{p+1}{n-p-2}\,\tfrac{1}{2}[-tr\Sigma_1^{-1}\Sigma_2 - \Delta_1^2 + p],\ x \text{ in } \Pi_2.$$

The coefficients $\frac{n_t-1}{n_t-p-2}$ in the expectations $E\{LO(E_u) \vdash x\}$ and $E\{LO(E_u)\}$ result from S_t^{-1}. Estimation of the covariance matrices is the main source of bias.

The derivation of the mean square errors seems in general to be interactable involving expectations like $E\{\ell n|S_t|(x-\bar{x}_t)' S_t^{-1}(x-\bar{x}_t)\}$. The conditional expectation of this term when p = 1 may be derived as follows:

Let $S_t = s_t$ and $\Sigma_t = \sigma_t$ when p = 1

then $\ell n|S_t|(x-\bar{x}_t)' S_t^{-1}(x-\bar{x}_t) = \ell n \; s_t \frac{(x-\bar{x}_t)^2}{s_t}$.

Conditional on x $\frac{n_t(x-\bar{x}_t)^2}{\sigma_t} \sim \chi^2 \left(1, \frac{n_t(x-\mu_t)^2}{\sigma_t}\right)$

and $n_t-1 \frac{s_t}{\sigma_t} \sim \chi^2_{n_t-1}$

and s_t and $\bar{x}_t$ are independently distributed. As $E[\chi^2(\nu,\lambda)] = \nu + \lambda$, $E[1/\chi^2_\nu] = \frac{1}{\nu-2}$ and $E[\frac{\ell n\chi^2 \nu}{\chi^2_\nu}] = \frac{1}{\nu-2}[\psi(\frac{\nu-2}{2}) + \ell n 2]$ the result follows. We have been able to extend this result to the unconditional case for p > 1 provided the covariance matrices are proportional, details are given in Appendix 7.B. There it is used to derive the unconditional mean square error of $LO(E_u)$.

7.10 The Predictive Method, $\Sigma_1 = \Sigma_2$, unknown.

For Σ unknown the predictive estimator of log-odds is from (6.3.1)

$$LO(P_e) = \tfrac{1}{2}p\ell n\ (c_1/c_2) - \tfrac{1}{2}(m+1)\ \ell n \left[\frac{\{1 + c_1/m\ w_1(x)\}}{\{1 + c_2/m\ w_2(x)\}}\right] \quad (7.10.1)$$

where $m = n_1+n_2-2$, $c_t = n_t/(n_t+1)$ and $w_t(x) = (x-x_t)'\ S^{-1}(x-\bar{x}_t)$, $t = 1$ and 2.

The expectations $E\{LO(P_e) \mid x\}$ and $E\{LO(P_e)\}$ involve the derivation of

$$E[\ell n\ \{\ 1 + c_t/m\ w_t(x)\}] \qquad t = 1 \text{ and } 2.$$

The unconditional case is easier and will be considered first.

Let $Z_t = c_t/m\ \ w_t(x) \qquad t = 1$ and 2

then $LO(Pe) = \tfrac{1}{2}p\ell n(c_1/c_2) - \tfrac{1}{2}(m+1)\ [\ell n(1+Z_1) - \ell n(1+Z_2)]$ (7.10.2)

and for x in Π_1 $\qquad Z_1 \sim \frac{p}{m-p+1} F(p,m-p+1) = \chi^2_p/\chi^2_{m-p+1}$

and $\qquad Z_2 \sim \frac{p}{m-p+1} F(p,m-p+1,\lambda_1) = \chi^2_p/\chi^2(p,m-p+1,\lambda_1)$

where $\qquad \lambda_1 = c_2\ \Delta^2.$

As Z_1 is a scalar multiple of an F variate the moment generating function of $\ell n(1+Z_1)$ is

$$E\{e^{s\ell n(1+Z_1)}\} = E\{(1+Z_1)^s\}$$

$$= \frac{1}{B(\tfrac{1}{2}p,\tfrac{1}{2}(m-p+1))} \int_0^\infty \frac{Z_1^{\frac{p}{2}-1}}{(1+Z_1)^{\frac{m+1}{2}-s}}\, dZ_1.$$

Substitution of $y = Z_1/(1+Z_1)$ where $y \sim$ Beta$(p/2, \frac{m-p+1}{2})$ gives

$$E\{(1+Z_1)^s\} = \frac{1}{B(\frac{p}{2}, \frac{m-p+1}{2})} \int_0^1 y^{\frac{p}{2}-1} (1-y)^{\frac{m-p+1}{2}-s-1} dy$$

$$= \frac{B(\frac{p}{2}, \frac{m-p+1}{2}-s)}{B(\frac{p}{2}, \frac{m-p+1}{2}).} \qquad (7.10.3)$$

The derivative of (7.10.3) with respect to s at s = o is

$$E\{\ell n(1+Z_1)\} = \frac{\Gamma'(\frac{m+1}{2})}{\Gamma(\frac{m+1}{2})} - \frac{\Gamma'(\frac{m-p+1}{2})}{\Gamma(\frac{m-p+1}{2})}$$

$$= \psi(\tfrac{m+1}{2}) - \psi(\tfrac{m-p+1}{2}) \qquad (7.10.4)$$

where $\psi(a) = \Gamma'(a)/\Gamma(a)$ is the digamma function.

For t = 2, Z_2 has a non-central F distribution and the density of $Z_2/(1+Z_2)$ is now a weighted sum of Beta densities with parameters $(p/2+j, \frac{m-p+1}{2})$ and Poisson probability weights with parameter $\frac{1}{2}\lambda_1$, Johnson and Kotz (1970, vol. 2, p191). The corresponding moment generating function of $\ell n(1+Z_2)$ is

$$E\{(1+Z_2)^s\} = \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \frac{B(\frac{p}{2}+j, \frac{m-p+1}{2}-s)}{B(\frac{p}{2}+j, \frac{m-p+1}{2})}$$

and the derivative at s = o is

$$E\{\ell n(1+Z_2)\} = [\sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \psi(\tfrac{m+1}{2}+j)] - \psi(\tfrac{m-p+1}{2}). \qquad (7.10.5)$$

Use of the digamma functions recurrence relationship $\psi(a+1) = \psi(a) + 1/a$, allows us to write

$$\psi(\tfrac{m+1}{2}+j) = \psi(\tfrac{m+1}{2}) + \sum_{i=o}^{j-1} \frac{2}{m+1+2i}, \; j \geq 1$$

and so

$$\sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \psi(\tfrac{m+1}{2}+j) = \psi(\tfrac{m+1}{2}) + \sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \sum_{i=o}^{j-1} \frac{2}{m+1+2i}. \qquad (7.10.6)$$

Combining the results (7.10.4), (7.10.5) and (7.10.6) in (7.10.2) gives

$$E\{LO(P_e)\} = \frac{p}{2}\ \ell n(c_1/c_2) + \sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!}\ e^{-\frac{1}{2}\lambda_1} \sum_{i=o}^{j-1} \frac{m+1}{m+1+2i} \tag{7.10.7}$$

where $c_t = n_t/(n_t+1)$, $m = n_1+n_2-2$, $\lambda_1 = c_2\Delta^2$ and x in Π_1.

Similarly for x in Π_2

$$E\{LO(P_e)\} = \frac{p}{2}\ \ell n(c_1/c_2) - \sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_2)^j}{j!}\ e^{-\frac{1}{2}\lambda_2} \sum_{i=o}^{j-1} \frac{m+1}{m+1+2i} \tag{7.10.8}$$

where $\lambda_2 = c_1\ \Delta^2$.

The unconditional bias of $LO(P_e)$ is therefore

$$B\{LO(P_e)\} = E\{LO(P_e)\} - \tfrac{1}{2}\Delta^2(-1)^{t-1},\ x \text{ in } \Pi_t,\ t = 1 \text{ and } 2.$$

For $n_1 = n_2 = n$, $c_1 = c_2$ and we see that the unconditional expectations and biases of $LO(P_e)$ are independent of the dimension parameter p and depend only on the separation Δ and sample size n.

Upper and lower bounds on $B\{LO(P_e)\}$ may be obtained as follows.

As $$\frac{j}{j+1} \leqslant \sum_{i=o}^{j-1} \frac{m+1}{m+1+2i} \leqslant j \qquad \text{for } j \geqslant 1$$

it follows that

$$0 < 1 + 2/\lambda_t[e^{-\frac{1}{2}\lambda_t} - 1] < \sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!}\ e^{-\frac{1}{2}\lambda_t} \sum_{i=o}^{j-1} \frac{m+1}{m+1+2i} < \tfrac{1}{2}\lambda_t\ .$$

Thus using (7.10.7), for x in Π_1

$$\frac{p}{2}\ \ln(c_1/c_2) < E\{LO(P_e)\} < \frac{p}{2}\ln(c_1/c_2) + \tfrac{1}{2}\lambda_1$$

and from (7.10.8), for x in Π_2

$$\frac{p}{2}\ \ln(c_1/c_2) > E\{LO(P_e)\} > \frac{p}{2}\ln(c_1/c_2) - \tfrac{1}{2}\lambda_2.$$

With equal sample sizes $c_1 = c_2$ and the unconditional bias of $LO(P_e)$ may be bounded for x in Π_1 as

$$-\tfrac{1}{2}\Delta^2 < B\{LO(P_e)\} < -\frac{1}{n+1}\ \tfrac{1}{2}\Delta^2$$

and for x in Π_2 as (7.10.9)

$$-(-\tfrac{1}{2}\Delta^2) > B\{LO(P_e)\} > -\frac{1}{n+1}(-\tfrac{1}{2}\Delta^2).$$

For equal sample sizes, therefore, the bias of $LO(P_e)$ is independent of the dimension parameter and P_e on average understates the true log-odds for observations from either population.

We now consider the derivation of the conditional expectation $E\{LO(P_e) \mid x\}$. Conditional on x we define

$$Z_t = n_t/m\ w_t(x) \qquad t = 1 \text{ and } 2$$

where $Z_t \sim \frac{p}{m-p+1} F(p,\ m-p+1,\ \lambda_t)$ and $\lambda_t = n_t\omega_t(x) = n_t(x-\mu_t)'\Sigma^{-1}(x-\mu_t)$.

The expression (7.10.1) for $LO(P_e)$ may now be written as

$$LO(P_e) = \tfrac{1}{2}\ p\ln\ (c_1/c_2) - \tfrac{1}{2}(m+1)[\ln(1+Z_1/k_1) - \ln(1+Z_2/k_2)] \qquad (7.10.10)$$

where $k_t = n_t+1$.

Noting that $Z_t/(1+Z_t)$ is again distributed as a weighted sum of Beta's with parameters

$$(\frac{p}{2} + j, \frac{m-p+1}{2})$$

and Poisson weights with parameter $\frac{1}{2}\lambda_t$. The transformation $y_t = Z_t/(1+Z_t)$ now gives

$$E\{(1+Z_t/k_t)^s\} = \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} e^{-\frac{1}{2}\lambda_t} \frac{1}{B(\frac{p}{2}+j, \frac{m-p+1}{2})} \{$$

$$\int_0^1 y_t^{\frac{p}{2}+j-1} (1-y_t)^{\frac{m-p+1}{2} - 1} (\frac{1-c_t y_t}{1-y_t})^s \; dy_t\}$$

and the derivative with respect to s at s = o is

$$E\{\ell n(1+Z_t/k_t)\} = \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} e^{-\frac{1}{2}\lambda_t} \frac{1}{B(\frac{p}{2}+j, \frac{m-p+1}{2})} \{$$

$$\int_0^1 y_t^{\frac{p}{2}+j-1} (1-y_t)^{\frac{m-p+1}{2} -1} \ell n(\frac{1-c_t y_t}{1-y_t}) \; dy_t.\}$$

But $-\ell n(1-y_t) = \ell n(1+Z_t)$ and the result (7.10.4) gives this part of the expectation. As $0 < y_t < 1$ and $c_t = n_t/n_t+1 < 1$ we may write

$$\ell n(1 - c_t y_t) = - \sum_{i=1}^{\infty} \frac{(c_t y_t)^i}{i}$$

and so

$$E\{\ell n(1 + Z_t/k_t)\} = \psi(\frac{m+1}{2}) - \psi(\frac{m-p+1}{2}) + \sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} e^{-\frac{1}{2}\lambda_t}\{\sum_{i=o}^{j-1} \frac{2}{m+1+2i}\}$$

$$- \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} e^{-\frac{1}{2}\lambda t} \sum_{i=1}^{\infty} \frac{c_t^i}{i} \frac{B(\frac{p}{2}+i+j, \frac{m-p+1}{2})}{B(\frac{p}{2}+j, \frac{m-p+1}{2})} . \qquad (7.10.11)$$

Combining (7.10.10) and (7.10.11) the conditional expectation $E\{LO(P_e) \mid x\}$ follows and the conditional bias is given by

$$B\{LO(P_e) \mid x\} = E\{LO(P_e) \mid x\} - LO(T_e).$$

From (7.10.11) it is clear that the bias $B\{LO(P_e) \mid x\}$ does depend on the dimension parameter p even when $n_1 = n_2$ unlike the unconditional bias $B\{LO(P_e)\}$.

Derivation of the mean square error of the predictive estimator when Σ is unknown proved intractable.

7.11 The Predictive Method, $\Sigma_1 \neq \Sigma_2$ Unknown

With $\Sigma_1 \neq \Sigma_2$ and unknown the predictive densities as given in Section 6.3 result in the predictive estimator $LO(P_u)$ where

$$LO(P_u) = \ln \left[\frac{\Gamma(\frac{m_1+1}{2})\ \Gamma(\frac{m_2-p+1}{2})}{\Gamma(\frac{m_2+1}{2})\ \Gamma(\frac{m_1-p+1}{2})} \right] + \tfrac{1}{2} p \ln\{\frac{c_1 m_2}{c_2 m_1}\}$$

$$- \tfrac{1}{2} \ln(|S_1|/|S_2|) - \tfrac{1}{2}(m_1+1) \ln\{1 + c_1/m_1\ w_1(x)\}$$

$$+ \tfrac{1}{2}(m_2+1) \ln\{1 + c_2/m_2\ w_2(x)\},$$

with $m_t = n_t - 1$, $c_t = n_t/(n_t+1)$ and $w_t(x) = (x-\bar{x}_t)'\ S_t^{-1}(x-\bar{x}_t)$, $t = 1$ and 2.

Conditional on x

$$\frac{n_t}{m_t} w_t(x] \sim \frac{p}{n_t-p}\ F(p,\ n_t-p,\ \lambda_t)$$

where $\lambda_t = n_t\ \omega_t(x) = n_t(x-\mu_t)'\ \Sigma_t^{-1}(x-\mu_t)$, $t = 1$ and 2.

With $n_1 = n_2 = n$ for simplicity, from the result (7.10.11) and $E\{\ell n|S_t|\}$ Section 7.9, the conditional expectation $E\{LO(P_u) \mid x\}$ may be written as

$$E\{LO(P_u) \mid x\} = -\tfrac{1}{2}\ell n(|\Sigma_1|/|\Sigma_2|) - \sum_{j=1}^{\infty} \{ \sum_{i=o}^{j-1} \frac{n}{n+2i} \}\{e^{-\frac{1}{2}\lambda_1} \frac{(\frac{1}{2}\lambda_1)^j}{j!}$$

$$- e^{-\frac{1}{2}\lambda_2} \frac{(\frac{1}{2}\lambda_2)^j}{j!}\} + \frac{n}{2} \sum_{j=o}^{\infty} \sum_{i=1}^{\infty} \frac{(\frac{n}{n+1})^i}{i} \Bigg[$$

$$\frac{B(\frac{p}{2}+i+j,\ \frac{n-p}{2})}{B(\frac{p}{2}+j,\ \frac{n-p}{2})} \{e^{-\frac{1}{2}\lambda_1} \frac{(\frac{1}{2}\lambda_1)^j}{j!} - e^{-\frac{1}{2}\lambda_2} \frac{(\frac{1}{2}\lambda_2)^j}{j!} \}\Bigg].$$

The unconditional expectation $E\{LO(P_u)\}$ can not be obtained from the results of Section 7.10 because, for x in Π_1, $w_2(x)$ is not distributed as a scalar multiple of an F variate. However one case which may be partially solved is that of proportional covariance matrices, i.e. $\alpha\Sigma_1 = \Sigma_2$, $\alpha > 0$ and $\alpha \neq 1$.

Now for x in Π_1

$$c_1/m_1\ w_1(x) \sim \frac{p}{n_1-p}\ F(p,\ n_1-p)$$

and for $\alpha\Sigma_1 = \Sigma_2$, $\alpha > 0$

$$\frac{\alpha n_2}{n_2+\alpha}\ \frac{1}{m_2}\ w_2(x) \sim \frac{p}{n_2-p}\ F(p,\ n_2-p,\ \lambda_1)$$

where $\lambda_1 = \frac{n_2}{n_2+\alpha}\ \Delta_1^2$ and $\Delta_1^2 = (\mu_1-\mu_2)'\ \Sigma_1^{-1}\ (\mu_1-\mu_2)$.

If $Z_1 = c_1/m_1 \; w_1(x)$ and $Z_2 = \frac{\alpha n_2}{n_2+\alpha} \; \frac{1}{m_2} \; w_2(x)$

then from (7.10.4)

$$E\{\ell n \; (1 + c_1/m_1 \; w_1(x)\} = E\{\ell n \; (1 + Z_1)\} = \psi(\tfrac{n_1}{2}) - \psi(\tfrac{n_1-p}{2}). \qquad (7.11.1)$$

Now $\ell n\{1 + c_2/m_2 \; w_2(x)\} \quad = \ell n(1 + Z_2/k)$

where $\quad k = \frac{\alpha(n_2+1)}{n_2+\alpha}$,

with the transformation $y = Z_2/(1+Z_2)$ and the result (7.10.11) it follows that

$$E\{\ell n(1 + Z_2/k)\} = \psi(\tfrac{n_2}{2}) - \psi(\tfrac{n_2-p}{2}) + \sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} \; e^{-\frac{1}{2}\lambda_1} \; \{\sum_{i=o}^{j-1} \frac{2}{n_2+2i}\}$$

$$- \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} \; e^{-\frac{1}{2}\lambda_1} \; \sum_{i=1}^{\infty} \frac{a_1^i}{i} \; \frac{B(\frac{p}{2}+i+j, \; \frac{n_2-p}{2})}{B(\frac{p}{2}+j, \; \frac{n_2-p}{2})} \qquad (7.11.2)$$

where $a_1 = 1 - 1/k = \frac{n_2}{n_2+1} \; (\frac{\alpha-1}{\alpha})$. However for this series expansion to be valid $-1 < a_1 \leq 1$ hence

$$\alpha > \frac{n_2}{2n_2+1} \; .$$

Combining (7.11.1) and (7.11.2) and the expectations $E\{\ell n|S_t|\}$ results in the unconditional expectations $E\{LO(P_u)\}$ when the covariance matrices are proportional. A similar result holds for x in Π_2 and with $n_1 = n_2 = n$ for simplicity we obtain

$$E\{LO(P_u)\} = (-1)^{t-1}[\sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} \; e^{-\frac{1}{2}\lambda_t} \sum_{i=o}^{j-1} \frac{n}{n+2i}$$

$$- \frac{n}{2} \; \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} \; e^{-\frac{1}{2}\lambda_t} \; \sum_{i=1}^{\infty} \frac{a_t^i}{i} \; \frac{B(\frac{p}{2}+i+j, \; \frac{n-p}{2})}{B(\frac{p}{2}+j, \; \frac{n-p}{2})}] + \tfrac{1}{2}p\ell n\alpha$$

x in Π_t $t = 1$ and 2, where

$$\lambda_1 = \frac{n}{n+\alpha}\,\Delta_1^2 \qquad a_1 = \frac{n}{n+1}\left(\frac{\alpha-1}{\alpha}\right)$$

and

$$\lambda_2 = \frac{n}{\alpha n+1}\,\Delta_1^2 \qquad a_2 = \frac{n}{n+1}\,(1-\alpha)$$

provided $-1 < a_t \leq 1$ i.e. $\dfrac{1}{2+\frac{1}{n}} < \alpha < 2 + \dfrac{1}{n}$.

The unconditional bias is now given by

$$B\{LO(P_u)\} = E\{LO(E_u)\} - E\{LO(T_u)\}$$

where from (7.3.3) with proportional covariance matrices

$$E\{LO(T_u)\} = \tfrac{1}{2}\{p\left(\frac{1-\alpha}{\alpha}\right) + \frac{1}{\alpha}\,\Delta_1^2 + p\ln\alpha\} \qquad x \text{ in } \Pi_1$$

and

$$E\{LO(T_u)\} = \tfrac{1}{2}\{p(1-\alpha) - \Delta_1^2 + p\ln\alpha\} \qquad x \text{ in } \Pi_2.$$

The mean square errors of $LO(P_u)$, $\Sigma_1 \neq \Sigma_2$ unknown have proved to be intractable.

7.12 Unbiased Estimation of True Log-Odds

As was noted in Section 6.3 $LO(E_e) = \hat{L}(x)$ is a biased estimator of true log-odds, Lachenbruch (1968). In Section 7.8 we noted that the main source of bias in $LO(E_e)$ when Σ is unknown is due to the multiplicative constant

$$\frac{n_1+n_2-2}{n_1+n_2-p-3}$$

in the expectation of S^{-1}. However an allocation rule with zero cut-off point is invarient to multiplication by a positive constant. For estimation however unbiased estimators seem worth considering.

Unbiased estimators of true log-odds are easily derived from the results on LO(E), Sections 7.4, 7.5, 7.8 and 7.9. From these sections it follows, with LO(U) denoting an unbiased estimator, that

for $\Sigma_1 = \Sigma_2$ known $\quad LO(U_e) = LO(E_e) - \frac{1}{2}p(\frac{n_1-n_2}{n_1n_2})$

$$(7.12.1)$$

and for $\Sigma_1 \neq \Sigma_2$ known $\quad LO(U_u) = LO(E_u) - \frac{1}{2}p(\frac{n_1-n_2}{n_1n_2})$.

With $n_1 = n_2$ the estimative and unbiased methods are identical and result in identical allocation.

From (7.12.1) $MSE\{LO(U) \mid x\} = MSE\{LO(E) \mid x\} - [\frac{1}{2}p(\frac{n_1-n_2}{n_1n_2})]^2$

for $\Sigma_1 = \Sigma_2$ and $\Sigma_1 \neq \Sigma_2$, known.

For $\Sigma_1 = \Sigma_2$ unknown

$$LO(U_e) = \frac{n_1+n_2-p-3}{n_1+n_2-2} \; LO(E_e) - \tfrac{1}{2}p(\frac{n_1-n_2}{n_1 n_2})$$

and with $n_1 = n_2$ allocation by $LO(E_e)$ and $LO(U_e)$ is identical.

For $\Sigma_1 \neq \Sigma_2$ unknown

$$LO(U_u) = \tfrac{1}{2}[-(\frac{n_1-p-2}{n_1-1})\; w_1(x) + (\frac{n_2-p-2}{n_2-1})\; w_2(x) - p(\frac{n_1-n_2}{n_1 n_2})$$

$$-p\ln(\frac{n_1-1}{n_2-1}) + \sum_{i=1}^{p} \{\psi(\frac{n_1-i}{2}) - \psi(\frac{n_2-i}{2})\} - \ln(|S_1|/|S_2|)].$$

The mean square errors of $LO(U_e)$ may for Σ unknown be obtained from Section 7.8 and Appendix 7.A. For $n_1 = n_2 = n$ they are

$$MSE\{LO(U_e) \mid x\} = (\frac{2n-p-3}{2n-2})^2 \; E\{LO^2(E_e) \mid x\} - LO^2(T_e)$$

$$MSE\{LO(U_e)\} \quad = (\frac{2n-p-3}{2n-2})^2 \; E\{LO \; (E_e)\} - (\Delta^2 + \tfrac{1}{4}\Delta^4)$$

$$x \text{ in } \Pi_t,\; t = 1 \text{ and } 2.$$

In Appendix 7.B the unconditional mean square error of $LO(E_u)$, Σ_1 and Σ_2, both unknown but proportional, is derived.

As $MSE\{LO(U_u)\} = E\{LO^2(U_u)\} - E\{LO^2(T_u)\}$ and with $n_1 = n_2$

$$E\{LO^2(U_u)\} = \tfrac{1}{4}E[(\frac{n-p-2}{n-1})^2 \; \{-w_1(x) + w_2(x)\}^2$$

$$+ (\frac{n-p-1}{n-1})-2\{-w_1(x) + w_2(x)\} \; \ln\{|S_1|/|S_2|\} + \{\ln|S_1|/|S_2|\}^2]$$

$$= \tfrac{1}{4}[(\frac{n-p-2}{n-1})^2 \; E(a) + (\frac{n-p-1}{n-1}) \; E(b) + E(c)]$$

where E(a), E(b), E(c) and $E\{LO^2(T_u)\}$ are given in Appendix 7.B, the unconditional mean square error of $LO(U_u)$ for $\alpha\Sigma_1 = \Sigma_2$ unknown is obtained.

As $(\bar{x}_t, S_t)$ is a complete sufficient statistic for (μ_t, Σ_t), it follows from the Rao-Blackwell theorem, that the unbiased estimator LO(U), being a function of these sample parameters is in all cases, the uniform minimum variance unbiased estimator.

Appendix 7.A

The Expectations $E\{LO^2(E_e) \mid x\}$ and $E\{LO^2(E_e)\}$, when Σ is unknown

With Σ unknown

$$LO(E_e) = (\bar{x}_1-\bar{x}_2)'\ S^{-1}\ \{x - \tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\}$$

$$= \nu'\ S^{-1}(x-\eta)$$

where $\nu = \bar{x}_1-\bar{x}_2$ and $\eta = \tfrac{1}{2}(\bar{x}_1+\bar{x}_2)$.

We assume without loss of generality the usual canonical form of the populations Π_1 and Π_2 i.e. y in Π_1 is distributed as $N_p(0,I)$ and in Π_2 as $N_p(\theta,I)$ where $\theta = (\Delta,0,0, \ldots\ 0)'$.

Thus $\quad mS \sim W(p,m,I) \quad$ where $m = n_1+n_2-2$

$$\nu \sim N_p(-\theta, \frac{n_1+n_2}{n_1n_2}\ I)$$

and $\quad \eta \sim N_p(\tfrac{1}{2}\theta, \dfrac{n_1+n_2}{4n_1n_2}\ I)$, both independent of S.

Now

$$E\{LO^2(E_e) \mid \nu,\eta,x\} = E[\{\nu'\ S^{-1}(x-\eta)\}^2]$$

$$= (x-\eta)'\ E[S^{-1}\ \nu\ \nu'\ S^{-1}]\ (x-\eta)$$

$$= \frac{m^2}{d}\ (x-\eta)'\ [\nu'\nu + (m-p-1)\ \nu\nu']\ (x-\eta)$$

$$= \frac{m^2}{d}\ [\nu'\nu\ \ (x-\eta)'(x-\eta) + (m-p-1)\{\nu'(x-\eta)\}^2]$$

where $\quad d = (m-p)\ (m-p-1)\ (m-p-3)$.

The expectation $E[S^{-1}\ \nu\nu'\ S^{-1} \mid \nu]$ is due to Das Gupta (1968) who derived the covariance matrix of the sample discriminant coefficients $S^{-1}\ (\bar{x}_1-\bar{x}_2)$.

