UNCERTAINTY IN DISCRIMINANT ANALYSIS

by

DAVID HIRST

A dissertation submitted to the

University of Glasgow

for the degree of

Doctor of Philosophy

1988

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

ABSTRACT

The aim of this thesis is to review and develop theory in discriminant analysis.

In chapter one an example of medical diagnosis is considered, and two types of uncertainty are illustrated. Firstly, the log odds ratio can be close to zero, and secondly there can be considerable uncertainty about its true value.

In chapter two we review existing methodology for constructing interval estimates for the log odds when the two populations are normal. Five different methods are considered for distributions with equal covariances, and three are generalised to the unequal covariance situation.

In chapter three these methods are investigated by simulation. It is seen that only two methods in the equal covariance case give intervals of reliable empirical confidence, and only one generalises successfully to the unequal covariance case.

In chapter four we go on to use the interval estimation methodology to assess a discriminant rule, suggesting some new ways of displaying the information available.

In chapter five we develop the methods of chapter four to construct an accurate error rate estimator, which is compared with standard techniques by simulation.

In chapter six the error rate estimator developed in chapter five is extended to the situation where there are more than two groups, and it is compared by simulation with generalisations of other standard techniques. The different methods are applied to a data set.

In chapter seven the limitations of the work are discussed, and possible developments suggested.

Acknowledgements

I would like to thank my supervisor Dr Ian Ford for his help and encouragement during the period of this research. This work was supported by a Science and Engineering Research Council grant.

TABLE OF CONTENTS

CHAPTER ONE

Introduction

This thesis is concerned with the problem of discriminant analysis, or statistical pattern recognition, which we will consider mostly in the context of medical diagnosis. In this chapter we use an example to illustrate what the discriminant analysis problem is, and to give an idea of the problems involved.

The example concerns Conn's Syndrome, a form of hypertension, which occurs in two distinct forms, which we will call type 1 and type 2. The two types need to be treated differently, but are difficult to distinguish. To aid the clinician in his diagnosis, measurements of eight variables have been made on 35 patients. The variables are:-

1 Age
2 Plasma concentration of sodium
3 " " " potassium
4 " " " carbon dioxide
5 " " " renin
6 " " " aldosterone
7 Systolic blood pressure
8 Diastolic blood pressure

Of the 35 patients, 20 are known to be of type 1, 11 are of type 2 and the other 4 are undiagnosed. The data are given in appendix one (reproduced from Aitchison and Dunsmore (1975)). We would like to decide on the type of the four undiagnosed patients, A, B, C and D.

A plot of two of the variables, 5 and 3 is shown in figure 1.1. The data have been log-transformed in order to make an assumption of multivariate normality plausible. We can see from

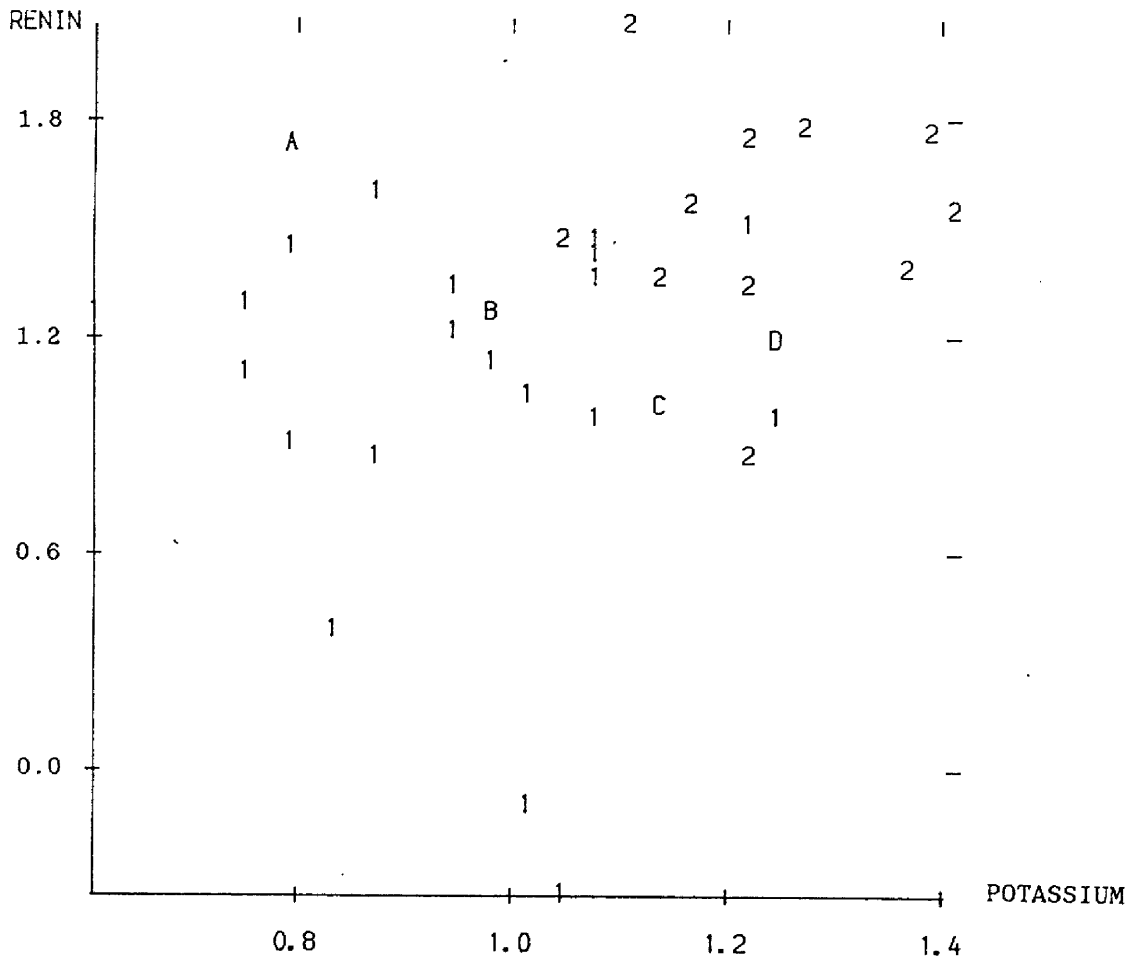

Figure 1.1
Conn's Syndrome:
Scatter plot for variables 3 and 5

the plot that A and B look like type 1s, C is a borderline case, and D looks like type 2. The aim of discriminant analysis is to formalise this judgement.

One approach is to estimate the log odds ratio $\theta(x)$ where

$$\theta(x)=\ln\left[\frac{p(\text{type 1} \mid x)}{p(\text{type 2} \mid x)}\right]$$

and x is the data on the undiagnosed patient. If we make the assumption of normality mentioned above, there are several possible estimators of $\theta(x)$. We will not assume that the covariance matrices of the two groups are equal, and so it will be neccessary to estimate them separately. This is clearly difficult with eight variables, and only 11 observations of type 2, and so the subset of the four variables 1,3,4 and 5 has been selected to reduce the dimensionality of the problem. Choosing a suitable subset of variables for dimensionality reduction is a complex problem, and ways of assessing and comparing different subsets are discussed in chapter four.

Using these four variables, and assuming equal prior probabilities for membership of the two types, the minimum variance unbiased estimator (Moran and Murphy (1979)), $\hat{\theta}(x)$, was used to estimate $\theta(x)$ for A,B,C and D. The estimates were:-

Patient	$\hat{\theta}(x)$
A	10.93
B	5.10
C	-0.28
D	-1.49

These are rather difficult to interpret, and so we use the transformation

$$\hat{p}(\text{type } 1|x) = \frac{\exp(\hat{\theta}(x))}{1+\exp(\hat{\theta}(x))}$$

in order to get the estimated probabilities below:-

Patient	$\hat{p}$(type 1|x)
A	1.00
B	0.99
C	0.43
D	0.18

These appear to suggest that patients A and B are almost certainly type 1, and C and D probably type 2. However, there are two sources of uncertainty. Firstly, the estimated probability for patient C is near to 0.5, implying that this is a borderline case (as the plot suggested) and that any diagnosis based on this data should be made with caution. Secondly, these are only estimates of the true probabilities p(type 1|x), and there may be some doubt as to the accuracy of the estimator.

Because of this second source of uncertainty, in chapters two and three we discuss ways of constructing interval estimates for the probabilities, but to give an idea of their importance the method due to Rigby (1982) was used to obtain approximate 95% confidence intervals for the four undiagnosed cases. They are as follows:-

Patient	Interval
A	(1.00,1.00)
B	(0.96,1.00)
C	(0.13,0.99)
D	(0.02,0.91)

These intervals show that we can be confident of our diagnoses for patients A and B, and they confirm our doubt about patient C. However, for patient D the point estimate of

p(type 1|x) appeared to indicate that the patient was probably of type 2, whereas the interval estimate shows that in actual fact we have insufficient evidence to make any such diagnosis.

This example illustrates some of the problems with discriminant analysis. In this case there were only two possible diagnoses, type 1 or type 2, though in general there could be any number. Most of the following work will only consider the two group case, though many of the results can be generalised. In chapter six we consider another example with more than two groups.

CHAPTER TWO

2.1 Inference for the Log Odds-Ratio

Consider the situation where a random vector X. of dimension p, arises from one of k multivariate Normal populations $N_p(\mu_i,\Omega_i)$, i=1,....,k with probability π_i.For any observed value x of X,it is desired to identify the population of origin of this observation.The posterior probabilities $\pi_i(x)$ that the observation is a member of group i are given by

$$\pi_i(x)=\pi_i f_i(x)/(\Sigma\pi_i f_i(x))$$

where $f_i(.)$ denotes the probability density function for population i. We will only be concerned with the case k=2,and so need only consider the posterior log-odds $\psi(x)$ in favour of population 1 (say).This is given by

$$\psi(x)=\psi+\theta(x)$$

where

$$\psi=\log(\pi_1/\pi_2)$$

and

$$\theta(x) = \tfrac{1}{2}\sum_{i=1}^{2}(-1)^i(\alpha_i(x) + \log\det\Omega_i)$$

Here $\alpha_i(x)$ denotes the squared Mahalanobis distance $(x-\mu_i)^T\Omega_i^{-1}(x-\mu_i)$, which measures the atypicality of x for population i. In the particular case $\Omega_1=\Omega_2=\Omega$, $\theta(x)$ is simply $\frac{1}{2}(\alpha_2(x)-\alpha_1(x))$.

We assume $\{\pi_i\}$ are known, and the $\{\mu_i,\Omega_i\}$ unknown.The training set T consists of independent random samples of size n_i from the two populations.Let $n=n_1+n_2$. and $\bar{X}_i$ and S_i denote the mean vector and corrected sums of squares and cross products matrix for the i^{th} population.or in the case $\Omega_1=\Omega_2$. S denotes the pooled matrix S_1+S_2. All inferences are made conditional on the sample sizes n_1 and n_2.

2.2 Point Estimation and Interval Estimation

There has been considerable discussion in the literature on the relative merits of several different 'predictive' and 'estimative' point estimates of $\theta(x)$.Following Geisser(1964) it has been suggested in Aitchison and Kay (1975) , Aitchison,Habbema and Kay (1977) , Geisser (1977) and M^{C}Lachlan (1977,1979) that predictive estimators are to be preferred.However. Moran and Murphy (1979) provide evidence that the bias corrected estimative estimators are essentially comparable.

Schaafsma (1984 section 2.1) points out that the difference between estimators will be of the order n^{-1} , whereas their standard deviations are of order $n^{-\frac{1}{2}}$. Therefore in the context of interval estimation the choice of estimator is less important.It is more important to derive and use an estimate of the variance of an estimator of $\theta(x)$.

We consider in detail several approaches to interval estimation for $\theta(x)$ in later sections,but the following is a brief historical summary of the work done so far:-

Much of the initial work was done by Schaafsma and colleagues in Holland, who derived methods based on asymptotic formulae (see for example Schaafsma and Van Vark (1979), Ambergen and Schaafsma (1984)).Schaafsma and others have also developed a computer package to aid the application of their approach, the POSCON project (Van der Sluis and Schaafsma (1984)). Critchley and Ford (1984,1985) discuss and derive approaches based on exact variance formulae, and Rigby (1982) describes a Bayesian approach to the problem. Critchley ,Ford and Rijal (1987) have developed methods based on the profile likelihood, and Davis (1987) uses moments and joint cumulants in another approach. We now look at some of

these methods in detail.

2.3 Equal Covariance Case

Consider the case $\Omega_1=\Omega_2=\Omega$. so $\theta(x)=\frac{1}{2}(\alpha_2(x)-\alpha_1(x))$.

2.3.1 Sampling Theory Approaches, E1 and E2

Moran and Murphy (1979) show that the minimum variance unbiased estimator for $\theta(x)$ is

$$\hat{\theta}(x)=(N-p-1)(\bar{X}_1-\bar{X}_2)^T S^{-1}(x-\tfrac{1}{2}(\bar{X}_1+\bar{X}_2))+\tfrac{1}{2}p(n_1^{-1}-n_2^{-1})$$

where $N=n_1+n_2-2$.

Schaafsma (1982 section 5) and Critchley and Ford(1985) show that the variance of $\hat{\theta}(x)$ is given by

$$\begin{aligned}(N-p)(N-p-3)\mathrm{var}(\hat{\theta}(x)) &= (N-p+1)\{\theta(x)-\tfrac{1}{2}(N-1)(n_1^{-1}-n_2^{-1})\}^2 \\ &\quad -(N-p-1)[\phi(x)\{(N-1)(n_1^{-1}+n_2^{-1})+\Delta^2\}-\tfrac{1}{4}\Delta^4] \\ &\quad +\tfrac{1}{4}(N-1)(N-p-1)\{2p(n_1^{-2}+n_2^{-2}) \\ &\quad -(N+1)(n_1^{-1}-n_2^{-1})^2\}\end{aligned}$$

where $\phi(x)=\frac{1}{2}(\alpha_1(x)+\alpha_2(x))$ measures the average atypicality of observation x from the two populations, and

$$\Delta^2=(\mu_1-\mu_2)^T\Omega^{-1}(\mu_1-\mu_2)$$

is the squared Mahalanobis distance between them.