It follows that

$$E\{LO^2(E_e) \mid x\} = E_{\nu,\eta}[E\{LO^2(E_e) \mid \nu,\eta,x\}] \tag{7.A.1}$$

$$= \frac{m^2}{d} E_{\nu,\eta}[\nu'\nu \;\; x'x - 2\nu'\nu \;\; x'\eta$$

$$+ \nu'\nu \;\; \eta'\eta + (m-p-1)\;\{\nu'(x-\eta)\}^2]$$

Four individual expectations are required. They are

$$E_{\nu,\eta}\{\nu'\nu \;\; x'x\} = x'x \; \mathrm{tr}E\{\nu\nu'\} = x'x \; \mathrm{tr}\{\frac{n_1+n_2}{n_1 n_2} I + \theta\theta'\}$$

$$= x'x \; \{p(\frac{n_1+n_2}{n_1 n_2}) + \Delta^2\}\;, \tag{7.A.2}$$

$$E_{\nu,\eta}\{\nu'\nu \;\; \eta'\eta\} = \tfrac{1}{4}E[(\bar{x}_1'\bar{x}_1)^2 - 4(\bar{x}_1'\bar{x}_2)^2 + (\bar{x}_2'\bar{x}_2)^2$$

$$+ 2\bar{x}_1'\bar{x}_1\bar{x}_2'\bar{x}_2]$$

$$= \tfrac{1}{4}[\frac{2p+p^2}{n_1^2} - 4(\frac{p}{n_1 n_2} + \frac{\Delta^2}{n_1}) + \frac{2p+4n_2\Delta^2}{n_2^2}$$

$$+ \frac{(p+n_2\Delta^2)^2}{n_2^2} + 2\frac{p}{n_1}(\frac{p+n_2\Delta^2}{n_2})]$$

$$= \tfrac{1}{4}[2p(\frac{1}{n_1} - \frac{1}{n_2})^2 - 4\Delta^2(\frac{1}{n_1} - \frac{1}{n_2}) + (\frac{p}{n_1} + \frac{p}{n_2} + \Delta^2)^2] \tag{7.A.3}$$

as $n_1\bar{x}_1'\bar{x}_1 \sim \chi^2{}_p$ independently of $n_2\bar{x}_2'\bar{x}_2 \sim \chi^2(p,\, n_2\Delta^2)$,

$$E_{\nu,\eta}\{2\ \nu'\nu\ \ x'\eta\} = E[x'\bar{x}_1\ \bar{x}_1'\bar{x}_1 - 2\ x'\bar{x}_1\ \bar{x}_1'\bar{x}_2 + x'\bar{x}_1\ \bar{x}_2'\bar{x}_2$$

$$+ x'\bar{x}_2\ \bar{x}_1'\bar{x}_1 - 2\ x'\bar{x}_2\ \bar{x}_1'\bar{x}_2 + x'\bar{x}_2\ \bar{x}_2'\bar{x}_2]$$

$$= [o - \frac{2x'\Theta}{n_1} + o + \frac{px'\Theta}{n_1} - o + x'\Theta(\frac{p+2}{n_2} + \Delta^2)]$$

$$= x'\Theta[\frac{p-2}{n_1} + \frac{p+2}{n_2} + \Delta^2] \qquad (7.A.4)$$

where the expectations $E\{x'\bar{x}_1\ \bar{x}_1'\bar{x}_1\}$ and $E\{x'\bar{x}_2\ \bar{x}_2'\bar{x}_2\}$ result from $E\{z'y\ y'y \mid z\} = z'\Theta\{(p+2) + \Theta'\Theta\}$ if $y \sim N_p(\Theta, I)$.

$$E_{\nu,\eta}[\{\nu'(x-\eta)\}^2] = \tfrac{1}{4}E[\{(x-\bar{x}_1)'\ (x-\bar{x}_1) - (x-\bar{x}_2)'\ (x-\bar{x}_2)\}^2]$$

as $n_1(x-\bar{x}_1)\ (x-\bar{x}_1) \sim \chi^2(p,\ n_1\ x'x)$ independently of $n_2(x-\bar{x}_2)'\ (x-\bar{x}_2) \sim \chi^2(p,\ n_2(x-\Theta)'\ (x-\Theta))$ it follows that

$$E_{\nu,\eta}[\{\nu'\ (x-\eta)\}^2] = \tfrac{1}{4}[\frac{2p+4n_1x'x}{n_1^2} + \frac{2p+4n_2(x-\Theta)'(x-\Theta)}{n_2^2}$$

$$+ (\frac{p+n_1x'x}{n_1} - \frac{p+n_2(x-\Theta)'(x-\Theta)}{n_2})^2]$$

$$= \tfrac{1}{4}[2p(\frac{1}{n_1^2} + \frac{1}{n_1^2}) + \frac{4x'x}{n_1} + \frac{4(x-\Theta)'(x-\Theta)}{n_2}$$

$$+ (p(\frac{n_1-n_2}{n_1n_2}) + 2\ LO(T_e))^2]. \qquad (7.A.5)$$

For the unconditional expectations

$$E\{LO^2(E_e)\} = E_x[E\{LO^2(F_e) \mid x\}]$$

and x in Π_t, t = 1 and 2, formulae (7.A.2), (7.A.3), (7.A.4) and (7.A.5) may be combined as in (7.A.1) noting that for t = 1, $x \sim N_p(0,I)$ and

$$E(x'x) = p, \quad E(x'\Theta) = 0, \quad E\{(x-\Theta)'\,(x-\Theta)\} = p + \Delta^2$$

$$E\{LO(T_e)\} = \tfrac{1}{2}\Delta^2 \quad \text{and} \quad E\{LO^2(T_e)\} = \Delta^2 + \tfrac{1}{4}\Delta^4,$$

for t = 2, $x \sim N_p(\Theta,I)$ and

$$E(x'x) = p + \Delta^2, \quad E(x'\Theta) = \Delta^2, \quad E\{(x-\Theta)'\,(x-\Theta)\} = p$$

$$E\{LO(Te)\} = -\tfrac{1}{2}\Delta^2 \quad \text{and} \quad E\{LO^2(Te)\} = \Delta^2 + \tfrac{1}{4}\Delta^4.$$

With $n_1 = n_2 = n$ the formulae for the expectations may be reduced considerably to

$$E\{LO^2(E_e) \mid x\} = \frac{m^2}{d}[\{\Delta^2 + \frac{2}{n}(m-1)\}\{x-\tfrac{1}{2}(\mu_1+\mu_2)\}'\Sigma^{-1}\{x-\tfrac{1}{2}(\mu_1+\mu_2)\}$$

$$+ (m-1)\{\frac{p}{n^2} + \frac{\Delta^2}{2n}\} + (m-p-1)\, LO^2(Te)]$$

and

$$E\{LO^2(E_e)\} = \frac{m^2}{d}[(m-1)\{p(\frac{2n+1}{n^2}) + \frac{n+1}{n}\Delta^2\} + (m-p)\,\Delta^4/4],$$

x in Π_t, t = 1 and 2.

Appendix 7.B

The Unconditional Mean Square Error of $LO(E_u)$ when the Covariance Matrices Σ_t are Proportional but Unknown.

With proportional covariance matrices $\alpha\Sigma_1 = \Sigma_2$, $\alpha > 0$, $\alpha \neq 1$ it may be assumed without loss of generality that $\mu_1 = 0$, $\mu_2 = \theta$ where $\theta = (\Delta_1, 0, 0 \text{-----} 0)'$ with $\Delta_1^2 = (\mu_1 - \mu_2)' \, \Sigma_1^{-1} \, (\mu_1 - \mu_2)$, $\Sigma_1 = I$ and $\Sigma_2 = \alpha I$. For Σ_t unknown

$$LO(E_u) = \tfrac{1}{2}\{-w_1(x) + w_2(x) - \ell n(|S_1|/|S_2|)\}$$

where $w_t(x) = (x-\bar{x}_t)' \, S_t^{-1} (x-\bar{x}_t) \qquad t = 1 \text{ and } 2,$

$$\text{and} \quad MSE\{LO(E_u)\} = E[\{LO(E_u) - LO(T_u)\}^2] \qquad (7.B.1)$$

$$= E\{LO^2(E_u)\} - 2E\{LO(E_u) \; LO(T_u)\} + E\{LO^2(T_u)\}.$$

Now $E\{LO^2(T_u)\}$ x in Π_t, t = 1 and 2 may be derived in general from the results of Section 7.3, for proportional covariance matrices the results are

$$E\{LO^2(T_u)\} = \tfrac{1}{2}p - p/\alpha + \tfrac{1}{2}p/\alpha^2 + \Delta_1^2/\alpha^2 + \tfrac{1}{4}\{p(\tfrac{1-\alpha}{\alpha})$$

$$+ \Delta_1^2/\alpha + p\ell n\alpha\}^2 \qquad x \in \Pi_1 \qquad (7.B.2)$$

$$\text{and} \quad E\{LO^2(T_u)\} = \tfrac{1}{2}p - p\alpha + \tfrac{1}{2}p\alpha^2 + \alpha\Delta_1^2 + \tfrac{1}{4}\{p(1-\alpha) - \Delta_1^2 + p\ell n\alpha\}^2$$

$$x \in \Pi_2$$

The expectation $E\{LO(E_u) \; LO(T_u)\}$ may be derived in general by considering $E_x[E\{LO(E_u) \mid x\} \; LO(T_u)]$, where the conditional expectation $E\{LO(E_u) \mid x\}$ is given in Section 7.9. Once again the results of Section 7.3 may be employed to obtain the unconditional expectation.

For proportional covariance matrices and equal sample sizes $n_1 = n_2 = n$ for simplicity, the results are

$$E\{LO(E_u)\ LO(T_u)\} = \tfrac{1}{4}[\frac{n-1}{n-p-2}\{2p + p^2 - 4p/\alpha - \frac{2p\Delta_1^2}{\alpha} - \frac{2p^2}{\alpha} + \frac{2p}{\alpha^2} + \frac{4\Delta_1^2}{\alpha^2} + (\frac{p+\Delta_1^2}{\alpha})^2\}$$

$$+ p\ln\alpha\ (\frac{2n-p-3}{n-p-2})\{-p + \frac{p}{\alpha} + \frac{\Delta_1^2}{\alpha}\} + (p\ln\alpha)^2] \qquad x \in \Pi_1 \qquad (7.B.3)$$

and
$$E\{LO(E_u)\ LO(T_u)\} = \tfrac{1}{4}[\frac{n-1}{n-p-2}\{2p\alpha^2 + 4\alpha\Delta_1^2 + (p\alpha+\Delta_1^2)^2 - 2p^2\alpha - 4p\alpha - 2p\Delta_1^2 + 2p + p^2\} + p\ln\alpha(\frac{2n-p-3}{n-p-2})\{-\alpha p - \Delta_1^2 + p\} + (p\ln\alpha)^2]$$

$$x \in \Pi_2 \; .$$

Now
$$E\{LO^2(E_u)\} = \tfrac{1}{4}E[\{-w_1(x) + w_2(x)\}^2 - 2\{-w_1(x) + w_2(x)\}\ln(|S_1|/|S_2|) + \{\ln(|S_1|/|S_2|)\}^2]$$

$$= \tfrac{1}{4}E[\ a + b + c]\ ,$$

the expectation of a, b and c are considered separately.

Consider $E\{a\} = E\{w_1^2(x) - 2w_1(x)\ w_2(x) + w_2^2(x)\}$

the expectation of the covariance may be derived conditional on x first and then unconditionallyusing the results of Section 7.3. The expectations are

$$E(a) = (\frac{n^2-1}{n})^2 [\frac{p^2 + 2p}{(n-p-2)\ (n-p-4)}] + (\frac{(n+\alpha)\ (n-1)}{\alpha n})^2 [\frac{(p+\lambda_1)^2 + 2(n+2\lambda_1)}{(n-p-2)(n-p-4)}]$$

$$-2(\frac{n-1}{n(n-p-2)})^2\ [p^2 + np(p + p/\alpha + \Delta_1^2/\alpha) + n^2(2p/\alpha + \frac{p^2}{\alpha} + \frac{p\Delta_1^2}{\alpha})]$$

for x in Π_1 and $\lambda_1 = \frac{n}{n+\alpha}\Delta_1^2$, (7.B.4)

and $$E(a) = (\frac{(\alpha n+1)(n-1)}{n})^2\ [\frac{(p+\lambda_2)^2 + 2(p+2\lambda_2)}{(n-p-2)(n-p-4)}] + (\frac{n^2-1}{n})^2 [\frac{p^2 + 2p}{(n-p-2)(n-p-4)}]$$

$$- 2(\frac{n-1}{n(n-p-2)})\ [p^2 + np(\alpha p + \Delta_1^2 + p) + n^2(2p\alpha + p^2\alpha + p\Delta_1^2)]$$

for x in Π_2 and $\lambda_2 = \frac{n}{\alpha n+1}\Delta_1^2$. (7.B.5)

Consider now $E\{c\} = E\{\ell n^2|S_1| - 2\ell n|S_1|\ell n|S_2| + \ell n^2|S_2|\}$

where S_1 and S_2 are independently distributed. Now $|S_t| = |A_t|/(n_t-1)^p$ where $A_t \sim W_p(n_t-1,\Sigma_t)$ and $|A_t|/|\Sigma_t| \sim \prod_{i=1}^{p} \chi^2_{n_t-i}$ where the χ^2's are independent, Anderson (1958, p171), hence $\ell n|A_t|/|\Sigma_t| \sim \sum_{i=1}^{p} \ell n\ \chi^2_{n_t-i}$.

Now $E\{\ell n\ \chi^2_{n_t-i}\} = \psi(\frac{n_t-i}{2}) + \ell n\ 2$ and

$$V\{\ell n\ \chi^2_{n_t-i}\} = \psi'(\frac{n_t-i}{2}) \quad \text{where} \quad \psi'(z) = \frac{d^2}{dz^2}\ \ell n\ \Gamma(z)$$

is usually referred to as the trigamma function, Abromowitz and Stegun (1965, p260). It follows with some manipulation that

$$E\{\ell n\ |S_t|\} = \ell n\ |\Sigma_t| - p\ell n\ (\frac{n_t-1}{2}) + \sum_{i=1}^{p} \psi(\frac{n_t-i}{2})$$

and

$$E\{\ell n^2|S_t|\} = \sum_{i=1}^{p} \psi'(\frac{n_t-i}{2}) + [E\{\ell n|S_t|\}]^2.$$

Hence with $n_1 = n_2$

$$\begin{aligned} E\{c\} &= 2 \sum_{i=1}^{p} \psi'(\tfrac{n-i}{2}) + [\ell n|\Sigma_1| - \ell n|\Sigma_2|]^2 \\ &= 2 \sum_{i=1}^{p} \psi'(\tfrac{n-i}{2}) + (p\ell n\alpha)^2. \end{aligned} \qquad (7.B.6)$$

Finally consider

$$E\{b\} = 2\ E\{w_1(x)\ell n|S_1| - w_2(x)\ell n|S_1| - w_1(x)\ell n|S_2| + w_2(x)\ell n|S_2|\}$$

the expectations $E\{w_1(x)\ell n|S_2|\}$ and $E\{w_2(x)\ell n|S_1|\}$ are easily derived by independence, the expectations $E\{w_1(x)\ell n|S_1|\}$ and $E\{w_2(x)\ell n|S_2|\}$ are more difficult and have only been derived when the covariance matrices are proportional. We give the derivation in some detail for x in Π_1, the results for x in Π_2 require only minor changes.

With x in Π_1

$$E\{\ell n|S_1|\ (x-\bar{x}_1)'\ S_1^{-1}(x-\bar{x}_1)\} = \frac{n_1+1}{n_1} E\{\ell n|S_1|trS_1^{-1}\}$$

$$E\{\ell n|S_2|\ (x-\bar{x}_2)'\ S_2^{-1}(x-\bar{x}_2) = \frac{n_2+\alpha}{n_2} E\{\ell n|S_2|trS_2^{-1} + \ell n|S_2|\theta_1^2\ S_2^{11}\}$$

where S_2^{11} is the first element in the trace of S_2^{-1}. The expectation $E\{\ell n|S_1|tr\ S_1^{-1}\}$ is required, a similar proof holds for t = 2.

Now $(n_1-1)\ S_1 \sim W_p(n_1-1,I)$, let $S_1 = A/(n_1-1)$ then

$$\ell n|S_1|tr\ S_1^{-1} = (n_1-1)\{\ell n|A|tr\ A^{-1} - p\ell n(n_1-1)\ tr\ A^{-1}\}$$

and as $E(A^{-1}) = \Sigma_1^{-1}/(n_1-p-2)$, $E\{tr\ A^{-1}\} = p/(n_1-p-2)$.

Consider now the expectation $E\{\ell n|A|\text{tr } A^{-1}\}$

if $A = (a_{ij})$ $i,j = 1,2,$ ----- p ,

$A_r = (a_{ij})$ $i,j = 1,2,$ ----- r, $1 \leqslant r \leqslant p$

and $A^{-1} = (a^{ij})$ $i,j = 1,2,$ ---- p then $\text{tr } A^{-1} = \sum_{i=1}^{p} a^{ii}$.

$$\text{Now} \quad |A| = \frac{|A_p|}{|A_{p-1}|} \times \frac{|A_{p-1}|}{|A_{p-2}|} \times \text{---------} \times \frac{|A_2|}{|A_1|} \times |A_1|$$

where $a^{pp} = \dfrac{|A_{p-1}|}{|A_p|}$ by definition of A^{-1}

and as the ordering of the elements is arbitrary, this approach applies to a^{ii} in general.

With $y_i = \dfrac{|A_i|}{|A_{i-1}|}$, $1 \leqslant i \leqslant p$ and $|A_o| = 1$ say, then

$$|A| = \prod_{i=1}^{p} y_i \ ;$$

where y_i, $1 \leqslant i \leqslant p$ are independently distributed a central χ^2 with n_1-i degrees of freedom, Rao (1973, p540), thus

$$E\{\ell n|A|\ a^{pp}\} = E\{(\sum_{i=1}^{p} \ell n\ y_i)\ \frac{1}{y_p}\}$$

$$= E\{\frac{\ell n y_p}{y_p}\} + E\{\sum_{i=1}^{p-1} y_i\}\ E\{\frac{1}{y_p}\}$$

$$= \frac{1}{n_1 - p - 2}[\psi(\frac{n_1 - p - 2}{2}) + \sum_{i=1}^{p-1} \psi(\frac{n_1 - 1}{2}) + p\ell n2]$$

and $E\{\ell n|A|\ \text{tr } A^{-1}\} = p\ E\{\ell n|A|\ a^{pp}\}$.

Hence $E\{\ell n\ |S_1|\ \text{tr}\ S_1^{-1}\} = \frac{(n_1-1)p}{n_1-p-2}[\psi(\frac{n_1-p-2}{2}) + \sum_{i=1}^{p-1} \psi(\frac{n_1-i}{2}) - p\ell n(\frac{n_1-1}{2})]$

similarly

$$E\{\ell n|S_2|\text{tr}\ S_2^{-1}\} = \frac{(n_2-1)p}{n_2-p-2}\,\frac{1}{\alpha}[\psi(\frac{n_2-p-2}{2}) + \sum_{i=1}^{p-1} \psi(\frac{n_2-i}{2}) - p\ell n(\frac{n_2-1}{2}) + p\ell n\alpha].$$

After some lengthymanipulation and $n_1 = n_2$ we have

$$E\{b\} = 2\,\frac{n^2-1}{n}\,\frac{p}{n-p-2}\,\{\frac{-2}{n-p-2} - p\ell n\alpha\}$$

$$+\ 2\,\frac{(n+\alpha)(n-1)}{\alpha n}\,\frac{(p+\lambda_1)}{n-p-2}\,\{-\,\frac{2}{n-p-2} + p\ell n\alpha\},\ x \text{ in } \Pi_1$$

and $$E\{b\} = 2\,\frac{n^2-1}{n}\,\frac{p}{n-p-2}\,\{-\,\frac{2}{n-p-2} + p\ell n\alpha\} \qquad (7.B.7)$$

$$+\ 2\,\frac{(\alpha n+1)(n-1)}{n}\,\frac{(p+\lambda_2)}{n-p-2}\{-\,\frac{2}{n-p-2} - p\ell n\alpha\},\ x \text{ in } \Pi_2$$

where λ_1 and λ_2 are as defined in (7.B.4) and (7.B.5). Combining the results (7.B.2), (7.B.3), (7.B.4), (7.B.5), (7.B.6) and (7.B.7) in (7.B.1) the unconditional mean square error of $LO(E_u)$, x in Π_t, $t = 1$ and 2, when $\alpha\Sigma_1 = \Sigma_2$, is obtained.

CHAPTER 8

COMPARISONS OF THE ESTIMATIVE AND PREDICTIVE ESTIMATORS OF LOG-ODDS

8.1 Introduction

The results of the previous chapter supplemented by simulation studies will be used to compare the estimative and predictive approaches to estimation of true log-odds. An adjusted predictive estimator which closely resembles the likelihood ratio approach of Section 6.3 is also considered.

8.2 Canonical Forms and True Log-Odds

Without loss of generality it is assumed that when $\Sigma_1 = \Sigma_2 = \Sigma$, the true densities f_t (6.2.4) are given by

$$f_1 = N_p(0,I) \quad \text{and} \quad f_2 = N_p(\theta,I)$$

where θ is a p-dimensional vector, the first element of which is Δ and the remaining elements zero. The true distances $\omega_t(x)$ of an observation x from a population mean as defined in (6.2.4) are therefore given by

$$\omega_1(x) = \sum_{i=1}^{p} x_i^2 \quad \text{and} \quad \omega_2(x) = (x_1-\Delta)^2 + \sum_{i=2}^{p} x_i^2 \qquad (8.2.1)$$

and the true log-odds may be written as

$$\begin{aligned} LO(T_e) &= \tfrac{1}{2}\{-\omega_1(x) + \omega_2(x)\} \\ &= -\Delta\{x_1 - \tfrac{1}{2}\Delta\}. \qquad (8.2.2) \end{aligned}$$

8.3 A Comparison of the Estimators for $\Sigma_1 = \Sigma_2$ known and Equal Sample Sizes.

In this and the following section we compare the estimative and predictive estimators when $\Sigma_1 = \Sigma_2 = \Sigma$ is known. Our reasons for considering the Σ known case are; exact expressions for the conditional and unconditional bias and mean square error of all the estimators were derived withthis assumption in Chapter 7. When Σ is unknown the mean square error of the predictive estimator is unavailable. In Chapter 2 with Σ known we have established the effect of unequal sample sizes on the classification behaviour of $\hat{L}(x) = LO(E_e)$ and $Z(x) \propto LO(P_e)$. Now we can investigate the interrelationship between bias and misclassification. Finally our study of the Σ known case indicates the form of our study when Σ is unknown.

For the moment we assume that the sample sizes are equal i.e. $n_1 = n_2 = n$. It follows from Sections 7.4,7.6 and 7.12 that

$$LO(E_e) = LO(U_e)$$

and (8.3.1)

$$LO(P_e) = \frac{n}{n+1} LO(E_e)$$

with unconditional bias and mean square errors

$$B\{LO(E_e)\} = B\{LO(U_e)\} = 0$$

$$B\{LO(P_e)\} = \frac{-1}{n+1} \; E\{LO(T_e)\} = -\frac{1}{n+1} \tfrac{1}{2}\Delta^2 \; (-1)^{t-1}$$

$$MSE\{LO(E_e)\} = p(\frac{2n+1}{n^2}) + \frac{\Delta^2}{n} \qquad (8.3.2)$$

and

$$MSE\{LO(P_e)\} = (\frac{n}{n+1})^2 \; MSE\{LO(E_e)\} + \frac{1}{(n+1)^2} E\{LO^2(T_e)\}$$

$$= \frac{1}{(n+1)^2} [p(2n+1) + \Delta^2(n+1) + \tfrac{1}{4}\Delta^2],$$

x in Π_t $t = 1$ and 2.

From (8.3.2) we note that the estimative method is (for $n_1 = n_2$) unbiased, while $LO(P_e)$ is negatively biased thus understating true log-odds. We also note that the relative bias of $LO(P_e)$ i.e.

$$B\{LO(P_e)\} \; / \; E\{LO(T_e)\}$$

is $-\frac{1}{n+1}$. This relative bias is small, independent of p and Δ and decreased as n increases.

Consider the difference

$$\begin{aligned} DMSE(\Delta) &= MSE\{LO(E_e)\} - MSE\{LO(P_e)\} \\ &= \frac{1}{(n+1)^2}\left[p\left(\frac{2n+1}{n}\right)^2 + \Delta^2\left(\frac{n+1}{n}\right) - \tfrac{1}{4}\Delta^4\right]. \end{aligned}$$

It is clear that it is a function of Δ and the following properties are easily deduced.

DMSE(Δ) has maximum positive difference when

$$\Delta = \sqrt{2\left(\frac{n+1}{n}\right)} = \Delta_{max} \text{ say}$$

$$DMSE(\Delta_{max}) = \frac{1}{(n+1)^2}\left[p\left(\frac{2n+1}{n}\right)^2 + \left(\frac{n+1}{n}\right)^2\right]$$

$DMSE(\Delta) > 0$ for $0 \leqslant \Delta < \Delta_o$ where

$$\Delta_o = \left[\frac{2}{n}\left\{(n+1) + \sqrt{p(2n+1)^2 + (n+1)^2}\right\}\right]^{\frac{1}{2}}$$

$DMSE(\Delta_o) = 0$ and $DMSE(\Delta) < 0$ for $\Delta > \Delta_o$

$DMSE(\Delta) \to 0$ as $n \to \infty$ and $DMSE(\Delta) \to \infty$ as $p \to \infty$.

In Table 8.3.1 we give for a range of values of n and p the maximum difference $DMSE(\Delta_{max})$, the value of $MSE\{LO(E_e)\}$ at Δ_{max} and the value Δ_o at which $MSE\{LO(E_e)\} = MSE\{LO(P_e)\}$.

Table 8.3.1

The Maximum Difference in the Unconditional Mean Square Error of the Estimators LO(E_e) and LO(P_e), Σ known and $n_1 = n_2$.

DMSE(Δ_{max})

n p	12	24	48
1	.03	.01	.00
4	.11	.03	.01
8	.21	.06	.01
16	.42	.11	.03

MSE{LO(E_e)} at Δmax

n p	12	24	48
1	0.35	0.17	0.08
4	0.87	0.43	0.21
8	1.59	0.77	0.38
16	2.96	1.45	0.72

Value at Δ_0 at which DMSE(Δ) = 0

n p	12	24	48
1	2.62	2.58	2.56
4	3.28	3.24	3.22
8	3.76	3.72	3.70
16	4.36	4.31	4.28

From Table 8.3.1 we see that the maximum difference in mean square error between the estimators is slight decreasing with increasing n but increasing as p increases. This slight difference in the estimators may be anticipated from (8.3.1). The value of Δ_o at which the mean square errors of the estimators are equal, indicates that the populations must be well separated for the predictive method to have the larger mean square error.

We now consider the estimators for fixed categories of true log-odds to see if the behaviour of the estimators is uniform over the range of possible $LO(T_e)$.

From Sections 7.4, 7.6 and 7.12 the conditional bias and mean square error of the estimators are

$$\begin{aligned} B\{LO(E_e)|x\} &= B\{LO(U_e)|x\} = 0 \\ B\{LO(P_e)|x\} &= -\frac{1}{n+1} LO(T_e) \qquad (8.3.3) \\ MSE\{LO(E_e)|x\} &= \frac{p}{n^2} + \frac{1}{n}\{\omega_1(x) + \omega_2(x)\} \\ MSE\{LO(P_e)|x\} &= (\frac{n}{n+1})^2 MSE\{LO(E_e)|x\} + [B\{LO(P_e)|x\}]^2. \end{aligned}$$

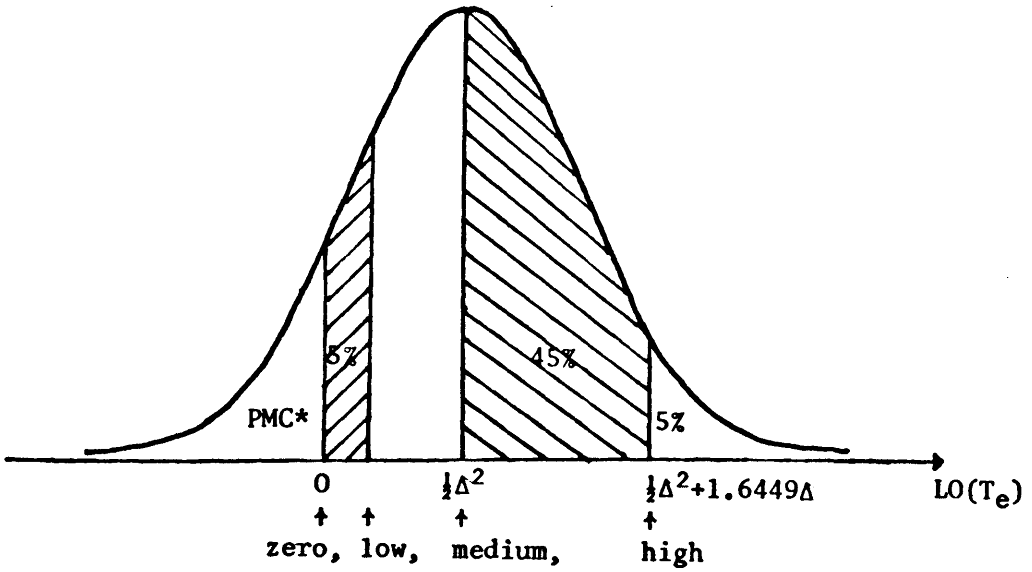
From (8.3.3) we see that the mean square errors of the estimators depend on the individual distances $\omega_t(x)$ and not simply their difference as in $LO(T_e)$. Choice of Δ and x_1 in (8.2.2) determines $LO(T_e)$. We must now specify reasonable values for the remaining p-1 elements of x, where x_i, $2<i<p$ are independent and identically distributed as $N_1(0,1)$ variates, Section 8.2. The values chosen were the expected values of the order statistics of a random sample of size p-1 from a $N_1(0,1)$ distribution. In retrospect it is clear from (8.2.1) that we need only have specified a reasonable value for

$$\sum_{i=2}^{p} x_i^2 \sim \chi^2_{p-1} ,$$

the mean p-1 being one possibility.

Figure 8.3.1

Distribution of $LO(T_e)$, x in Π_1 as $N_1(\frac{1}{2}\Delta^2,\Delta^2)$.



*Optimal probability of misclassification given by $\Phi(-\frac{1}{2}\Delta)$.

In fixing true log-odds for x from Π_1 we wished to investigate the behaviour of the estimators for high, medium, low and zero log-odds. Now for x in Π_1, $LO(T_e) \sim N_1(\frac{1}{2}\Delta^2, \Delta^2)$ and so a medium $LO(T_e)$ say varies with Δ. Thus we set $LO(T_e)$ equal to the 95th, 50th, $(\Phi(-\frac{1}{2}\Delta) + .05)$ x 100th and $\Phi(-\frac{1}{2}\Delta)$ x 100th percentiles for high, medium, low and zero log-odds, where $\Phi(-\frac{1}{2}\Delta)$ is the optimal probability of misclassification (PMC), Figure 8.3.1.

In Table 8.3.2 we tabulate the conditional mean square error of $LO(E_e)$ and $LO(P_e)$ for $n = 12$, $p = 1,4,8$ and 16, $\Delta = 1.049$ and 3.290. As well as indicating the category of true log-odds we also give in brackets the value of $LO(T_e)$.

For high true log-odds and as might be anticipated from the unconditional results, with Δ large, the conditional mean square error of $LO(P_e)$ exceeds that of $LO(E_e)$, Table 8.3.2. Overall the difference in mean square errors are slight and will decrease with increasing sample size n. As the bias of $LO(P_e)$ (8.3.3) depends on the size of $LO(T_e)$ it accounts for the larger mean square error of $LO(P_e)$ for high true log-odds when Δ is large.

We conclude that with Σ known and $n_1 = n_2$ there is little to choose in bias and mean square error between the estimators, which have identical allocation. Given the conservative bias of $LO(P_e)$ and its smaller mean square error for reasonable values of Δ, and so of $LO(T_e)$, the predictive approach is preferable.

Table 8.3.2

The Conditional Mean Square Error of $LO(E_e)$ and $LO(P_e)$, Σ known and $n_1 = n_2$.