The exact distribution of $\hat{\theta}(x)$ appears to be intractable although invariance arguments, given in Critchley and Ford (1985) show that it depends only θ,ϕ and Δ^2. Their argument is based on a simplifying transformation:-

Let P be a matrix such that $\Omega^{-1}=P^TP$, let Q be an orthogonal

matrix whose first column is proportional to $P(\mu_1-\mu_2)$ and let $\delta^T=(\Delta,0,\ldots,0)$. Now transform each vector in the training set, and the observed vector x, according to $t\to t^*=At-b$, where $A=QP$ and $b=A\mu_2+\frac{1}{2}\delta$. Then

$$\bar{X}_1\to\bar{X}_1^*\sim N_p(\tfrac{1}{2}\delta,n_1^{-1}I)$$

$$\bar{X}_2\to\bar{X}_2^*\sim N_p(-\tfrac{1}{2}\delta,n_2^{-1}I)$$

$$S\to S^*\sim W_p(N,I)$$

where $W_p(N,I)$ denotes the Wishart distribution with N degrees of freedom and parametric matrix I.

$\alpha_1(x)$, $\alpha_2(x)$ and Δ are all invariant to this transformation, as is $\hat{\theta}(x)$. If x_1^* is the first component of x^*, then we also have rotational symmetry about the x_1^* axis, and so the distribution must depend only upon Δ, x_1^* and the length of x^*, apart from the known constants n_1, n_2 and p. Alternatively, since $\phi(x)=x^{*T}x^*+\frac{1}{4}\Delta^2$, the dependence is only upon Δ, $\theta(x)$ and $\phi(x)$.

Being an affine transformation of the maximum likelihood estimator for $\theta(x)$, $\hat{\theta}(x)$ is asymptotically normally distributed, with mean and variance as given. It also depends quadratically on $\theta(x)$, increasing as $\theta(x)^2$ when $n_1=n_2$. This allows the construction of two different interval estimates.

Method E1, called PLUGALL by Critchley and Ford (1985) uses $\hat{\theta}\pm1.96(\hat{v})^{\frac{1}{2}}$ as an approximate 95% confidence interval, where $\hat{v}$ is obtained by plugging in estimates for all the unknown parameters in the above expression for $\mathrm{var}(\hat{\theta}(x))$. The estimators used are unbiased for $\theta(x)$, $\phi(x)$ and Δ. They are $\hat{\theta}(x)$ as given, and

$$\hat{\phi}(x)=\tfrac{1}{2}(\hat{\alpha}_1+\hat{\alpha}_2-p(n_1^{-1}+n_2^{-1}))$$

where $\hat{\alpha}_i=(N-p-1)(x-\bar{X}_i)^TS^{-1}(x-\bar{X}_i)$, $i=1,2$

$$\hat{\Delta}=(N-p-1)(\bar{X}_1-\bar{X}_2)S^{-1}(\bar{X}_1-\bar{X}_2)$$

Method E2, called PLUGPART by Critchley and Ford (1985),

substitutes estimates $\hat{\phi}(x)$ and $\hat{\Delta}$ for $\phi(x)$ and Δ to obtain $\mathrm{var}(\hat{\theta}(x)|\theta(x),\hat{\phi}(x),\hat{\Delta})$, and so gets an approximate interval of the form

$$\{\theta(x):(\theta(x)-\hat{\theta}(x))^2<(1.96)^2\mathrm{var}(\hat{\theta}(x)|\theta(x),\hat{\phi}(x),\hat{\Delta})\}$$

If we write $\mathrm{var}(\hat{\theta}(x)|\theta(x),\hat{\phi}(x),\hat{\Delta})=a\theta(x)^2+b\theta(x)+c(\hat{\phi}(x),\hat{\Delta})$, then the endpoints of the interval can be found by solving for θ the quadratic equation

$$(\theta(x)-\hat{\theta}(x))^2=(1.96)^2(a\theta(x)^2+b\theta(x)+c(\hat{\phi}(x),\hat{\Delta}))$$

2.3.2 The Bayesian Approach E3

Rigby (1982) considers a Bayesian approach to the problem. In his paper he considers only the unequal covariance situation, but he has unpublished work on the equal covariance case. His approach involves a rather different philosophy to that of the previous section, and he defines the problem as follows:-

Prior to being observed, X is assumed to be drawn from the combined distribution $\pi_1 p_1(x|\xi)+\pi_2 p_2(x|\xi)$, where $p_i(x|\xi)$ is the i^{th} population density at x given the population parameters ξ. In the Bayesian approach due to Rigby, ξ is regarded as a random variable. We are interested in the distribution of the log odds-ratio $L=l_1-l_2$ given the feature x, where $l_i=\log(p_i(x|\xi))$ and L is now an induced random variable since it is a function of ξ. If the moment generating function of the posterior distribution of L, $\Phi_L(t)$, given the training data can be calculated, then the distribution can be approximated by finding the first four moments and fitting Pearson curves (Elderton and Johnson 1969, chapter five).

$\Phi_L(t)$ is defined by

$$\Phi_L(t)=E(e^{tL})=\int\left[\frac{p_1(x|\xi)}{p_2(x|\xi)}\right]^t p(\xi|T)d\xi$$

where $p(\xi|T)$ is the posterior distribution of ξ given the training data T. If vague prior information is assumed, ie $p(\xi)\propto|\Omega^{-1}|^{-\frac{1}{2}(p+1)}$ then $(\xi|T)$ has a Normal-Normal-Wishart distribution

$$p(\xi|T)=\text{NoNoWi}_p(\bar{X}_1,n_1,\bar{X}_2,n_2,n_1+n_2-2,S)$$

Rigby then shows that the moment generating function of L is given by

$$\Phi_L(t)=\left[\frac{n_1}{n_1+t}\right]^{\frac{1}{2}p}\left[\frac{n_2}{n_2+t}\right]^{\frac{1}{2}p}\frac{|S|^{\frac{1}{2}N}}{|H|^{\frac{1}{2}N}}$$

where $|H|=|S|\left[1+\frac{n_1t}{n_1+t}D_1-\frac{n_2t}{n_2-t}D_2-\frac{n_1n_2t^2}{(n_1+t)(n_2-t)}(D_1D_2-D_{12}^2)\right]$

and $D_i=(\bar{X}_i-x)^TS^{-1}(\bar{X}_i-x)$, $i=1.2$, $D_{12}=(\bar{X}_1-x)S^{-1}(\bar{X}_2-x)$

Hence the first four moments, denoted by m_i, $i=1,\ldots,4$, are given by

$$m_1 = \tfrac{1}{2}(N(D_2-D_1)+p(n_2^{-1}-n_1^{-1}))$$

$$m_2 = \tfrac{1}{2}[(N(D_2^2+D_1^2)+2N(D_2/n_2+D_1/n_1)+p(n_2^{-2}+n_1^{-2})] -ND_{12}^2$$

$$m_3=N(D_2^3-D_1^3)+3N(D_2^2/n_2-D_1^2/n_1) +3N(D_2/n_2^2-D_1/n_1^2)+p(1/n_2^3-1/n_1^3) +3ND_{12}^2((D_2-D_1)+(1/n_2-1/n_1))$$

$m_4=A+3m_2^2$ where

$$A = 3[N(D_1^4+D_2^4)+4N(D_1^3/n_1+D_2^3/n_2) +6N(D_1^2/n_1^2+D_2^2/n_2^2)+4N(D_1/n_1^3+D_2/n_2^3) +p(1/n_1^4+1/n_2^4)] +6ND_{12}^4+12ND_{12}^2((D_1-D_2)+(1/n_1-1/n_2))^2 -6ND_{12}^2(D_1+1/n_1)(D_2+1/n_2)$$

2.3.3 The Profile Likelihood Approach E4

An alternative approach is to base the interval estimate for θ on the profile of the log likelihood function (Kalbfleisch and Sprott (1970),Kalbfleisch (1979)).The profile likelihood is defined as follows:-

Let Ξ be the parameter space for the model, and let ξ denote the full set of unknown parameters. Then given the training data the log odds-ratio θ is a function of ξ, say $\theta=h(\xi)$ where $h:\Xi\to\mathbb{R}$. Let $\Theta=h(\Xi)$. Let $\ell:\Xi\to\mathbb{R}$ be the log likelihood function. For given $\theta\in\Theta$ consider the problem $P(\theta)$:

Maximise $\ell(\xi)$ over $\xi\in\Xi$ subject to $h(\xi)=\theta$

Where it exists, denote the maximal value by $p(\theta)$. $p(\theta)$ is the profile log likelihood function for θ.

Critchley, Ford and Rijal (1988) show how strong Lagrangian theory can be used to simplify the construction of the profile likelihood, and to obtain interval estimates for θ. We give their results in some detail, as they can be used for distributions other than the normal (see later).

Consider the unconstrained problem $Q(\lambda)$, $\lambda\in\mathbb{R}$:

Maximise $\ell(\xi)-\lambda h(\xi)$ over $\xi\in\Xi$

Let Λ be the set of λ for which a solution to $Q(\lambda)$ exists, and let $\hat{\Theta}$ be the set of $h(\hat{\xi}_\lambda)$ such that $\hat{\xi}_\lambda$ is a solution to $Q(\lambda)$ for some $\lambda\in\Lambda$. Consider the following conditions:-

(i) There exists some $\hat{\xi}\in\Xi$ such that for all $\xi\in\Xi$,

$$\ell(\xi)\leqslant\ell(\hat{\xi})<\infty$$

(ii) $\Theta=\hat{\Theta}$

(iii) The interior of Θ, denoted Θ°, is convex.

(iv) For all $\lambda\in\Lambda$, $\hat{\xi}_\lambda$ is unique, and the unique $h(\hat{\xi}_\lambda)$ is denoted by $\hat{\theta}_\lambda$.

(v) Θ is open

(vi) $p(.)$ has a derivative, denoted by $p'(.)$, in Θ°.

If these six conditions hold, then Critchley, Ford and Rijal (1988) show that

(1) $p(.)$ has domain Θ.

(2) For each $\lambda\in\Lambda$, a solution $\hat{\xi}_\lambda$ to $Q(\lambda)$ is also a solution to $P(\theta)$ for $\theta=h(\hat{\xi}_\lambda)$, and so $p(\theta)=\ell(\hat{\xi}_\lambda)$. Therefore the entire function $p(.)$ is obtained by letting λ vary through Λ.

(3) $p(.)$ is strictly concave on Θ.

(4) The family of interval estimates for θ based on the profile likelihood is $\{I_\alpha:\alpha\geqslant 0\}$ where $I_\alpha=\{\theta\in\Theta:p(\hat{\theta})-p(\theta)\leqslant\alpha\}$, $\hat{\theta}=h(\hat{\xi})$. These intervals are convex.

(5) For all $\theta\in\Theta^\circ$ there is a unique $\lambda\in\Lambda$, written $\lambda(\theta)$, for which $\theta=h(\hat{\xi}_\lambda)$

(6) $\lambda(\theta)=p'(\theta)$

(7) $\theta(.)$ is a strictly decreasing function on Θ°.

These results mean that in order to find an interval estimate for θ, it is neccessary only to find Λ, $\hat{\theta}_\lambda$ and $p(\hat{\theta}_\lambda)$. The maximum of $p(.)$, $p(\hat{\theta})$ will be at $\lambda=0$ and, since asymptotically $-2(p(\hat{\theta}_\lambda)-p(\hat{\theta}))$ has a chi-squared distribution with one degree of freedom, it is only necessary to find the two values of λ such that

$$p(\hat{\theta}_\lambda)-p(\hat{\theta})=-\tfrac{1}{2}x^2(1;0.95)$$

This is most easily done by drawing the $(\theta,p(\theta))$ graph, or it can be done numerically.

In the case of normally distributed populations with equal covariances we have, ignoring constant terms,

$$2\{\ell(\xi)-\lambda h(\xi)\}=-(n_1+n_2)\ln|\Omega|$$
$$-\{\Sigma(x_{1i}-\mu_1)^T\Omega^{-1}(x_{1i}-\mu_1)-\lambda(x-\mu_1)^T\Omega^{-1}(x-\mu_1)\}$$
$$-\{\Sigma(x_{2i}-\mu_2)^T\Omega^{-1}(x_{2i}-\mu_2)+\lambda(x-\mu_2)^T\Omega^{-1}(x-\mu_2)\}$$

Here λ can be regarded as the weight with which x is added to

population 2 and subtracted from population 1. Hence the unique $\hat{\xi}_\lambda$ can be written down from standard maximum likelihood theory, and conditions (i) and (iv) must hold.

$$\hat{\mu}_1(\lambda)=(n_1\bar{X}_1-\lambda x)/(n_1-\lambda)$$

$$\hat{\mu}_2(\lambda)=(n_2\bar{X}_2+\lambda x)/(n_2+\lambda)$$

$$\hat{\Omega}(\lambda)=(n_1+n_2)^{-1}[S+(n_1^{-1}-\lambda^{-1})^{-1}(x-\bar{X}_1)(x-\bar{X}_1)^T + (n_2^{-1}+\lambda^{-1})^{-1}(x-\bar{X}_2)(x-\bar{X}_2)^T]$$

The condition $|\hat{\Omega}|>0$ gives Λ as

$$\Lambda=\{\lambda:(n_1-\lambda)(n_2+\lambda)-(n_2+\lambda)n_1D_1\lambda+(n_1-\lambda)n_2D_2\lambda - n_1n_2\lambda^2(D_1D_2-D_{12}^2)>0\}$$

where $D_i=(x-\bar{X}_i)^TS^{-1}(x-\bar{X}_i)$, $i=1,2$, $D_{12}=(x-\bar{X}_1)^TS^{-1}(x-\bar{X}_2)$

The relevant formulae for $\hat{\theta}_\lambda$ and $p(\hat{\theta}_\lambda)$ are:-

$$\hat{\theta}_\lambda=-\tfrac{1}{2}(n_1+n_2)u/v$$

where

$$u=D_1[n_1/(n_1-\lambda)]^2\{1+n_2D_2\lambda/(n_2+\lambda)\} -D_2[n_2/(n_2+\lambda)]^2\{1-n_1D_1\lambda/(n_1-\lambda)\} -D_{12}^2\lambda\{n_1/(n_1-\lambda)\}\{n_2/(n_2+\lambda)\}\{n_1/(n_1-\lambda)+n_2/(n_2+\lambda)\}$$

$$v=\{1-n_1D_1\lambda/(n_1-\lambda)\}\{1+n_2D_2\lambda/(n_2+\lambda)\} +n_1n_2D_{12}^2\lambda^2/\{(n_1-\lambda)(n_2+\lambda)\}$$

$$2p(\hat{\theta}_\lambda)=(n_1+n_2)p\{\log(n_1+n_2)-1\}-(n_1+n_2)\log\det S -(n_1+n_2)\log(w)+2\lambda\hat{\theta}_\lambda$$

where

$$w=1-n_1D_1\lambda/(n_1-\lambda)+n_2D_2\lambda/(n_2+\lambda) -n_1n_2\lambda^2(D_1D_2-D_{12}^2)/\{(n_1-\lambda)(n_2+\lambda)\}$$

Conditions (iii) and (iv) hold since $\Theta=\mathbb{R}$, and (vi) is clearly true. Also $\hat{\Theta}=\mathbb{R}$ and so all the conditions are satisfied. Hence it is easy to obtain an interval estimate for θ.

2.3.4 An Approach Based On Sampling Cumulants Of $\tilde{\theta}$, E5

Davis (1987) has developed a technique for evaluating the exact first four moments and asymptotic expansions of the cumulants of statistics of the form W where

$$W=\nu tr(zY^{T}S^{-1}Y)$$

where $S\sim W_p(n_1+n_2-2,\Omega)$, $\nu=n_1+n_2-p-3$, Z is a constant 2x2 matrix and Y has the multivariate normal distribution given by

$$(2\pi)^{-p}|A|^{-\frac{1}{2}}|\Omega|^{-1}\exp[-tr\{\tfrac{1}{2}A^{-1}(Y-M)^{T}\Omega^{-1}(Y-M)\}]$$

where A is a symmetric 2x2 matrix and the mean vector M is px2

Davis considers the (biased) estimators

$$\tilde{\theta}(x)=\tfrac{1}{2}(\tilde{\alpha}_2(x)-\tilde{\alpha}_1(x))$$

$$\tilde{\phi}(x)=\tfrac{1}{2}(\tilde{\alpha}_2(x)+\tilde{\alpha}_1(x))$$

$$\tilde{\Delta}^2(x)=\nu(\bar{X}_1-\bar{X}_2)^{T}S^{-1}(\bar{X}_1-\bar{X}_2)$$

where

$$\tilde{\alpha}_i(x)=\nu(x-\bar{X}_i)^{T}S^{-1}(x-\bar{X}_i)\ ,\ i=1,2$$

$t_1\tilde{\theta}(x)+t_2\tilde{\phi}(x)+t_3\tilde{\Delta}^2$ has the form W with

$$Z=\begin{bmatrix} t_2 & \tfrac{1}{2}t_1 \\ \tfrac{1}{2}t_1 & \tfrac{1}{4}t_2+t_3 \end{bmatrix}$$

$$M=(x-\tfrac{1}{2}(\mu_1+\mu_2),\mu_1-\mu_2)$$

$$A=\begin{bmatrix} \tfrac{1}{4}(n_1^{-1}+n_2^{-1}) & \tfrac{1}{2}(n_2^{-1}-n_1^{-1}) \\ \tfrac{1}{2}(n_2^{-1}-n_1^{-1}) & n_1^{-1}+n_2^{-1} \end{bmatrix}$$

He then uses the method of Peers and Iqbal (1985) to construct approximate interval estimates.This method is based on the construction of a series of functions $\{h_r\}$ which has the property

$$p\{m^{\frac{1}{2}}(\tilde{\gamma}_1-\gamma_1)<\tilde{h}_r|\gamma\} = \alpha+O_p(m^{-r/2})$$

where a confidence interval is required for γ_1 , the first element of $\gamma=(\gamma_1,\ldots,\gamma_q)$, the other γ_i being nuisance parameters. It is assumed that an estimator $\tilde{\gamma}=(\tilde{\gamma}_1,\ldots,\tilde{\gamma}_q)$ is available for γ

such that $v=m^{1/2}(\tilde{\gamma}-\gamma)$ has moment generating function with asymptotic expansion of the form

$$M_v(t)=\exp\{(1/2!)\lambda_{ij}t_it_j+m^{-1/2}\lambda_it_i+(m^{-1/2}/3!)\lambda_{ijk}t_it_jt_k$$
$$+(m^{-1}/4!)\lambda_{ijkl}t_it_jt_kt_l+O(m^{-3/2})\}$$

where $t=(t_1,\ldots,t_q)$ is a vector of scalars and m is a parameter which increases with sample size. The λs are the cumulants of v scaled by the associated powers of m, and the tensor convention of summing over repeated subscripts from 1 to q has been adopted. $\tilde{h}_r$ is obtained by substituting the estimators $\tilde{\gamma}$ into an expression for h_r.

Setting $\tilde{\gamma}=(\tilde{\theta}(x),\tilde{\phi}(x),\tilde{\Delta}^2)$ and $m=\nu$, $m^{1/2}(\tilde{\gamma}-\gamma)$ has the required form and so Davis uses h_3 in order to get intervals of confidence $\alpha+O_p(m^{-3/2})$. From Peers and Iqbal (1985),

$$h_r=\sum_{i=1}^{r} m^{-1/2(i-1)}g_i$$

where if η denotes the α percent point of the standard normal distribution,

$$g_1 = \eta(\lambda_{11})^{1/2}$$

$$g_2 = \lambda_1-\tfrac{1}{2}\eta^2\frac{\lambda_{ij}}{\lambda 11}\frac{\partial\lambda_{11}}{\partial\theta_j}+\frac{(\eta^2-1)}{6}\frac{\lambda_{111}}{\lambda_{11}}$$

$$g_3 = -\lambda_jg_{1(j)}-\tfrac{1}{2}\lambda_{ij}g_{1(ij)}+\frac{(\eta^2-1)}{2\lambda_{11}}\left[\left[\frac{2\lambda_{1j}}{3\lambda_{11}}\lambda_{111}-\lambda_{11j}\right]g_{1(j)}\right.$$
$$\left.-\lambda_{1i}\lambda_{1j}g_{1(ij)}\right]$$
$$+\tfrac{1}{2}\eta\lambda_{11}^{1/2}\left[\frac{-2\lambda_{1j}}{\lambda_{11}}g_{2(j)}+\frac{\lambda_{ij}}{\lambda_{11}}g_{1(i)}g_{1(j)}-\right.$$
$$\left.-\left[\frac{\lambda_{ij}}{\lambda_{11}}g_{1(j)}\right]^2\right]$$

$$+\frac{(n^3-3n)\lambda_{1111}}{24\lambda_{11}^{3/2}}-\frac{(2n^3-5n)\lambda_{111}^2}{36\lambda_{111}^{5/2}}$$

where $g_{r(i)}\equiv\frac{\partial g_r}{\partial\gamma_i}$

The approprate asymptotic cumulants are, where $\rho_i=\nu/n_i, i=1,2$, $\psi(x)=\phi(x)-\frac{1}{2}\Delta^2$ and omitting the argument x for brevity

$$\lambda_1 = \tfrac{1}{2}p(\rho_2-\rho_1)$$

$$\lambda_{11} = \theta^2+\psi\Delta^2+\rho_1(\phi-\theta)+\rho_2(\theta+\phi)$$
$$+\nu^{-1}[3\theta^2+\psi\Delta^2+\rho_1\{(p+1)(\phi-\theta)-2\theta\}$$
$$+\rho_2\{(p+1)(\theta+\phi)+2\theta\}+\tfrac{1}{2}p(\rho_1^2+\rho_2^2)]+O(\nu^{-2})$$

$$\lambda_{22} = \theta^2+2\psi^2+(1/8)\Delta^4+\rho_1(\phi-\theta)+\rho_2(\phi+\theta)+O(\nu^{-1})$$

$$\lambda_{33} = 2\Delta^4+4\Delta^2(\rho_1+\rho_2)+O(\nu^{-1})$$

$$\lambda_{12} = 2\theta\phi+\rho_1(\theta-\phi)+\rho_2(\theta+\phi)+O(\nu^{-1})$$

$$\lambda_{13} = 2\theta\Delta^2+\rho_1(2\theta-\Delta^2)+\rho_2(2\theta+\Delta^2)+O(\nu^{-1})$$

$$\lambda_{23} = 2\theta^2+\tfrac{1}{2}\Delta^4-\rho_1(2\theta-\Delta^2)+\rho_2(2\theta+\Delta^2)+O(\nu^{-1})$$

$$\lambda_{111} = 4\theta(\theta^2+3\psi\Delta^2)+6\rho_1\{-(\theta^2+\psi\Delta^2)+2\theta\phi\}+6\rho_2\{\theta^2+\psi\Delta^2+2\theta\phi\}$$
$$+3\{\rho_1^2(\theta-\phi)+\rho_2^2(\theta+\phi)+O(\nu^{-1})$$

$$\lambda_{112} = 4\phi(3\theta^2+\psi\Delta^2)+\rho_1(3\theta^2-6\theta\phi+2\phi^2+\psi\Delta^2)$$
$$+2\rho_2(3\theta^2+6\theta\phi+2\phi^2+\psi\Delta^2)$$
$$+3\{\rho_1^2(\phi-\theta)+\rho_2^2(\phi+\theta)\}+O(\nu^{-1})$$

$$\lambda_{113} = 4\Delta^2(3\theta^2+\psi\Delta^2)+4\rho_1\{3\theta^2-3\theta\Delta^2+(\phi+\psi)\Delta^2\}$$
$$+4\rho_2\{3\theta^2+3\theta\Delta^2+(\phi+\psi)\Delta^2\}$$
$$+2\{\rho_1^2(-3\theta+\phi+\Delta^2)+\rho_1\rho_2(2\phi-\Delta^2)$$
$$+\rho_2^2(3\theta+\phi+\Delta^2)\}+O(\nu^{-1})$$

$$\lambda_{1111} = 30\{(\theta^4+6\theta^2\psi\Delta^2+\psi^2\Delta^4)+2\rho_1(-\theta^3+3\theta^2\phi-3\theta\psi\Delta^2+\phi\psi\Delta^2)$$
$$+2\rho_2(\theta^3+3\theta^2\phi+3\theta\psi\Delta^2+\phi\psi\Delta^2)$$
$$+12\{\rho_1^2(5\theta^2-10\theta\phi+2\phi^2+3\psi\Delta^2)+\rho_2^2(5\theta^2+10\theta\phi+2\phi^2+3\psi\Delta^2)$$
$$+\rho_1^3(\phi-\theta)+\rho_2^3(\phi+\theta)\}+O(\nu^{-1})$$

The estimators $\tilde{\theta}(x)$, $\tilde{\phi}(x)$ and $\tilde{\Delta}$ are plugged into these equations, which are then used in the construction of $\tilde{h}_3$, and so

an interval estimate for $\theta(x)$ is obtained.

2.4 Unequal Covariance Case

Three of the methods of section 2.3 can be generalised to the case $\Omega_1 \neq \Omega_2$. The exceptions are the PLUGPART approach of Critchley and Ford (1985), which relied upon writing the variance of $(\hat{\theta}(x)|\theta(x),\hat{\phi}(x),\hat{\Delta})$ as a quadratic function of $\theta(x)$, and the method of Davis (1987), since the distribution of $\theta(x)$ is not dependent on only three variables as it is in the equal covariance case. In order to use Davis' method it would be neccessary to find the joint cumulants of many more parameters, and they would include terms involving $\log |\Omega_i|$.

2.4.1 The Sampling Theory Approach, U1

It is straightforward to construct interval estimates for $\theta(x)$ by a method analogous to E1. Critchley, Ford and Rijal (1987) give the best unbiased estimator of $\theta(x)$ as

$$\hat{\theta}_G(x) = \tfrac{1}{2}\sum_{i=1}^{2}(-1)^i[\hat{\alpha}_{iG}(x)+\log\det S_i - \sum_{j=1}^{p}\gamma\{\tfrac{1}{2}(n_i-j)\}]$$

where

$$\hat{\alpha}_{iG}(x)=(n_i-p-2)(x-\bar{X}_i)^T S_i^{-1}(x-\bar{X}_i)-p/n_i$$

and $\gamma(.)$ denotes the digamma function, defined in Abramowitz and Stegun (1965,p258). It may be evaluated efficiently using the algorithm in Bernado (1976).

The distribution of $\hat{\theta}_G(x)$ appears to be intractable, but it is asymptotically normal, and Critchley, Ford and Rijal (1987) give its approximate variance as $\tilde{v}_G$ where

$$\tilde{v}_G = \sum_{i=1}^{2} \left[\frac{\alpha_{iG}^2(x)}{2(n_i-p-4)} + \left[\frac{1}{n_i} - \frac{(n_i-2)}{(n_i-p-1)(n_i-p-4)} \right] \alpha_{iG}(x) \right.$$

$$\left. + \frac{p(n_i-2)}{2(n_i-p-1)(n_i-p-4)} \right]$$

where $\alpha_{iG}(x)=(x-\mu_i)\Omega_i^{-1}(x-\mu_i)$, $i=1,2$.

Substituting the estimate $\hat{\alpha}_{iG}(x)$ of $\alpha_{iG}(x)$ into $\tilde{v}_G$ to get $\hat{v}_G$, they obtain the approximate 95% confidence interval for $\theta_G(x)$ given by

$$\hat{\theta}_G(x) \pm 1.96(\hat{v}_G)^{\frac{1}{2}}$$

2.4.2 The Bayesian Approach, U2

In section 2.3.2 we gave a Bayesian approach to the problem in the equal covariance situation, due to Rigby (1982). This paper in fact only considers the unequal covariance case, the methods being very similar in the two contexts.