Sample Sizes $n_1 = n_2 = 12$

Method		E_e & U_e				P_e			
Category of $LO(T_e)$		zero	low	med.	high	zero	low	med.	high
Size of $LO(T_e)$		(0)	(0.15)	(0.55)	(2.28)	(0)	(0.15)	(0.55)	(2.28)
Δ & PMC	p								
	1	0.05	0.06	0.10	0.84	0.04	0.05	0.09	0.74
$\Delta = 1.049$	4	0.31	0.32	0.36	1.10	0.27	0.27	0.31	0.97
.30	8	0.94	0.95	0.99	1.73	0.81	0.81	0.84	1.50
	16	2.29	2.29	2.33	3.07	1.95	1.95	1.99	2.65
Size of $LO(T_e)$		(0)	(1.19)	(5.41)	(10.82)	(0)	(1.19)	(5.41)	(10.82)
	1	0.46	0.48	0.91	2.26	0.39	0.42	0.95	2.62
$\Delta = 3.290$	4	0.72	0.74	1.17	2.52	0.61	0.64	1.17	2.84
.05	8	1.35	1.37	1.80	3.15	1.15	1.18	1.71	3.38
	16	2.69	2.71	3.14	4.49	2.29	2.32	2.85	4.52

8.4 A Comparison of the Estimators for Σ known and Unequal Sample Sizes

With unequal sample sizes the situation is more complicated and more interesting. The estimators are now,

$$LO(E_e) = \tfrac{1}{2}\{-w_1(x) + w_2(x)\}$$

$$LO(U_e) = LO(E_e) - \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right)$$

$$LO(P_e) = \tfrac{1}{2}p\ell n\ \left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\} + \tfrac{1}{2}\left\{-\frac{n_1}{n_1+1}w_1(x) + \frac{n_2}{n_2+1}w_2(x)\right\}.$$

With $\hat{L}(x)$ and $Z(x)$ denoting the sample linear discriminant function and Z statistic of Chapter 2, we note that

$$LO(E_e) = \hat{L}(x)$$

$$LO(U_e) = \hat{L}(x) - \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right) \qquad (8.4.1)$$

$$LO(P_e) = \tfrac{1}{2}Z(x) + \tfrac{1}{2}p\ \ell n\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\}$$

and so the unconditional distribution of the estimators may be obtained from the results of Chapter 2.

Given the connection between bias and allocation i.e. persistent understatement of true log-odds may be so severe that the sign of the estimator is incorrect and so misclassifies, we investigated the effect of unequal sample sizes on the misclassification performance of the estimators. With $n_1 = 8$, $n_2 = 40$, assuming that in all cases allocation is based on the sign of the estimators, evaluation of their exact expected actual PMC's, Section 2.5, indicates that in contrast to LO(E_e) = $\hat{L}(x)$, Tables 2.10.1 and 2.10.2, the unbiased rule

$$LO(U_e) = \hat{L}(x) - \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right)$$

almost equates the expected actual PMC's. In fact LO(U_e) is in this regard almost as good as the Z statistic, Section 2.10, whereas LO(P_e) a linear function of Z(x), (8.4.1) is almost as sensitive to $n_1 \neq n_2$ as is LO(E_e). The above suggest that an adjusted predictive estimator which we will denote by LO(PA_e), where

$$LO(PA_e) = LO(P_e) - \tfrac{1}{2}p\ell n\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\} = \tfrac{1}{2}\, Z(x) \tag{8.4.2}$$

is worthy of consideration. This adjusted predictive estimator is in fact the likelihood ratio estimator of Section 6.3. We also conclude that adjustment of the cut-off point from zero to

$$\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right)$$

in the standard linear allocation rule $\hat{L}(x)$, will improve its classification behaviour when $n_1 \neq n_2$.

With $n_1 \neq n_2$ the unconditional bias of the four estimators are

$$B\{LO(E_e)\} = \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right)$$

$$B\{LO(U_e)\} = 0 \tag{8.4.3}$$

$$B\{LO(P_e)\} = \tfrac{1}{2}p\ell n\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\} - \frac{1}{n_{3-t}+1}\,\tfrac{1}{2}\Delta^2(-1)^{t-1}$$

$$B\{LO(PA_e)\} = B\{LO(P_e)\} - \tfrac{1}{2}p\ell n\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\}$$

$x \in \Pi_t$, t=1 and 2.

We note that the relative biasi.e. $B\{LO(M)\}/E\{LO(T_e)\}$- where $E\{LO(T_e)\} = \frac{1}{2}\Delta^2(-1)^{t-1}$ $t = 1$ and 2,if negative denotes understatement, if positive overstatement, of true log-odds for Π_1 with x in Π_1 or Π_2. From (8.4.3) we see that with $n_1 \neq n_2$ the bias of $LO(E_e)$ depends on p but not on Δ. However its relative bias does and decreases as Δ increases. If $n_1 < n_2$ the relative bias of $LO(E_e)$ is negative in Π_1 but positive in Π_2 and vica versa if $n_1 > n_2$. The bias and relative bias of $LO(P_e)$ now depends on p as well as Δ. If $n_1 < n_2$ the relative bias of $LO(P_e)$ is negative in Π_1, for Π_2 true log-odds may be understated or overstated depending on n_1, n_2, p and Δ. The bias of the adjusted predictive estimator $LO(PA_e)$ is independent of p, its relative bias is

$$-\frac{1}{n_{3-t}+1} \quad t = 1 \text{ and } 2$$

which is always negative, independent of Δ and so $LO(PA_e)$ understates true log-odds in both populations.

The unconditional mean square errors of the estimators are

$$MSE\{LO(E_e)\} = \tfrac{1}{2}p\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right) + p\left(\frac{n_1+n_2}{n_1 n_2}\right) + \Delta^2/n_2 + [B\{LO(E_e)\}]^2, \quad x \text{ in } \Pi_1$$

where, for x in Π_2 Δ^2/n_2 is replaced by Δ^2/n_1.

$$MSE\{LO(U_e)\} = MSE\{LO(E_e)\} - [B\{LO(E_e)\}]^2 \quad x \text{ in } \Pi_t,\ t = 1 \text{ and } 2. \qquad (8.4.4)$$

$$\text{MSE}\{\text{LO}(P_e)\} = \tfrac{1}{2}p\{\frac{1}{(n_1+1)^2} + \frac{1}{(n_2+1)^2}\} + p\{\frac{n_1}{(n_1+1)^2} + \frac{n_2}{(n_2+1)^2}\} \qquad (8.4.4)$$

$$+ [\tfrac{1}{2}p \, \ell n\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\}]^2 + \tfrac{1}{2}p\{\frac{1}{(n_1+1)} - \frac{1}{(n_2+1)}\}^2$$

$$+ \frac{\Delta^2}{n_2+1} + \frac{\frac{1}{4}\Delta^4}{(n_2+1)^2} - \tfrac{1}{2}p \, \ell n\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\} \frac{\Delta^2}{n_2+1} (-1)^{t-1}$$

$$x \text{ in } \Pi_1$$

where, for x in Π_2 Δ^2/n_2+1 is replaced by Δ^2/n_1+1.

$$\text{MSE}\{\text{LO}(PA_e)\} = \text{MSE}\{\text{LO}(P_e\} - \tfrac{1}{2}p \, \ell n\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\} [B\{\text{LO}(P_e)\}]$$

$$x \text{ in } \Pi_t, \ t = 1 \text{ and } 2.$$

From (8.4.4) we note that if $n_1 < n_2$ the mean square errors of LO(E_e), LO(U_e) and LO(PA_e) for Π_1 are less than their counterparts in Π_2. With $n_1 \neq n_2$ the mean square error of LO(U_e) is always less than that of LO(E_e). With $n_1 < n_2$,

$$\tfrac{1}{2}p \, \ell n\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\}$$

is negative as is the bias of LO(P_e) in Π_1 hence the mean square error of LO(PA_e) (8.4.4), is less than that of LO(P_e) in Π_1. For Π_2 however the mean square error of LO(PA_e) may be larger than that of LO(P_e) depending on the relative size of n_1 to n_2, and the size of p and Δ.

In Table 8.4.1 we tabulate some unconditional biases and mean square errors of the four estimators for both populations. The sample sizes were chosen so that $n_1 + n_2 = 48$ and evaluations were carried out for various ratios of n_1 to n_2, the particular case listed is $n_1 = 8$, $n_2 = 40$. Results for other ratios were similar but less marked.

From the biases Table 8.4.1, we note as shown the understatement of true log-odds by $LO(E_e)$ and $LO(P_e)$ in Π_1 and the overstatment by $LO(E_e)$ in Π_2. This corresponds with the misclassification behaviour noted in these estimators. The size of the bias of $LO(PA_e) = \frac{1}{2}Z(x)$ and its relationship with $LO(P_e)$ (8.4.2) indicates the amount of understatement that may occur in the predictive approach without seriously imbalancing the misclassification rates. We also note in Table 8.4.2 that there is little difference between the mean square errors of the estimators, with $LO(E_e)$ having the largest mean square error in both populations.

We conclude that for Σ known and $n_1 \neq n_2$, the unbiased estimator $LO(U_e)$ is superior to $LO(E_e)$, with smaller mean square error, zero bias and good classification behaviour. The adjusted predictive estimator $LO(PA_e)$ even though it may have slightly larger mean square error than $LO(P_e)$, has good classification behaviour, understates $LO(T_e)$ in both populations and is considered the better predictive approach. The choice between $LO(U_e)$ and $LO(PA_e)$ is more difficult as they have similar classification behaviour. The slight conservative bias of $LO(PA_e)$ and its somewhat smaller mean square error give it an edge here.

Table 8.4.1

The Unconditional Bias and Mean Square Error of LO(M), M in $\{E_e, U_e, P_e, PA_e\}$, Σ known and $n_1 \neq n_2$.

Sample Sizes

$n_1 = 8$, $n_2 = 40$

Population		Π_1				Π_2			
Method		E_e	U_e	P_e	PA_e	E_e	U_e	P_e	PA_e
Δ & PMC	p	Bias				Bias			
$\Delta = 1.049$	1	-.05	0	-.06	-.01	-.05	0	.01	.06
.30	4	-.20	0	-.20	-.01	-.20	0	-.13	.06
	8	-.40	0	-.39	-.01	-.40	0	-.31	.06
	16	-.80	0	-.76	-.01	-.80	0	-.68	.06
$\Delta = 3.290$	1	-.05	0	-.18	-.13	-.05	0	.55	.60
.05	4	-.20	0	-.32	-.13	-.20	0	.42	.60
	8	-.40	0	-.50	-.13	-.40	0	.23	.60
	16	-.80	0	-.88	-.13	-.80	0	-.14	.60
Method		E_e	U_e	P_e	PA_e	E_e	U_e	P_e	PA_e
Δ & PMC	p	MSE				MSE			
$\Delta = 1.049$	1	0.19	0.19	0.16	0.16	0.30	0.30	0.26	0.26
.30	4	0.70	0.66	0.60	0.56	0.81	0.77	0.67	0.66
	8	1.45	1.29	1.24	1.09	1.56	1.40	1.28	1.19
	16	3.20	2.56	2.73	2.15	3.31	2.67	2.71	2.25
$\Delta = 3.290$	1	0.43	0.43	0.43	0.41	1.51	1.51	1.64	1.70
.05	4	0.94	0.90	0.90	0.81	2.03	1.99	1.91	2.10
	8	1.70	1.54	1.58	1.34	2.73	2.62	2.32	2.63
	16	3.44	2.80	3.16	2.41	4.52	3.88	3.35	3.69

8.5 A Comparison of the Estimators when Σ is unknown and $n_1 = n_2$.

We begin our investigation of the three estimators $LO(E_e)$, $LO(U_e)$ and $LO(P_e)$ when the covariance matrix Σ is unknown.by considering their behaviour in the unconditional case. This will give us an overview of their performance. In subsequent sections their conditional behaviour and their behaviour when $n_1 \neq n_2$ is considered.

Here the three estimators will be compared in terms of unconditional bias and mean square error where the unconditional mean square error of $LO(P_e)$ is estimated by simulation.

From Sections 7.8, 7.10 and 7.12

$$LO(E_e) = \tfrac{1}{2}\{-w_1(x) + w_2(x)\}$$

$$LO(U_e) = \frac{2n-p-3}{2n-2} \; LO(E_e)$$

$$LO(P_e) = -\tfrac{1}{2}(2n-1) \;\; \ell n \; \left[\frac{\{1 + \frac{n}{(n^2-1)} \; \tfrac{1}{2}w_1(x)\}}{\{1 + \frac{n}{(n^2-1)} \; \tfrac{1}{2}w_2(x)\}}\right]$$

and all three estimators provide identical classification. The unconditional bias of the estimators have already been derived as

$$B\{LO(E_e)\} = \frac{p+1}{2n-p-3} \;\; \tfrac{1}{2}\Delta^2 \; (-1)^{t-1}$$

$$B\{LO(U_e)\} = 0 \qquad (8.5.1)$$

$$B\{LO(P_e)\} = [\sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} \; e^{-\frac{1}{2}\lambda}\{\sum_{i=o}^{j-1} \frac{2n-1}{2n-1+2i}\} - \tfrac{1}{2}\Delta^2 \,] \;\; (-1)^{t-1}$$

where $\lambda \,\hat{=}\, \frac{n}{n+1} \;\; \Delta^2$ and x in Π_t, $t = 1$ and 2.

It is clear from (8.5.1) that the relative bias of $LO(E_e)$ i.e.

$$\frac{p+1}{2n-p-3}$$

is positive in both populations, thus $LO(E_e)$ overstates $LO(T_e)$ in both populations. Also its relative bias is independent of Δ, increases with increasing p and decreases as n increases. The bias and relative bias of $LO(P_e)$ are independent of p when $n_1 = n_2$. From (7.10.9) the relative bias of $LO(P_e)$ is bounded above by $-\frac{1}{n+1}$ and below by -1, x in Π_t, t = 1 and 2 and so $LO(P_e)$ on average understates $LO(T_e)$ in both populations. The relative bias of $LO(P_e)$ decreases as n increases and increases with increasing Δ.

The relative biases of $LO(E_e)$ and $LO(P_e)$, multiplied by 10^2 for convenience, are tabulated in Table 8.5.1 for various combinations of n, p and Δ. The infinite series in $B\{LO(P_e)\}$ was summed by recursivly evaluating the terms $(\frac{1}{2}\lambda)^j/j!$ and

$$\sum_{i=o}^{j-1} \frac{2n-1}{2n-1+2i} .$$

The residual sum at any stage j = r may be bounded above as follows

$$\sum_{j=r+1}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} e^{-\frac{1}{2}\lambda} \{\sum_{i=o}^{j-1} \frac{2n-1}{2n-1+2i}\} < \sum_{j=r+1}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} e^{-\frac{1}{2}\lambda} \{j\}$$

$$= \tfrac{1}{2}\lambda \sum_{j-1=r}^{\infty} \frac{(\frac{1}{2}\lambda)^{j-1}}{(j-1)!} e^{-\frac{1}{2}\lambda}$$

$$= \tfrac{1}{2}\lambda\{1 - \sum_{j=o}^{r-1} \frac{(\frac{1}{2}\lambda)^j}{j!} e^{-\frac{1}{2}\lambda}\}.$$

Evaluation of the series was terminated when the residual was less than 10^{-4}.

Table 8.5.1

The Relative Unconditional Biases of $LO(E_e)$ and $LO(P_e) \times 10^2$, Σ unknown and $n_1 = n_2$.

	Method		E_e	
	p \ n	12	24	48
all Δ	1	10	5	2
	4	29	12	6
	8	69	24	11
	16	340	59	22

	Method		P_e		$\frac{\lvert E\{LO(T_e)\}\rvert}{\lvert \frac{1}{2}\Delta^2(-1)^{t-1}\rvert}$
	Δ \ n	12	24	48	
all p	1.049	-9	-5	-2	0.55
	1.683	-12	-6	-4	1.42
	2.563	-17	-10	-5	3.28
	3.290	-23	-13	-7	5.41

In Table 8.5.1 we note the large positive relative bias of LO(E_e) when p is large and n is small regardless of the separation of the populations. The size of this relative bias is directly attributable to estimating Σ^{-1} by the inverse sample covariance matrix S^{-1}. The unbiased method adjusts for this. The negative relative bias of LO(P_e) is small even for well separated populations.

From Sections 7.8 and 7.12 the unconditional mean square errors of LO(E_e) and LO(U_e) are given by

$$\text{MSE}\{\text{LO}(E_e)\} = \text{E}\{\text{LO}^2(E_e)\} - \frac{2n+p-1}{2n-p-3}\ \text{E}\{\text{LO}^2(T_e)\}$$

$$\text{MSE}\{\text{LO}(U_e)\} = \left(\frac{2n-p-3}{2n-2}\right)^2 \text{E}\{\text{LO}^2(E_e)\} - \text{E}\{\text{LO}^2(T_e)\} \qquad (8.5.2)$$

where $\text{E}\{\text{LO}^2(E_e)\} = \frac{(2n-2)^2}{d}\left[(2n-3)\ \{(\frac{2n+1}{n^2})p + (\frac{n+1}{n})\ \Delta^2\} + (2n-p-2)\ \tfrac{1}{4}\Delta^4\right]$

$$d = (2n-p-2)\ (2n-p-3)\ (2n-p-5)$$

and $\text{E}\{\text{LO}^2(T_e)\} = \Delta^2 + \tfrac{1}{4}\Delta^4$ $\qquad$ x in Π_t, t = 1 and 2.

The conditional and hence the unconditional mean square errors of LO(U_e) are always less than those of LO(E_e); this follows from the elementary result, if X = LO(U_e) and $E(X) = \alpha$ = LO(T_e), Y = LO(E_e) = cX where

$$c = \frac{2n-2}{2n-p-3} > 1$$

then $\text{E}\{(Y-\alpha)^2\} = c^2\ \text{E}\{(X-\alpha)^2\} + \alpha^2(c-1)^2 \qquad (8.5.3)$

$$> \text{E}\{(X-\alpha)^2\} \qquad \text{if } c > 1.$$

From (8.5.2) we note that both unconditional mean square errors increase with increasing p and Δ, they decrease as n increases.

A simulation study was undertaken to estimate the unconditional mean square error of $LO(P_e)$, the details are as follows. For each n,p combination, 2n observations were generated from Π_1, a $N_p(0,I)$ distribution. Details of the random number generator used are given in Appendix 8.A. By an additive transformation the sample mean $\bar{x}_2$ was calculated for various values of Δ i.e. population Π_2. Also generated and stored were 100 test observations from Π_1 and the mean, bias, variance and mean square error given a particular set of sample parameters was calculated for these test observations. By repeating the sample generation process 100 times estimates of the unconditional mean, bias, variance and mean square error of LO(M), M in $\{T_e, E_e, U_e, P_e\}$ were obtained. Estimates were also obtained of the misclassification rates of $LO(T_e)$ and $LO(E_e)$ i.e. all three estimators. The test observations were stratified by size of $LO(T_e)$, four strata corresponding to the quartiles of the distribution were used and the unconditional mean, bias and mean square error of the estimators recorded for each strata. Using standard statistical tests the estimated mean, variance and error rate of $LO(T_e)$, the number in each strata of $LO(T_e)$, and the unconditional mean of $LO(E_e)$ and $LO(U_e)$ were compared with their exact values. At the 5% level no significant results were recorded.

The exact and estimated unconditional mean square error of $LO(E_e)$, $LO(U_e)$ and $LO(P_e)$ are given in Table 8.5.2 for a range of n, p and Δ. At the base of Table 8.5.2 we indicated for $LO(E_e)$ and $LO(U_e)$ the minimum and maximum ratios of their estimated mean square errors to their exact values.

Table 8.5.2

The Unconditional Mean Square Errors of $LO(E_e)$, $LO(U_e)$ and $LO(P_e)$, Σ unknown and $n_1 = n_2$.

Method		E_e			U_e			P_e		
Δ & PMC	p	Sample Sizes $n_1 = n_2$			Sample Sizes $n_1 = n_2$			Sample Sizes $n_1 = n_2$		
		12	24	48	12	24	48	12	24	48
Δ = 1.049 .30	1	0.56	0.23	0.10	0.46	0.20	0.10	0.32	0.16	0.07
	4	2.52	0.76	0.31	1.43	0.59	0.27	1.38	0.46	0.23
	8	10.78	1.94	0.66	3.53	1.20	0.52	2.46	1.01	0.51
	16	447.10	7.93	1.72	22.26	4.97	1.11	9.45	2.98	1.04
Δ = 1.683 .20	1	1.25	0.49	0.22	0.99	0.44	0.21	0.63	0.33	0.16
	4	4.46	1.30	0.52	2.41	0.98	0.45	2.03	0.78	0.38
	8	17.88	3.11	1.02	5.43	1.83	0.79	3.51	1.44	0.78
	16	687.32	12.39	2.60	32.61	4.26	1.59	12.41	4.11	1.43
Δ = 2.563 .10	1	3.48	1.35	0.60	2.73	1.20	0.57	1.80	0.96	0.45
	4	10.21	2.87	1.11	5.20	2.08	0.95	3.77	1.85	0.85
	8	38.31	6.38	2.02	10.47	3.47	1.49	5.89	2.61	1.45
	16	1330.18	24.70	4.95	58.34	7.45	2.75	18.55	6.46	2.34
Δ = 3.290 .05	1	7.24	2.78	1.23	5.65	2.47	1.16	4.36	2.29	1.04
	4	19.38	5.33	2.03	9.50	3.76	1.71	6.92	3.73	1.67
	8	70.05	11.36	3.50	17.75	5.82	2.49	9.30	4.54	2.44
	16	2269.69	43.22	8.37	93.29	11.70	4.31	26.05	9.32	3.51
Min. Ratio		0.70	0.60	0.75	0.67	0.66	.75			
Max. Ratio		1.43	1.20	1.19	1.40	1.21	1.14			

The most notable feature of Table 8.5.2 is that the predictive estimator LO(P_e) has the smallest mean square error for all combinations of n, p and Δ. This was also true of the simulation results. Like the mean square errors of LO(E_e) and LO(U_e) those of LO(P_e) increase with increasing p, decrease as n increases and increase with increasing Δ. The difference between the mean square errors of LO(E_e) and LO(U_e) is substantial when p is large and n is small. Similarly for p large and n small there is a noticeable difference between the mean square errors of LO(U_e) and LO(P_e).

Overall for Σ unknown and $n_1 = n_2$ the unbiased estimator is the superior frequentist approach, given LO(E_e)'s overstatement of LO(T_e) and its larger mean square error. However LO(P_e)'s understatement of LO(T_e) and smaller mean square error especially for large p and n small make the predictive approach preferable.

8.6 Comparing the Estimators $LO(E_e)$, $LO(U_e)$ and $LO(P_e)$ for fixed True Log-odds, Σ unknown and Equal Sample Sizes.

In this section we address the problem of where $LO(P_e)$ derives its superiority by investigating the behaviour of the estimators for fixed true log-odds for x from Π_1. The three estimators are compared in terms of their conditional bias and mean square error.

From equations (7.8.1), (7.10.10) and (7.10.11) the conditional bias of the estimators are

$$B\{LO(E_e)|x\} = \frac{p+1}{2n-p-3} \; LO(T_e)$$

$$B\{LO(U_e)|x\} = 0$$

$$B\{LO(P_e)|x\} = E\{LO(P_e)|x\} - LO(T_e)$$

where

$$E\{LO(P_e)|x\} = -\sum_{j=1}^{\infty} \sum_{i=o}^{j-1} \frac{2n-1}{2n-1+2i} \{e^{-\frac{1}{2}\lambda_1} \frac{(\frac{1}{2}\lambda_1)^j}{j!} - e^{-\frac{1}{2}\lambda_2} \frac{(\frac{1}{2}\lambda_2)^j}{j!}\} \qquad (8.6.1)$$

$$+ \frac{2n-1}{2} \sum_{j=o}^{\infty} \sum_{i=1}^{\infty} \frac{(\frac{n}{n+1})^i}{i} \prod_{k=1}^{i} (\frac{\frac{p}{2}+i+j-k}{\frac{2n-1}{2}+i+j-k}) \{$$

$$e^{-\frac{1}{2}\lambda_1} \frac{(\frac{1}{2}\lambda_1)^j}{j!} - e^{-\frac{1}{2}\lambda_2} \frac{(\frac{1}{2}\lambda_2)^j}{j!} \}$$

and $\lambda_t = n \, \omega_t(x)$, $t = 1$ and 2.

As in Section 8.5 the relative conditional bias of $LO(E_e)$ i.e. $B\{LO(E_e)|x\}/LO(T_e)$ is

$$\frac{p+1}{2n-p-3}$$

and so independent of the size of true log-odds. The size of this relative bias indicates the inadvisability of using $LO(E_e)$ as an estimator of $LO(T_e)$. Considering the conditional bias of LO(Pe) we note that it is not just a function of LO(Te) but of thc individual distances $\omega_t(x)$ and unlike its unconditional bias it now depends on the dimension parameter p. For zero true log-odds all estimators are unbiased.

In Table 8.6.1 we list the relative conditional bias x 10^2 of $LO(P_e)$ for low, medium and high true log-odds percentiles as specified in Section 8.3. The relative bias of $LO(E_e)$ is as given in Table 8.5.1. The first infinite series in $E\{LO(P_e)|x\}$ (8.6.1) was evaluated as described in Section 8.5. The second and double infinite series is

$$\sum_{j=o}^{\infty} e^{-\frac{1}{2}\lambda_t} \frac{(\frac{1}{2}\lambda_t)^j}{j!} \sum_{i=1}^{\infty} \frac{(\frac{n}{n+1})^i}{i} \prod_{k=1}^{i} \left(\frac{\frac{p}{2}+i+j-k}{\frac{2n-1}{2}+i+j-k}\right)$$

$$= \sum_{j=o}^{\infty} e^{-\frac{1}{2}\lambda_t} \frac{(\frac{1}{2}\lambda_t)^j}{j!} I_j \quad \text{say}$$

where the inner series I_j is independent of λ_t. If evaluation of I_j is terminated when $i = r \geqslant 1$ the residual sum may be bounded above by

$$\sum_{i=r+1}^{\infty} \frac{(\frac{n}{n+1})^i}{i} \prod_{k=1}^{i} \left(\frac{\frac{p}{2}+i+j-k}{\frac{2n-1}{2}+i+j-k}\right) < \sum_{i=r+1}^{\infty} \frac{(\frac{n}{n+1})^i}{i} \quad \text{as } p < 2n-1$$

$$= -\ln\left(1 - \frac{n}{n+1}\right) - \sum_{i=1}^{r} \frac{(\frac{n}{n+1})^i}{i}$$

$$= \ln(n+1) - \sum_{i=1}^{r} \frac{(\frac{n}{n+1})^i}{i} .$$

This upper bound depends only on n. Hence for given n the value of r at which evaluation of I_j may be terminated for a pre-set tolerence of 10^{-4} is the same for all p and j. Further as I_j is independent of λ_t the values of I_j were stored to avoid unnecessary computation. If evaluation of the double infinite series is terminated when $j = m > o$ the residual may be bounded above by

$$\sum_{j=m+1}^{\infty} e^{-\frac{1}{2}\lambda_t} \frac{(\lambda_t)^j}{j!} I_j < \ell n(n + 1) \{1 - e^{-\frac{1}{2}\lambda_t} \sum_{j=o}^{m} \frac{(\frac{1}{2}\lambda_t)^j}{j!} \} \qquad (8.6.2)$$

which may be summed as the series is summed. Evaluation of the double infinite series was terminated when the upper bound (8.6.2) was less than 10^{-4}.

Noting that negative relative bias implies understatement and positive relative bias overstatement of true log-odds, we see from Table 8.6.1 that $LO(P_e)$'s tendency to understate $LO(T_e)$, Section 8.5 is due to its understatement of high $LO(T_e)$ regardless of n, p and Δ. Overstatement of $LO(T_e)$ by $LO(P_e)$ is confined to low and medium log-odds and increases as Δ decreases, it is however small. With increasing Δ the understatement of $LO(T_e)$ spreads to medium ($\Delta = 2.563$) and low ($\Delta = 3.290$) true log-odds. The influence of increasing p is small and is not consistent. As anticipated the relative bias of LO(Pe) decreases as n increases.

Table 8.6.1

The Relative Conditional Bias of $LO(P_e) \times 10^2$, Σ unknown and $n_1 = n_2$.

Method		P_e					
Sample Size		$n_1 = n_2 = 12$			$n_1 = n_2 = 24$		
Category of $LO(T_e)$		low	med.	high	low	med.	high
Size of $LO(T_e)$		0.15	0.55	2.28	0.15	0.55	2.28
Δ & PMC	p						
Δ = 1.049 .30	1	7	2	-15	0	2	-8
	4	13	9	-11	7	4	-5
	8	13	9	-11	7	5	-5
	16	7	7	-11	7	5	-5
Size of $LO(T_e)$		0.28	1.42	4.19	0.28	1.42	4.19
Δ = 1.683 .20	1	11	-2	-21	0	-1	-11
	4	7	4	-17	4	2	-9
	8	7	4	-16	4	3	-8
	16	7	3	-17	4	3	-8
Size of $LO(T_e)$		0.63	3.28	7.50	0.63	3.28	7.50
Δ = 2.563 .10	1	-3	-9	-28	-2	-5	-16
	4	3	-4	-25	2	-2	-14
	8	3	-3	-25	2	-1	-13
	16	2	-5	-25	2	-1	-14
Size of $LO(T_e)$		1.19	5.41	10.82	1.19	5.41	10.82
Δ = 3.290 .05	1	-8	-16	-34	-4	-9	-21
	4	-3	-11	-31	-1	-6	-18
	8	-3	-11	-31	0	-5	-18
	16	-3	-12	-31	-1	-5	-18

We now consider the conditional mean square errors of the estimators, those of $LO(E_e)$ and $LO(U_e)$ were derived in Chapter 7 as

$$MSE\{LO(E_e)|x\} = E\{LO^2(E_e)|x\} - \frac{2n+p-1}{2n-p-3} LO^2(T_e)$$

and

$$MSE\{LO(U_e)|x\} = \left(\frac{2n-p-3}{2n-2}\right)^2 E\{LO^2(E_e)|x\} - LO^2(T_e)$$

where

$$E\{LO^2(E_e)|x\} = \frac{(2n-2)^2}{d} [\{\Delta^2 + \frac{2}{n}(2n-3)\} (x-\eta)' \Sigma^{-1}(x-\eta)$$

$$+ (2n-3) \{\frac{p}{n^2} + \frac{\Delta^2}{2n}\} + (2n-p-3) LO^2(T_e)]$$

$$d = (2n-p-2)(2n-p-3)(2n-p-5) \text{ and } \eta = \tfrac{1}{2}(\mu_1+\mu_2).$$

From the canonical forms of Section 8.2

$$(x-\eta)' \Sigma^{-1} (x-\eta) = (x-\tfrac{1}{2}\theta)' (x-\tfrac{1}{2}\theta)$$

$$= (x_1-\tfrac{1}{2}\Delta)^2 + \sum_{i=2}^{p} x_i^2$$

$$\text{and } LO^2(T_e) = \Delta^2(x_1-\tfrac{1}{2}\Delta)^2.$$

As noted in Section 8.5 the conditional mean square error of $LO(U_e)$ is always less than that of $LO(E_e)$. Both mean square errors increase with increasing p and decrease with increasing n, for all $LO(T_e)$. The relative conditional mean square error i.e. $MSE\{LO(M)|x\} / LO^2(T_e)$ of $LO(E_e)$ and $LO(U_e)$ decrease with increasing $LO(T_e) > 0$ for all fixed $\Delta > 0$. This indicates that $LO(E_e)$ and $LO(U_e)$ give better estimates of higher rather than lower true log-odds.