Let $l_i=\{\log p_i(x|\xi)\}$ as before where $\xi=\{\mu_1,\mu_2,\Omega_1,\Omega_2\}$, the set of population parameters. Here the posterior distributions of l_1 and l_2 are independent and so the posterior moments of $L(=l_1-l_2)$ can be obtained by finding the posterior moments of l_1 and l_2 separately. Let $z_i=p_i(x|\xi)=p_i(x|\mu_i,\Omega_i)$. If the moment generating function of l_i is $\Phi_{li}(t)$ then

$$\Phi_{li}(t)=E(\exp(tl_i))=E(z_i^t)$$

If vague prior information is assumed, ie $p(\mu_i,\Omega_i) \propto |\Omega_i^{-1}|^{-\frac{1}{2}(p+1)}$, then $p(\mu_i,\Omega_i|T)$ has a Normal-Wishart distribution

$$p(\mu_i,\Omega_i|T)=\text{NoWi}_p(X_i,n_i,n_i-1,S)$$

and Rigby (1982) shows that

$$E(z_i^t)=\frac{1}{|S_i|^{\frac{1}{2}t}\pi^{\frac{1}{2}pt}}\left[\frac{n_i}{n_i+t}\right]^{\frac{1}{2}p} \frac{\Gamma_p(\frac{1}{2}(n_i-1+t))}{\Gamma_p(\frac{1}{2}(n_i-1))} \Big/ \left[\frac{n_i t}{1+\frac{n_i t}{n_i+t}D_i}\right]^{\frac{1}{2}(n_i-1+t)}$$

where here $D_i=(x-X_i)^T S_i^{-1}(x-X_i)$, $i=1,2$

Rigby goes on to calculate the first four moments of l_1 and l_2, and then it is easy to obtain the moments of L using

$$E(L)=E(l_1)-E(l_2)$$

$$V(L)=V(l_1)+V(l_2)$$

$$E(L-E(L))^3=E(l_1-E(l_1))^3-E(l_2-E(l_2))^3$$

$$E(L-E(L))^4=E(l_1-E(l_1))^4+6V(l_1)V(l_2)+E(l_2-E(l_2))^4$$

They are m_{Gi}, i=1,...,4, where

$$m_{G1} = \tfrac{1}{2} \sum_{i=1}^{2} (-1)^i[(n_i-1)D_i+\log\det S_i + p/n_i - \sum_{j=1}^{p} \gamma(\tfrac{1}{2}(n_i-j))]$$

where γ denotes the digamma function (Abramowitz and Stegun (1965))

$$m_{G2} = \tfrac{1}{2} \sum_{i=1}^{2} \{(n_i-1)D_i^2-2D_i/n_i+p/n_i^2 + \tfrac{1}{2} \sum_{j=1}^{p} \gamma'(\tfrac{1}{2}(n_i-j))]$$

where γ' denotes the trigamma function (Abramowitz and Stegun (1965))

$$m_{G3} = \sum_{i=1}^{2} (-1)^i[-3D_i/n_i^2+1.5D_i^2(1-2/n_i) + D_i^3(n_i-1) + p/n_i^3 + (1/8) \sum_{j=1}^{p} \gamma^{(2)}(\tfrac{1}{2}(n_i-j))]$$

where $\gamma^{(2)}$ denotes the tetragamma function (Abramowitz and Stegun (1965))

$$m_{G4} = \sum_{i=1}^{2} [-12D_i n_i^{-3}+6(n_i-3)D_i^2 n_i^{-2} + 4(2n_i-3)D_i^3 n_i^{-1}+3D_i^4(n_i-1) + 3pn_i^{-4} + (1/16) \sum_{j=1}^{p} \gamma^{(3)}(\tfrac{1}{2}(n_i-j))] + 3m_{G3}^2$$

where $\gamma^{(3)}$ denotes the pentagamma function (Abramowitz and Stegun (1965)). Pearson curves are then fitted to approximate the distribution of L as in section 2.3.2.

2.4.3 The Profile Likelihood Approach, U3

Critchley, Ford and Rijal (1987) obtain approximate intervals

in the same manner as that used for the equal covariance case. The appropriate formulae are:-

$$2\ell(\xi)-\lambda h(\xi)=-(n_1-\lambda)\log|\Omega_1|-(n_2+\lambda)\log|\Omega_2|$$
$$-\Sigma(x_{1j}-\mu_1)^T\Omega_1^{-1}(x_{1j}-\mu_1)-\lambda(x-\mu_1)^T\Omega_1^{-1}(x-\mu_1)$$
$$-\Sigma(x_{2j}-\mu_2)^T\Omega_2^{-1}(x_{2j}-\mu_2)+\lambda(x-\mu_2)^T\Omega_2^{-1}(x-\mu_2)$$

Again λ can be regarded as the weight with which x is added to population 2 and subtracted from population 1. Hence $\hat{\xi}_\lambda$ is given by

$$\hat{\mu}_1(\lambda)=(n_1\bar{X}_1-\lambda x)/(n_1-\lambda)$$
$$\hat{\mu}_2(\lambda)=(n_2\bar{X}_2+\lambda x)/(n_2+\lambda)$$
$$\hat{\Omega}_1(\lambda)=[S_1+(n_1^{-1}-\lambda^{-1})^{-1}(x-\bar{X}_1)(x-\bar{X}_1)^T]/(n_1-\lambda)$$
$$\hat{\Omega}_2(\lambda)=[S_2+(n_2^{-1}+\lambda^{-1})^{-1}(x-\bar{X}_2)(x-\bar{X}_2)^T]/(n_2+\lambda)$$

$\hat{\Omega}_i(\lambda)$ must be positive definite for i=1,2, and so

$$\Lambda = \{\lambda: -n_2/(n_2D_2+1) < \lambda < n_1/(n_1D_1+1)\}$$

where

$$D_i = (x-\bar{X}_i)^T S_i^{-1}(x-\bar{X}_i),\ i=1,2$$

This gives:-

$$2\hat{\theta}_\lambda = \Sigma(-1)^i[\log\det S_i + \log\{n_i+(-1)^i\lambda(1+D_in_i)\}$$
$$+n_i^2D_i/\{n_i+(-1)^i\lambda(1+D_in_i)\}$$
$$-(p+1)\log(n_i+(-1)^i\lambda)]$$
$$2p(\hat{\theta}_\lambda) = \Sigma n_i[(p+1)\log(n_i+(-1)^i\lambda)-\log\det S_i$$
$$-\log\{n_i+(-1)^i\lambda(1+D_in_i)\}-p$$
$$+(-1)^i\lambda D_in_i/\{n_i+(-1)^i\lambda(1+D_in_i)\}]$$

Conditions (i) to (vi) hold as before, and so interval estimates for θ can be found.

2.5 Non Normal Data

2.5.1 Logistic Discrimination

If the distributions of interest are clearly non-normal, for example if they include discrete variates, one possible approach

to the interval estimation problem is through logistic discrimination (Anderson 1982). This is a 'partially distributional' method in that the only assumption made is that the log-odds ratio $\theta(x)$ is linear in the x_j (or sometimes simple functions of the x_j), where $x=(x_1,....x_p)^T$. ie

$$\theta(x)=\beta_0+\beta^T x \qquad (*)$$

where $\beta^T=(\beta_1,....\beta_p)$.

Anderson (1982) gives a large number of distributional families satisfying (*). They are (i) Multivariate normal distributions with equal covariances, (ii) Multivariate discrete distributions following the log-linear model with equal interaction terms, (iii) Joint distributions of continuous and discrete variables following (i) and (ii), (iv) Selective and truncated versions of the foregoing, (v) Versions of the foregoing with any specified functions of the x_j. Kay and Little (1987) show that, under suitable transformations of the x_j, any member of the exponential family also satisfies (*).

β is estimated by maximum likelihood. This is straightforward if the training data are sampled from the mixture distribution of the two populations. Let $n_i(x)$ be the number of points from population i at x. Day and Kerridge (1967), show that β can be evaluated by maximising L, where

$$L=\Pi(p_1(x))^{n_1(x)}(p_2(x))^{n_2(x)}$$

where $p_i(x)=p(\text{population } i|x)$ and

$$p_1(x)=\frac{\exp(\beta_0+\beta^T x)}{1+\exp(\beta_0+\beta^T x)}$$

$$p_2(x)=\frac{1}{1+\exp(\beta_0+\beta^T x)}$$

This is generally maximised by a Newton-Raphson type procedure.

The asymptotic variances and covariances of the parameters are found by inverting the sample information matrix $I=(-\partial^2 \ln L/\partial\beta_j\partial\beta_1)$ at $\beta=\hat{\beta}$.

It is easily shown that

$$\partial^2 \ln L/\partial\beta_j\partial\beta_1=-\Sigma n(x)p_1(x)p_2(x)x_jx_1$$

where $n(x)=n_1(x)+n_2(x)$. Maximum likelihood estimators are asymptotically normal, and so an approximate 95% confidence interval for the log-odds is $\hat{\theta}(x)\pm 1.96(\hat{var}(\theta(x)))^{\frac{1}{2}}$, where

$$\hat{var}(\theta(x))=x^T I^{-1} x$$

The situation is more complicated if the training data are drawn as separate samples from each population, but Anderson (1979) suggests that β (but not β_0) can be estimated in the same way. He proves this in the case where the variates are all discrete. In Anderson (1982) it is suggested that the procedure (ie maximising L) is also valid for continuous variates. The appropriate estimate of β_0 is $\beta_0'=\beta_0-\ln(\pi_1/\pi_2)$ where π_i is the proportion of population i in the mixture distribution. This must be estimated separately. The variance is estimated in the same way, but there is an additional error of $o(1/n)$ introduced into the variance of β_0.

Efron (1975) compared the asymptotic relative efficiency of logistic and normal discrimination when the populations are normal, and found that the logistic procedure is between 1/2 and 1/3 as efficient as the normal one. Amemiya and Powell (1983) compare the two methods when the independent variables are binary and independent, (here the logistic method is correct), using asymptotic theory. They concluded that the normal method did quite well in terms of correct classification probability, and

also in terms of the mean squared error of the log-odds ratio. These results might suggest that the normal procedure is worth using even if there is doubt about the normality of the populations, but in neither paper was interval estimation considered.

2.5.2 The Profile Likelihood

As mentioned in section 2.3.3 the profile likelihood can be used for non-normal distributions, it is only neccessary to show that conditions (i) to (vi) hold. Take for example two populations with exponential distributions:-

$$f_i(x)=\gamma_i^{-1}\exp(-x\gamma_i^{-1}),\ \gamma_i>0,\ i=1,2$$

Then

$$\Theta=\ln(\gamma_2)-\ln(\gamma_1)+x\gamma_2^{-1}-x\gamma_1^{-1}$$

$$\Theta=\mathbb{R}$$

$$\Phi=(\gamma_1,\gamma_2)$$

$$\Phi=\mathbb{R}^+\times\mathbb{R}^+$$

$$\ell(\phi)=-n_1\ln(\gamma_1)-n_2\ln(\gamma_2)-\gamma_1^{-1}\Sigma x_{1i}-\gamma_2^{-1}\Sigma x_{2i}$$

The problem $Q(\lambda)$ is

Maximise $\ell(\phi)-\lambda\Theta$ over $\phi\in\Phi$ where

$$\ell(\phi)-\lambda\Theta=-(n_1-\lambda)\ln(\gamma_1)-(n_2+\lambda)\ln(\gamma_2)$$
$$-\gamma_1^{-1}(\Sigma x_{1i}-\lambda x)-\gamma_2^{-1}(\Sigma x_{2i}+\lambda x)$$

Hence $\hat{\phi}_\lambda=(\hat{\gamma}_1(\lambda),\hat{\gamma}_2(\lambda))$ where

$$\hat{\gamma}_1(\lambda)=(\Sigma x_{1i}-\lambda x)/(n_1-\lambda)$$

$$\hat{\gamma}_2(\lambda)=(\Sigma x_{2i}+\lambda x)/(n_2+\lambda)$$

The constraints $\hat{\gamma}_i(\lambda)>0$, and the definition of $\hat{\gamma}_i(\lambda)$ as maximum (rather than minimum) likelihood estimators leads to the definition of Λ as

$$\Lambda=\{\lambda:\max(-n_2,n_2\bar{X}_2/x)<\lambda<\min(n_1,n_1\bar{X}_1/x)$$

Now

$$\theta_\lambda = \ln\hat{\gamma}_2(\lambda) - \ln\hat{\gamma}_1(\lambda) + x\hat{\gamma}_2(\lambda)^{-1} - x\hat{\gamma}_1(\lambda)^{-1}$$

and

$$p(\theta_\lambda) = \ell(\hat{\phi}_\lambda)$$

and it is only neccessary to find the two λ such that

$$(p(\hat{\theta}) - p(\theta_\lambda)) = -\tfrac{1}{2}\chi^2(1;.95)$$

and these λ will correspond to an approximate 95% confidence interval for θ.

CHAPTER THREE

A Simulation Study

All the approaches to interval estimation given in the previous chapter involve some form of approximation. Therefore, in order to assess and compare them it is neccessary to perform an extensive simulation study. We wish to investigate how close to the nominal confidence level of 95% each method gets, for differing sample sizes, population parameters μ_1,μ_2,Ω_1 and Ω_2, and values of the observed value x. It is clearly impossible to examine every possible combination of the above factors, but simplifying transformations mean that we can restrict our attention to certain subsets. Details are given in later sections.

2.1 The Equal Covariance Case

Using the transformation given in the previous chapter,Critchley and Ford (1985) show that the distribution of $\hat{\theta}(x)$ depends only on the three parameters θ,ϕ and Δ, where Δ is the Mahalanobis distance between the populations, and

$$\phi=\tfrac{1}{2}(\alpha_1(x)+\alpha_2(x)).$$

Therefore we need only consider the case $\Omega_1=\Omega_2=I_p$, $\mu_1=-\delta$, $\mu_2=\delta$ where $\delta=(\tfrac{1}{2}\Delta,0,\ldots,0)$, and we lose nothing by assuming $n_1\leqslant n_2$.

We follow the scheme of Critchley and Ford (1985), who examined the behaviour of PLUGALL and PLUGPART. Their results suggested that changing Δ or setting $n_1\neq n_2$ had little effect, so we only consider $n_1=n_2=n$, and $\Delta=1.6832$. This gives an optimal misclassification probability of 0.2. The x values we examine are shown in figure 3.1, their values being:-

$$A=(2.4866,0,\ldots,0)$$

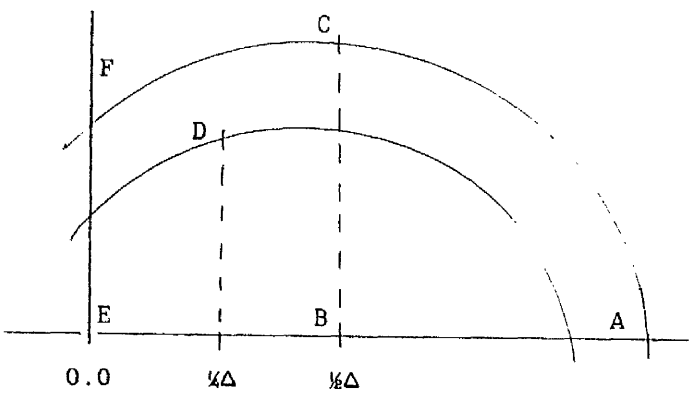

Figure 3.1

Location of points A to F

as used in the simulation study.

Outer semicircle indicates 90% contour

of underlying normal distribution, inner

semicircle indicates 50% contour.

B=(0.8416,0,...,0)

C=(0.8416,1.6450,0,...,0)

D=(0.4208,0.5265,0,...,0)

E=(0,...,0)

F=(0,1.4134,0,...,0)

We consider n=10,30 and 100, representing small, medium and large sample sizes. Initially we only consider dimension p=2.

The method of simulation was as follows:-

(1) Generate data from appropriate distributions

(2) Obtain an interval estimate estimate for $\theta(x)$ by whatever method is being considered.

(3) Repeat from (1) a large number NREP of times, and count the total number n_t of times the interval contains the true (known) value of $\theta(x)$.

An estimate of the true confidence, c, of the interval is then $\hat{c}=n_t/NREP$. n_t has a binomial distribution with parameter c. Hence $var(n_t)=c(1-c)NREP$ and so $var(\hat{c})=c(1-c)/NREP$. $\hat{c}$ is asymptotically normal, and so an approximate 95% confidence interval for c would be $\hat{c}\pm 1.96(var(\hat{c}))^{\frac{1}{2}}$. In order to get the estimate of c to within 1% we therefore want $1.96(var(\hat{c}))^{\frac{1}{2}}\approx .005$. $\hat{c}\approx .95$, and so the condition is

$$1.96(.95(1-.95)/NREP)^{\frac{1}{2}}\approx .005.$$

hence NREP=10000 is a reasonable sample size.

For the details of the random number generation see appendix two. To obtain the percentage points of Pearson curves for methods E3 and U2 an algorithm due to Davis and Stephens (1983) was used. For an algorithm to greatly speed up evaluation of methods E4 and U3, see appendix three. We also recorded the number of times the interval was wholly greater than the true

value $\theta(x)$, and the number of times it was wholly less than $\theta(x)$.

3.2 Results of Equal Covariance Simulations

The results for the equal covariance simulations are given in table 3.1.

3.3 Summary of Equal Covariance Results

As would be expected, all methods perform well when n=100. For n=10 and n=30, methods E3 and E5 both perform very well, consistantly getting very close to the nominal 95% confidence level, with perhaps E3, the Bayesian approach, being slightly better than E5, Davis' method. E2 and E4 are both rather less good,with E4 (profile likelihood) consistantly undershooting the target value, and E2 (PLUGPART), consistantly overshooting it. E1 (PLUGALL) is the poorest performer. The results for PLUGALL and PLUGPART are consistent with those of Critchley and Ford (1985).

The greater than/less than $\theta(x)$ figures show that most methods are not symmetrical, ie they are more likely to produce intervals that are greater than $\theta(x)$ than less than $\theta(x)$, or vice versa, and that different methods are asymmetrical in different ways. As might be expected for points E and F, where the true value of $\theta(x)$ is zero, all methods are approximately symmetrical. For the other points $\theta(x)$ is negative, and methods E1, E2 and E3 tend to give intervals greater than $\theta(x)$, ie nearer to zero. E1 is very bad in this respect, whilst E3 is nearly symmetrical. E4 and E5 show the opposite trend, with E4 being particularly bad. The worst offenders in respect of non-symmetry, E1, E2 and E4, are also the worst as regards empirical confidence. It is possible that a Bartlett type correction could improve methods E2 and E4 since they give intervals of consistantly high or low

Table 3.1 Results for the Equal Covariance Simulations

The table shows the empirical confidence of approximate 95% confidence intervals, obtained by the five methods E1,...,E5, for equal sample sizes of 10, 30 and 100, at the six observation points A to F. The figures in parentheses are the percentages of simulations in which the interval was wholly greater than/ less than $\theta(x)$.

(i) Sample Size = 10

	Method				
Point	E1	E2	E3	E4	E5
A	94 (6.3/0.0)	98 (1.1/0.7)	95 (2.6/2.0)	91 (1.3/7.2)	96 (1.3/2.3)
B	91 (8.5/0.0)	96 (3.1/0.5)	95 (2.9/2.4)	93 (1.5/6.0)	96 (1.0/2.5)
C	98 (2.2/0.0)	98 (1.6/0.5)	95 (3.0/2.2)	92 (2.5/5.2)	96 (1.9/2.5)
D	95 (5.3/0.0)	96 (3.4/0.5)	95 (3.1/2.1)	93 (2.4/4.5)	96 (1.5/2.6)
E	99 (0.4/0.3)	98 (1.0/0.9)	97 (1.6/1.8)	93 (3.5/3.1)	96 (2.2/2.2)
F	100 (0.0/0.0)	98 (0.9/0.7)	95 (2.7/2.4)	93 (3.8/3.7)	95 (2.6/2.3)

(ii) Sample Size = 30

	Method				
Point	E1	E2	E3	E4	E5
A	95 (4.3/0.7)	96 (2.4/1.8)	95 (2.7/2.7)	93 (1.6/5.0)	95 (2.2/2.6)
B	94 (5.1/0.6)	96 (3.1/1.4)	95 (2.9/2.6)	94 (1.8/4.1)	95 (2.1/2.5)
C	96 (3.1/0.9)	96 (2.4/1.7)	95 (2.7/2.4)	94 (2.2/3.6)	95 (2.1/2.7)
D	95 (4.2/0.7)	96 (3.0/1.4)	95 (3.0/2.3)	94 (2.8/2.8)	95 (2.0/2.6)
E	97 (1.3/1.4)	96 (2.0/1.8)	95 (2.2/2.5)	94 (2.8/2.8)	94 (2.9/2.8)
F	97 (1.5/1.7)	96 (1.8/2.1)	95 (2.7/2.6)	95 (2.7/2.7)	95 (2.5/2.7)

table 3.1(continued)

(iii) Sample Size = 100

	Method				
Point	E1	E2	E3	E4	E5
A	95 (3.6/1.4)	95 (2.4/2.4)	95 (2.6/2.6)	95 (1.6/3.4)	95 (2.4/2.6)
B	95 (3.8/1.3)	95 (3.3/2.1)	95 (2.7/2.6)	95 (2.1/3.2)	95 (2.4/2.5)
C	95 (3.0/1.7)	95 (2.7/2.2)	95 (2.7/2.7)	94 (2.5/3.1)	95 (2.4/2.7)
D	95 (3.5/1.4)	95 (3.1/1.8)	95 (2.7/2.5)	95 (2.4/2.9)	95 (2.3/2.6)
E	96 (2.1/2.0)	95 (2.5/2.2)	95 (2.5/2.4)	95 (2.7/2.4)	95 (2.2/2.6)
F	96 (1.9/2.0)	95 (2.4/2.1)	95 (2.6/2.8)	95 (2.9/2.6)	95 (2.3/2.4)

confidence. E1 probably would not benefit from such a correction.

3.4 Unequal Covariance Simulations

A similar simplifying transformation used in section 3.1 means we only need consider the case $\mu_1=\underline{0}p$, $\Omega_1=I_p$, and $\Omega_2=D$ where $D=diag(d_1,\ldots,d_p)$. Here we consider n=10, n=30 and n=100, again only for p=2 as in section 3.1.

We let D=diag(2,1) and consider the three cases

(a) $\mu_2=(1.6832,0)$

(b) $\mu_2=(1.1902,1.1902)$

(c) $\mu_2=(0,1.6832)$

so $(\mu_1-\mu_2)(\mu_1-\mu_2)^T=(1.6832)^2$ in each case.The x values considered for case (a) are those used in section 3.1, with 0.8416 added to the first coordinate in each case, since here $\mu_1=(0,0)$ rather than (-0.8416,0).For cases (b) and (c) we rotate these points through 45° and 90° respectively, so that the line $\mu_1\mu_2$ always corresponds to the line EBA. Note that the optimal misclassification probability is no longer 0.2, and is in fact different in each case. The three cases were chosen to be equivalent to changing only the variance-covariance matrix of population 2 in section 3.1, keeping the group means and points A to F constant.

3.5 Results of Unequal Covariance Simulations

The results for the unequal covariance simulations are given in table 3.2.

3.6 Summary of Unequal Covariance Results

Again, all methods performed well when n=100.For n=30 and n=10 U2, the Bayesian approach, is very good, with empirical

Table 3.2 Results for the Unequal Covariance Simulations

The tables show the empirical confidence of approximate 95% confidence intervals obtained by the three methods U1, U2 and U3, for equal sample sizes of 10, 30 and 100, at the six observation points A,...,F, in cases (a), (b) and (c). The figures in parentheses are the percentage of simulations in which the interval was wholly greater than/less than $\theta(x)$

CASE(a)

		Method		
Sample size	Point	U1	U2	U3
10	A	91 (9.3/0.0)	93 (2.5/4.3)	90 (1.1/8.6)
	B	98 (1.7/0.0)	94 (1.2/4.6)	90 (0.8/9.6)
	C	98 (1.6/0.0)	94 (1.2/4.6)	91 (3.0/6.3)
	D	99 (0.8/0.0)	94 (1.8/3.9)	91 (2.1/7.4)
	E	100 (0.0/0.0)	95 (1.9/2.9)	92 (3.1/5.4)
	F	100 (0.0/0.0)	94 (3.0/3.4)	91 (4.1/5.0)
30	A	93 (6.4/0.2)	94 (2.3/3.3)	94 (1.3/4.7)
	B	96 (4.1/0.4)	95 (1.6/3.3)	94 (1.4/5.1)
	C	96 (3.7/0.3)	95 (2.3/2.7)	94 (2.2/3.9)
	D	96 (3.4/0.9)	95 (1.8/3.1)	94 (1.8/4.6)
	E	96 (1.9/1.8)	95 (2.0/2.8)	94 (2.4/3.4)
	F	98 (1.6/0.8)	95 (2.7/2.7)	94 (2.8/3.5)

CASE(a)(continued)

sample size	point	Method U1	U2	U3
100	A	95 (4.3/1.0)	95 (2.2/2.5)	95 (1.8/3.7)
	B	96 (3.3/1.1)	95 (2.1/2.7)	95 (1.8/3.7)
	C	96 (3.3/1.2)	95 (2.3/2.6)	95 (2.3/3.1)
	D	96 (3.0/1.4)	95 (2.0/2.6)	95 (2.2/3.3)
	E	95 (2.4/2.3)	95 (2.2/2.4)	94 (2.6/3.0)
	F	96 (2.3/2.1)	95 (2.2/2.6)	95 (2.5/3.0)

CASE (b)

Sample size	Point	Method U1	U2	U3
10	A	94 (6.4/0.0)	93 (2.9/3.9)	91 (1.9/7.3)
	B	99 (1.4/0.0)	94 (2.4/4.0)	91 (2.6/6.5)
	C	100 (0.0/0.0)	93 (3.3/3.3)	91 (4.4/4.4)
	D	100 (0.0/0.2)	94 (3.3/3.1)	91 (5.0/4.1)
	E	99 (0.0/0.0)	94 (3.4/2.2)	91 (6.2/3.1)
	F	98 (0.0/1.6)	93 (3.5/3.3)	91 (6.4/2.9)
30	A	94 (5.9/0.2)	95 (2.4/2.9)	94 (1.6/4.6)
	B	96 (4.0/0.4)	95 (2.3/3.0)	94 (1.9/4.3)
	C	99 (0.8/0.7)	95 (2.7/2.7)	94 (3.2/3.0)
	D	97 (1.1/1.5)	95 (2.6/2.8)	94 (3.2/3.0)
	E	96 (1.2/2.6)	95 (2.7/2.4)	94 (3.7/2.6)
	F	96 (0.2/3.4)	95 (2.8/2.7)	94 (3.9/2.2)
100	A	95 (3.8/1.3)	95 (2.6/2.7)	95 (2.3/3.2)
	B	95 (3.2/1.6)	95 (2.5/2.7)	95 (2.2/3.0)
	C	96 (2.1/2.1)	95 (2.6/2.5)	94 (2.7/2.8)
	D	96 (1.8/2.4)	95 (2.5/2.5)	95 (2.8/2.5)
	E	96 (1.9/2.6)	95 (2.7/2.4)	95 (2.9/2.2)
	F	96 (1.3/3.2)	95 (2.7/2.4)	95 (2.9/2.4)

CASE(c)

Sample size	Point	Method U1	U2	U3
10	A	94 (6.3/0.0)	93 (3.1/3.9)	91 (2.1/7.3)
	B	98 (1.7/0.0)	94 (1.3/4.7)	90 (1.2/9.3)
	C	94 (5.9/0.0)	94 (2.6/3.9)	90 (2.2/7.6)
	D	99 (0.8/0.0)	94 (1.3/4.4)	91 (2.2/7.1)
	E	99 (0.2/0.3)	95 (2.6/2.6)	92 (4.2/4.0)
	F	99 (0.6/0.0)	95 (2.4/3.6)	91 (3.7/5.6)
30	A	95 (5.2/0.2)	95 (2.4/3.0)	94 (1.7/4.5)
	B	95 (4.2/0.5)	95 (1.7/3.6)	93 (1.4/5.4)
	C	95 (5.3/0.2)	95 (2.6/2.8)	94 (1.7/4.7)
	D	96 (3.0/0.7)	95 (1.9/3.2)	94 (1.7/4.7)
	E	96 (1.9/2.0)	95 (2.3/2.5)	94 (2.9/3.0)
	F	97 (2.2/0.7)	95 (2.6/2.7)	94 (2.4/3.7)
100	A	95 (4.2/1.1)	95 (2.5/2.9)	95 (1.8/3.2)
	B	95 (3.8/1.2)	96 (2.2/3.3)	95 (1.6/3.5)
	C	95 (4.2/1.2)	95 (2.4/3.0)	95 (2.0/3.2)
	D	95 (3.3/1.6)	95 (2.2/3.1)	95 (2.0/3.1)
	E	95 (2.4/2.1)	96 (2.2/2.3)	95 (2.6/2.3)
	F	95 (2.8/1.9)	95 (2.5/2.6)	95 (2.5/2.9)

confidence levels ranging from .93 to .95 for n=10. Next best is U3, the profile likelihood approach, with values around .90, and worst is U1. These results are consistant across the three different cases. The asymmetries noted in the equal covariance situation are also apparent here. U1 tends to give intervals too close to zero and the other two methods have the opposite tendency. Again the worst asymmetry corresponds to the worst empirical confidence.