In Table 8.6.2 we list the conditional mean square errors of $LO(U_e)$ and $LO(P_e)$, the latter derived from the simulation study of Section 8.5. The estimator $LO(E_e)$ is omitted as $LO(U_e)$ is clearly superior. At the base of Table 8.6.3 we indicate the minimum and maximum ratio of estimated to exact mean square error of $LO(U_e)$.

In Table 8.6.2 we note that the conditional mean square errors of $LO(P_e)$ are always less than or equal to the corresponding mean square errors of $LO(U_e)$, in the simulation study they were always less than. The one exception is for $\Delta = 1.044$, $p = 8$, $n = 12$ and low $LO(T_e)$, however allowance must be made for the sampling error in the predictive results. For high $LO(T_e)$ and for all p, n and Δ, $LO(P_e)$ has smaller mean square error than $LO(U_e)$. The mean square errors of $LO(P_e)$ display similar behaviour to those of $LO(U_e)$ as regards increasing n, p, Δ and $LO(T_e)$. As for $LO(U_e)$ the relative conditional mean square error of $LO(P_e)$ indicates better estimation of higher rather than lower true log-odds.

We conclude that for Σ unknown and $n_1 = n_2$, when p is large and n is small, the predictive estimator LO(Pe) is superior to the unbiased estimator LO(Ue) regardless of size of $LO(T_e)$ or Δ. This preference for $LO(P_e)$ is due to its considerably smaller mean square error, Table 8.6.2, rather than the size of its bias, Table 8.6.1. For n or p moderately large the difference in the methods is slight.

Table 8.6.2

The Conditional Mean Square Errors of LO(U_e) and LO(P_e), Σ unknown and $n_1 = n_2$.

Sample Sizes		$n_1 = n_2 = 12$						$n_1 = n_2 = 24$					
Method		U_e			P_e			U_e			P_e		
Category LO(T_e)		low	med.	high	low	med.	high	low	med.	high	low	med.	high
Δ & PMC	p												
Δ = 1.049 .30	1	0.06	0.14	1.51	0.06	0.14	1.00	0.03	0.06	0.68	0.03	0.06	0.58
	4	0.52	0.61	2.24	0.51	0.61	1.41	0.21	0.25	0.91	0.20	0.24	0.70
	8	2.15	2.28	4.47	2.16	2.20	2.52	0.73	0.77	1.51	0.62	0.71	1.27
	16	17.27	17.70	25.26	9.11	9.71	9.91	2.30	2.36	3.31	2.17	2.20	2.52
Δ = 1.683 .20	1	0.15	0.49	3.23	0.12	0.44	2.09	0.07	0.22	1.44	0.06	0.20	1.19
	4	0.78	1.19	4.46	0.77	0.81	2.38	0.32	0.48	1.79	0.29	0.46	1.16
	8	3.03	3.58	8.02	3.01	3.04	3.54	1.02	1.21	2.66	1.00	1.21	1.99
	16	23.88	25.83	41.49	10.03	10.55	10.60	3.16	3.39	5.28	2.92	3.04	3.50
Δ = 2.563 .10	1	0.37	1.81	8.15	0.26	1.25	6.45	0.17	0.80	3.57	0.16	0.74	3.35
	4	1.37	3.11	10.70	1.38	1.48	6.98	0.56	1.24	4.22	0.51	0.93	3.68
	8	4.97	7.32	17.64	4.41	4.36	7.30	1.67	2.43	5.75	1.62	2.10	4.19
	16	38.29	46.76	83.80	15.76	16.23	17.76	5.02	6.00	10.30	4.48	4.79	6.10
Δ = 3.290 .05	1	0.69	4.26	15.53	0.43	2.66	14.98	0.32	1.87	6.76	0.30	1.61	7.75
	4	2.14	6.42	19.93	1.69	2.67	15.90	0.87	2.54	7.81	0.80	1.60	6.28
	8	7.31	13.13	31.50	5.57	5.69	20.73	2.44	4.30	10.17	2.31	3.28	8.42
	16	55.17	76.24	142.70	19.92	18.62	29.99	7.19	9.60	17.20	6.14	6.73	10.54
Min. Ratio		0.65	0.66	0.70				.67	.67	.71			
Max. Ratio		1.40	1.25	1.19				1.05	1.07	1.10			

8.7 A Comparison for the Estimators for Σ unknown and Unequal Sample Sizes

The Comparison of Section 8.5 is extended here to consider the effect of unequal sample sizes on the estimators when the covariance matrix Σ is unknown. Included in the comparison is an adjusted predictive estimator which is denoted by LO(PA$_e$). As before the comparison is based on evaluation of unconditional biases and mean square errors with estimation of expected actual probabilities of misclassification now included.

The four estimators of true log-odds are now given by

$$LO(E_e) = \tfrac{1}{2}\{-w_1(x) + w_2(x)\} = \hat{L}(x)$$

$$LO(U_e) = \frac{n_1+n_2-p-3}{n_1+n_2-2}\, LO(E_e) - \tfrac{1}{2}p\,\left(\frac{n_1-n_2}{n_1 n_2}\right) \qquad (8.7.1)$$

$$LO(P_e) = \tfrac{1}{2}p\ \ell n\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\} - \tfrac{1}{2}(n_1+n_2-1)\,\ell n\left[\frac{\left\{1 + \frac{n_1}{n_1+1}\,\frac{w_1(x)}{n_1+n_2-2}\right\}}{\left\{1 + \frac{n_2}{n_2+1}\,\frac{w_2(x)}{n_1+n_2-2}\right\}}\right]$$

$$LO(PA_e) = LO(P_e) - \tfrac{1}{2}p\ \ell n\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\}.$$

In Section 8.4 with Σ known we noted the connection between bias and misclassification performance when $n_1 \neq n_2$. From (8.7.1) we see that if we classify observations using the sign of the estimators, then classifying by sign of LO(E$_e$) = $\hat{L}(x)$ is equivalent to using the sample linear discrimination function with zero cut-off point. Classifying by sign of LO(U$_e$) is equivalent to

$$\hat{L}(x) \gtrless \frac{n_1+n_2-2}{n_1+n_2-p-3}\ \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right) .$$

Noting that the likelihood ratio estimator $LO(LR_e)$ of Section 6.3 is related to the Z statistic of Chapter 2 with Σ now unknown i.e.

$$LO(LR_e) = \tfrac{1}{2}\, Z(x),$$

classification by sign of $LO(P_e)$ is equivalent to

$$Z(x) \gtrless \frac{n_1+n_2+1}{n_1+n_2-1}\ p\ \ln\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\}$$

and by $LO(PA_e)$ is equivalent to

$$Z(x) \gtrless 0.$$

Hence, our investigation of the estimators when $n_1 \neq n_2$ will enable us to see whether as in Section 8.4 for Σ known, adjusting the cut-off point of $\hat{L}(x)$ from zero to

$$\frac{n_1+n_2-2}{n_1+n_2-p-3}\ \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right),$$

will result in the expected actual probabilities of misclassification of $\hat{L}(x)$ being equated. For this reason we retain $LO(E_e)$ in our comparison despite its clear inferiority to $LO(U_e)$ as an estimator of log-odds. The inclusion of $LO(PA_e)$ will enable us to see whether the equation of expected actual probabilities of misclassification of $Z(x)$, Σ known Section 2.10, persists when Σ is unknown.

The unconditional bias of the estimators, x in Π_t, t = 1 and 2 are from Sections 7.8 and 7.10,

$$B\{LO(E_e)\} = \frac{p+1}{n_1+n_2-p-3}\ \tfrac{1}{2}\Delta^2(-1)^{t-1} + \frac{n_1+n_2-2}{n_1+n_2-p-3}\ \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1n_2}\right)$$

$$B\{LO(U_e)\} = 0 \qquad (8.7.2)$$

$$B\{LO(P_e)\} = \left\{ \sum_{j=1}^{\infty} \frac{(\tfrac{1}{2}\lambda_t)^j}{j!}\ e^{-\frac{1}{2}\lambda_t} \sum_{i=o}^{j-1} \frac{n_1+n_2-1}{n_1+n_2-1+2i} - \tfrac{1}{2}\Delta^2\right\}(-1)^{t-1}$$

$$+ \tfrac{1}{2}p\ \ln\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\} \text{ where } \lambda_t = \frac{n_{3-t}}{n_{3-t}+1}\,\Delta^2, \text{ and}$$

$$B\{LO(PA_e)\} = B\{LO(P_e)\} - \tfrac{1}{2}p\ \ln\left\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\right\}.$$

We define as before the relative unconditional bias of LO(M) to be

$$B\{LO(M)\} \; / \; E\{LO(T_e)\} \qquad (8.7.3)$$

where $E\{LO(T_e)\} = \frac{1}{2}\Delta^2 \; (-1)^{t-1}$ $t = 1$ and 2,

and if the relative bias is negative interpret this as LO(M) understating true log-odds for Π_1, with x in Π_1 or Π_2. Positive relative bias is interpreted as overstatement of LO(Te), with x in Π_1 or Π_2.

From (8.7.2) and (8.7.3) we see that the relative unconditional bias of LO(Ee) is

$$\frac{p+1}{n_1+n_2-p-3} + \frac{n_1+n_2-2}{n_1+n_2-p-3} \frac{1}{2}p(\frac{n_1-n_2}{n_1n_2}) \; / \; \frac{1}{2}\Delta^2 \; (-1)^{t-1}, \; t = 1 \text{ and } 2. \qquad (8.7.4)$$

The first term in (8.7.4) is the relative bias of LO(Ee) for $n_1 = n_2$ and with $n_1 \neq n_2$ we see from the second term of (8.7.4) that the relative bias of LO(Ee) now depends on Δ. With $n_1 < n_2$ the relative bias of LO(Ee) may be positive or negative in Π_1 but will be positive in Π_2. Thus unlike the $n_1 = n_2$ case where LO(Ee) overstates LO(Te) in both populations with $n_1 < n_2$ it may understate LO(Te) for x in Π_1. We conclude that with $n_1 \neq n_2$ LO(Ee) will overstate true log-odds for Π_1, with x in the population corresponding to the larger sample size.

From (8.7.2) we see that the bias and relative bias of LO(Pe) now depend on the dimension parameter p unlike the $n_1 = n_2$ case. From the bounds on E{LO(Pe)}, derived in Section 7.10 we see that the relative bias of LO(Pe) may be bounded above by

$$\frac{1}{2}p \; \ell n\{\frac{n_1(n_2+1)}{n_2(n_1+1)}\} \; / \; \frac{1}{2}\Delta^2(-1)^{t-1} - \frac{1}{n_{3-t}+1} \quad x \text{ in } \Pi_t, \; t = 1 \text{ and } 2. \qquad (8.7.5)$$

With $n_1 < n_2$ and x in Π_1 this bound is negative and so LO(Pe) understates true log-odds in Π_1, with x in Π_2 however the upper bound (8.7.5) may be positive and LO(Pe) may overstate LO(Te) in Π_2. Thus with $n_1 \nmid n_2$, LO(Pe) understates true log-odds for Π_1 with x in the population corresponding to the smaller sample size.

For the adjusted predictive estimator LO(PA_e) from (8.7.2), (8.7.5) and (7.10.9) its relative bias may be bounded above by

$$- \frac{1}{n_{3-t}+1}$$

and below by -1, for x in Π_t, t = 1 and 2. Thus with $n_1 \nmid n_2$, LO(PA_e) understates true log-odds in Π_1, with x in Π_1 or Π_2.

The relative unconditional bias x10^2 of the estimators are listed in Table 8.7.1 for various values of p and Δ when $n_1 = 8$, $n_2 = 40$. Several ratios of n_1 to n_2 were considered i.e. 1:2, 1:3 and 1:5, where $n_1+n_2 = 48$. The latter total of $n_1 + n_2$ was decided on so as to allow a marked imbalance in the sample sizes without either sample size becoming too small and to facilitate a comparison with previous tabulations for $n_1 = n_2 = 24$. The evaluation of the infinite series in B{LO(P_e)} (8.7.2) was as described in Section 8.5.

The expectation of the actual probability of misclassification of LO(M), M$\varepsilon\{E_e,U_e,P_e,PA_e\}$ were estimated from a simulation study similar to that detailed in Section 8.5 but with 100 test observations from each population. These estimated expected actual PMC's x10^2 are listed in Table 8.7.2 as is their maximum standard error. Also included in this table are the estimated expected actual PMC's for $n_1 = n_2 = 24$, which were obtained from the simulation study of Section 8.5.

Table 8.7.1

The Relative Unconditional Bias x10^2 of LO(E_e), LO(P_e) and LO(PA_e), Σ unknown and $n_1 \neq n_2$.

Population		Π_1			Π_2		
Method		E_e	P_e	PA_e	E_e	P_e	PA_e
Δ, PMC & $\lvert E\{LO(T_e)\}\rvert = \frac{1}{2}\Delta^2$	p						
	1	- 5	- 13	- 4	15	- 4	- 13
Δ = 1.049	4	- 29	- 45	- 4	53	31	- 13
.30	8	- 65	- 80	- 4	115	64	- 13
$\frac{1}{2}\Delta^2$ = 0.55	16	- 173	- 147	- 4	289	133	- 13
	1	1	- 8	- 5	8	- 10	- 13
Δ = 1.683	4	- 4	- 21	- 5	28	3	- 13
.20	8	- 11	- 35	- 5	59	16	- 13
$\frac{1}{2}\Delta^2$ = 1.42	16	- 31	- 61	- 5	148	42	- 13
	1	3	- 10	- 8	6	- 15	- 16
Δ = 2.563	4	5	- 16	- 8	19	- 9	- 16
.10	8	9	- 21	- 8	40	- 3	- 16
$\frac{1}{2}\Delta^2$ = 3.28	16	20	- 32	- 8	97	8	- 16
	1	4	- 13	- 12	6	- 18	- 19
Δ = 3.290	4	8	- 16	- 12	16	- 15	- 19
.05	8	15	- 19	- 12	33	- 11	- 19
$\frac{1}{2}\Delta^2$ = 5.41	16	35	- 26	- 12	82	- 4	- 19

Table 8.7.2

Estimated Expected Actual Probabilities of Misclassification x10^2 of LO(E_e), LO(U_e), LO(P_e) and LO(PA_e), Σ unknown, $n_1 = n_2$ and $n_1 \neq n_2$.

Sample Sizes		$n_1 = 8$, $n_2 = 40$								$n_1 = n_2 = 24$
Population		Π_1				Π_2				Π_1 & Π_2
Method		E_e	U_e	P_e	PA_e	E_e	U_e	P_e	PA_e	All
Δ & PMC	p									
	1	26	24	26	25	27	29	26	28	27
Δ = 1.049	4	39	33	40	35	32	39	31	37	33
.30	8	43	33	43	33	26	35	26	34	35
	16	54	39	55	41	28	42	28	41	38
	1	18	17	18	18	20	20	20	20	16
Δ = 1.683	4	27	24	28	25	22	27	22	26	23
.20	8	28	22	28	22	18	24	18	23	24
	16	41	31	41	31	22	32	22	32	28
	1	10	10	10	10	10	10	10	10	9
Δ = 2.563	4	15	13	15	13	10	12	10	12	11
.10	8	15	12	14	11	9	12	10	12	14
	16	27	20	27	20	15	21	15	21	18
	1	6	5	5	5	4	5	5	5	5
Δ = 3.290	4	8	7	8	7	5	5	5	6	5
.05	8	8	6	7	6	5	6	5	7	8
	16	18	14	17	13	9	13	10	14	12
Max. S.E.*		.012				.007				.008

S.E. = Standard Error

Considering the results in Tables 8.7.1 and 8.7.2 we note immediately the differing classification behaviour of LO(E_e) and LO(P_e) as compared to LO(U_e) and LO(PA_e). We see from Table 8.7.2 that the estimated expected actual PMC's of LO(E_e) and LO(P_e) display considerable imbalance especially for large p, with the larger expected actual PMC's corresponding to the population with the smaller sample size. For LO(U_e) and LO(PA_e) this is not the case, with the expected actual PMC's of LO(PA_e) almost equal and those of LO(U_e) reasonably balanced, with a tendency for the slightly larger expected actual PMC of both to correspond to the population with the larger sample size. Relating this classification behaviour to the unconditional bias of the estimators, Table 8.7.1, we note that LO(E_e)'s large overstatement of LO(T_e) in Π_2, which is always greater than any overstatement by it in Π_1, coincides with its classification behaviour. Similarly LO(P_e)'s large understatement of LO(T_e) in Π_1 which is usually less than any understatement by it in Π_2 coincides with its classification behaviour. Finally the slight understatement of LO(T_e) by LO(PA_e) in both population with the larger understatement in Π_2 and the unbiasedness of LO(U_e) coincide with their noted classification behaviour. For the other ratios of n_1 to n_2 considered, similar but less marked behaviour was noted.

These classification results enable us to confirm that with $n_1 \neq n_2$ the Z statistic's ability to equate expected actual PMC's when Σ is known persists when Σ is unknown. Also, that the classification behaviour of the sample linear discriminant function can be substantially improved by alternating its cut-off point from 0 to

$$\frac{n_1+n_2-2}{n_1+n_2-p-3} \ \tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right).$$

We now consider the unconditional mean square errors of the estimators. With $n_1 \neq n_2$, the derivations of Sections 7.8, 7.12 and Appendix 7.A give

$$\mathrm{MSE}\{\mathrm{LO}(E_e)\} = E_t\{\mathrm{LO}^2(E_e)\} - \left(\frac{n_1+n_2+p-1}{n_1+n_2-p-3}\right)(\Delta^2+\tfrac{1}{4}\Delta^4)$$

$$- \left(\frac{n_1+n_2-2}{n_1+n_2-p-3}\right) p\left(\frac{n_1-n_2}{n_1 n_2}\right) \tfrac{1}{2}\Delta^2(-1)^{t-1}$$

and

$$\mathrm{MSE}\{\mathrm{LO}(U_e)\} = E_t\{\mathrm{LO}^2(E_e)\} - [\tfrac{1}{2}p\left(\frac{n_1-n_2}{n_1 n_2}\right) + \tfrac{1}{2}\Delta^2(-1)^{t-1}]^2 - \Delta^2$$

$$x \text{ in } \Pi_t, \; t = 1 \text{ and } 2.$$

The expression for $E_t\{\mathrm{LO}^2(E_e)\}$ is very lengthy and is given in full in Appendix 7.A. In a similar manner to (8.5.3) it may be shown that

$$\mathrm{MSE}\{\mathrm{LO}(E_e)\} = \left(\frac{n_1+n_2-2}{n_1+n_2-p-3}\right)^2 \mathrm{MSE}\{\mathrm{LO}(U_e)\} + E_x[B\{\mathrm{LO}(E_e)\,|\,x\}^2]$$

$$x \text{ in } \Pi_t, \; t = 1 \text{ and } 2 \tag{8.7.6}$$

and as is the case for $n_1 = n_2$, the mean square error of $\mathrm{LO}(U_e)$ is for $n_1 \neq n_2$ always less than that of $\mathrm{LO}(E_e)$, for observations from either population. With some algebraic manipulation it may be shown that if $n_1 < n_2$ the unconditional mean square errors of $\mathrm{LO}(E_e)$ and $\mathrm{LO}(U_e)$ for x in Π_1 are less than their corresponding values in Π_2. From (8.7.6) and the results of Section 7.8

$$\mathrm{MSE}\{\mathrm{LO}(E_e)\} = \left(\frac{n_1+n_2-2}{n_1+n_2-p-3}\right)^2 \mathrm{MSE}\{\mathrm{LO}(U_e)\} + \left(\frac{n_1+n_2-2}{n_1+n_2-p-3} - 1\right)^2 V_x\{\mathrm{LO}(T_e)\}$$

$$+ [B\{\mathrm{LO}(E_e)\}]^2 \tag{8.7.7}$$

where $V_x\{\mathrm{LO}(T_e)\} = \Delta^2$ for x in Π_t, t = 1 and 2,

and this expression (8.7.7) allows us to assess the contribution of bias in $\mathrm{LO}(E_e)$ to its unconditional mean square error.

The unconditional mean square errors of the four estimators are listed in Table 8.7.3, those of LO(P_e) and LO(PA_e) were estimated from the simulation study previously cited in this section. From (8.7.1) and (8.7.2)

$$MSE\{LO(PA_e)\} = MSE\{LO(P_e)\} - 2c\ B\{LO(PA_e)\} - c^2$$

where $c = \frac{1}{2}p\ \ell n\ \{\frac{n_1(n_2+1)}{n_2(n_1+1)}\}$.

With $n_1 < n_2$, $c < 0$ and for x in Π_1, $B\{LO(PA_e)\} < 0$ hence

$$MSE\{LO(PA_e)\} < MSE\{LO(P_e)\},\ n_1 < n_2 \text{ and } x \text{ in } \Pi_1, \text{ and}$$

we see that the mean square error of LO(PA_e) is always less than that of LO(P_e) for the population corresponding to the smaller sample size. The precise contribution of bias to the unconditional mean square error of LO(P_e) is unclear. At the base of Table 8.7.3 we indicate for LO(E_e) and LO(U_e) the minimum and maximum ratio of their estimated mean square errors as given by the simulation, to their exact values.

In Table 8.7.3 we note that the estimated unconditional mean square errors of LO(P_e) and LO(PA_e) for x in Π_1 exceed their corresponding values in Π_2. The reverse is true of the exact mean square errors of LO(E_e) and LO(U_e) and was also the case in the simulation study. From (8.7.2), (8.7.7), and the bias of the estimators Table 8.7.1 we conclude that LO(E_e) will have its larger mean square error for that population in which it has the larger absolute bias. This is also the case for LO(P_e) but not for LO(PA_e). In Table 8.7.3 we note for population Π_1 that the estimated mean square errors of LO(P_e) and LO(PA_e) exceed those of LO(U_e) for p = 1 and 4 and for all Δ. The differences are however small and this behaviour was not noted to hold in general in the simulation study. Any noticeable reduction in mean

Table 8.7.3

The Unconditional Mean Square Error of LO(M), M in $\{E_e, U_e, P_e, PA_e\}$, Σ unknown and $n_1 \neq n_2$.

Sample Sizes
$n_1 = 8$, $n_2 = 40$

Population		Π_1				Π_2			
Method		E_e	U_e	P_e	PA_e	E_e	U_e	P_e	PA_e
Δ & PMC	p								
Δ = 1.049 .30	1	0.29	0.26	0.35	0.34	0.42	0.38	0.29	0.30
	4	1.17	0.90	1.01	0.97	1.39	1.03	1.03	1.02
	8	3.15	1.91	1.62	1.50	3.64	2.06	1.46	1.38
	16	13.38	4.81	4.72	4.17	15.50	5.00	4.42	4.01
Δ = 1.683 .20	1	0.50	0.46	0.65	0.64	0.87	0.77	0.55	0.57
	4	1.61	1.25	1.51	1.45	2.19	1.58	1.46	1.52
	8	4.04	2.49	2.24	2.07	5.31	2.87	1.82	1.84
	16	16.41	6.06	5.98	5.34	21.85	6.55	5.17	4.98
Δ = 2.563 .10	1	1.28	1.15	1.46	1.43	2.10	1.87	1.39	1.43
	4	2.98	2.26	2.65	2.53	4.31	3.04	2.66	2.86
	8	6.72	4.04	3.78	3.44	9.65	4.91	2.85	3.12
	16	25.64	9.12	8.94	8.03	38.26	10.25	7.05	7.39
Δ = 3.290 .05	1	2.61	2.33	2.80	2.73	3.96	3.52	3.01	3.09
	4	5.21	3.86	4.29	4.04	7.41	5.14	4.71	5.10
	8	11.03	6.29	6.03	5.40	15.85	7.71	4.64	5.26
	16	40.65	13.24	12.80	11.34	61.44	15.10	9.67	10.75
Min Ratio		0.90	0.92			0.75	0.74		
Max Ratio		1.45	1.46			1.27	1.29		

square error by the predictive methods as compared with $LO(U_e)$ occurs for p large. Similar but less marked behaviour in the estimators was noted for the other ratios of n_1 to n_2 considered.

We conclude that with Σ unknown and $n_1 \neq n_2$ the predictive method $LO(P_e)$ has deficiencies as an estimator of true log-odds. Its understatement of $LO(T_e)$ may be so severe in the population with the smaller sample size that it misclassifies considerably more observations from this population than the other. Adjusted for bias the predictive method $LO(PA_e)$ is a good estimator of $LO(T_e)$ with small conservative bias, good classification behaviour and slightly smaller mean square than $LO(U_e)$, especially for p large. As such it has a slight edge over $LO(U_e)$ as an estimator of true log-odds.

8.8 A Comparison of the Estimators $LO(E_u)$, $LO(U_u)$ and $LO(P_u)$ for Σ_1 and Σ_2 Unknown but Proportional.

In this final comparison of the estimators the effect of unequal but proportional covariance matrices i.e. $\alpha\Sigma_1 = \Sigma_2$ where $0 < \alpha < 1$, on their preformance as estimators of true log-odds is considered. As in previous sections bias, mean square error and misclassification rates are the criteria of comparison. In the interest of simplicity it is assumed that the sample sizes are equal.

With $\Sigma_1 \neq \Sigma_2$ the true log-odds in favour of x from Π_1 is

$$LO(T_u) \quad = \quad \tfrac{1}{2}\{-\omega_1(x) + \omega_2(x)\} - \tfrac{1}{2}\ \ln(|\Sigma_1|/|\Sigma_2|)$$

where

$$\omega_t(x) \quad = \quad (x-\mu_t)'\ \Sigma_t^{-1}\ (x-\mu_t), \quad t = 1 \text{ and } 2.$$

With $n_1 = n_2 = n$ the three estimators of $LO(T_u)$ are from Sections 7.9, 7.11 and 7.12

$$LO(E_u) \quad = \quad \tfrac{1}{2}\{-w_1(x) + w_2(x)\} - \tfrac{1}{2}\ \ln(|S_1|/|S_2|)$$

$$LO(U_u) \quad = \quad \frac{n-p-2}{n-1}\ \tfrac{1}{2}\{-w_1(x) + w_2(x)\} - \tfrac{1}{2}\ \ln(|S_1|/|S_2|)$$

$$LO(P_u) \quad = \quad -\frac{n}{2}\ \ln\ \left[\frac{\{1 + \frac{n}{n^2-1}\ w_1(x)\}}{\{1 + \frac{n}{n^2-1}\ w_2(x)\}}\right] - \tfrac{1}{2}\ \ln(|S_1|/|S_2|)$$

where $w_t(x) \quad = \quad (x-\bar{x}_t)'\ S_t^{-1}\ (x-\bar{x}_t), \ t = 1$ and 2.

Here with $\Sigma_1 \neq \Sigma_2$ and $n_1 = n_2$ the predictive estimator is not identical to the likelihood ratio estimator of Section 6.3. However the difference in the estimators is slight, as

$$LO(LR_u) = \frac{n+1}{n}\ \ LO(P_u) + \frac{1}{n}\ \tfrac{1}{2}\ \ln(|S_1|/|S_2|),$$

and $LO(LR_u)$ is not considered further.

The usual canonical forms are assumed without loss of generality: viz.

$$f_1 = N_p(0,I) \quad , \quad f_2 = N_p(\theta,\alpha I)$$

$$\text{where } \theta = (\Delta_1,0,0,\ldots\ldots,0)$$

$$\text{and } \Delta_t^2 = (\mu_1-\mu_2)' \, \Sigma_t^{-1} \, (\mu_1-\mu_2) \;,\; t = 1 \text{ and } 2.$$

We take α in the range $0 < \alpha < 1$, noting that if we reverse the role of the populations a result for α may be interpreted as a result for $\frac{1}{\alpha}$. The true log-odds may now be written as

$$\begin{aligned} LO(T_u) &= \tfrac{1}{2}\{-x'x + \tfrac{1}{\alpha}(x-\theta)'(x-\theta) + p\,\ell n\alpha\} \\ &= \tfrac{1}{2}\{\tfrac{1-\alpha}{\alpha}\,(x-\tfrac{\theta}{1-\alpha})'\,(x-\tfrac{\theta}{1-\alpha}) - \tfrac{\theta'\theta}{1-\alpha} + p\,\ell n\alpha\}. \end{aligned}$$

Now for x in Π_1

$$(x - \frac{\theta}{1-\alpha})'\;(x - \frac{\theta}{1-\alpha}) \sim \chi^2(p, \frac{\theta'\theta}{(1-\alpha)^2})$$

while for x in Π_2 $\quad \frac{1}{\alpha}\,(x-\frac{\theta}{1-\alpha})'\,(x-\frac{\theta}{1-\alpha}) \sim \chi^2(p, \frac{\alpha\theta'\theta}{(1-\alpha)^2})$

hence the unconditional mean and variance of $LO(T_u)$ are

$$\left.\begin{aligned} E\{LO(T_u)\} &= \tfrac{1}{2}\{p(\tfrac{1-\alpha}{\alpha}) + \Delta_1^2/\alpha + p\,\ell n\alpha\} \\ V\{LO(T_u)\} &= \tfrac{1}{2}p(\tfrac{1-\alpha}{\alpha})^2 + \Delta_1^2/\alpha^2 \end{aligned}\right\} \; x \text{ in } \Pi_1 \qquad (8.8.1)$$

$$\left.\begin{aligned} E\{LO(T_u)\} &= \tfrac{1}{2}\{p(1-\alpha) - \Delta_1^2 + p\,\ell n\alpha\} \\ V\{LO(T_u)\} &= \tfrac{1}{2}p(1-\alpha)^2 + \alpha\Delta_1^2 \end{aligned}\right\} \; x \text{ in } \Pi_2$$

As $\frac{\alpha-1}{\alpha} < \ell n\alpha < \alpha-1$ for $0 < \alpha < 1$, $E\{LO(T_u)\}$ is positive in Π_1 and negative in Π_2. We also note that the mean and variance of $LO(T_u)$ increase as p increases and the larger mean and variance of $LO(T_u)$ correspond to the population with the larger variance i.e. Π_1.

For allocation purposes the rule $LO(T_u) \gtrless 0$ is equivalent to

$$\frac{1-\alpha}{\alpha}\left(x - \frac{\theta}{1-\alpha}\right)'\left(x - \frac{\theta}{1-\alpha}\right) \gtrless \frac{\theta'\theta}{1-\alpha} - p\ \ell n\alpha$$

$$\gtrless H$$

and hence the optimal probabilities of misclassification are

$$Q_1 = \Pr\{\chi^2(p, \frac{\Delta_1^2}{(1-\alpha)^2}) < \frac{\alpha}{1-\alpha} H\} \qquad x \text{ in } \Pi_1$$

$$Q_2 = \Pr\{\chi^2(p, \frac{\alpha\Delta_1^2}{(1-\alpha)^2}) > \frac{1}{1-\alpha} H \qquad x \text{ in } \Pi_2 .$$

These probabilities are readily evaluated for odd degrees of freedom, Section 3.6.

The unconditional bias of $LO(E_u)$ was derived in Section 7.9 as

$$B\{LO(E_u)\} = \frac{p+1}{n-p-2} \tfrac{1}{2}\{p(\frac{1-\alpha}{\alpha}) + \Delta_1^2/\alpha\} \qquad x \text{ in } \Pi_1$$

and

$$B\{LO(E_u')\} = \frac{p+1}{n-p-2} \tfrac{1}{2}\{p(1-\alpha) - \Delta_1^2\} \qquad x \text{ in } \Pi_2.$$

Hence the relative unconditional bias of $LO(E_u)$ i.e. $B\{LO(E_u)\} / E\{LO(T_u)\}$, is

$$\frac{p+1}{n-p-2} (1 - \tfrac{1}{2}\, p\ \ell n\alpha/E\{LO(T_u)\}). \qquad (8.8.2)$$

As $0 < \alpha < 1$ we see from (8.8.1) and (8.8.2) that the relative bias of $LO(E_u)$ is positive for x in Π_1, but may, depending on the size of n, p and Δ_1, be negative for x in Π_2. Hence while LO(E_u) always overstates true log-odds for Π_1 with x in Π_1, it may understate true log-odds for Π_1 with x in Π_2. The relative bias of LO(E_u) decreases as n increases with x in Π_1 or Π_2 and with x in Π_1 increases as p increases.

The unconditional mean of LO(P_u) was derived in Section 7.11 as

$$E\{LO(P_u)\} = (-1)^{t-1} \Big[\sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} e^{-\frac{1}{2}\lambda_t} \sum_{i=o}^{j-1} \frac{n}{n+2i} \tag{8.8.3}$$

$$- \frac{n}{2} \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_t)^j}{j!} e^{-\frac{1}{2}\lambda_t} \sum_{i=1}^{\infty} \frac{c_t^i}{i} \prod_{k=1}^{i} \left(\frac{\frac{p}{2}+i+j-k}{\frac{n}{2}+i+j-k} \right) \Big] + \tfrac{1}{2}\, p \ell n \alpha$$

where $\lambda_1 = \frac{n}{n+\alpha} \Delta_1^2$ $\qquad c_1 = \frac{n}{n+1} (\frac{\alpha-1}{\alpha})$

and

$\lambda_2 = \frac{n}{\alpha n+1} \Delta_1^2$ $\qquad c_2 = \frac{n}{n+1} (1-\alpha)$,

x in Π_t, t = 1 and 2.

For these series expansions to be valid $-1 < c_t \leq 1$, hence

$$\frac{1}{2+\frac{1}{n}} < \alpha < 2 + \frac{1}{n} . \tag{8.8.4}$$

The bias and relative bias of LO(P_u) are clearly dependent on the dimension parameter p. With x in Π_1 i.e. the population with the larger variance, the bias and relative bias of LO(P_u) are negative. Thus LO(P_u) understates true log-odds for Π_1 with x in Π_1. However with x in Π_2 it may overstate LO(T_u) depending on the size of n, p and Δ_1.

That $B\{LO(P_u)\} < 0$ for x in Π_1 follows by noting that its mean $E\{LO(P_u)\}$, Section (8.8.3), may be bounded above, since

$$\sum_{j=1}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \sum_{i=o}^{j-1} \frac{n}{n+2i} < \tfrac{1}{2}\lambda_1 \quad \text{as in Section 7.10}$$

$$\frac{n}{2} \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \sum_{i=1}^{\infty} - \frac{c_1^i}{i} \prod_{k=1}^{i} \left(\frac{\frac{p}{2}+i+j-k}{\frac{n}{2}+i+j-k} \right) < \frac{n}{2} \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \{-c_1 \frac{\frac{p}{2}+j}{\frac{n}{2}+j}\}$$

(8.8.5)

$$< -c_1 \sum_{j=o}^{\infty} \frac{(\frac{1}{2}\lambda_1)^j}{j!} e^{-\frac{1}{2}\lambda_1} \{\tfrac{p}{2}+j\}$$

$$= -c_1 \{\tfrac{p}{2} + \tfrac{1}{2}\lambda_1\}.$$

The intermediate results (8.8.5) follow as $-c_1 > 0$, $p < n$ and

$$\sum_{i=1}^{\infty} \frac{-c_1^i}{i} \prod_{k=1}^{i} \left(\frac{\frac{p}{2}+i+j-k}{\frac{n}{2}+i+j-k} \right)$$

is a convergent alternating series. Hence with x in Π_1, from (8.8.1), (8.8.3) and (8.8.5)

$$B\{LO(P_u)\} < [-\tfrac{1}{2}p(\frac{1-\alpha}{\alpha}) - \tfrac{1}{2}\Delta_1^2/\alpha](1 - \frac{n}{n+1})$$

$$< 0 \quad \text{as} \quad 0 < \alpha < 1.$$

The unconditional mean of LO(T_u) and relative unconditional bias of LO(E_u) and LO(P_u) x 10^2 are given in Table 8.8.1. Their evaluation requires specification of α, Δ_1, p and n. We have assumed $0 < \alpha < 1$. Given the constraint (8.8.3) on α, our previous studies on the quadratic discriminant function Chapters 3, 4 and 5 and our desire to introduce reasonable imbalance in the covariance matrices we set $\alpha = 0.5$. The evaluation of the infinite series in E{LO(P_u)} was as described in Sections 8.5 and 8.6 with a minor modification when $t = 1$ as $c_1 < 0$. Following our approach in Chapter 5, we define Σ as

$$\Sigma = \tfrac{1}{2}(\Sigma_1 + \Sigma_2) = \frac{1+\alpha}{2} \Sigma_1$$

thus

$$\Delta^2 = \frac{2}{1+\alpha} (\mu_1 - \mu_2)' \Sigma_1^{-1} (\mu_1 - \mu_2)$$

$$= \frac{2}{1+\alpha} \Delta_1^2 . \qquad (8.8.6)$$

The separation parameter Δ was chosen as before to give optimal linear PMC's when $\alpha = 1$ of .3, .2, .1 and .05. The values of Δ_1 follow from the relationship (8.8.6). The dimension parameter p was set at 1, 3, 9 and 15 to allow exact evaluation of the optimal PMC's, Q_t. To ensure positive definiteness of the sample covariance matrices S_t i.e. $n_t - p > 0$, and to facilitate comparison with the results of the previous sections, we set $n_1 = n_2 = 24$.

Given the relationship between bias and classification performance a simulation study was undertaken similar to that detailed in Section 8.5, with 100 test observations from Π_1 and Π_2 and with 200 sample iterations. From this estimates of the expected actual PMC's of LO(E_u), LO(U_u) and LO(P_u) were obtained. These estimated expected actual PMC's x 10^2, together with the optimal PMC's, Q_t x 10^2 are listed in Table 8.8.2. Also included are the maximum standard errors of the estimates.

Table 8.8.1

The Unconditional Mean of LO(T_u) and Relative Bias of LO(E_u) and LO(P_u) x 10^2,

Σ_t Unknown and Proportional

Sample Sizes

$n_1 = n_2 = 24$

Proportionality $\alpha = 0.5$

Population		Π_1			Π_2		
Δ & PMC	p	Mean T_u	Relative Bias E_u	 P_u	Mean T_u	Relative Bias E_u	 P_u
	1	0.98	13	- 17	- 0.51	4	0
Δ = 1.049	3	1.29	38	- 26	- 0.70	- 10	1
.30	9	2.21	186	- 51	- 1.28	- 110	- 7
	15	3.13	608	- 72	- 1.86	- 410	- 23
	1	2.28	11	- 19	- 1.16	7	- 2
Δ = 1.683	3	2.58	29	- 26	- 1.35	5	1
.20	9	3.51	145	- 46	- 1.93	- 47	- 1
	15	4.43	497	- 64	- 2.51	- 245	- 9
	1	5.08	10	- 24	- 2.56	8	- 7
Δ = 2.563	3	5.39	25	- 29	- 2.75	13	- 4
.10	9	6.31	115	- 44	- 3.33	5	1
	15	7.23	393	- 58	- 3.91	- 75	1
	1	8.27	10	- 30	- 4.15	9	- 12
Δ = 3.290	3	8.58	24	- 34	- 4.35	16	- 9
.05	9	9.50	102	- 45	- 4.91	28	- 2
	15	10.42	343	- 57	- 5.51	13	1

Table 8.8.2

The Optimal Probabilities of Misclassification of LO(T_u) x 10^2 and the Estimated Expected Actual Probability of Misclassification of LO(E_u), LO(U_u) and LO(P_u) x 10^2, Σ_t Unknown and Proportional

Sample Sizes $n_1 = n_2 = 24$

Proportionality $\alpha = 0.5$

Population		Π_1				Π_2			
Method		T_u	E_u	U_u	P_u	T_u	E_u	U_u	P_u
Δ & PMC	p	Q_1				Q_2			
	1	39	45	46	45	19	23	23	22
Δ = 1.049	3	33	39	43	42	18	26	22	23
.30	9	23	25	39	39	15	50	25	29
	15	18	23	46	45	11	61	28	38
	1	24	30	30	30	14	17	17	17
Δ = 1.683	3	22	27	30	29	13	19	17	17
.20	9	16	18	29	30	10	36	20	23
	15	12	20	41	40	8	55	26	33
	1	12	12	12	12	7	9	9	9
Δ = 2.563	3	11	12	13	13	7	11	10	10
.10	9	8	10	17	17	5	24	13	14
	15	7	15	31	32	4	45	21	25
	1	6	6	6	6	4	5	5	5
Δ = 3.290	3	5	6	6	7	3	6	5	5
.05	9	4	5	9	10	3	15	8	8
	15	3	11	23	25	2	36	17	18
Max. S.E.			.008				.011		

S.E. = Standard Error

In Tables 8.8.1 and 8.8.2 we note that the optimal PMC's Q_t decrease with increasing p for all Δ i.e. as $\hat{E}\{LO(T_u)\}$ increases and that Q_1 is always greater than Q_2. However when we consider the estimated expected actual PMC's, Table 8.8.2 which are in all cases in excess of Q_t we see that in general they increase with increasing p. We also note that while the expected actual PMC's of $LO(U_u)$ and $LO(P_u)$ in Π_1 are, like Q_1, greater than their corresponding values in Π_2, this is not the case for $LO(E_u)$ when p = 9 and 15. This classification behaviour of $LO(E_u)$ when related to the size and direction of its relative bias Table 8.8.1 indicates that it is only when the relative bias of $LO(E_u)$ in Π_1 exceeds 1 that there is a definite relationship with classification performance. The reversal in the size of the expected actual PMC's of $LO(E_u)$ when $p > 9$ is also apparent in the asymptotic expansions of its expected actual PMC as derived by Mc Lachlan (1975). The predictive method's understatement of $LO(T_u)$ in Π_1 for all p is clearly reflected in Table 8.8.2 where its larger expected actual PMC's are in Π_1. Another surprising result in Table 8.8.2 is the differing classification behaviour of $LO(U_u)$ and $LO(P_u)$ in Π_2 when p = 15 and Δ = 1.049 and 1.683, where we see that $LO(P_u)$ misallocates substantially more observations than does $LO(U_u)$. The size of $LO(P_u)$'s relative bias in Table 8.8.1 would not indicate this and we conclude that the relationship between bias and misallocation is not as simple as in the case of equal covariance matrices. Overall $LO(U_u)$ has the smaller total expected actual PMC.

We now consider the unconditional mean square errors of the estimators, that of LO(E_u) was derived in Appendix 7.B. As

$$MSE\{LO(U_u)\} = E\{LO^2(U_u)\} - E\{LO^2(T_u)\}$$

where

$$E\{LO^2(U_u)\} = \tfrac{1}{4}\left[\left(\frac{n-p-2}{n-1}\right)^2 E(a) + \frac{n-p-2}{n-1} E(b) + E(c)\right]$$

with the expectations of a, b and c given in Appendix 7.B, the result for LO(U_u) follows. In Table 8.8.3 we list the unconditional mean square errors of LO(E_u), LO(U_u) and LO(Pu), the latter estimated from the simulation study previously cited in this section. At the base of Table 8.8.3 we give the maximum and minimum ratios of the estimated mean square errors of LO(E_u) and LO(U_u) to their exact values.

In Table 8.8.3 we note that the mean square errors of all the estimators in Π_1 are greater than their corresponding values in Π_2, as were the simulation estimates. This was to be expected as Π_1 has the greater population variance. All mean square errors increase with increasing p and Δ. The mean square error of LO(U_u) is always less than that of LO(E_u), the difference being substantial for $p \geqslant 9$. We were unable to prove this result analytically. However it is once again apparent that the correction for bias is vital in the estimative approach. The mean square error of LO(P_u) is in general less than that of LO(U_u), especially in Π_1 for $p \geqslant 9$. The three exceptions are when p = 1, Δ = 2.563 and p = 1 and 3, Δ = 3.290, these occurred in the simulation also. It is interesting to note that the larger expected actual PMC's of LO(P_u) as compared with LO(U_u) in Π_2 when

Table 8.8.3

The Unconditional Mean Square Errors of LO(E_u), LO(U_u) and LO(P_u),

Σ_t Unknown and Proportional

Sample Sizes

$n_1 = n_2 = 24$

Proprotionality $\alpha = 0.5$

Population		Π_1			Π_2		
Δ & PMC		Method			Method		
	p	E_u	U_u	P_u	E_u	U_u	P_u
Δ = 1.049 .30	1	1.30	1.03	0.76	0.28	0.23	0.20
	3	5.52	3.41	2.21	1.27	0.84	0.63
	9	106.62	26.59	12.39	24.20	6.64	5.69
	15	1991.24	141.91	48.11	467.10	35.32	27.87
Δ = 1.683 .20	1	2.88	2.26	1.87	0.54	0.43	0.33
	3	9.04	5.39	3.78	1.77	1.14	0.80
	9	137.67	32.50	15.46	26.16	7.48	6.14
	15	2351.63	161.23	53.33	480.00	37.97	28.81
Δ = 2.563 .10	1	7.84	6.11	6.44	1.50	1.17	0.76
	3	19.11	11.02	9.37	3.45	2.14	1.34
	9	217.95	47.34	25.33	33.73	9.79	7.09
	15	3238.28	207.47	68.70	535.22	44.82	30.64
Δ = 3.290 .05	1	16.07	12.51	16.30	3.23	2.52	1.55
	3	34.71	19.67	20.35	6.41	3.84	2.27
	9	331.51	67.70	43.20	47.89	13.30	8.16
	15	4429.80	267.75	93.86	643.67	54.53	39.45
Min Ratio		0.87	0.87		0.85	0.27	
Max Ratio		1.03	1.01		1.20	1.21	

$p = 15$ and $\Delta = 1.049$ and 1.683 are not reflected in larger mean square errors. This indicates that here the size of the log-odds of the misclassified observations by $LO(P_u)$ are small.

When estimating true log-odds, given that the populations are multinormally distributed with unknown proportional covariance matrices, we conclude that the unbiased method is the better estimative approach. It has smaller mean square errors, zero bias and smaller total expected actual PMC than $LO(E_u)$. Also its individual expected actual PMC's do not reverse direction for moderately large p as is the case for $LO(E_u)$. The clear superiority of the predictive approach when Σ_t were equal and p large does not persist. The predictive estimator usually understates true log-odds. This understatement can on occasions be severe. With small Δ and large p, $LO(P_u)$ misallocates substantially more observations than does $LO(U_u)$ in Π_2, the population with the smaller variance, otherwise their misclassification performance is comparable. Even though the mean square errors of $LO(P_u)$ are in general less than those of $LO(U_u)$ this is not enough to unreservedly recommend the predictive approach. Here we prefer the unbiased estimative approach as the better all round estimator of true log-odds.

APPENDIX 8A

RANDOM NUMBER GENERATOR

The following subroutine was used to generate standard normal random variates. It is based on the suggestions of Chen (1971) for computers with 32-bit words and was tested as outlined by Chen and Newman and Odell (1971, ch 9), before routine use with an IBM 370/138. The uniform random number generators are of the multiplicative congruential type and the Box-Muller(1958) transformations are used to obtain standard normal random variates.

```
  SUBROUTINE RAND32(X,J)
  REAL*8 XN1,XN2
  DIMENSION X(30)
  COMMON / SEED / XN1,XN2
C GENERATES NORMAL OBSERVATIONS
C SEE CHEN J.A.S.A. 1971 P400-403
C SEEDS XN1,XN2 MUST BE ODD INTEGERS
  IF (XN1 .EQ. 0.0D0)XN1=16387.0D0
  IF (XN2 .EQ. 0.0D0)XN2=262147.0D0
  XN1=DMOD (XN1*16387.0D0,2147483647.0D0)
  XN2=DMOD (DMOD(XN2*262147.0D0,2147483647.0D0) *262147.0D0,2147483647.0D0)
  A=DSQRT (-2.0*DLOG(XN1/2147483647.0D0))
  B=6.283185*(XN2/2147483647.0D0)
  X(J)=A*COS(B)
  X(J+1)=A*SIN(B)
  RETURN
  END
```

CHAPTER 9

KERNEL DENSITY ESTIMATION AND LOG ODDS

9.1 Introduction

In previous chapters estimators of true log-odds assumed the populations were multinormally distributed. A comparison with an estimator which does not require this assumption, difficult to ensure in practice, is undertaken here. Many non-parametric methods of discrimination are available in the literature. Of these we have chosen to consider kernel based methods, since the kernel method provides an estimator of the true density which is itself a density. As such, kernel based discrimination, unlike other non-parametric methods is undertaken by estimating true log-odds.

Given the weaker assumptions of the kernel method, namely that the true densities f_t (t = 1 and 2) are continuous, one would expect it to yield inferior estimates of true log-odds than parametric methods which assume the correct form of f_t. However for two multinormally distributed populations with equal covariance matrices, Van Ness and Simpson (1976) found that kernel based methods had superior allocation ability to the estimative method for small sample sizes and for relatively large dimension sizes. This surprising result stimulated our interest in kernel based estimation of true log-odds.

With the true probability density functions $f_t(t=1$ and $2)$ given by $N_p\ (\mu_t,\Sigma_t)$ and with $\Sigma_1 = \Sigma_2 = \Sigma$, we extend our comparison of the parametric estimators $LO(E_e)$, $LO(U_e)$ and $LO(P_e)$ of Chapter 8, to include a kernel based estimator of true log-odds. The notation of previous chapters is retained and will be supplemented for the kernel estimator as necessary.

9.2 Kernel Density Estimation

The kernel method is only one of many non-parametric density estimation methods proposed in the literature. The principal alternatives are spline, orthogonal series and histogram-type estimators. Reviews of these methods including the kernel method are given by Cover (1972), Wegman (1972a) and Fryer (1977) and compared by means of simulation in Anderson (1969) and Wegman (1972b). For estimation of log-odds these alternative methods were rejected since their density estimates are not themselves densities, with spline and orthogonal series methods giving negative values and the histogram type density of Loftsgaarden and Quesenberry (1965) integrating to infinity, Wagner (1975).

Kernel density estimation in the univariate case is considered first, followed by its extension to the multivariate case. Let $\{x_i\}$ $1 \leqslant i \leqslant n$ be a sequence of one dimensional independent identically distributed random variables with continuous probability density function f. Then the kernel density estimate $\hat{f}$ of f based on x_i is given by

$$\hat{f}(x) = \frac{1}{n}\frac{1}{h} \sum_{i=1}^{n} K(\frac{x-x_i}{h}) \qquad (9.2.1)$$

where K(z) is the kernel function and h a function of n is the smoothing parameter. For various constraints on h and K the density estimate $\hat{f}$ may be shown to be consistent, asymptotically Normal and unbiased. With $K(z) \geqslant 0$ and

$$\int K(z)\, dz = 1,$$

$\hat{f}$ is a density in its own right.

The origins of the kernel method may be traced back to Fix and Hodges (1951) who proposed a "naive estimator" or "running-histogram" approach to density estimation. By choosing an interval of width h, Fix and Hodges estimated the density f at any point as being proportional to the number of observations falling within an interval of width h centered at the point under consideration. It was this naive estimator which led Rosenblatt (1956) to define the univariate kernel or window estimator (9.2.1). For the naive estimator Rosenblatt showed that the asymptotic conditional mean square error MSE(x) where

$$\mathrm{MSE}(x) = E[\{f(x) - \hat{f}(x)\}^2]$$

approaches zero if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, thus establishing pointwise consistency of $\hat{f}$. He also obtained the value of h which minimises the asymptotic MSE(x). However this value of h(n) depends on the true density at the point x. By minimising the expanded integrated mean square error IMSE, where

$$\mathrm{IMSE} = \int \mathrm{MSE}(x)\, dx \qquad (9.2.2)$$

Rosenblatt obtained a value of h which only depends on n, however a knowledge of the true form of f is still required. The integrated mean square error (9.2.2) has become the principal measure of closeness of the kernel density estimate $\hat{f}$ to the true density $\tilde{f}$, Anderson (1969), Wegman (1972b), Fryer (1976,1977).

Parzen (1962) extended Rosenblatt's earlier work by defining the general kernel density estimate as in (9.2.1) i.e.

$$\hat{f}(x) = \frac{1}{n}\frac{1}{h} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

with the kernel function K(z) satisfying the conditions

$$\begin{array}{ll}
K(z) \text{ is a Borel function} & \lim\limits_{z\to\infty} \mid K(z)\, z \mid \, = 0 \\[2ex]
\sup\limits_{|z|<\infty} \mid K(z) \mid \, < \infty & \displaystyle\int K(z)\, dz = 1 \\[2ex]
\displaystyle\int \mid K(z) \mid dz < \infty & K(z) \geqslant 0 \\[2ex]
\displaystyle\int z^2 \mid K(z) \mid dz < \infty & \displaystyle\int z\, K(z)\, dz = 0.
\end{array} \qquad (9.2.3)$$

Parzen showed that if $h \to 0$ as $n \to \infty$, $\hat{f}$ is asymptotically unbiased, if $nh \to \infty$ as $n \to \infty$, $\hat{f}$ is point-wise consistent in mean square error. He also gave conditions for uniform consistency and asymptotic Normality of $\hat{f}$. From the conditions (9.2.3) it is clear that the kernel function K must be a symmetric density function. Included in Parzen's paper is a list of suitable kernel functions among which are the rectangular, triangular, Cauchy, Normal and Laplace densities. With more stringent conditions on h and K stronger consistency properties may be obtained, Wegman (1972a).

Cacoullos (1966) extended kernel density estimation to the multivariate case. It is now assumed that the true continuous probability density function f is p-dimensional as are the observations $\{x_i\}$, $1 \leqslant i \leqslant n$. The multivariate kernel density estimate is now defined as

$$f(x) = \frac{1}{n} \frac{1}{h^p} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \qquad (9.2.4)$$

where K is a p-dimensional density function satisfying similar conditions on R^p to those of Parzen (9.2.3) on R^1.

Cacoullos also defined a more general multivariate kernel density estimate f* where

$$f^*(x) = \frac{1}{n}\, \frac{1}{\prod_{j=1}^{p} h_j} \sum_{i=1}^{n} K\left(\frac{x_1-x_{i1}}{h_1}, \frac{x_2-x_{i2}}{h_2}, \ldots\ldots, \frac{x_p-x_{ip}}{h_p}\right)$$

and $x = (x_1, x_2, \ldots., x_p)'$

but restricted his attention to $h_1 = h_2 = \ldots.. = h_p = h$. In a similar manner to Parzen for p = 1, Cacoullos showed that if $h \to 0$ as $n \to \infty$, $\hat{f}$ is asymptotically unbiased and if $nh^p \to \infty$ as $n \to \infty$, $\hat{f}$ is point-wise consistent in mean square error. Uniform consistency and asymptotic Normality were also shown to hold for various conditions on h and K. Van Ryzin (1969) gives stronger consistency results for $\hat{f}$. The value of h(n) which minimises asymptotic MSE(x) was also found but as in Rosenblatt (1956) requires a knowledge of f(x).

While not considering the general kernel density estimate f*, Cacoullos considered a particular case which he called the product kernel case. Here

$$K(x) = K(x_1)\, K(x_2) \,\ldots\ldots..\, K(x_p)$$

$$= \prod_{j=1}^{p} K(x_j) \qquad (9.2.5)$$

where $K(x_j)$ is a univariate kernel. As K(x) is a multivariate density the condition (9.2.5) implies that K is a product of p univariate densities and so requires independence of the kernel function. The resulting product kernel density estimate f** is given by

$$f^{**}(x) = \frac{1}{n} \sum_{i=I}^{n} \left\{ \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{x_j-x_{ij}}{h_j}\right)\right\}. \qquad (9.2.6)$$

Cacoullos showed that the optimal choice of h_j $1 \leq j \leq p$ to minimise the asymptotic MSE(x) is to take $h_1 = h_2 = \ldots = h_p = h$ and if $h \to 0$ as $n \to \infty$ and $nh^p \to \infty$ as $n \to \infty$, unbiased and point-wise consistency of f** hold. The advantage of the product kernel density f** over the non-product kernel density $\hat{f}$ is, as Cacoullos notes, that the former is invariant under different scale transformations in each dimension, a desirable property in practice since the components of x may represent incommensurable characteristics.

9.3 The Choice of Kernel Function K and Smoothing Parameter h

We have seen in Section 9.2 that the only restriction on the kernel function K is that it must be a symmetric multivariate density. Which function to pick and how to obtain its associated smoothing parameter h for fixed n are the problems considered here.

Epanechnikov (1969) derived for the product kernel density f** (9.2.6) the optimal kernel function which minimises the asymptotic relative IMSE i.e. the IMSE divided by $\int f^2(x)\,dx$, as

$$K(z) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{z^2}{5}) & \text{if } |z| \leq 5 \\ 0 & \text{otherwise} \end{cases} \qquad (9.3.1)$$

which is independent of the true density f, the sample size n and dimension p. He also showed that the relative efficiency of the univariate Normal, rectangular and Laplace kernel functions to this optimal kernel function are .95, .93 and .76 respectively. Deheuvels (1977) has extended Epanechnikov's work to the non-product kernel density $\hat{f}$ (9.2.4).

As a kernel based estimator of true log-odds will be the log-ratio of the estimated densities, kernel density functions such as the optimal one (9.3.1) are to be avoided due to the zeros in their definitions. In Anderson (1969) it is shown that the exact form of K is not critical provided the smoothing parameter h is properly chosen. For these reasons and to facilitate comparison with the study of Van Ness and Simpson (1976), we will use a multivariate Normal density function for K, with covariance matrix hM i.e.

$$K(z) = N_p\,(0,hM)$$

The choice of M is considered in Section 9.5. This choice of K results in the kernel density estimate

$$\hat{f}(x) = (2\pi)^{-\frac{p}{2}}\,|hM|^{-\frac{1}{2}}\,\frac{1}{n}\sum_{i=1}^{n}\exp\{-\tfrac{1}{2}\,\frac{1}{h}\,(x-x_i)'\,M^{-1}(x-x_i)\}.$$

Having decided on the form of K we must now consider how to choose its associated smoothing parameter h. As we will specify that the true density f is $N_p(\mu,\Sigma)$, h may be chosen to minimise the IMSE. While this approach is not possible in practice it will allow us to see how the kernel method performs under what we might term optimal conditions. With $f(x) = N_p(\mu,\Sigma)$ and $K(z) = N_p(0,hM)$ the exact IMSE (9.2.2) is given by

$$\text{IMSE} = \frac{1}{(4\pi)^{\frac{p}{2}}}\left[\frac{1}{|\Sigma|^{\frac{1}{2}}} - \frac{2^{\frac{p}{2}+1}}{|hM+2\Sigma|^{\frac{1}{2}}} + \frac{1}{n}\,\frac{1}{|hM|^{\frac{1}{2}}} + \frac{n-1}{n}\,\frac{1}{|hM+\Sigma|^{\frac{1}{2}}}\right] \qquad (9.3.2)$$

details of its derivation are given in Appendix 9.A. Besides IMSE other measures of closeness of $\hat{f}$ to f available in the literature are the expected mean square error i.e. $E_x\{MSE(x)\}$, Specht (1971), Wegman (1972b), Fryer (1976) and the Kullback and Liebler (1951) information measure i.e.

$$\int f(x)\ \ell n\{\hat{f}/f\}\ dx,$$

Bryan (1971), Wegman (1972b). An exact expression for $E_x\{MSE(x)\}$ with $f(x) = N_p(\mu,\Sigma)$ and $K(z) = N_p(0,hM)$ is derived in Appendix 9.A, the values of h which minimise the expected mean square error are similar to those that minimise the IMSE. No exact expression for the Kullback and Liebler measure was obtained but the values of h obtained by Herman and Habbema's (1975) sample based "modified maximum likelihood" method of estimating it, were similar to those that minimise the IMSE. Other methods used in practice to choose h when the form of f is unknown are reviewed in Fryer (1977).

9.4 Use of Kernel Density Estimation in Discriminant Analysis

We do not attempt here an exhaustive list of the application of kernel density estimation in discriminant analysis, a bibliography by Wertz and Schneider (1979) may be consulted for this. Instead we describe the principal studies which are in the main concerned with allocation performance rather than estimation of log-odds.