3.7 Simulations in Higher Dimensions

The Bayesian approach performed very well in two dimensions, giving confidence levels very close to the nominal 95% level. Here we examine its performance for dimension p=5, assuming both equal and unequal covariances. All parameters are essentially as before, with three 'uninformative' variables being added to each point. ie A becomes (2.4866,0,0,0,0), μ_1 becomes (-0.8416,0,0,0,0) etc. We only consider case (c) in the unequal covariance case (D=(2,1,1,1,1)).

3.8 Results For Higher Dimensional Simulations

The results of the higher dimensional simulations are given in table 3.3.

3.9 Summary of Higher Dimensional Results

In the equal covariance case the empirical confidence levels are still very close to .95 for n=100 and n=30, and for n=10 they vary only from .92 to .96. In the unequal covariance case the performance is worse, with values from .82 to .91 for n=10. though for n=100 they are still close to .95, and for n=30 only range from .93 to .95. The asymmetry is still apparent, being

Table 3.3 Results of Simulations in Higher Dimensions

The table gives the empirical confidence of approximate 95% confidence intervals obtained by Rigby's method, for equal sample sizes of 10, 30 and 100, at the six observation points A,...,F, in both the equal and unequal covariance cases. The figures in parentheses are the percentage of simulations in which the interval was wholly greater than/less than $\theta(x)$.

Equal Covariance Case

	Sample size		
Point	10	30	100
A	93 (1.2/6.3)	94 (1.5/4.3)	95 (1.9/3.3)
B	93 (1.2/5.9)	95 (1.5/3.8)	95 (1.8/3.0)
C	92 (2.6/4.9)	94 (2.4/3.8)	95 (2.3/2.9)
D	94 (1.9/4.3)	95 (2.1/3.1)	95 (2.2/2.8)
E	96 (2.2/2.2)	96 (2.3/1.9)	95 (2.1/2.4)
F	93 (3.8/3.7)	94 (2.6/3.0)	95 (2.6/2.6)

Unequal Covariance Case

	Sample size		
Point	10	30	100
A	82 (2.4/15.3)	93 (1.5/5.8)	94 (1.9/3.9)
B	84 (0.7/15.8)	93 (1.0/6.5)	94 (1.6/4.4)
C	84 (2.1/14.0)	93 (1.4/5.8)	94 (2.0/3.6)
D	87 (1.6/11.3)	94 (1.4/5.1)	95 (1.9/3.6)
E	91 (4.6/4.9)	95 (2.4/2.4)	95 (2.4/2.6)
F	87 (4.4/8.6)	94 (2.3/3.7)	95 (2.3/3.0)

worst for small samples, and unequal covariances.

3.10 Conclusions from Simulation Results

The Bayesian approach due to Rigby (1982) performs consistently well, giving confidence levels close to the nominal 95% level in all but the extreme case of two samples of size 10 in five dimensions. Of the other methods, that due to Davis (1987) is comparable with Rigby's for the equal covariance case, but the other methods are all relatively poor. The greater than/less than $\theta(x)$ figures show that all the methods are asymmetrical, the profile likelihood and Davis' methods tending to give intervals too close to zero, and the other methods having the opposite tendency. It is noticeable that the poorer the method is, the worse the asymmetry appears to be.

In all that follows, we use Rigby's method for constructing confidence intervals for $\Theta(x)$

CHAPTER FOUR

Evaluation of a Discriminant Rule

4.1 Introduction

Having constructed a discriminant rule it is often neccessary to assess its value.For example, we may wish to compare one method of discrimination with another, or we may want to know if a particular variable is worth measuring. In the case of Conn's Syndrome, where we have eight variables measured on only 31 patients, it is neccessary to reduce the dimensionality of the problem. Here we will need to select a subset of the variables which perform 'adequately' and so will need some measure of the value of possible subsets.

There are several approaches to this problem in the literature. Here we review some of them and suggest some new ideas based on the concept of interval estimation for the log-odds. In section 4.2 we look at graphical methods, section 4.3 reviews so called 'discrete' methods such as error rate estimation, and in section 4.4 we look at 'continuous' methods. Section 4.5 contains a discussion of all the methods considered.

4.2 Graphical Methods

The simplest and most obvious method of assessing variables is to draw scatter plots of the training data, using different symbols for the groups. For example in figure 4.1 two of the Conn's Syndrome variables, sodium and potassium, are plotted. (for future reference we will call this subset 1). Here a '1' represents type 1 and a '2' type 2, with A,B,C,D being the four unknown cases. This type of plot does not actually show how well a discriminant rule is performing, but it gives some idea of the potential of the variables. Clearly there is quite good

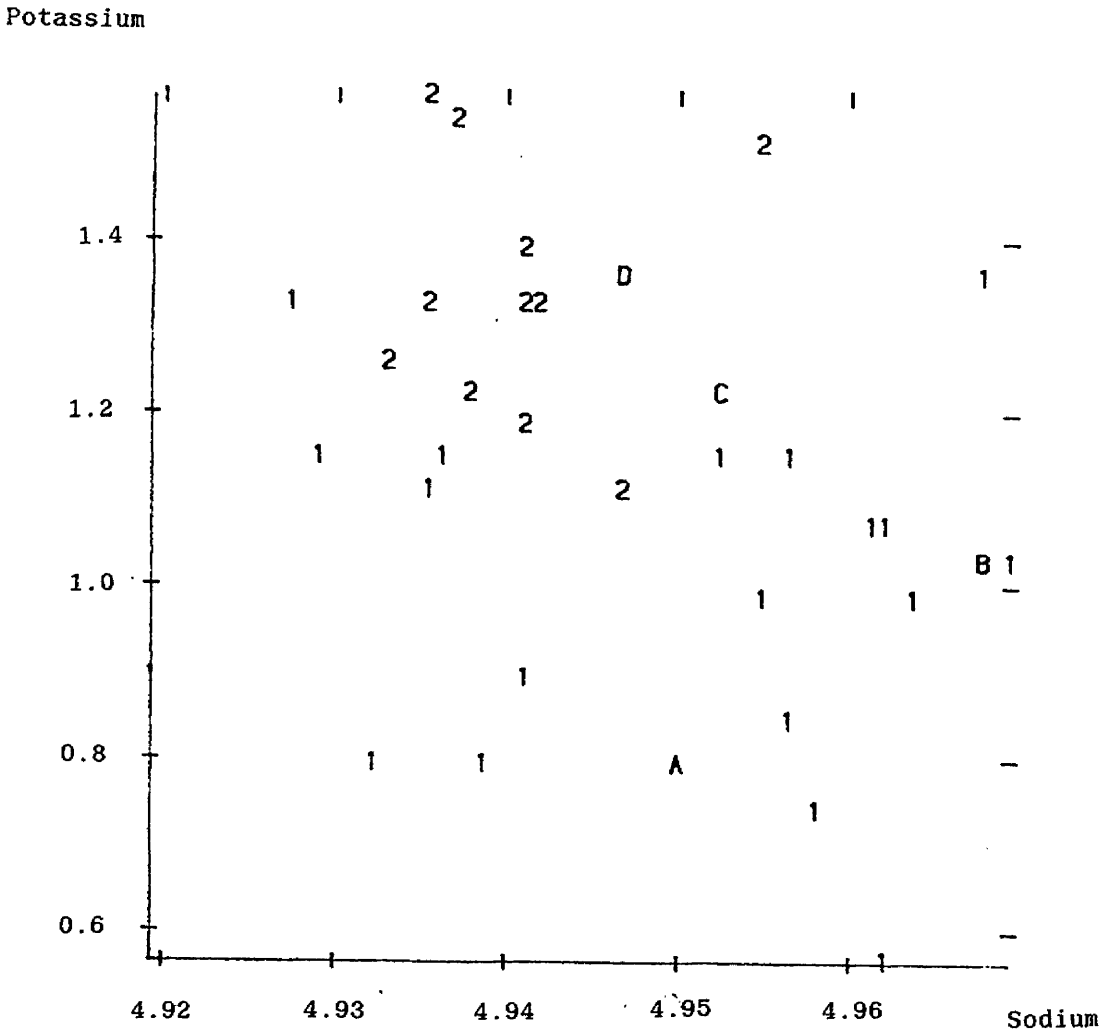

Figure 4.1
Conn's Syndrome
Scatter plot for subset 1

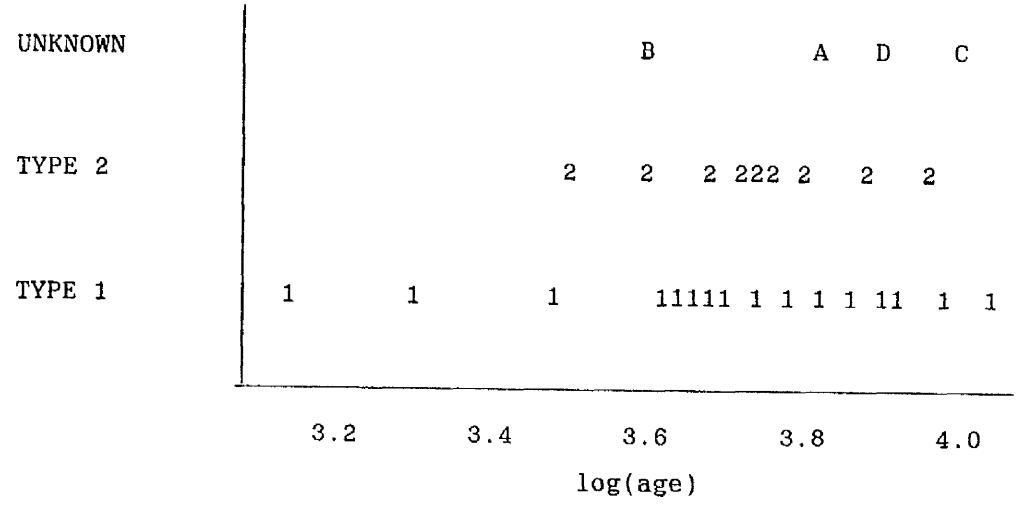

Figure 4.2

Conn's Syndrome: plot of age against type

separation of the groups and therefore it could be expected that a discriminant analysis using only subset 1 would be quite successful. A similar type of plot is shown in figure 4.2 for the single variable age (subset 2). Here it can be seen that there is considerable overlap between the groups, and so subset 2 will not provide a useful discriminant rule.

These types of plot are obviously only useful for subsets of 1 or 2 variables, but Sammon (1970) and Foley and Sammon (1975) consider a similar type of graph for higher dimensional data. Here the first coordinate displays the separation of the groups based on the linear discriminant function, and the second is orthogonal to the first, maximising the difference between the means of the samples subject to the orthogonality constraint.

Critchley and Ford (1985) suggest a similar plot in which the first coordinate is also proportional to the linear discriminant score, and the second gives a measure of the atypicality of each observation. Hence outliers can be easily picked out, and they suggest dividing their plot into the three areas where interval estimates for $\theta(x)$ are wholly positive, wholly negative or contain zero.

Chang (1987) extends the ideas of Sammon (1970) and Foley and Sammon (1975) to the case of unequal covariance matrices. His first coordinate displays separation mainly due to the sample means, and his second coordinate displays differences soley due to differences in the covariances. He shows that a straight line can be determined visually to show the degree and nature of group separation. Chang's plot for the four variables age, potassium, carbon dioxide and aldosterone is shown in figure 4.3. This shows good separation, though Chang points out that some of the apparent separation is due to sampling variation.

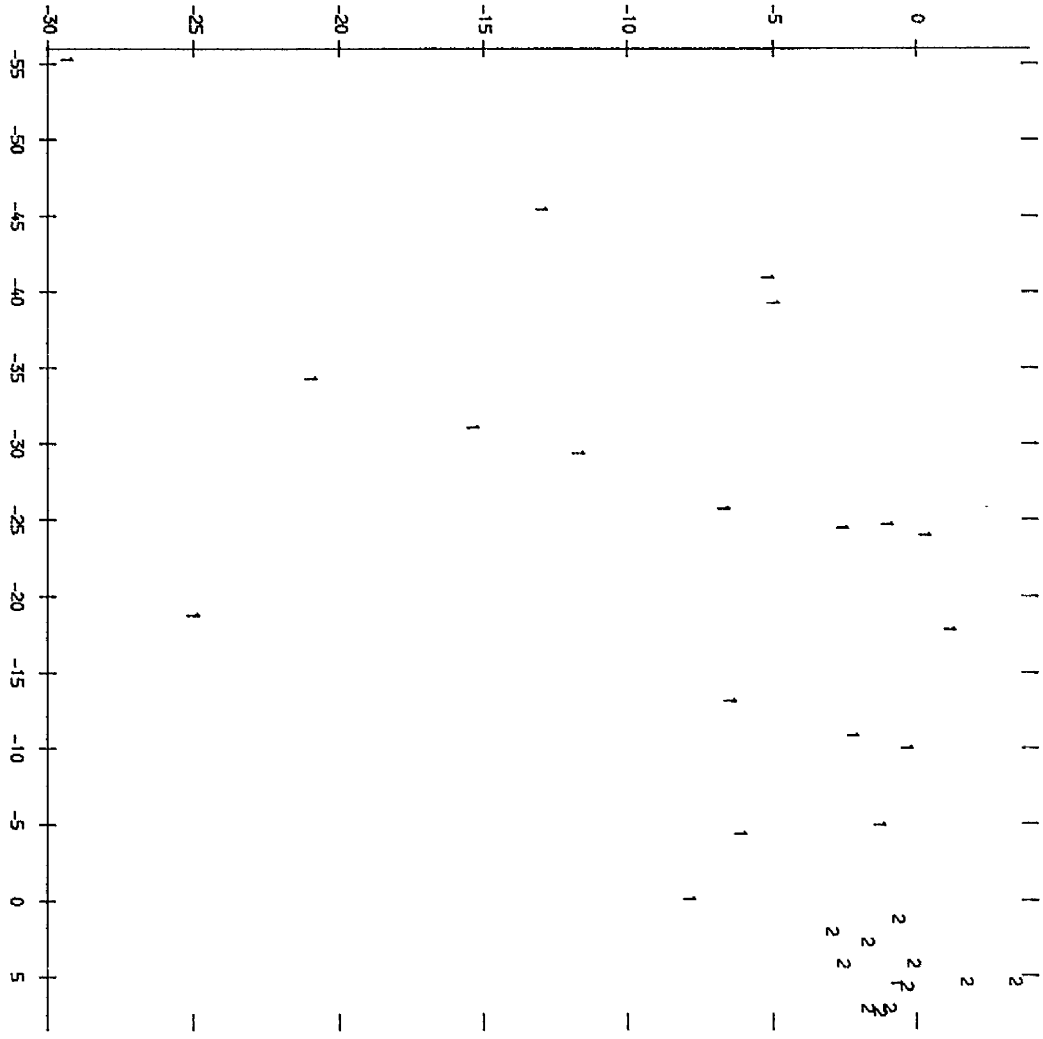

Figure 4.3
Conn's Syndrome
Chang's plot for subset 3

All of the plots mentioned so far are only capable of showing potential for discriminant analysis. In order to evaluate how well a rule performs in practice, Habbema et al (1978) suggest a plot of the estimated probabilities of group membership. A modification of their plot is shown in figures 4.4 (a),(b) and (c) for subsets 1,2 and 3. The minimum variance unbiased estimator of the probability is used here. If preferred its Jackknifed equivalent (see appendix four) can be used instead. These plots confirm our earlier impressions, that subset 2 is of little use, and that the other two appear to be quite good.