Van Ness and Simpson (1976) investigated the effect of increasing dimension size p, with $p \leqslant 20$ and small sample sizes $n_1 = n_2 = 10, 20$, on the allocation ability of the kernel and estimative methods. Their study involved the simulation of sample and test observations from two multinormal populations with equal covariance matrices and equal prior-probabilities. The test observations were then allocated by five different rules (i) estimative with Σ known (ii) estimative with Σ unknown (iii) estimative with $\Sigma_1 \neq \Sigma_2$ unknown (iv) product Normal kernel with joint covariance matrix hI and (v) product Cauchy kernel with joint covariance matrix hI. Van Ness and Simpson assumed the standard canonical form for the population densities f_t i.e. $f_1(x) = N_p(0,I)$ and $f_2(x) = N_p(\theta,I)$ where $\theta = (\Delta,0,0,\ldots 0)'$. The joint smoothing parameter h was chosen by generating additional observations from Π_1 and Π_2 and picking that h which maximised correct allocation of these observations. For the Normal kernel with $p = 1$ and $n = 10$ they quote the value of h obtained as $h = 2.25$ and for the product Cauchy kernel when $p = 20$ and $n = 20$, $h = 49.0$. These values of h are exceedingly large as compared with those that minimise the IMSE, Appendix 9.B. With these population assumptions and values of h, Van Ness and Simpson found that the kernel methods allocation ability was distinctly superior to the estimative method with Σ unknown for $p > 2$ and was in fact comparable to the estimative method with Σ known!

Van Ness (1979) extended the investigation to the case of Normal populations with unequal covariance matrices. The allocation performance of the rules (o) estimative μ_t and $\Sigma_1 \neq \Sigma_2$ known (i) estimative $\Sigma_1 \neq \Sigma_2$ known (ii) estimative $\Sigma_1 \neq \Sigma_2$ unknown (iii) estimative $\Sigma_1 = \Sigma_2 = \Sigma$ unknown (iv) product Normal kernel with joint covariance matrix hI (v) product Normal kernel with covariance matrices $h_t I$ where $h_1 = \alpha h_2$ and (vi) average linkage method, were compared. Once again a standard canonical form for f_t was assumed, namely, $f_1(x) = N_p(0,I)$ and $f_2(x) = N_p(\theta,\frac{1}{2}I)$. However the method of choosing the smoothing parameters was changed, a jacknife approach being used and the values of h now quoted appear to be more reasonable. Van Ness found that the kernel allocation rule (v) whose covariance matrices have similar structure to the populations was superior for $p > 5$ to all other methods except rules (o) and (i). The standard estimative rule (ii) and the cluster method (vi) were found to be the worst when $p \geqslant 2$, while rules (iii) and (iv) were comparable and better than rules (ii) and (vi).

A search of the literature to find confirmation of these studies was undertaken. In a paper by Koffler and Penfield (1979) the classification behaviour of the estimative rule for bivariate Normal populations with $\Sigma_1 = \Sigma_2$ unknown, was compared by simulation with the product Normal kernel with joint covariance matrix hI. The standard canonical form for the populations was assumed. With sample sizes $n_1 = n_2 = 64$, 200 and 729, 300 observations from each population and 5 to 10 sample and test iterations, the classification performance of the rules was found to be comparable. However the large sample sizes may account for this. In a similar simulation study by Gessaman and Gessaman (1972) for untransformed bivariate Normal populations with equal covariance matrix $\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$; means $\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, the same sample and test sizes but only one iteration, the product Normal kernel method had inferior classification performance. Both studies used the same fixed value for h of $n^{-\frac{1}{4}}$.

Specht (1966) developed a polynominal discriminant function based on the estimation of the true densities by product Normal kernels with covariance matrix hI. By simulation he compared the estimative rule $\Sigma_1 \neq \Sigma_2$ unknown and his polynominal discriminant function with joint covariance matrix on two five-dimensional Normal populations with densities $N_5(0,100^2I)$ and $N_5(\theta,50^2I)$ where $\theta = (100,0,0....0)'$. With 15 repetitions of sample sizes $n_1 = n_2 = 8$ and 450 test observations from each population, the polynomial discriminant function was distinctly superior to the standard estimative rule, a similar result to that of Van Ness (1979).

The question of whether the product kernel method is given an inherent advantage by the assumption of independence with the standard canonical form of the population is raised by these studies, as is the effect of large values of h as used by Van Ness and Simpson (1976).

In the literature comparative studies of the non-product Normal kernel and the estimative allocation rule are few. Byran (1971) for two bivariate Normal populations with equal covariance matrices compared the estimative allocation rule with a rule based on non-product Normal kernel densities with equal covariance matrix hS. With the standard canonical form of the populations, sample sizes of 61, test sizes of 500 from each population, one iteration and h chosen to minimise the Kullback and Liebler (1951) information measure, the non-product Normal kernel methods allocation performance was comparable to that of the estimative rule. Byran also shows that with the sample covariance matrix S in the Normal kernel density estimate the standard canonical form of the populations may be adopted without loss of generality. A similar study by Fukunaga and Kessell (1971) with p = 8 indicates that the non-product Normal kernel methods allocation performance is somewhat inferior to the estimative rule, however their choice of h may account for this.

While there is no published study of kernel based estimation of log-odds, kernel based estimation of posterior probabilities and odds on specific examples has been considered by Hermans and Habbema (1975) and Aitchison and Aitken (1976). Hermans and Habbema (1975) compared five estimators of posterior probabilities and odds on two practical examples. The methods compared were (i) estimative $\Sigma_1 = \Sigma_2$ and (ii) $\Sigma_1 \neq \Sigma_2$ unknown, (iii) predictive $\Sigma_1 = \Sigma_2$ and (iv) $\Sigma_1 \neq \Sigma_2$ unknown and (v) product Normal kernel with distinct covariance matrices $h_t\ D_t$, where D_t were the diagonal matrices of estimated variances, thus allowing for differing units of measurement in the components of x. They also proposed a sample based "modified maximum likelihood method" for choosing the smoothing parameters h_t. Bryan (1971) had proposed and implemented the same method. In both examples p = 2 with sample sizes of $n_1 = 40$, $n_2 = 35$ in one and $n_1 = 22$ and $n_2 = 33$ in the other. The kernel and appropriate parametric method provided comparable estimates of posterior odds.

Aitchison and Aitken (1976) considered the application of kernel methods to the estimation of posterior odds in the p-dimensional binary case. They established a suitable product kernel function the smoothing parameter of which was chosen as in Hermans and Habbema by the method of modified maximum likelihood, which in this case gives a consistent density estimate. Aitchison and Aitken applied their kernel method to a set of data previously analysed by Anderson (1972) using a logistic approach and found that the kernel method gave comparable allocation rates. The similarity of the odds given by the kernel and logistic methods was reflected in similar doubtful allocations based on the size of the estimated odds.

In concluding Aitchison and Aitken note the ease with which the product kernel method may be extended to the awkward yet common problem of discrimination with both discrete and continuous variables.

9.5 Standard Canonical Forms and Their Implications

The distribution of x in Π_t (t = 1 and 2) has been taken as $f_t = N_p\ (\mu_t, \Sigma)$ and we define our kernel density estimates $\hat{f}_t$ as

$$\hat{f}_t(x) = (2\pi)^{-\frac{p}{2}}\ |h_t M_t|^{-\frac{1}{2}}\ n_t^{-1} \sum_{i=1} \exp\{-\tfrac{1}{2}\ \frac{1}{h_t}\ (x-x_{it})'\ M_t^{-1}(x-x_{it})\} \tag{9.5.1}$$

The usual canonical form adopted in this case is to take the density f_t of Π_t as

$$f_1(x) = N_p(0,I) \quad \text{and} \quad f_2(x) = N_p(\theta,I) \tag{9.5.2}$$

respectively, where I is a pxp unit matrix and $\theta = (\Delta,0,0,\ldots,0)'$.

A transformation which accomplishes this is

$$y = AB(x-\mu_1) \quad = \quad T(x-\mu_1) \qquad (9.5.3)$$

where $B = D^{-\frac{1}{2}}\,\Gamma'$ is a matrix such that $B\,\Sigma\,B' = I$, with D the diagonal matrix of eigen values λ_j $1 \leqslant j \leqslant p$ of Σ, Γ an orthogonal matrix of eigen vectors of Σ and A an orthogonal matrix with first row $(\nu_1/\sqrt{\nu'\nu},\ \nu_2/\sqrt{\nu'\nu},\ \ldots\ldots,\ \nu_p/\sqrt{\nu'\nu})$ where $\nu = B(\mu_2-\mu_1)$ and $\nu'\nu = \Delta^2$ the canonical form (9.5.2) results. We note that the transformation

$$y = B(x-\mu_1) \qquad (9.5.4)$$

allows the canonical form

$$f_1(x) = N_p(0,I) \quad , \quad f_2(x) = N_p(\nu,I). \qquad (9.5.5)$$

Now whereas the parametric estimators of log-odds i.e. $LO(E_e)$, $LO(U_e)$ and $LO(P_e)$ are invariant to the transformation (9.5.3) the kernel density estimate may not be, depending on the choice of its covariance matrix h_tM_t. In the literature the two choices of h_tM_t are Case (a) $h_tM_t = h_tI$ and Case (b) $h_tM_t = h_tS$. Case (a) corresponds to standardising all the variables to unit variances and will be treated as such. We note that Case (a) results in a product kernel density as defined by Cacoullos (9.2.6), Case (b) in a non-product kernel density. We now show that Case (a) product kernels are not invariant to the transformation (9.5.3) whereas Case (b) non-products kernels are.

Case (a) $M_t = I$. Here on the original scale x

$$\hat{f}_t(x) = (2\pi)^{-\frac{p}{2}} |h_t I|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t}(x-x_{it})'(x-x_{it})\}$$

with the transformation (9.5.4), $y = B(x-\mu_1)$

$$\hat{f}_t(y) = (2\pi)^{-\frac{p}{2}} |h_t B\ B'|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t}(y-y_{it})'(B\ B')^{-1} (y-y_{it})\}$$

but $B\ B' = D^{-\frac{1}{2}}\ \Gamma'\ \Gamma\ D^{-\frac{1}{2}} = D^{-1}$ as Γ orthogonal, hence

$$\hat{f}_t(y) = (2\pi)^{-\frac{p}{2}} |h_t D^{-1}|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t}(y-y_{it})'\ D(y-y_{it})\}$$

i.e. $K_t(z) = N_p(0,h_t D^{-1})$. Thus one may assume without loss of generality the canonical form (9.5.5) for f_t and the kernel density estimate is still a product kernel as D is a diagonal matrix. However with the full transformation (9.5.3), $y = T(x-\mu_1)$

$$\hat{f}_t(y) = (2\pi)^{-\frac{p}{2}} |h_t T\ T'|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t}(y-y_{it})'\ (T\ T')^{-1}(y-y_{it})\}$$

and $T\ T' = A\ B\ B'\ A' = A\ D^{-1}\ A'$, hence

$$\hat{f}_t(y) = (2\pi)^{-\frac{p}{2}} |h_t A\ D^{-1}\ A'|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t}(y-y_{it})'\ A\ D\ A'(y-y_{it})\} \tag{9.5.6}$$

i.e. $K_t(z) = N_p(0,h_t A\ D^{-1}\ A')$, a non-product kernel unless $\lambda_j = \alpha$ for all j i.e. $\Sigma = \alpha I$.

Case (b) $M_t = S_x$. Here on the original scale x

$$\hat{f}_t(x) = (2\pi)^{-\frac{p}{2}} |h_t S_x|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t} (x-x_{it})' S_x^{-1} (x-x_{it})\}$$

with the full transformation (9.5.3), $y = T(x-\mu_1)$

$$\hat{f}_t(y) = (2\pi)^{-\frac{p}{2}} |h_t T S_x T'|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t} (y-y_{it})'(T S_x T')^{-1}(y-y_{it})\}$$

But $(n_1+n_2-2) S_y = \sum_{i=1}^{n_1} (y_{i1}-\bar{y}_1)(y_{i1}-\bar{y}_1)' + \sum_{i=1}^{n_2} (y_{i2}-\bar{y}_2)(y_{i2}-\bar{y}_2)'$

$$= (n_1+n_2-2) T S_x T'$$

hence $\hat{f}_t(y) = (2\pi)^{-\frac{p}{2}} |h_t S_y|^{-\frac{1}{2}} n_t^{-1} \sum_{i=1}^{n_t} \exp\{-\tfrac{1}{2} \frac{1}{h_t} (y-y_{it})' S_y^{-1} (y-y_{it})\}$

and so the standard canonical form (9.5.2) may be adopted without loss of generality.

While this invariance property of the Normal kernel density estimate when $M_t = S$ is attractive,we concentrate our attention on the product Normal kernel where $M_t = I$. The reasons for this decision are; we have demonstrated that Van Ness and Simpson's (1976) assumption of $M_t = I$ and the standard canonical form (9.5.2) for f_t implies from (9.5.6) that $A D^{-1} A' = I$ i.e. $\lambda_j=1$ for all j $\Sigma=I$ ab-inito, and so their results and conclusions do not apply in general. Even so the distinct superiority of their kernel method when $\Sigma = I$ and $p \geqslant 2$ remains to be explained. We are also interested in seeing what happens when $\Sigma \neq I$ and whether the good classification behaviour of the product kernel method is accompanied by good estimation of true log-odds.

One possible explanation of the good classification behaviour of the product kernels of Van Ness and Simpson when $\Sigma = I$ is the large values of the smoothing parameter $h_1 = h_2 = h$ used by them. Their quoted value of $h = 2.25$ for the Normal kernel when $p = 1$ and $n_1 = n_2 = 10$ contrasts with $h = .575$ which minimises the IMSE, Appendix 9.B. We now show that with their population assumptions and method of choosing h, large values of h were inevitable and this resulted in comparable classification performance with the estimative rule when Σ is known and equal to I.

With the standard canonical form (9.5.2) for f_t, $K_t(z) = N_p(0,hI)$ and $n_1 = n_2 = n$, Van Ness and Simpson's (1976) kernel log-odds is

$$\ell n \left[\frac{\sum_{i=1}^{n} \exp\{-\tfrac{1}{2} \frac{1}{h} (x-x_{i1})' (x-x_{i1})\}}{\sum_{i=1}^{n} \exp\{-\tfrac{1}{2} \frac{1}{h} (x-x_{i2})' (x-x_{i2})\}} \right] \qquad (9.5.7)$$

with allocation to Π_1 or Π_2 according as (9.5.7) $\gtrless 0$. If h is sufficiently large to allow the expansion of the exponential terms this would approximate to the rule, allocate to Π_1 or Π_2 according as

$$\sum_{i=1}^{n} (x-x_{i2})' (x-x_{i2}) \gtrless \sum_{i=1}^{n} (x-x_{i1})' (x-x_{i1})$$

or $$(x-\bar{x}_2)' (x-\bar{x}_2) + \frac{n-1}{n} \text{tr } S_2 \gtrless (x-\bar{x}_1) (x-\bar{x}_1) + \frac{n-1}{n} \text{tr } S_1$$

or $$(\bar{x}_1-\bar{x}_2)' \{x-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\} \gtrless \tfrac{1}{2} \frac{n-1}{n} \{\text{tr } S_1 - \text{tr} S_2\}. \qquad (9.5.8)$$

This behaviour was noted by Specht (1966). The left hand side of (9.5.8) corresponds to the classical rule

$$LO(E_e) = (\bar{x}_1-\bar{x}_2)' S^{-1} \{x-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\} \gtrless 0$$

with Σ known and replaced by I, while the right hand side corresponds to the difference in trace between the estimated covariance matrices of the two populations. The latter will vary about zero the optimal cut-off point. Ironically the comparable performance of the kernel and classical rules for Σ known was noted by Van Ness and Simpson but no explanation was given. With their trial and error method of selecting h so as to maximise correct allocation on additional observations from Π_1 and Π_2 large values of h were inevitable. A similar argument may be carried out for their Cauchy kernel.

9.6 Interclass Correlation Matrix Models

To investigate the questions raised in Section 9.5 we choose to look at the particular case of equi-correlated variables and assume that $\Sigma = R$ where

$$R = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots \; 1 \; \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

is the pxp interclass correlation matrix. Thus we have

$$f_1(x) = N_p(\mu_1, R), \quad f_2(x) = N_p(\mu_2, R) \quad \text{and} \quad K_t(z) = N_p(0, h_t I)$$

and with the transformation $y = (x-\mu_1)$ we may assume without loss of generality that

$$f_1(x) = N_p(0, R), \quad f_2(x) = N_p(\eta, R) \quad \text{and} \quad K_t(z) = N_p(0, h_t I) \tag{9.6.1}$$

where $\eta = (\mu_2 - \mu_1) > 0$. Mahalonabis' squared distance Δ^2 between the populations is given by

$$\Delta^2 = \eta' R^{-1} \eta = \frac{1}{1-\rho}\left[\sum_1^p \eta_j^2 - \frac{\rho(\sum_1^p \eta_j)^2}{1+(p-1)\rho}\right] . \tag{9.6.2}$$

The eigen values of R are $\lambda_1 = 1+(p-1)\rho$, $\lambda_2 = \lambda_3 = \text{-----}\ \lambda_p = 1-\rho$ and for R positive definite

$$\frac{1}{1-p} < \rho < 1.$$

Now from Section 9.5 we see that with the transformation $y = B(x-\mu_1)$ we may assume without loss of generality that

$$f_1(x) = N_p(0,I), \quad f_2(x) = N_p(\nu,I) \quad \text{and } K_t(z) = N_p(0,h_t D^{-1}) \qquad (9.6.3)$$

where $\nu = B\eta = D^{-\frac{1}{2}}\ \Gamma'\ (\mu_2-\mu_1)$ and $\nu'\nu = \Delta^2$. Here D is the diagonal matrix of eigen values λ_j of R and the orthogonal matrix Γ is taken to be the pxp Helmert matrix

$$\Gamma = \begin{pmatrix} \frac{1}{\sqrt{p}} & \frac{1}{\sqrt{1.2}} & \frac{1}{\sqrt{2.3}} & \cdots & \frac{1}{\sqrt{(p-1)p}} \\ \frac{1}{\sqrt{p}} & \frac{-1}{\sqrt{1.2}} & \frac{1}{\sqrt{2.3}} & \cdots & \vdots \\ \vdots & 0 & \frac{-2}{\sqrt{2.3}} & \cdots & \vdots \\ \vdots & \vdots & 0 \ \cdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{p}} & 0 & \cdots & \cdots & \frac{-(p-1)}{\sqrt{(p-1)p}} \end{pmatrix}.$$

For convenience we will assume that the sample sizes n_t are equal i.e. $n_1 = n_2 = n$ and with $LO(K_e)$ denoting our kernel estimator of true log-odds from (9.5.1) and (9.6.3) it follows that

$$LO(K_e) = \ln\{\frac{\hat{f}_1(x)}{\hat{f}_2(x)}\} = \ln \left[\frac{\sum_{i=1}^{n} \exp\{-\frac{1}{2}\ \frac{1}{h}\ (x-x_{i1})'\ D(x-x_{i1})\}}{\sum_{i=1}^{n} \exp\{-\frac{1}{2}\ \frac{1}{h}\ (x-x_{i2})'\ D(x-x_{i2})\}}\right].$$

The single smoothing parameter h, as $n_1 = n_2 = n$, was chosen to minimise the integrated mean square error (9.3.2) and is both a function of n and ρ i.e. the population covariance matrix. Details of how this was accomplished are given in Appendix 9.B.

We must now specify the parameters $\{n,p,\rho,\Delta,\eta_j\}$. The parameters n,p and Δ were chosen to coincide with the studies of Chapter 8, while being similar to the range of parameters considered by Van Ness and Simpson (1976), i.e. $n = 12$, $p = 1,4,8,16$ and Δ so as to give optimal PMC's of 30, 20, 10 and 5%. To guide our choice of ρ and the mean parameters η_j or ν_j, $1 \leqslant j \leqslant p$, we first look at the easier case of $p = 2$.

From (9.6.1) with $p = 2$

$$R' = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \ -1 < \rho < 1, \ \lambda_1 = 1+\rho, \ \lambda_2 = 1-\rho, \ \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \text{ and}$$

$$K_t(z) = N_2\left(\binom{0}{0}, hI\right).$$

Model I. Let $\eta_1 = \eta_2 = c > 0$ then from (9.6.2)

$$\Delta^2 = \frac{1}{1-\rho} \quad [2c^2 - \frac{\rho 4c^2}{1+\rho}] = \frac{2c^2}{1+\rho}$$

hence
$$c = \Delta\sqrt{\frac{1+\rho}{2}} = \Delta\sqrt{\frac{\lambda_1}{2}} .$$

With fixed Δ, as $\rho \to -1$, $\lambda_1 \to 0$ and $c \to 0$ and the populations become indistinguishable, and we show that $LO(K_e) \to 0$.

$$\text{For } \nu = D^{-\frac{1}{2}}\,\Gamma'\,\eta = D^{-\frac{1}{2}}\begin{pmatrix} c\sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} c\sqrt{2/\lambda_1} \\ 0 \end{pmatrix} = \begin{pmatrix} \Delta \\ 0 \end{pmatrix}$$

hence from (9.6.3) the transformed densities are

$$f_1(x) = N_2\left(\binom{0}{0}, I\right), \ f_2(x) = N_2\left(\binom{\Delta}{0}, I\right), \ K_t(z) = N_2\left(\binom{0}{0}, hD^{-1}\right)$$

$$\text{and} \quad LO(K_e) = \ln\left[\frac{\sum_{i=1}^{n} \exp\{-\frac{1}{2}\,\frac{1}{h}\,(x-x_{i1})'\,D(x-x_{i1})\}}{\sum_{i=1}^{n} \exp\{-\frac{1}{2}\,\frac{1}{h}\,(x-x^*_{i1}-\nu)'\,D(x-x^*_{i1}-\nu)\}}\right] \qquad (9.6.4)$$

with $x_{i2} = x^*_{i1} + \nu$, where x^*_{i1}, $1 \leqslant i \leqslant n$ are distributed as x_{i1} i.e. $N_2(0,I)$.

As $\nu = \binom{\Delta}{o}$ the quadratic forms

$$\sum_{j=1}^{2} (x_j - x_{i1j})^2 \lambda_j \quad \text{and} \quad \sum_{j=1}^{2} (x_j - x^*_{i1j} - \nu_j)^2 \lambda_j \qquad (9.6.5)$$

in the exponential arguments of $LO(K_e)$ (9.6.4), are only distinguished in their first term. As $\rho \to -1$, $\lambda_1 \to 0$ and the quadratic forms (9.6.5) become identically distributed with $LO(K_e)$ varying about zero. Thus we would anticipate very poor classification behaviour by $LO(K_e)$ as the first eigen values approaches zero i.e. as R becomes singular.

A possible criticism of Model I is that it only behaves badly as $\lambda_1 \to 0$ that is for ρ negative whereas in practice ρ is usually positive. This led us to consider Model II.

Model II. Let $\nu = \binom{o}{\Delta}$ and as $\eta = \Gamma D^{\frac{1}{2}} \nu$, $\eta = \binom{c}{-c}$ with

$$c = \Delta\sqrt{\frac{\lambda_2}{2}} \ .$$

Now as $\lambda_2 = 1-\rho$, $\lambda_2 \to 0$ as $\rho \to 1$ and similar classification behaviour in $LO(K_e)$ as occurred in Model I is to be expected. A possible criticism of Model II is that while ρ is positive it may well be large i.e. close to 1, when $p > 2$, before the anticipated classification breakdown materialises. What we then sought was an intermediate model that would display the behaviour of both Models I and II.

With p = 2 from (9.6.2)

$$\Delta^2 = \frac{1}{\lambda_2}[(\eta_1^2+\eta_2^2) - \frac{\rho(\eta_1+\eta_2)^2}{\lambda_1}]$$

in Model I $\eta_1 = \eta_2 = c > o$, $\Delta^2 = \frac{2c^2}{\lambda_1}$, $c \to o$ as $\lambda_1 \to o$ i.e. as $\rho \to -1$

in Model II $\eta_1 = -\eta_2 = c > o$, $\Delta^2 = \frac{2c^2}{\lambda_2}$, $c \to o$ as $\lambda_2 \to o$ i.e. as $\rho \to 1$.

Now for Model III we wish Δ to depend on λ_1 and λ_2 and one possibility is to take

Model III $\eta_1 = c > o$, $\eta_2 = o$, then $\Delta^2 = \frac{c^2}{\lambda_2}(\frac{\lambda_1-\rho}{\lambda_1})$, $\nu = \begin{pmatrix} \Delta\sqrt{\frac{\lambda_2}{2(\lambda_1-\rho)}} \\ \Delta\sqrt{\frac{\lambda_1}{2(\lambda_1-\rho)}} \end{pmatrix}$

and $c \to o$ as $\lambda_1 \to o$ or $\lambda_2 \to o$ i.e. as $\rho \to \pm 1$.

These models generalised to the p > 2 case are, with $\eta = (\mu_2-\mu_1) > 0$, $\Sigma = R$ and their transformed values $\nu = B(\mu_2-\mu_1)$, $B\Sigma B' = I$.

Model I

$$\eta = (c,c,-----,c)' \quad , \quad \nu = (\Delta,0,0,-----0)'$$

Model II

$$\eta = (c,c,-----c,-(p-1)c)', \quad \nu = (0,0,-------0,\Delta)'$$

(9.6.6)

Model III

$$\eta = (c,0,0,-----,0)' \quad , \quad \nu = (\nu_1,\nu_2,-----\nu_j,----\nu_p)'$$

where $\nu_1 = \Delta\sqrt{\frac{\lambda_2}{p(\lambda_1-\rho)}}$, $\nu_j = \sqrt{\frac{\lambda_1}{(j-1)j\ (\lambda_1-\rho)}}$

$2 \leqslant j \leqslant p$

and $\lambda_1 = 1+(p-1)\rho$, $\lambda_j = 1-\rho \quad 2 \leqslant j \leqslant p$.

As $\frac{1}{1-p} < \rho < 1$ and the maximum value of p is 16 we took $\rho = \{-.066, 0, .2, .3, .4, .5, .6, .8\}$. The extreme values of $\{-.066, .8\}$ were chosen to introduce when p = 16 near singularity in the models. The value $\rho = 0$ was chosen to reproduce the case of $\Sigma = I$ of Van Ness and Simpson (1976), we also note that with $\rho = 0$ all three models coincide. The intermediate values $\rho = \{.2, .3, .4, .5, .6\}$ were chosen to investigate where the break-down in classification behaviour of the models would occur.

9.7 A Simulation Study

For the three models adopted in Section 9.6 a simulation study was undertaken to investigate the estimation and classification behaviour of the kernel method K_e as compared with the parametric methods E_e, U_e and P_e. The study was similar in form to that detailed in Section 8.5 and only a brief description is given here.

With fixed n and p, 100 test observations were generated from Π_1 i.e. a $N_p(0,I)$ distribution, and stored. The sample parameters $\bar{x}_1$, $\bar{x}_2$ and S were then generated and for each Δ and ρ i.e. each of the three models (9.6.6), the unconditional mean, bias, variance, mean square error, expected actual PMC and standard error of the means, mean square errors and expected actual PMC's of LO(M), $M \varepsilon \{E_e, U_e, P_e, K_e\}$ calculated. The mean, variance and PMC of LO(T_e) were also estimated. This sample generation process was repeated 100 times and suitable tests carried out to check estimated values with their exact values. To allow comparisons with the studies of Chapter 8, the same seeds for the random number generator, Appendix 8.A, were used here.

The estimated expected actual PMC's x10^2 obtained from the simulation study are listed in Table 9.7.1. As $n_1 = n_2$ only one result is listed for the parametric methods $\{E_e, U_e, P_e\}$ under the heading E_e. The K_e result for $\rho = 0$ is separated from the three models since as noted they coincide in this instance. With $p = 1$ the case of $\rho \neq 0$ does not arise. Maximum standard errors are given at the base of the table.

From the results of Table 9.7.1 we see that with $\rho = 0$ i.e. $\Sigma = I$ the results of Van Ness and Simpson (1976) hold for high dimensions when reasonable values of the smoothing parameter h are used. Contrary to their results however, for $\rho = 0$ and $p \leqslant 8$ the estimated expected actual PMC's of LO(K_e) are in all cases greater than or equal to those of LO(E_e). With increasing p they improve and for $p = 16$ they are smaller, but the superiority is not as marked as in Van Ness and Simpson's study. Further the expected actual PMC's of LO(K_e) do not display the pattern noted by Van Ness and Simpson where they paralleledthose of LO(E_e) with Σ known. This phenomenon is explained by Van Ness and Simpson's choice of large values for the smoothing parameters.

For the results of Model I we note as anticipated that with $\rho = -.066$ the estimated expected actual PMC's of LO(K_e) are larger than those of LO(E_e) and LO(K_e) with $\rho = 0$. With $p = 16$ and $\rho = -.066$ as demonstrated in Section 9.6 the estimated PMC's of LO(K_e) are close to .50 for all Δ. With increasing $\rho > 0$ the estimated expected actual PMC's of LO(K_e) decrease and for $\Delta \geqslant 2.563$ they are smaller than the corresponding values for LO(E_e) in all dimensions.

Table 9.7.1

Estimated Expected Actual Probabilities of Misclassification x 10^2

of LO(M), M $\varepsilon\{E_e, U_e, P_e, K_e\}$.

Sample Sizes $n_1 = n_2 = 12$

		Parametric Methods	K_e all models	K_e Model I				K_e Model II				K_e Model III			
Δ & PMC	p	E_e	ρ = 0	-.066	.2	.4	.8	-.066	.2	.4	.8	-.066	.2	.4	.8
Δ = 1,049 .30	1	37.5	37.2	-	-	-	-	-	-	-	-	-	-	-	-
	4	31.5	35.8	36.9	34.0	33.3	33.3	37.7	38.5	39.1	41.7	36.1	36.6	37.9	41.3
	8	38.7	43.1	45.5	40.7	39.7	39.1	44.3	45.0	45.5	46.4	43.0	42.8	43.7	45.4
	16	42.7	42.5	49.9	38.7	37.8	36.9	45.4	44.7	45.1	46.5	49.1	43.5	44.6	45.8
Δ = 1.683 .20	1	25.1	25.8	-	-	-	-	-	-	-	-	-	-	-	-
	4	20.8	24.0	25.3	22.2	21.6	21.0	24.6	25.3	26.8	30.9	24.7	25.3	26.4	31.2
	8	27.5	32.1	37.0	27.7	27.2	26.8	33.0	34.4	35.2	39.0	31.9	32.2	33.4	38.1
	16	35.9	34.7	49.7	28.3	27.9	27.7	37.9	38.0	39.6	42.4	47.8	36.6	38.2	41.1
Δ = 2.563 .10	1	10.2	11.3	-	-	-	-	-	-	-	-	-	-	-	-
	4	11.2	11.2	12.3	10.1	9.3	8.9	11.8	12.4	13.3	17.3	14.0	14.2	15.2	18.8
	8	15.3	17.4	22.6	13.7	12.6	12.4	18.4	19.9	21.1	26.5	18.1	17.9	19.8	25.7
	16	26.9	22.4	49.4	14.7	13.7	13.4	24.9	26.0	28.1	33.6	45.2	24.4	26.9	32.9
Δ = 3.290 .05	1	3.9	4.3	-	-	-	-	-	-	-	-	-	-	-	-
	4	5.6	5.3	5.8	4.2	3.9	4.0	5.8	6.2	6.5	9.8	7.9	8.1	8.5	11.5
	8	8.8	9.1	13.2	6.3	5.9	5.5	9.5	10.4	11.8	16.8	10.0	10.3	11.4	16.6
	16	20.4	13.5	49.0	7.8	6.8	6.5	14.0	16.2	18.8	25.3	42.2	15.3	17.8	25.0
Maximum S.E.		.011	.013	.011	.012	.011	.010	.012	.012	.010	.010	.012	.011	.011	.010

In Model II where we anticipated that the classification performance of LO(K_e) would disimprove as $\rho \to 1$ we see from Table 9.7.1 that this disimprovement as compared with LO(E_e) sets in at the early stage of $\rho = .2$ except for $p = 16$ and $\Delta > 2.563$ where it occurs at $\rho = .3$ when $\Delta = 2.563$ and $\rho = .6$ when $\Delta = 3.290$.

For the intermediate model, Model III, where we anticipated poor classification performance as $\rho \to 1$ and $1+(p-1)\ \rho \to 0$, we see from Table 9.7.1 that this is indeed the case. As in Model II the inferior classification performance as compared with LO(E_e) sets in at $\rho = .2$ except for $p = 16$ and $\Delta \geqslant 2.563$ where it occurs at $\rho = .4$ when $\Delta = 2.563$ and $\rho = .6$ when $\Delta = 3.290$.

We conclude from these classification results that for variables which are close to independence, product kernel methods can outperform conventional parametric methods, particularly at higher dimensions and for well separated populations if the parametric methods make no effort to incorporate the independence assumption.

Consider now the performance of the product kernel method as an estimator of true log-odds. In Table 9.7.2 we list the estimated unconditional means of LO(K_e) for the three models, also listed are the exact means of LO(U_e) and LO(P_e) derived and evaluated in Chapters 7 and 8. The exact unconditional mean square error of LO(U_e) and the estimated unconditional mean square errors of LO(P_e) and LO(K_e) are listed in Table 9.7.3. Included at the base of each table are the minimim and maximum ratios of the estimated to exact values. The estimative method E_e is omitted as it has been shown in Chapter 8, to be a poor estimator of true log-odds. The appropriate results for LO(E_e) may be found in Table 8.5.1 and 8.5.2.

Table 9.7.2

Exact and Estimated Unconditional Mean of LO(M)

$M\varepsilon\{U_e, P_e, K_e\}$ and x in Π_1.

Sample Sizes $n_1 = n_2 = 12$

		Parametric Methods		K_e all models	K_e Model I				K_e Model II				K_e Model III			
Δ & PMC	p	U_e	P_e	ρ = 0	-.066	.2	.4	.8	-.066	.2	.4	.8	-.066	.2	.4	.8
	1	0.55	0.50	0.35	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 1.049	4	0.55	0.50	0.53	0.47	0.73	0.95	2.08	0.47	0.41	0.38	0.36	0.47	0.47	0.45	0.43
.30	8	0.55	0.50	0.31	0.14	0.54	0.82	2.46	0.27	0.24	0.23	0.22	0.31	0.33	0.31	0.31
	16	0.55	0.50	0.40	0.02	0.97	1.59	5.58	0.26	0.28	0.30	0.47	0.05	0.38	0.39	0.59
	1	1.42	1.25	1.00	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 1.683	4	1.42	1.25	1.30	1.13	1.80	2.39	5.67	1.22	1.05	0.97	0.84	1.20	1.18	1.12	0.97
.20	8	1.42	1.25	0.91	0.52	1.57	2.37	7.14	0.89	0.78	0.73	0.67	0.91	0.92	0.88	0.81
	16	1.42	1.25	0.92	0.03	2.41	4.12	14.55	0.77	0.71	0.71	0.85	0.13	0.87	0.87	1.04
	1	3.28	2.71	2.58	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 2.563	4	3.28	2.71	2.97	2.56	4.22	5.85	15.48	2.89	2.46	2.26	1.88	2.82	2.77	2.60	2.14
.10	8	3.28	2.71	2.22	1.33	4.03	6.37	20.83	2.27	1.96	1.85	1.64	2.22	2.22	2.09	1.87
	16	3.28	2.71	2.05	0.04	5.85	10.60	36.98	1.93	1.69	1.65	1.71	0.28	1.94	1.91	2.00
	1	5.41	4.19	4.63	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 3.290	4	5.41	4.19	4.93	4.23	7.21	10.31	29.14	4.91	4.14	3.78	3.07	4.76	4.65	4.35	3.49
.05	8	5.41	4.19	3.75	2.27	7.19	11.87	40.39	3.91	3.34	3.15	2.76	3.75	3.72	3.50	3.08
	16	5.41	4.19	3.35	0.06	10.20	19.31	63.66	3.30	2.84	2.76	2.71	0.46	3.18	3.10	3.09
Min.Ratio		0.81	0.80													
Max.Ratio		1.42	1.38													

Table 9.7.3

Exact and Estimated Unconditional Mean Square Error of LO(M) $M\varepsilon\{U_e, P_e, K_e\}$ and x in Π_1.

Sample Sizes $n_1 = n_2 = 12$

		Parametric Methods		K_e all models	K_e Model I				K_e Model II				K_e Model III			
Δ & PMC	p	U_e	P_e	$\rho = 0$	-.066	.2	.4	.8	-.066	.2	.4	.8	-.066	.2	.4	.8
	1	0.46	0.32	0.37	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 1.049	4	1.43	1.38	1.64	1.72	1.84	2.79	20.42	1.77	1.68	1.87	4.22	1.66	1.64	1.88	4.42
.30	8	3.53	2.46	2.38	2.65	3.02	5.66	64.62	2.50	2.66	3.30	10.93	2.59	2.77	3.48	11.39
	16	22.26	9.45	4.20	5.98	7.20	20.26	362.05	4.58	5.25	8.57	75.98	5.81	5.44	8.93	77.68
	1	0.99	0.63	0.83	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 1.683	4	2.41	2.03	2.21	2.36	2.92	6.28	79.31	2.37	2.35	2.63	5.28	2.22	2.27	2.67	5.73
.20	8	5.43	3.51	3.03	3.81	4.48	12.79	232.25	3.13	3.45	4.19	12.12	3.47	3.70	4.57	13.04
	16	32.61	12.41	5.05	9.86	11.83	50.79	989.00	5.14	5.91	9.28	76.69	9.27	6.31	9.98	79.65
	1	2.73	1.80	2.11	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 2.563	4	5.20	3.77	3.64	4.17	6.48	21.29	380.57	3.80	4.13	4.82	8.71	3.54	3.80	4.66	9.42
.10	8	10.47	5.89	4.93	7.89	9.72	44.70	996.15	4.82	5.84	6.93	15.80	5.92	6.31	7.66	17.59
	16	58.34	18.55	7.71	23.41	28.97	178.29	2971.99	6.87	8.25	11.81	79.41	21.16	9.09	13.16	84.46
	1	5.65	4.36	4.00	-	-	-	-	-	-	-	-	-	-	-	-
Δ = 3.290	4	9.50	6.92	5.49	6.77	13.14	54.44	1088.94	5.56	6.58	8.06	14.36	5.12	5.64	7.27	14.91
.05	8	17.75	9.30	7.81	15.12	20.11	117.57	2521.13	7.14	9.51	11.28	21.79	9.45	10.10	12.29	24.51
	16	93.29	26.05	12.02	47.40	63.50	448.97	5812.84	9.65	12.38	16.31	84.49	41.99	13.68	18.35	91.61

Min. Ratio	0.67
Max. Ratio	1.40

Considering the unconditional means and mean square errors of Tables 9.7.2 and 9.7.3 we see that with $\rho = 0$, $E\{LO(K_e)\}$ is for all p and Δ less than $E\{LO(U_e)\} = E\{LO(T_e)\}$ and so K_e on average understates true log-odds. For $p \geqslant 8$ this understatement of true log-odds by $LO(K_e)$ is in excess of $LO(P_e)$'s. This is also reflected in Table 9.7.3 where for $\rho = 0$ and $p \geqslant 8$ the mean square errors of $LO(K_e)$ are less than those of $LO(P_e)$.

The unconditional means of $LO(K_e)$ in Model I display behaviour consistent with the classification behaviour of K_e noted in Table 9.7.1 . With $\rho = -.066$ and for all Δ, $LO(K_e)$ understates true log-odds increasingly so as p increases. As demonstrated in Appendix 9.C, where the asymptotic unconditional mean of $LO(K_e)$ is derived, for $p = 16$ and $\rho = -.066$ $E\{LO(K_e)\}$ is almost zero. However with $\rho > 0$ we note that the improving classification performmance of $LO(K_e)$ Table 9.7.1 is due to its increasing overstatement of true log-odds. This overstatement and understatement of true log-odds is reflected in the mean square errors of $LO(K_e)$ Table 9.7.3 . With $\rho = -.066$, $MSE\{LO(K_e)\}$ is usually less than $MSE\{LO(U_e)\}$ but greater than $MSE\{LO(P_e)\}$. With increasing $\rho > 0$ the mean square errors of $LO(K_e)$ increase and are larger in all cases than those of $LO(Ue)$ for $\rho \geqslant .3$.

For Model II we note Table 9.7.2 that $LO(K_e)$ understates on average true log-odds for all p, Δ and ρ and that for $\rho \geqslant .2$ this understatement is in all cases greater than that of $LO(P_e)$. We also note that $E\{LO(K_e)\}$ usually decreases with increasing p for all Δ and ρ. From the mean square errors of Model II, Table 9.7.3 we note that ρ must be as large as .8 for the mean square errors of $LO(K_e)$ to be greater in all cases than those of $LO(U_e)$ and ρ must be equal to .6 before they are larger in all cases than those of $LO(P_e)$.

For Model III similar comments as made for Model II apply. With $p = 16$ and $\rho = -.066$ and .8 we note the similar size of $E\{LO(K_e)\}$ and $MSE\{LO(K_e)\}$ as compared with their corresponding results for Model I ($\rho = -.066$) and Model II ($\rho = .8$). This was to be expected from the construction of Model III, Section 9.6.

These results on the mean and mean square error of $LO(K_e)$ with $\rho = 0$ confirm that for variables which are independent and for large dimension sizes relative to sample sizes the product kernel method is a superior estimator of true log-odds as compared with the parametric methods E_e, U_e and P_e. However the results of Model I show that with $\rho \geqslant .2$, $LO(K_e)$ becomes progressively poorer as an estimator of true log-odds while improving its classification performance. It must however be admitted that Model I and indeed Model II are somwhat extreme cases. From the behaviour of $LO(K_e)$ in Model III, especially its classification behaviour, we conclude that for n small, p large and multinormally distributed populations with equal covariance matrices the product kernel method is a good estimator of true log-odds provided the variables are moderately correlated.

APPENDIX 9.A.

DERIVATION OF THE INTEGRATED MEAN SQUARE ERROR WITH $f(x) = N_p(\mu,\Sigma)$ and $K(z) = N_p(0,hM)$

We assume that the true probability density function $f(x)$ is $N_p(\mu,\Sigma)$ and that the kernel function $K(z)$ is $N_p(0,hM)$, where $|hM| > 0$ and M is independent of the sample observations $\{x_i\}$, $1 \leqslant i \leqslant n$. We define the conditional mean square error, MSE(x) as

$$\mathrm{MSE}(x) = E[\{f(x) - \hat{f}(x)\}^2],$$

where $\hat{f}(x) = (2\pi)^{-\frac{p}{2}} |hM|^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} \exp\{-\frac{1}{2} \frac{1}{h} (x-x_i)' M^{-1}(x-x_i)\}$,

and the expectation is taken over the sample observations x_i. The integrated mean square error IMSE is then defined as

$$\mathrm{IMSE} = \int \mathrm{MSE}(x)\, dx. \qquad (9.A.1)$$

As

$$\mathrm{MSE}(x) = f^2(x) - 2\, f(x)\, E\{\hat{f}(x)\} + E\{\hat{f}^2(x)\}$$

we require the expectations $E\{\hat{f}(x)\}$ and $E\{\hat{f}^2(x)\}$.

Let $\Sigma = B\, B'$ and $y_i = B^{-1}(x-x_i)$, then $y_i \sim N_p(\nu,I)$ where $\nu = B^{-1}(x-\mu)$ and

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |hM|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\, y_i'\, B'\, (hM)^{-1}\, B\, y_i\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |hM|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\, y_i'\, A\, y_i\}$$

where $A = \frac{1}{h}\, B'\, M^{-1}\, B$.

Now as the x_i and so the y_i are independent identically distributed it follows that

$$E\{\hat{f}(x)\} = (2\pi)^{-\frac{p}{2}} \, |hM|^{-\frac{1}{2}} \; E[\exp\{-\tfrac{1}{2}\, y'\, Ay\}]$$

and $$\qquad (9.A.2)$$

$$E\{\hat{f}^2(x)\} = (2\pi)^{-p} \, |hM|^{-1} \, \frac{1}{n^2} \Big(nE[\exp\{-y'\, Ay\}] + n(n-1)\, E^2[\exp\{-\tfrac{1}{2}y'\, Ay\}]\Big)$$

where $y \sim N_p(\nu, I)$.

But $y'\, Ay$ is a non-central positive definite quadratic form with characteristic function

$$E[\exp\{\imath t\; y'\, Ay\}] = |I-2\imath t\; A|^{-\frac{1}{2}} \, \exp\{\imath t\; \nu'\, A(I-2\imath t\; A)^{-1}\, \nu\} \qquad (9.A.3)$$

and with $\imath t = -\frac{1}{2}$

$$E[\exp\, -\tfrac{1}{2}y'\, Ay\}] = |I+A|^{-\frac{1}{2}} \, \exp\{-\tfrac{1}{2}\, \nu'\, A(I+A)^{-1}\, \nu\}$$

$$= |I+\frac{1}{h}\, B'\, M^{-1}\, B|^{\frac{1}{2}} \, \exp\{-\tfrac{1}{2}\, \frac{1}{h}(x-\mu)'\, M^{-1}B(I+\frac{1}{h}\, B'\, M^{-1}B)^{-1}B'(x-\mu$$

$$= |I+\frac{1}{h}\, M^{-1}\, \Sigma|^{\frac{1}{2}} \, \exp\{-\tfrac{1}{2}(x-\mu)'\, (hM+\Sigma)^{-1}\, (x-\mu)\},$$

thus from (9.A.2)

$$E\{\hat{f}(x)\} = (2\pi)^{-\frac{p}{2}} \, |hM+\Sigma|^{-\frac{1}{2}} \, \exp\{-\tfrac{1}{2}(x-\mu)'\, (hM+\Sigma)^{-1}\, (x-\mu)\} \qquad (9.A.4)$$

$$= N_p(\mu, hM+\Sigma).$$

The expectation (9.A.4) was derived by Specht (1966) and Anderson (1969) using different approaches to that used here.

With ιt =-lin(9.A.3) we see that

$$E[\exp\{-y' Ay\}] = \left|I+\frac{2}{h} M^{-1} \Sigma\right|^{-\frac{1}{2}} \exp\{-(x-\mu)' (hM+2\Sigma)^{-1} (x-\mu)\}$$

and so from (9.A.2)

$$E\{\hat{f}^2(x)\} = (2\pi)^{-p} |hM|^{-\frac{1}{2}} |hM+2\Sigma|^{-\frac{1}{2}} \frac{1}{n} \exp\{-(x-\mu)' (hM+2\Sigma)^{-1}(x-\mu)\} \qquad (9.A.5)$$

$$+ (2\pi)^{-p} |hM+\Sigma|^{-1} \frac{n-1}{n} \exp\{-(x-\mu)' (hM+\Sigma)^{-1} (x-\mu)\}.$$

This expectation (9.A.5) was also studied by Anderson (1969) for the case M = I. His result however is incorrect as he writes $1/\sqrt{2}\ |\frac{1}{2}hM+\Sigma|^{\frac{1}{2}}$ for $1/|hM+2\Sigma|^{\frac{1}{2}}$.

For the integrated mean square error (9.A.1) we will require three integrals. These are easily derived from the integral of the Multivariate Normal density.

$$\text{Now} \int f^2(x)\, dx = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} 2^{-\frac{p}{2}} \int (2\pi)^{-\frac{p}{2}} 2^{\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x-\mu)' 2\Sigma^{-1}(x-\mu)\}\, dx$$

$$= \frac{1}{(4\pi)^{\frac{p}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}}, \qquad (9.A.6)$$

from (9.A.4)

$$\int f(x)\ E\{\hat{f}(x)\}dx = (2\pi)^{-\frac{p}{2}} \frac{|\Sigma|^{-\frac{1}{2}}|hM+\Sigma|^{-\frac{1}{2}}}{|\Sigma^{-1}+(hM+\Sigma)^{-1}|^{\frac{1}{2}}} \int (2\pi)^{-\frac{p}{2}} |\Sigma^{-1}+(hM+\Sigma)^{-1}|^{\frac{1}{2}} [$$

$$\exp\{-\tfrac{1}{2}(x-\mu)'(\Sigma^{-1}+(hM+\Sigma)^{-1})\ (x-\mu)\}]\, dx$$

$$= (2\pi)^{-\frac{p}{2}} \frac{1}{|hM+2\Sigma|^{\frac{1}{2}}} = \frac{1}{(4\pi)^{\frac{p}{2}}} \frac{2^{\frac{p}{2}}}{|hM+2\Sigma|^{\frac{1}{2}}} \qquad (9.A.7)$$

and from (9.A.5)

$$\int E\{\hat{f}^2(x)\}dx = (2\pi)^{-\frac{p}{2}} |hM|^{-\frac{1}{2}} \frac{1}{n} 2^{-\frac{p}{2}} [$$

$$\int (2\pi)^{-\frac{p}{2}} 2^{\frac{p}{2}} |hM+2\Sigma|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x-\mu)'\ 2(hM+2\Sigma)^{-1}\ (x-\mu)\}dx]$$

$$+\ (2\pi)^{-\frac{p}{2}} |hM+\Sigma|^{-\frac{1}{2}} \frac{n-1}{n}\ 2^{-\frac{p}{2}} [$$

$$\int (2\pi)^{-\frac{p}{2}} 2^{\frac{p}{2}} |hM+\Sigma|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x-\mu)'\ 2(hM+\Sigma)^{-1}\ (x-\mu)\}dx]$$

$$= \frac{1}{(4\pi)^{\frac{p}{2}}} \frac{1}{n} \frac{1}{|hM|^{\frac{1}{2}}} + \frac{1}{(4\pi)^{\frac{p}{2}}} \frac{n-1}{n} \frac{1}{|hM+\Sigma|^{\frac{1}{2}}} \ .$$

(9.A.8)

As $\quad \text{IMSE} = \int f^2(x)dx - 2\int f(x)E\{\hat{f}(x)\}dx + \int E\{\hat{f}^2(x)\}dx$

it follows from (9.A.6), (9.A.7) and (9.A.8) that

$$\text{IMSE} = \frac{1}{(4\pi)^{\frac{p}{2}}} \left[\frac{1}{|\Sigma|^{\frac{1}{2}}} - \frac{2^{\frac{p}{2}+1}}{|hM+2\Sigma|^{\frac{1}{2}}} + \frac{1}{n} \frac{1}{|hM|^{\frac{1}{2}}} + \frac{n-1}{n} \frac{1}{|hM+\Sigma|^{\frac{1}{2}}} \right] .$$

(9.A.9)

The latter result coincides with the result for $f(x) = N_p(0,I)$ and $K(z) = N_p(0,hI)$ given by Epanechnikov (1969).

The formula for the expected mean square error i.e.

$$E_X\{MSE(x)\} = \int MSE(x)\; f(x)dx$$

iseasilyderived from the above results and is given by

$$E_X\{MSE(x)\} = \frac{1}{(2\pi)^p}\left[\frac{(\frac{1}{3})^{\frac{p}{2}}}{|\Sigma|} - \frac{2}{|2hM+3\Sigma|^{\frac{1}{2}}\;|\Sigma|^{\frac{1}{2}}} + \frac{1}{n}\;\frac{1}{|hM|^{\frac{1}{2}}\;|hM+4\Sigma|^{\frac{1}{2}}} + \frac{n-1}{n}\;\frac{1}{|hM+\Sigma|^{\frac{1}{2}}\;|hM+3\Sigma|^{\frac{1}{2}}}\right].$$

Specht (1971) derived $E_X\{MSE(x)\}$ when $p = 1$ and $f(x) = N_1(0,1)$, it coincides with the above result in this instance.

APPENDIX 9.B

CHOICE OF h TO MINIMISE THE IMSE

In the models of Section 9.6 we let $M = D^{-1}$ where D is the diagonal matrix λ_j $1 \leqslant j \leqslant p$ of eigen values of Σ, with $\lambda_1 = 1+(p-1)\rho$, $\lambda_2 = \lambda_3 = \text{-----} = \lambda_p = 1-\rho$ and now $\Sigma = I$. From (9.A.9) the IMSE for these models is

$$\text{IMSE} = \frac{1}{(4\pi)^{\frac{p}{2}}}\left[1 - \frac{2^{\frac{p}{2}+1}}{|hD^{-1}+2I|^{\frac{1}{2}}} + \frac{1}{n}\frac{1}{|hD^{-1}|^{\frac{1}{2}}} + \frac{n-1}{n}\frac{1}{|hD^{-1}+I|^{\frac{1}{2}}}\right]$$

$$= \frac{|D|^{\frac{1}{2}}}{(4\pi)^{\frac{p}{2}}}\left[\frac{1}{|D|^{\frac{1}{2}}} - \frac{2^{\frac{p}{2}+1}}{|hI+2D|^{\frac{1}{2}}} + \frac{1}{n}\frac{1}{|hI|^{\frac{1}{2}}} + \frac{n-1}{n}\frac{1}{|hI+D|^{\frac{1}{2}}}\right]$$

$= g(h)$ say.

We note that the IMSE g(h), is independent of the mean of the true populations, hence the value of h, for given n, p and ρ, that minimises g(h) will be the same for all three models. This was to be expected as h is a smoothing parameter and all three models require the same amount of smoothing as they differ only in their mean vectors. We also note as $|D| = |\Sigma|$, $|hI+2D| = |hI+2\Sigma|$ and $|hI+D| = |hI+\Sigma|$ that $g(h) = |\Sigma|^{\frac{1}{2}}x\text{IMSE}$ where $f_t(x) = N_p(\mu_t,\Sigma)$ and $K_t(z) = N_p(0,hI)$, thus the values of h that minimise g(h) also minimise the IMSE for the untransformed models where $\Sigma = R$.

We now require for given n, p and ρ that value of $h > 0$ which minimises $g(h)$, this is equivalent to finding $h > 0$ so that $g'(h) = 0$ and $g''(h) > 0$. Now

$$g'(h) = \frac{|D|^{\frac{1}{2}}}{(4\pi)^{\frac{p}{2}}}\left[\frac{p}{2}\,\frac{2^{\frac{p}{2}+1}\{h+2+(p-2)2\rho\}}{\{h+2+(p-1)2\rho\}^{\frac{3}{2}}\,(h+2-2\rho)^{\frac{p-1}{2}+1}} - \frac{p}{2}\,\frac{1}{n}\,\frac{1}{h^{\frac{p}{2}+1}} - \frac{p}{2}\,\frac{n-1}{n}\left[\frac{\{h+1+(p-2)\rho\}}{\{h+1+(p-1)\rho\}^{\frac{3}{2}}\,(h+1-\rho)^{\frac{p-1}{2}+1}}\right]\right]$$

$$= \frac{|D|^{\frac{1}{2}}}{(4\pi)^{\frac{p}{2}}}\left[\frac{p}{2}\,2^{\frac{p}{2}+1}\,\alpha - \frac{p}{2}\,\frac{1}{n}\,\frac{1}{h^{\frac{p}{2}+1}} - \frac{p}{2}\,\frac{n-1}{n}\,\beta\right]$$

and $g'(h) = 0$ when

$$u(h) = -\,n\,2^{\frac{p}{2}+1} + \frac{1}{\alpha h^{\frac{p}{2}+1}} + (n-1)\,\frac{\beta}{\alpha} = 0.$$

It is easily shown that $\frac{1}{h^{\frac{p}{2}+1}} > \beta > 0$ and as $\alpha > 0$ it follows that

$$u(h) < -\,n\,2^{\frac{p}{2}+1} + \frac{n}{\alpha h^{\frac{p}{2}+1}}$$

$$= -\,n\,2^{\frac{p}{2}+1} + n\,\frac{\{h+2+(p-1)2\rho\}^{\frac{3}{2}}\,(h+2-2\rho)^{\frac{p-1}{2}+1}}{\{h+2+(p-2)2\rho\}\;h^{\frac{p}{2}+1}}$$

$= w(h)$ say.

Now if $\rho \geqslant 0$

$$\begin{aligned} h+2+(p-2)2\rho &\geqslant h+2-2\rho > 0 \\ h+2+(p-1)2\rho &\geqslant h+2-2\rho > 0 \end{aligned}$$

hence

$$w(h) \leqslant -n\,2^{\frac{p}{2}+1} + n\,\frac{\{h+2+(p-1)2\rho\}^{\frac{p}{2}+1}}{h^{\frac{p}{2}+1}}$$

$$= 0 \text{ when } h = 2+(p-1)2\rho = 2\lambda_1.$$

If $\rho < 0$ then

$$(p-1)2\rho < (p-2)2\rho$$

and

$$\begin{aligned} 0 < h+2+(p-1)2\rho &< h+2+(p-2)2\rho \\ 0 < h+2+(p-1)2\rho &< h+2-2\rho \end{aligned}$$

hence

$$w(h) < -n\,2^{\frac{p}{2}+1} + n\,\frac{(h+2-2\rho)^{\frac{p}{2}+1}}{h^{\frac{p}{2}+1}}$$

$$= 0 \text{ when } h = 2-2\rho = 2\lambda_2.$$

Thus the value of $h > 0$ that minimises $g(h)$ must be in the interval

$$\begin{array}{llll} (0,2\lambda_1) & \text{for } \rho > 0 & \text{where } \lambda_1 = 1+(p-1)\rho & \\ (0,2] & \text{for } \rho = 0 & \text{where } \lambda_1 = \lambda_2 = 1 & (9.B.1) \\ (0,2\lambda_2) & \text{for } \rho < 0 & \text{where } \lambda_2 = 1-\rho \text{ and } \lambda_j=\lambda_2, & \\ & & \qquad\qquad 2 \leqslant j \leqslant p. & \end{array}$$

Using the search intervals (9.B.1) and the IBM SSP routine RTMI (1970) i.e. a regula falsi approach, those values of $h > 0$ which minimise the IMSE $g(h)$ for given n, p and ρ were obtained. The complete search interval was scanned in each case to ensure that the global minimum was found. The resulting values of h are listed in Table 9.B.1.

Table 9.B.1

Values of the Smoothing Parameter h which minimise the IMSE when $f(x) \doteq N_p(\nu,I)$ and $K(z) = N_p(0,hD^{-1})$, based on a sample of size n = 12.

p \ ρ	-.066	0	.2	.3	.4	.5	.6	.8
1	.526	.526	.526	.526	.526	.526	.526	.526
4	.722	.727	.689	.648	.595	.532	.459	.281
8	.927	.943	.871	.805	.725	.636	.536	.309
16	1.180	1.234	1.112	1.011	.897	.773	.641	.351

We see from Table 9.B.1, that with fixed p, as ρ increases i.e. as the shape of the kernel alters, the amount of smoothing required decreases. We also noted in obtaining these values of h that the corresponding values of g(h) = IMSE were for fixed p almost constant with increasing ρ. This coincides with the result of Anderson (1969) and Epanechnikov (1969), that the shape of the kernel is not critical provided the appropriate smoothing value is used.

APPENDIX 9.C

AN ASYMPTOTIC EXPANSION FOR THE UNCONDITIONAL MEAN OF $LO(K_e)$

Now $f_t(x) = N_p(\mu_t, \Sigma)$ $t = 1$ and 2, $n_1 = n_2 = n$ and

$$LO(K_e) = \ln\{\frac{\hat{f}_1(x)}{\hat{f}_2(x)}\}$$

where $\hat{f}_t(x) = \frac{1}{n} \sum_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |hM|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \frac{1}{h} (x-x_{it})' M^{-1} (x-x_{it})\}$.