We have seen in the introduction that point estimates of these probabilities are not neccessarily reliable, and so we suggest a new type of plot which displays more fully the information we have on the discriminatory power of a subset of variables. Three of these plots are given in figures 4.5, 4.6, and 4.7. for subsets 1, 2 and 3 respectively.The 'x axis' on these plots represents the probability that an observation is of type 1. The 'y axis' is case number, so 1 to 20 (the solid lines) are type 1, and 21 to 32 (the broken lines) are type 2. The x's are the minimum variance unbiased estimators of the probability of type 1, the j's their jackknifed equivalents. The lines represent 95% confidence intervals for the true probabilities. Again we see that subset 2 is of little use, but perhaps it is now easier to choose between the other two subsets. The degree of uncertainty in classifying observations of type 2 seems to be greater with subset 3, and the estimated probability for case number 5 is badly wrong. This would suggest that of these three subsets, subset 1 is the best for discrimination.

All of the graphical methods discussed above are useful summaries of the data, but they need subjective judgement and

Figure 4.4

Conn's Syndrome

Probability plots

x=type 1 o=type 2

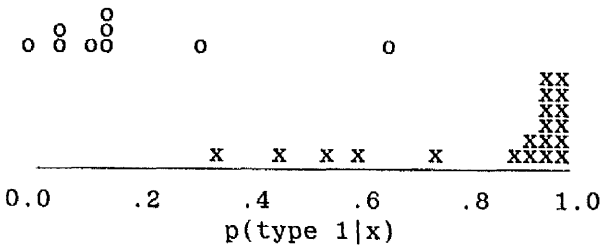

fig. 4.4(a): subset 1

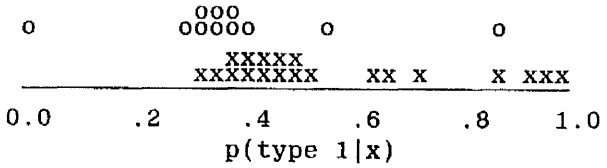

fig 4.4(b): subset 2

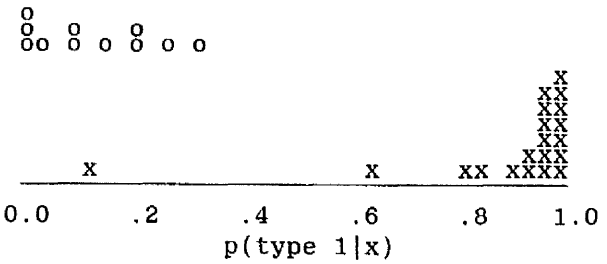

fig 4.4(c): subset 3

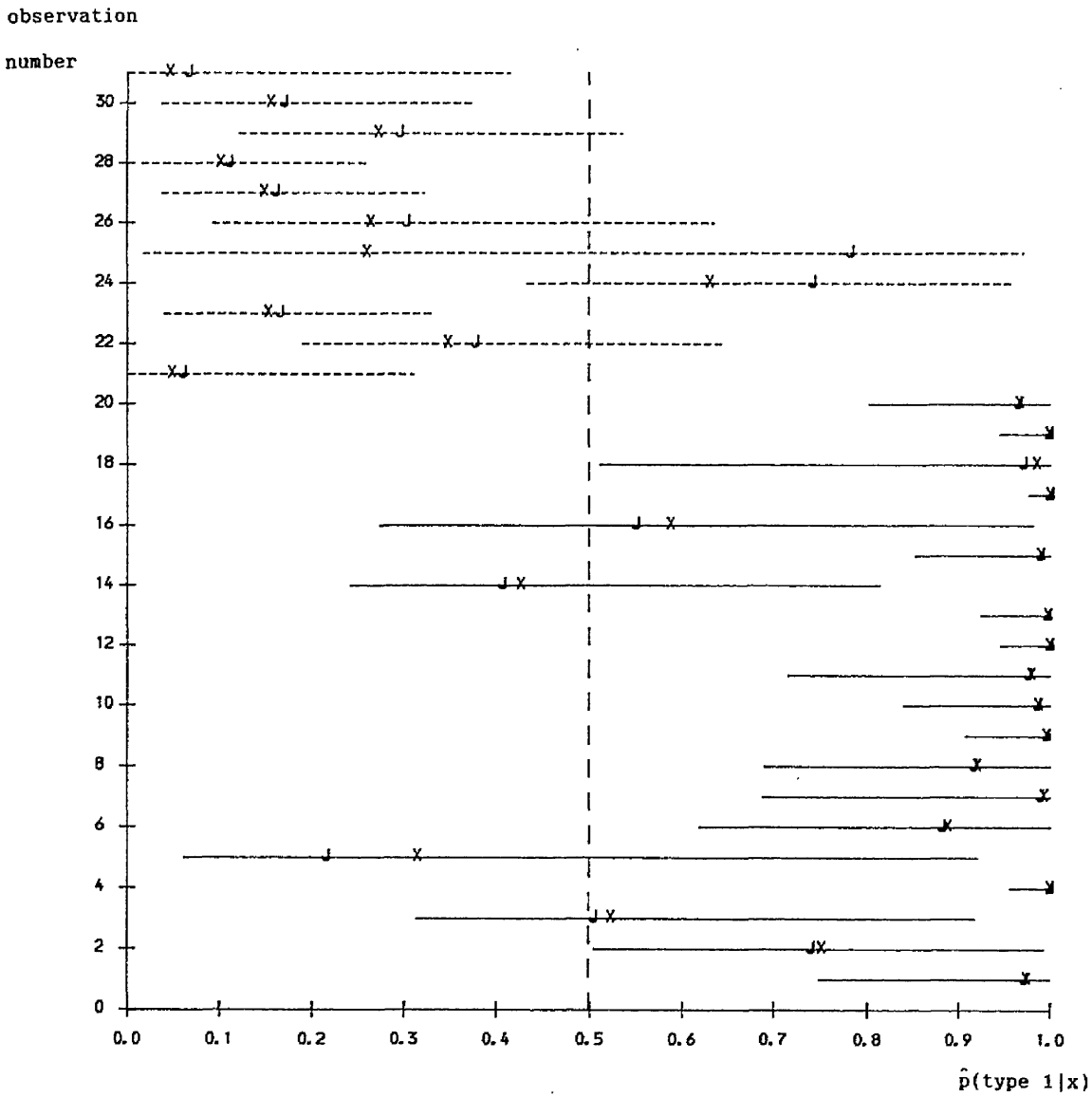

Figure 4.5
Interval plot for subset 1

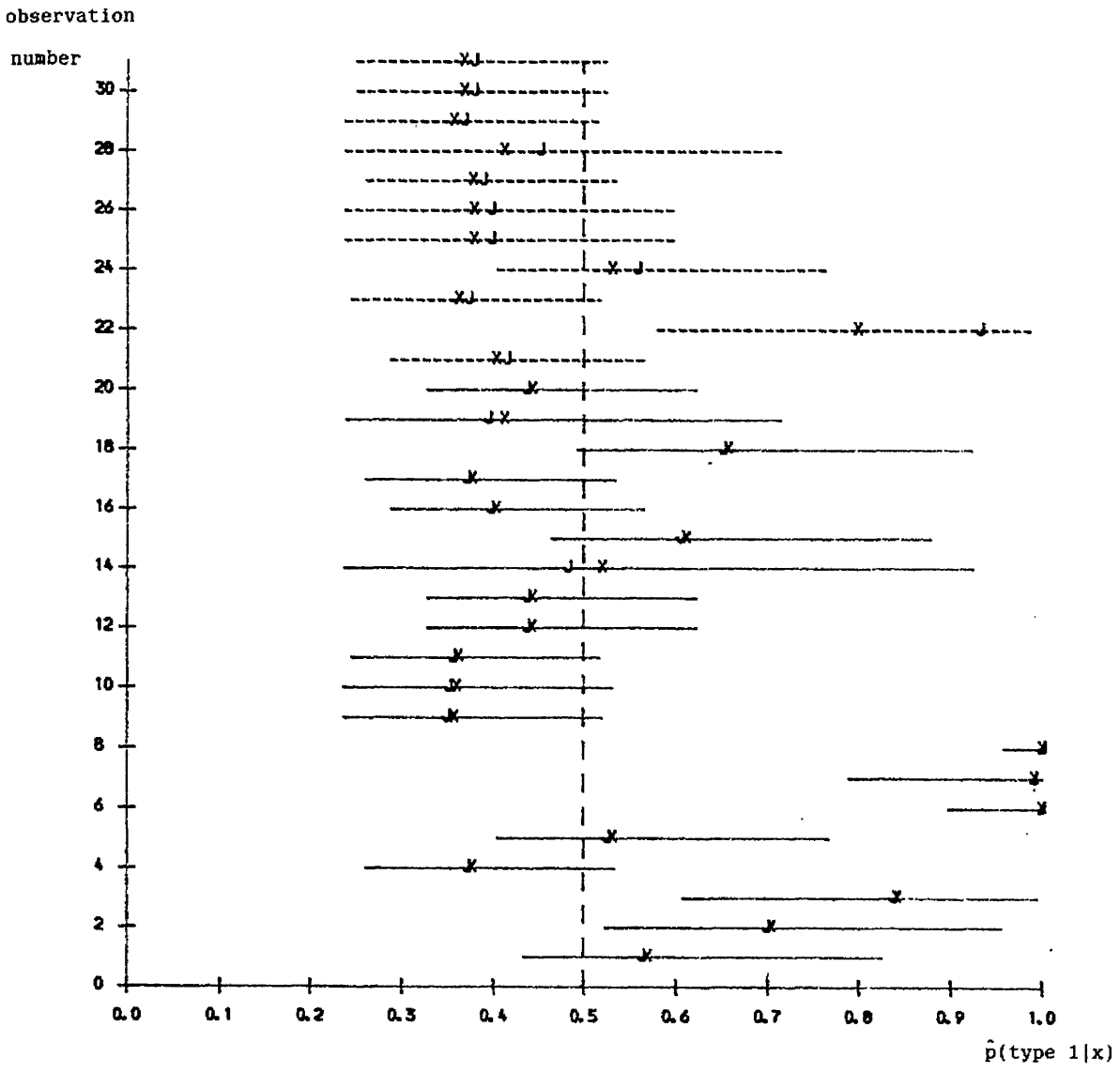

Figure 4.6
Interval plot for subset 2

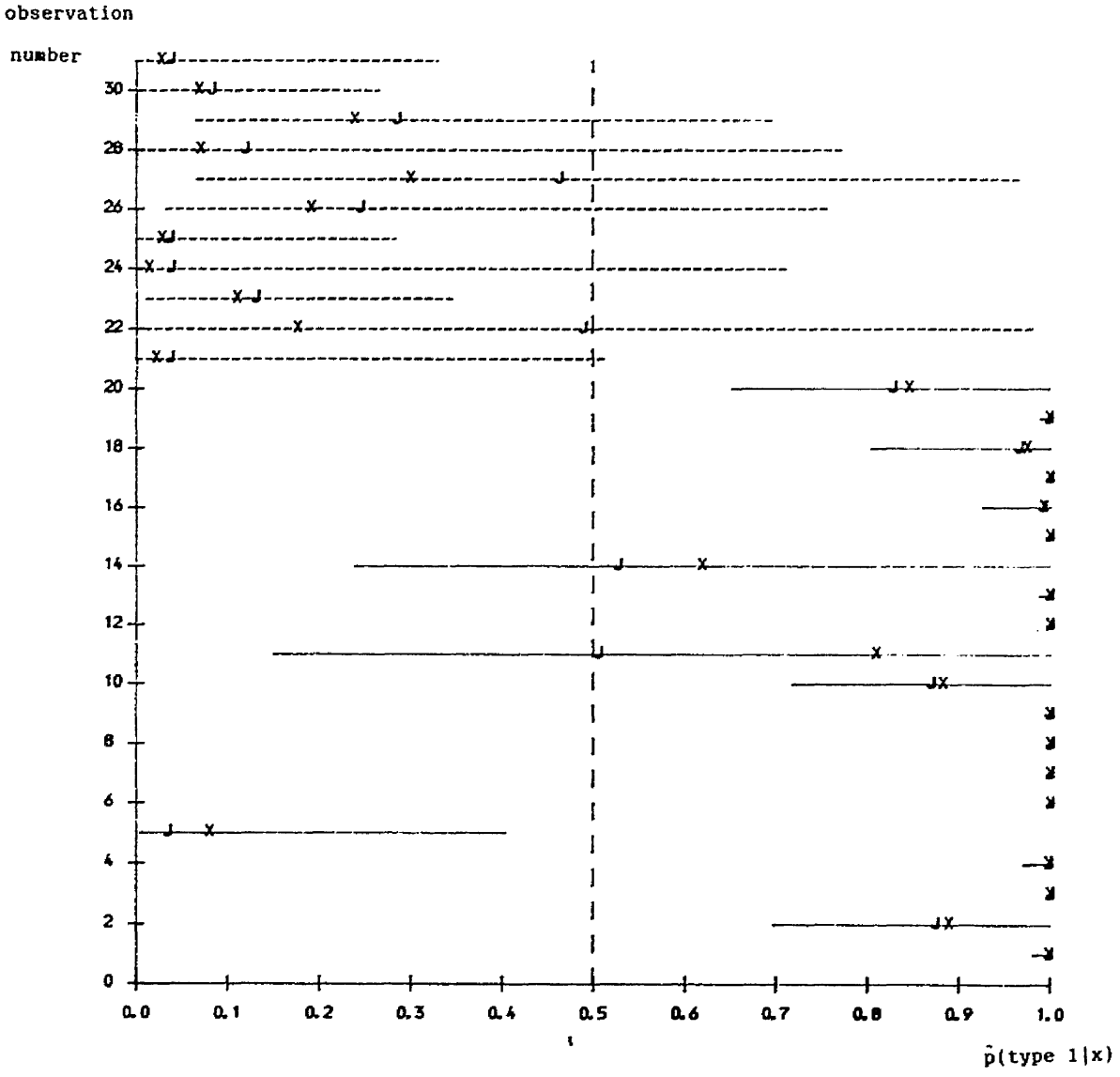

Figure 4.7
Interval plot for subset 3

would be of no use in an automated procedure. or in a manual procedure if a large number of comparisons had to be made. We therefore go on to quantitative rather than qualitative methods.