By a Taylor series expansion and taking expectations with respect to the sample values x_{it} we have the asymptotic conditional expectation

$$E\{LO(K_e)|x\} = E\left[\ln\{\frac{\hat{f}_1(x)}{\hat{f}_2(x)}\}\middle| x\right]$$

$$\simeq \ln\left[\frac{E\{\hat{f}_1(x)\}}{E\{\hat{f}_2(x)\}}\right] + \left[-\frac{1}{2}\frac{V\{\hat{f}_1(x)\}}{E^2\{\hat{f}_1(x)\}} + \frac{1}{2}\frac{V\{\hat{f}_2(x)\}}{E^2\{\hat{f}_2(x)\}}\right]$$

$$= T_1 + T_2$$

As $V\{\hat{f}_t(x)\} = E\{\hat{f}_t^2(x)\} - E^2\{\hat{f}_t(x)\}$

and the results of Appendix 9.A where it is shown that

$$E\{\hat{f}_t(x)\} = N_p(\mu_t, hM+\Sigma)$$

$$E\{\hat{f}_t^2(x)\} = \frac{1}{n}(2\pi)^{-p} |hM|^{-\frac{1}{2}} |hM+2\Sigma|^{-\frac{1}{2}} \exp\{-(x-\mu_t)'(hM+2\Sigma)^{-1}(x-\mu_t)\} + \frac{n-1}{n} E^2\{\hat{f}_t(x)\}$$

it follows that

$$T_1 = (\mu_1-\mu_2)' (hM+\Sigma)^{-1} \{x-\frac{1}{2}(\mu_1+\mu_2)\}$$

and

$$T_2 = -\tfrac{1}{2}\,\frac{1}{n}\,\frac{|hM+\Sigma|}{|hM|^{\frac{1}{2}}\;|hM+2\Sigma|^{\frac{1}{2}}}\left[\exp\{(x-\mu_1)'C(x-\mu_1)\} - \exp\{(x-\mu_2)'C(x-\mu_2)\}\right] \tag{9.C.1}$$

where $C = (hM+\Sigma)^{-1} - (hM+2\Sigma)^{-1}$

$$= (hM+\Sigma)^{-1}\;\Sigma\;(hM+2\Sigma)^{-1}. \tag{9.C.2}$$

The asymptotic unconditional expectation of $LO(K_e)$, x in Π_1 is given by

$$E\{LO(K_e)\} \simeq E_x(T_1) + E_x(T_2)$$

where (9.C.3)

$$E_x(T_1) = \tfrac{1}{2}(\mu_1-\mu_2)'\;(hM+\Sigma)^{-1}\;(\mu_1-\mu_2).$$

For $E_x(T_2)$ we require from (9.C.1)

$$E_x[\;\exp\{(x-\mu_t)'\;C(x-\mu_t)\}] \qquad t = 1 \text{ and } 2 \tag{9.C.4}$$

where $x \sim N_p(\mu_1,\Sigma)$. These expectations (9.C.4) may be obtained as in Appendix 9.A. As before we let $y_t = B^{-1}(x-\mu_t)$ where $\Sigma = B\,B'$, then $y_1 \sim N_p(0,I)$ and $y_2 \sim N_p(\nu,I)$ where $\nu = B^{-1}(\mu_2-\mu_1)$. From (9.A.3) for $t = 1$

$$\begin{aligned} E_x[\exp\{(x-\mu_1)'\;C(x-\mu_1)\}] &= E_{y_1}[\exp\{y_1\;H\;y_1\}] \text{ where } H = B'CB \\ &= |I-2H|^{-\frac{1}{2}} = |I-2BB'C|^{-\frac{1}{2}} \\ &= |I-2\Sigma C|^{-\frac{1}{2}} = |C|^{-\frac{1}{2}}|C^{-1}-2\Sigma|^{-\frac{1}{2}}, \end{aligned} \tag{9.C.5}$$

for $t = 2$

$$\begin{aligned} E_x[\exp\{(x-\mu_2)'\;C(x-\mu_2)\}] &= E_{y_2}[\exp\{y_2\;H\;y_2\}] \\ &= |I-2H|^{-\frac{1}{2}}\exp\{\nu'H(I-2H)^{-1}\;\nu\} \\ &= |C|^{-\frac{1}{2}}|C^{-1}-2\Sigma|^{-\frac{1}{2}}\exp\{(\mu_1-\mu_2)'(C^{-1}-2\Sigma)^{-1}(\mu_1-\mu_2)\}. \end{aligned} \tag{9.C.6}$$

From (9.C.2)

$$C^{-1} = (hM+2\Sigma)\ \Sigma^{-1}(hM+\Sigma)$$

giving $$C^{-1}-2\Sigma = h^2\ M\ \Sigma^{-1}\ M + 3hM$$

hence

$$\frac{|hM+\Sigma|}{|hM|^{\frac{1}{2}}\ |hM+2\Sigma|^{\frac{1}{2}}}\ \frac{|C|^{-\frac{1}{2}}}{|C^{-1}-2\Sigma|^{\frac{1}{2}}} = \frac{|hM+\Sigma|}{|hM|^{\frac{1}{2}}\ |hM+2\Sigma|^{\frac{1}{2}}}\ \frac{|hM+2\Sigma|^{\frac{1}{2}}\ |\Sigma|^{-\frac{1}{2}}\ |hM+\Sigma|^{\frac{1}{2}}}{|hM+3\Sigma|^{\frac{1}{2}}\ |\Sigma|^{-\frac{1}{2}}\ |hM|^{\frac{1}{2}}}$$

$$= \frac{|hM+\Sigma|^{\frac{3}{2}}}{|hM|\ |hM+3\Sigma|^{\frac{1}{2}}}\ . \qquad (9.C.7)$$

Combining the results (9.C.5), (9.C.6), (9.C.7) and (9.C.1) gives

$$E_x(T_2) = -\tfrac{1}{2}\ \frac{1}{n}\ \frac{|hM+\Sigma|^{\frac{3}{2}}}{|hM|\ |hM+3\Sigma|^{\frac{1}{2}}}\ [1 - \exp\{(\mu_1-\mu_2)'(h^2M\Sigma^{-1}M+3hM)^{-1}(\mu_1-\mu_2)\}]. \qquad (9.C.8)$$

Thus from (9.C.3) and (9.C.8)

$$E\{LO(K_e)\} \simeq \tfrac{1}{2}(\mu_1-\mu_2)'\ (hM+\Sigma)^{-1}\ (\mu_1-\mu_2)$$
$$+ \tfrac{1}{2}\ \frac{1}{n}\ \frac{|hM+\Sigma|^{\frac{3}{2}}}{|hM|\ |hM+3\Sigma|^{\frac{1}{2}}}\ [\exp\{(\mu_1-\mu_2)'\ P^{-1}(\mu_1-\mu_2)\} - 1] \qquad (9.C.9)$$

where $P=h^2\ M\ \Sigma^{-1}M + 3hM$.

In the models of Section 9.6 $\mu_1 = 0$, $\mu_2 = \nu$, $\Sigma = I$ and $M = D^{-1}$ where D is the diagonal matrix of eigen values λ_j, $1 \leq j \leq p$, of R. For these models, from (9.C.9)

$$E\{LO(K_e)\} \simeq \tfrac{1}{2}\ \nu'\ (hD^{-1}+I)^{-1}\ \nu$$
$$+ \tfrac{1}{2}\ \frac{1}{n}\ \frac{|hD^{-1}+I|^{\frac{3}{2}}}{|hD^{-1}|\ |hD^{-1}+3I|^{\frac{1}{2}}}\ [\exp\{\nu'P^{-1}\nu\} - 1]$$

where $P = h^2\ D^{-2} + 3h\ D^{-1}$.

Now

$$(hD^{-1}+I)^{-1} = \begin{pmatrix} \frac{\lambda_1}{h+\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\lambda_p}{h+\lambda_p} \end{pmatrix} \text{ and } P^{-1} = \begin{pmatrix} \frac{\lambda_1^2}{h^2+3h\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\lambda_p^2}{h^2+3h\lambda_p} \end{pmatrix}$$

and so for Model I, (9.6.6), as $\nu = (\Delta,0,0,-----,0)'$

$$E\{LO(K_e)\} \approx \tfrac{1}{2}\Delta^2 \frac{\lambda_1}{h+\lambda_1} + \tfrac{1}{2}\,\frac{1}{n}\,\frac{|hI+D|^{\frac{3}{2}}}{|hI|\;|hI+3D|^{\frac{1}{2}}}\left[\exp\{\Delta^2 \frac{\lambda_1^2}{h^2+3h\lambda_1}\} - 1\right]$$

(9.C.10).

As $\lambda_1 = 1+(p-1)\rho \rightarrow 0$ i.e. $\rho \rightarrow \frac{-1}{1-p}$, $E\{LO(K_e)\} \rightarrow 0$ as anticipated. However as $\rho \rightarrow 1$, $\lambda_1 \rightarrow p$ and the second term in the expansion of $E\{LO(K_e)\}$ (9.C.10) becomes extremely large, rendering the expansion worthless. Unfortunately the size of the second term is still quite large for $\rho = 0$ and so the asymptotic expansion does not provide reliable information except in near singular cases.

CHAPTER 10

POSSIBLE EXTENSIONS

We conclude by considering possible extensions and development of the work undertaken in the previous chapters.

We have seen in Chapters 7 and 8 that allowing for the bias in estimating Σ^{-1} by S^{-1} results in a considerable improvement in estimating log-odds. However the biased predictive method was superior. The use of alternative biased estimators of Σ^{-1} and their consequences for estimation of log-odds have yet to be investigated. There have been some recent studies of biased allocation rules; Di Pillo (1976, 1977, 1979) introduced a ridge adjustment in the estimative allocation rule i.e. he replaced S^{-1} by $(S + kI)^{-1}$, $k > 0$. His simulation study with n small indicates considerable classification improvement when the population covariance matrices are poorly conditioned. Similar investigations have been undertaken by Smidt and Mc Donald (1976a,b). Given the arithmetic connection between multiple regression and the estimative allocation rule Anderson (1958,p140), the improvement in estimating the regression coefficients with inclusion of a ridge adjustment, Hoerl and Kennard (1970), might be reflected in a better estimate of true log-odds especially for highly correlated population covariance matrices. Efron and Morris (1976) have shown that

$$\hat{\Sigma}^{-1} = \frac{n_1+n_2-p-3}{n_1+n_2-2} S^{-1} + \frac{p^2+p-2}{(n_1+n_2-2)\ \mathrm{tr}\ S} I ,$$

is for $p \geqslant 2$ a uniformly better estimator of Σ^{-1} than the unbiased estimator

$$\frac{n_1+n_2-p-3}{n_1+n_2-2} S^{-1}$$

for a particular loss function. With $n_1 = n_2$ use of $\hat{\Sigma}^{-1}$ would result in overstatement of true log-odds. The implications of this bias for mean square error and classification are not known. Some analytic mean square error results may be possible here. Haff (1979) has extended the work of Efron and Morris and proposed modifications to $\hat{\Sigma}^{-1}$ which are optimum for alternative loss functions.

The logistic approach to estimating log-odds is one important area not investigated in this thesis. The logistic formulation is valid for a range of distribution assumptions and its economy of parameters make it an attractive method. With normality assumptions however the logistic method is unlikely to perform as well as some of the parametric methods considered here. Some support for this view is given by the result of Efron (1975) who derived the asymptotic relative classification efficiency of the logistic to the estimative allocation rule when the populations were multinormally distributed, these relative efficiencies were quite low for reasonable population separations Δ. However Mc Lachlan and Byth (1979) showed that the asymptotic expected actual probabilities of misclassification of the logistic and estimative allocation rules are comparable for reasonable Δ and n moderately large. It is likely however that as an estimator of log-odds the logistic method has a substantial bias. Expressions in Mc Lachlan and Byth will facilitate calculation of the asymptotic bias and mean square error of the logistic method when the populations are normally distributed. The presence of a substantial bias in the logistic method is also supported by the study of Anderson and Richardson (1979) who derived in general the bias corrections when estimating the logistic parameters. Their simulation study, on univariate normal distribution with small sample sizes, indicated that the logistic parameters adjusted for bias were closer to the true parameters than the unadjusted logistic parameters·and that this bias reduction was not accompanied by increased variability of the estimators.

Turning now to the estimation of log-odds by the kernel method, the obvious extension to the study of Chapter 9 is to consider the behaviour of product kernels in the presence of correlated variables when the covariance matrices are unequal. For allocation, the case of zero correlation and proportional covariance matrices has been considered by Van Ness (1979). Use of the full sample covariance matrices in the kernel functions also needs to be considered. Preliminary investigations of this when $\Sigma_1 = \Sigma_2$ indicate that the kernel method is now a poor estimator of log-odds. It is also inferior for classification purposes unless the dimension size p is large relative to the sample size n, when it is comparable to the parametric methods. Recently the behaviour of the product kernel method when the populations are continuous but non-normal has been studied by Koffler and Penfield (1979) and Remme, Habbema and Hermans (1980). In the latter reference it is noted that on the basis of previous work, Habbema, Hermans and Remme (1978), the product kernel with fixed smoothing parameter may have difficulties when the population distributions are skewed. The variable kernel modification of Brieman, Meisel and Purcell (1977) addresses this problem.

REFERENCES

Abdel-Aty, S.H. (1954). Approximate formulae for the percentage points and the probability integral of the non-central χ^2-distribution. Biometrika 41, 538-540.

Abramowitz, M and Stegun, I.A. (Eds) (1965). Handbook of Mathematical Functions. Dover Publications.

Aitchison, J. (1975). Goodness of prediction fit. Biometrika 62, 547-554.

Aitchison, J. and Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. Biometrika 63, 413-420.

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis. Cambridge Univ. Press.

Aitchison, J., Habbema, J.D.F. and Kay, J.W. (1977). A critical comparison of two methods of statistical discrimination. Appl. Statist., 26, 15-25.

Aitchison, J. and Kay, J.W. (1973). Principles, practice and performance in decision problems of categorisation. Paper presented at the NATO conference on the role and the effectiveness of decision theories in practice. Luxembourg.

Anderson, G.D. (1969). A comparison of the methods for estimating a probability density. Ph.D. Diss. Univ. of Washington.

Anderson, J.A. (1972). Separate sample logistic discrimination. Biometrika 59, 19-35.

Anderson, J.A. (1975). Quadratic logistic discrimination. Biometrika 62, 149-154.

Anderson, J.A. and Richardson, S.C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. Technometrics 21, 71-78.

Anderson, T.W. (1951). Classification by multivariate analysis. Psychometrika 16, 31-50.

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley.

Anderson, T.W. (1973). An asymptotic expansion of the distribution of the Studentized classification statistic W. Ann. Statist. 1, 964-972.

Anderson, T.W. and Bahadur, R.R. (1962). Classification into two multivariate Normal distributions with different covariance matrices. Ann. Math. Statist. 33, 420-431.

Bartlett, M.S. and Please, N.W. (1963). Discrimination in the case of zero-mean differences. Biometrika 50, 17-21.

Bowker, A.H. (1961). A representation of Hotelling's T^2 and Anderson's classification statistic in terms of simple statistics. Studies in Item Analysis and Prediction. Stanford Univ. Press. 285-292.

Bowker, A.H. and Sitgreaves, R. (1961). An asymptotic expansion for the distribution of the W-classification statistic. Studies in Item Analysis and Prediction. Stanford Univ. Press. 293-310.

Box, G.E.P. and Muller, M.E. (1958). A note on the generation of random normal deviates. Ann. Math. Statist. 29, 610-611.

Brieman, L., Meisel, W. and Purcell, E. (1977). Variable kernel estimates of multivariate densities and their calibration. Technometrics 19, 135-144.

Broffitt, J.D. and Williams, J.S. (1973). Minimum variance estimators for misclassification probabilities in discriminant analysis. J. Mult. Analy. 3, 311-327.

Bryan, J.K. (1971). Classification and clustering using density estimation. Ph.D. Diss. Univ. of Missouri.

Cacoullos, T. (1966). Estimation of a multivariate density. Ann. Inst. Statist. Math. 18, 179-189.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesteral and syptolic blood pressure: a discriminant function analysis.
Proc. Fed. Amer. Soc. Exp. Biol. 21(2), 58-61.
*Chen (1971) see below
Chou, C.and Siotani, M. (1974). Asymptotic expansion of the non-null distribution of the ratio of two conditionally independent Hotelling's T_0^2-statistic.
Ann. Inst. Statist. Math. 26, 277-288.

Cover, T.M. (1972). A hierarchy of probability density function estimates. In, Frontiers of Pattern Recognition. 83-98 Ed., Wantanabe, S. Academic Press.

Cox, D.R. (1966). Some procedures associated with the logistic qualitative response curve. Research Papers in Statistics: Fetschrift for J. Neyman. 55-71.Wiley.

Das Gupta, S. (1965). Optimum classification rules for classification into two multivariate normal populations. Ann. Math. Statist. .36, 1174-1184.

Das Gupta, S. (1968). Some aspects of discrimination function coefficients. Sankhya A. 30, 387-400.

*Chen, E.H. (1971). A random normal number generator for 32-bit-word computers. J. Am. Statist. Assoc. 66, 400-403.

Das Gupta, S. (1974). Probability inequalities and errors in classification. Ann. Statist. 2, 751-762.

Day, N.E. and Kerridge, D.F. (1967). A general maximum likelihood discriminant. Biometrics 23, 313-323.

Deheuvels, P. (1977). Estimation non parametrique de la densite par histogrammes generalises II.
Pub. Inst. Stat. Univ. Paris 22, 1-23.

Desu, M.M. and Geisser, S. (1973). Methods and applications of equal-mean discrimination. Discriminant Analysis and Applications 139-159. Ed. T. Cacoullos. Academic Press.

Di Pillo, P.J. (1976). The application of bias to discriminant analysis. Comm. Statist. - Theor. Meth. A5, 843-854.

Di Pillo, P.J. (1977). Further applications of bias to discriminant analysis. Comm. Statist. - Theor. Meth. A6, 933-943.

Di Pillo, P.J. (1979). Biased discriminant analysis: Evaluation of the optimum probability of misclassification.
Comm. Statist. - Theor. Meth. A8, 1447-1457.

Edwards, A.W.F. (1972). Likelihood. Cambridge Univ. Press.

Efron, B. (1975). The efficiency of logistic regression compared to Normal discriminant analysis. J. Am. Statist. Assoc. 70, 892-898.

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. Ann. Statist. 4, 22-32.

Ellison, B.E. (1962). A classification problem in which information about alternative distributions is based on samples.
Ann. Math. Statist. 33, 213-223.

Enis, P. and Geisser, S. (1970). Sample discriminants which minimize posterior squared error loss. S. Afr. Statist. J. 4, 85-93.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory. Prob. Appl. U.S.S.R. 14, 153-158.

Fisher, R.A. (1936). Use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179-188.

Fix, E and Hodges, J.L. (1951). Discriminatory analysis, non-parametric discrimination: consistency properties. Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Brooks Air Force Base, Texas.

Fryer, M.J. (1976). Some errors associated with non-parametric estimation of density functions. J. Inst. Math. Applics.18, 371-380.

Fryer, M.J. (1977). A review of some non-parametric methods of density estimation. J. Inst. Math. Applics. 20, 335-354.
*Fukunaga and Kessell (1971) see below.
Geisser, S. (1964). Posterior odds for multivariate normal classification. J.R..Statist. Soc. B, 26, 69-76.

Geisser, S. (1967). Estimation associated with linear discriminants. Ann. Math. Statistc. 38, 807-817.

Gessaman, M.P. and Gessaman, P.H. (1972). A comparison of some multivariate discrimination procedures. J. Am. Statist. Assoc. 67, 468-472.

Ghosh, B.K. (1973). Some monotonicity theorems for χ^2, F and t distributions with applications. J.R. Statist. Soc. B. 35, 480-492.

*Fukunaga, K. and Kessell, D.L. (1971). Estimation of classification error. IEEE. Trans. Comput. C-20, 1521-1527.

Ghurye, S.G. and Olkin, I. (1969). Unbiased estimation of some multivariate probability densities and related functions. Ann. Math. Statist. 40, 1261-1271.

Gilbert, E.S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. Biometrics 25, 505-516.

Glick, N. (1972). Sample-based classification procedures derived from density estimators. J. Am. Statist. Assoc. 67, 116-122.

Habbema, J.D.F., Hermans, J. and Remme, J. (1978). Variable kernel density estimation in discriminant analysis. Compstat 1978, 178-185, Eds. L. Corsten and J. Hermans, Physica-Verlag, Vienna.

Haff, L.R. (1979). Estimation of the inverse covariance matrix; random mixtures of the inverse Wishart matrix and the identity. Ann.of Statist. 7, 1264-1276.

Han, C.P. (1968). A note on discrimination in the case of unequal covariance matrices. Biometrika 55, 586-587.

Han, C.P. (1969). Distribtuion of discriminant function when covariance matrices are proportional. Ann. Math. Statist. 40, 979-985.

Han, C.P. (1970). Distribution of discriminant function in circular models. Ann. Inst. Stat. Math. 22, 117-125.

Han, C.P. (1974). Asymptotic distribution of discriminant function when covariance matrices are proportional and unknown. Ann. Inst. Stat. Math. 26, 127-133.

Han, C.P. (1975). Some relationships between non-central chi-squared and normal distributions. Biometrika 62, 213-214.

Han, C.P. (1978). On the computation of non-central chi-squared distributions. J. Statist. Comput. Simul. 6, 207-210.

Hermans, J. and Habbema, J.D.F. (1975). Comparisons of five methods to estimate posterior probabilities.
E.D.V. in Med. und. Biol. 6, 14-19.

Hildebrandt, B. Michaelis, J. and Koller, S. (1973).
Die haufigkreit der fehlklassifikation bei der quadratischen diskriminonzanalyse. Biom. Z. 15, 3-12.

Hills, M. (1966). Allocation rules and their error rates.
J.R. Statist. Soc. B. 28, 1-32.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics 12, 55-67.

I.B.M. (1970). System/360 Scientific Subroutines Package.
Version III. Programmer's Manual. 5th Ed. I.B.M.

Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. Biometrika 48, 419-429.

Jeffreys, H. (1961). Theory of Probability. 3rd edn.,
Oxford Univ. Press.

Jensen, D.R. and Solomon, H. (1972). A Gaussian approximation to the distribution of a definite quadratic form.
J. Am. Statist. Assoc. 67, 898-902.

John, S. (1960a). On some classification problems I.
Sankhya, 22, 301-308.

John, S. (1960b). On some classification statistics.
Sankhya, 23, 309-316.

John, S. (1961). Errors in discrimination.
Ann. Math. Statist. 32, 1125-1144.

John, S. (1963). On classification by the statistics R˙ and Z. Ann. Inst. Statist. Math. 14, 237-246.

Johnson, N.L. and Kotz, S. (1970). Continuous Univariate Distributions -I and II. Wiley.

Kendall, M and Stuart, A. (1976). The Advanced Theory of Statistics Vol. 3. Design and Analysis, and Time Series. 3rd Edn, Griffen.

Koffler, S. and Penfield, D. (1979). Nonparametric discrimination procedures for non-normal distributions. J. Statist. Comput. Simul. 8, 281-299.

Kullback, S. and Liebler, R.A. (1951). On information and sufficiency. Ann. Math. Statist. 22, 79-86.

Lachenbruch, P.A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics 23, 639-645.

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, 828-834.

Lachenbruch, P.A. and Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. Technometrics 10, 1-11.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. Ann. Math. Statist. 36, 1049-1051.

Marks, S andDunn, O.J. (1974). Discriminant functions when covariance matrices are unequal. J. Am. Statist. Assoc. 69, 555-559.

Memon, A.Z. (1970). Distribution of the classification statistic Z when covariance matrix is known. Punjab. Univ. J. Math. 3, 59-67.

Memon, A.Z. and Okamoto, M. (1971). Asymptotic expansion of the distribution of the Z statistic in discriminant analysis. J. Mult. Analysis 1, 294-307.

Michaelis, J. (1973). Simulation experiments with multiple group linear and quadratic discriminant analysis. Discriminant Analysis and Applications, 225-238. Ed. T. Cacoullos. Academic Press.

Moran, M.A. (1974). The performance of the linear discriminant function with and without selection of variables. Ph.D. Thesis. Univ. of Reading.

Mosteller, F and Wallace, D.L. (1963). Inference in an authorship problem. J. Am. Statist. Assoc. 58, 275-309.

Mosteller, F and Wallace, D.L. (1964). Inference and Disputed Authorship : The Federalist. Addison-Wesley.

Murray, G.D. (1977). A note on the estimation of probability density functions. Biometrika 64, 150-152.

Murray, G.D. (1979). The estimation of multivariate Normal density functions using incomplete data. Biometrika 66, 375-380.

Mc Lachlan, G.J. (1972). An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function. Austral. J. Statist. 14, 68-72.

Mc Lachlan, G.J. (1973). An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. Austral. J. Statist. 16, 210-214.

Mc Lachlan, G.J. (1974). The asymptotic distributions of the conditional error rate and risk in discriminant analysis. Biometrika 61, 131-135.

Mc Lachlan, G.J. (1975). Some expected values for the error rates of the sample quadratic discriminant function. Austral. J. Statist. 17, 161-165.

Mc Lachlan, G.J. (1976). The bias of the apparent error rate in discriminant analysis. Biometrika 63, 239-244.

Mc Lachlan, G.J. (1977). The bias of sample based posterior probabilities. Biometrical. J. 6, 421-426.

Mc Lachlan, G.J. (1979). A comparison of the estimative and predictive methods of estimating posterior probabilities. Comm. Statist. - Theor. Meth. A8, 919-929.

Mc Lachlan, G.J. and Byth, K. (1979). Expected error rates for logistic regression versus normal discriminant analysis. Biometrical J. 21, 47-56.

Newman, T.G. and Odell, P.L. (1971). The Generation of Random Variates. Griffin's Statistical Monographs and Courses, No. 29.

Okamoto, M. (1961). Discrimination for variance matrices. Osaka Math. J. 13, 1-39.

Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. Ann. Math. Statist. 34, 1286-1301.

O'Neill, T. (1980). The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. J. Am. Statist. Assoc. 75, 154-160.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist. 33, 1065-1073.

Patnaik, P.B. (1949). The noncentral chi-square and F-distributions and their approximations. Biometrika 36, 202-232.

Pearson, E.S. (1959). Note on an approximation to the distribution of a non-central χ^2. Biometrika 46, 364.

Pearson, K. (1934). Tables of the Incomplete Beta Function. Cambridge Univ. Press.

Penrose, L.S. (1947). Some notes on discrimination. Ann. Eugen. 13, 228-237.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. Biometrika 66, 403-411.

Press, S.J. (1966). Linear combinations of non-central chi-square variates. Ann. Math. Statist. 37, 480-487.

Press, S.J. (1972). Applied Multivariate Analysis. Holt, Rinehart and Winston.

Price, R. (1964). Some non-central F distributions expressed in closed form. Biometrika 51, 107-122.

Ralston, A. (1965). A First Course in Numerical Analysis. Mc Graw-Hill.

Rao, C.R. (1954). A general theory of discrimination when the information about alternative population distributions is based on samples. Ann. Math. Statist. 25, 651-670.

Rao, C.R. (1973). Linear Statistical Inference and its Applications. 2nd edn. Wiley.

Remme, J. Habbema, J.D.F. and Hermans, J. (1980). A simulative comparison of linear, quadratic and kernel discrimination. J. Statist. Comput. Simul. 11, 87-106.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. Ann. Math. Statist. 27, 832-837.

Schaafsma, W. (1973). Classifying when populations are estimated. Discriminant Analysis and Applications. 339-364 Ed. T. Cacoullos. Academic Press.

Schaafsma, W. and Van Vark, G.N. (1977). Classification and discrimination problems with applications, part I. Statist. Neerlandica 31, 25-45.

Schaafsma, W. and Van Vark, G.N. (1979). Classification and discrimination problems with applications, part IIa. Statist. Neerlandica 33, 91-126.

Seber, G.A.F. (1963). The non-central chi-squared and beta distributions. Biometrika 50, 542-544.

Sedrask, N. and Okamoto, M. (1971). Estimation of the probabilities of misclassification for a linear discriminant function in the univariate case. Ann. Inst. Math. Statist. 23, 419-435.

Shah, B.K. (1963). Distribution of definite and indefinite quadratic forms from a non-central normal distribution. Ann. Math. Statist. 34, 186-190.

Siotani, M. (1956). On the distributions of Hotelling's T^2-Statistics. Ann. Inst. Statist. Math. 8, 1-14.

Siotani, M. and Wang, R.H. (1977). Asymptotic expansions for error rates and comparisons of the W-procedure and the Z-procedure in discriminant analysis. Multivariate Analysis IV, 523-545. Ed. P.R. Krishnaiah, North-Holland.

Sitgreaves, R. (1952). On the distribution of two random matrices used in classification procedures. Ann. Math. Statist. 23, 263-270.

Smidt, R.K. and Mc Donald, L.L. (1976a). Ridge estimation of the inverse of a covariance matrix.
Research Paper 106, Stat. Lab. S-1976-548. Univ. of Wyoming.

Smidt, R.K. and Mc Donald, L.L. (1976b). Ridge discriminant analysis. Research Paper 108, Stat. Lab. S-1976-549.
Univ. of Wyoming.

Smith, C.A.B. (1947). Some examples of discrimination.
Ann. Eugen. 13, 272-282.

Solomon, H and Stephens, M.A. (1977). Distribution of a sum of weighted chi-square variables. J. Am. Statist. Assoc. 72, 881-885.

Solomon, H and Stephens, M.A. (1978). Approximation to density functions using Pearson curves. J. Am. Statist. Assoc. 73, 153-160.

Sorum, M. (1968). Estimating the probability of misclassification.
Ph.D. Diss. Univ. of Minnesota, Minneapolis.

Sorum, M. (1971). Estimating the conditional probability of misclassification. Technometrics 13, 333-343.

Sorum, M. (1972). Three probabilities of misclassification.
Technometrics 14, 309-316.

Sorum, M. (1973). Estimating the expected probability of misclassification for a rule based on a linear discriminant function : univariate normal case. Technometrics 15, 329-339.

Specht, D.F. (1966). Generation of polynominal discriminant functions for pattern recognition. Ph.D. Diss. Stanford Univ.

Specht, D.F. (1971). Series estimation of a probability density function. Technometrics 13, 409-424.

Tiku, M.L. (1970). Tables of the double non-central F-distribution. Technical Report, Dept. of Appl. Math. Mc Master Univ.

Van Ness, J. (1979). On the effects of dimension in discriminant analysis for unequal covariance populations. Technometrics 21, 119-127.

Van Ness, J. and Simpson, C. (1976). On the effects of dimension in discriminant analysis. Technometrics 18, 175-187.

Van Ryzin, J. (1969). On strong consistency of density estimates. Ann. Math. Statist. 40, 1765-1772.

Wagner, T.J. (1975). Non-parametric estimates of probability densities. IEEE. Trans. Infor. Theory IT-21, 438-440.

Wahl, P.A. and Kronmal, R.A. (1977). Discriminant functions when covariances are unequal and sample sizes are moderate. Biometrics 33, 479-484.

Wang, Y.Y. (1967). A comparison of several variance component estimators. Biometrika 54, 301-305.

Wegman, E.J. (1972a). Non-parametric probability density estimation I. A summary of available methods. Technometrics 14, 533-546.

Wegman, E.J. (1972b). Non-parametric probability density estimation II : A comparison of density estimation methods. J. Statist. Comput. Simul. 1, 225-245.

Welch, B.L. (1939). Note on discriminant functions. Biometrika 31, 218-220.

Wertz, W. and Schneider, B. (1979). Statistical density estimation : A bibliography. Inter. Statist. Review 47, 155-175.