4.3 Discrete Methods

'Discrete' methods are so called because they generally summarise performance by classification of the data into type 1 or type 2, rather than using the actual estimated probabilities. They usually take the form of classification matrices and error rates. See Habbema et al (1978) and Hilden et al (1978a) for several useful methods. We go on to discuss the most important.

4.3.1 Forced Classification Matrices

A forced classification matrix is formed by allocating each observation in the training data to the type for which it has the highest estimated probability, regardless of the value of this probability. ie if $\hat{p}(\text{type } 1|x)>0.5$ then x is assigned to type 1, otherwise it is assigned to type 2. The forced classification matrix for subset 3 is given below :-

		assigned type	
		1	2
true type	1	19	1
	2	0	11

It can be seen that the discriminant rule is correctly classifying most of the training data.The version of this table obtained by using the jackknifed probabilities (appendix four) is in this case identical, but in general it will give a less optimistically biased evaluation of the rule. This is illustrated in the matrices for subset 1. The unjackknifed matrix is:-

		assigned type 1	2
true type	1	18	2
	2	1	10

The jackknifed version is

		assigned type 1	2
true type	1	18	2
	2	2	9

This suggests that this subset is not quite so good. For comparison the matrix for subset 2 is

		assigned type 1	2
true type	1	10	10
	2	9	2

and its jackknifed equivalent:-

		assigned type 1	2
true type	1	9	11
	2	9	2

The numbers in all of these tables can be given as percentages of the total from each type, if preferred.

4.3.2 Classification Matrices With Doubt

Habbema and Hermans (1974) suggest that if the estimated probability is near 0.5, the observation could be classified to another 'doubt' category. Exactly how close to 0.5 the probability needs to be is calculated on the basis of a loss function, where the loss for a 'doubt' assignment must be determined, and it must be less than the loss for an incorrect

assignment. Arbitrarily assigning an observation to the doubt category if its estimated probability is between 0.3 and 0.7, the classification matrix with doubt for subset 1 is :-

		assigned type		
		1	2	doubt
true type	1	16	0	2
	2	0	9	2

For subset 2 it is:-

		assigned type		
		1	2	doubt
true type	1	5	0	15
	2	1	0	10

For subset 3 it is:-

		assigned type		
		1	2	doubt
true type	1	18	1	1
	2	0	11	0

The large amount of doubt in classifying with subset 2 can clearly be seen here, but it is not easy to choose between subsets 1 and 3.

4.3.3 The Uncertainty Matrix

We suggest a matrix similar in form to that in the previous section, where an observation is classified to the 'uncertain' type if a 95% confidence interval for the probability of group membership contains 0.5. The matrix for subset 1 is:-

		assigned type		
		1	2	uncertain
true type	1	16	0	4
	2	0	6	5

The reasoning behind this type of matrix is rather different to that in section 4.3.2. There we were saying that in four cases out of 20 from type 1, using subset 1 the diagnosis was sufficiently unclear to defer making a decision. Here the interpretation is that in four out of 20 cases there is insufficient information on which to base any diagnosis at all.

The table for subset 3 is:-

		assigned type		
		1	2	uncertain
true type	1	17	1	2
	2	0	4	7

and for subset 2 it is

		assigned type		
		1	2	uncertain
true type	1	5	0	15
	2	1	0	10

We feel that these matrices give a useful indication of the value of a discriminant rule, by providing a good summary of the interval plots described earlier. In this case they indicate that subset 1 is rather better than subset 3, given the uncertainty in classifying observations from group 2 using subset 3, and the totally incorrect classification of observation 5

4.3.4 Error Rate Estimation

A frequently used method of evaluating a discriminant rule is the error rate, or more often the non-error rate (NER). This is simply the proportion of observations correctly classified, assuming that each observation is classified into type 1 or type 2. It can be regarded as an estimate of the probability of correctly classifying a future observation drawn at random from

one of the two populations. A fuller discussion of error rate estimation is given in chapter 5, here we give a method based on classification matrices for completeness.

The most obvious estimate of the NER is the proportion of the training data correctly classified, ie the trace of the forced classification matrix divided by the total sample size. For subset 1 it is 28/31, for subset 2 30/31 and for subset 3 12/31. These are optimistically biased and the less biased jackknifed equivalents are 27/31, 30/31 and 11/31. In chapter 5 we suggest a more reliable estimator based on interval estimation, along with a discussion of other methods.

4.4 Continuous Methods

Continuous methods are those where the actual estimated probabilities are taken into account. For example a correct diagnosis with an estimated probability of 0.99 is given more credit than one with an estimated probability of 0.51. There are many possible statistics used to evaluate discriminant rules in a continuous manner, and Hilden et al (1978b) give a good discussion of several.

Perhaps the most popular is the Briers score, B, where

$$B=(2/N)\Sigma(1-p_i)^2$$

in the case of two groups, and p_i is the estimated probability that observation i belongs to its true group. Clearly $0 \leq B \leq 2$, and the nearer B is to 0, the better the discriminant rule. A similar statistic is the logarithmic score L, where

$$L=(1/N)\Sigma\ln(p_i)$$

4.5 Discussion

As mentioned earlier, the graphical procedures outlined in section 4.1 are useful for getting subjective impressions of the value of a set of variables for discrimination, but in most situations it will be more useful to have a quantitative measure of discriminatory power. It will therefore be necessary to choose between a discrete or continuous method.

Shapiro (1977) gives a good comparison of the two types of procedure. The classification matrices and error rates have the advantage of simplicity and ease of interpretation, and are very popular for that reason. Their main drawback is lack of sensitivity, they do not take into account the magnitude of an error. In the situation where a decision has to be made regardless of the doubt, then it does not matter whether the estimated probability is .51 or .99, but this is unusual in practice. However, these methods will continue to be used, and so we feel it is useful to acknowledge the doubt about the true value of the probabilities, and would recommend the uncertainty matrix proposed in section 4.3.3.

If the discriminant rule is only to be used as a guide, for example to aid a clinician's diagnosis, then it is more important to consider the actual estimated probabilities. Here a continuous method of evaluation is appropriate, and for a full discussion see Hilden et al (1978b). It should be noted that none of the statistics they suggest take into account the uncertainty in estimating the parameters involved.

CHAPTER FIVE

Error Rate Estimation

5.1 Introduction

Error rate estimation was mentioned in the previous chapter as a method of evaluating a discriminant rule. Although it is not always the most appropriate method, it is an often useful, and certainly very common guide to how well a rule is performing. Therefore, in this chapter we review some of the many possible error-rate estimators, and suggest some new ones. We will restrict our attention to the case of two group linear discrimination with multivariate normal populations, since most of the literature considers this situation. Toussaint (1974) gives an extensive bibligraphy of work done up to that date and Hand (1986) updates this. There have been many comparative studies of error rate estimators in recent years (Lachenbruch and Mickey (1968), Sedransk and Okamoto (1971), Sorum (1971,1972,1973), McLachlan (1974), Glick (1978), McLachlan (1980), Snapinn and Knoke (1984,1985)). Unfortunately they differ from each other both in terms of the types of eror rate they are estimating, and the criteria by which they are assessed. In this chapter we have decided to follow the work of Snapinn and Knoke (1984,1985), in order to make some comparisons with earlier work possible. Therefore, the error rate we will be concerned with is that which is usually termed the conditional error rate, and our criterion for assessing an estimator is its unconditional mean squared error (UMSE).

5.2 Notation

Following the definitions of chapter two, if x is our unknown observation from population pop_1 or pop_2, and $\theta(x)$ is the (true) log odds ratio,then we classify x as follows:-

If $\theta(x)>0$ classify x into pop_1

If $\theta(x)\leqslant 0$ classify x into pop_2

We denote this rule by r.

In general $\theta(x)$ is unknown, and is estimated by $\hat{\theta}(x)$, giving the rule $\hat{r}$:-

If $\hat{\theta}(x)>0$ classify x into pop_1

If $\hat{\theta}(x)\leqslant 0$ classify x into pop_2

There are three distinct error rates associated with these rules. Let $e_i(r)$ be the probability that an observation drawn at random from pop_i is misallocated by rule r:-

$$e_1(r)=p(\theta(x)\leqslant 0 | x\epsilon pop_1)$$

$$e_2(r)=p(\theta(x)>0 | x\epsilon pop_2)$$

These are the optimal error rates, ie the error rates that would occur if $\theta(x)$ were known. From now on we will assume x is drawn from pop_1, and only consider e_1, dropping the subscript. The optimal error rate therefore becomes

$$e(r)=p(\theta(x)\leqslant 0)$$

The conditional error rate is defined as the probability that x is misclassified when the rule $\hat{r}$ is used,

$$e(\hat{r})=p(\hat{\theta}(x)\leqslant 0 | T)$$

Where T is the training data. This is the error rate conditional on the parameters, ie the one that will occur in practice using the rule $\hat{r}$ defined by the training data.

The expected error rate is the expectation of the conditional error rate over all training samples, defined as

$$E(e(\hat{r}))=E(p(\hat{\theta}(x)\leqslant 0 | T))$$

Note that the optimal error rate is a function only of the population distributions. The expected error rate is a function of the population distributions and the training sample sizes, and the conditional error rate is a function of the population

distributions and the particular training samples chosen. It is this error rate that we will be interested in.

Let $\hat{e}$ be an arbitrary estimator of the conditional error rate. Our criterion for assessing $\hat{e}$ is its UMSE, defined as

$$UMSE=E(\hat{e}-e(\hat{r}))^2$$

We assume multinormal populations $N(\mu_i,\Omega)$ (i=1,2). If we estimate $\Theta(x)$ by its minimum variance unbiased estimator (chapter two) then (assuming equal prior probabilities and equal training sample sizes $n_1=n_2=n$ for simplicity) our rule is equivalent to Anderson's (1958) linear discriminant rule:-

assign x to pop_1 if $W(x)>0$

" " " pop_2 otherwise

where $W(x)=(n_2+n_2-p-3)(x-\frac{1}{2}(\bar{X}_1+\bar{X}_2))^TS^{-1}(\bar{X}_1-\bar{X}_2)$,

Conditional on the training samples, W(x) has a univariate normal distribution, and the conditional error rate is the probability that $W(x)<0$ given by

$$e(\hat{r})=\Phi\left[\frac{-W(\mu_1)}{(n_1+n_2-p-3)((\bar{X}_1-\bar{X}_2)^TS^{-1}\Omega S^{-1}(\bar{X}_1-\bar{X}_2))^{\frac{1}{2}}}\right] \qquad (5.1)$$

where $\Phi(t)$ is the standard normal distribution function evaluated at t.

If we have unequal prior probabilities π_1 and π_2 and unequal sample sizes, then the rule is

assign x to pop_1 if $U(x)>0$

" " " pop_2 otherwise

where $U(x)=(n_1+n_2-p-3)(\bar{X}_1-\bar{X}_2)^TS^{-1}(x-\frac{1}{2}(\bar{X}_1+\bar{X}_2))$

$$+\tfrac{1}{2}p(n_1^{-1}-n_2^{-1})+\log(\pi_1/\pi_2)$$

This is equivalent to the rule:-

assign x to pop_1 if $W(x)>-\frac{1}{2}p(n_1^{-1}-n_2^{-1})+\log(\pi_1/\pi_2)$

" " " pop_2 otherwise

and the obvious alterations to the conditional error rate should

be made.

5.3 Estimators of $e(\hat{r})$

5.3.1 The Resubstitution Estimator $\hat{e}^R$

Let x_j $(j=1,\ldots,n)$ be the training sample from pop_1, and define a counting function $h_R(.)$ where

$h_R(x_j)=1$ if $W(x)\leqslant 0$

" $=0$ otherwise.

The resubstitution estimator is defined as

$$\hat{e}^R=(1/n)\Sigma h_R(x_j).$$

It measures how well the rule performs on the training data. It is well known that $\hat{e}^R$ has an optimistic bias due to the fact that it tests the rule on the data from which it was derived.

5.3.2 The Leave One Out Estimator $\hat{e}^L$

In order to reduce the bias of the resubstitution estimator, Lachenbruch (1967) suggested the leave one out estimator, sometimes called the jackknifed estimator. Here each observation is omitted in turn and the discriminant rule is calculated using the remaining data. A new counting function h_J is defined in terms of this rule, and the estimator is defined as

$$\hat{e}^L=(1/n)\Sigma h_J(x_j)$$

This is less biased than $\hat{e}^R$ but has a large variance (Glick (1978)). See appendix four for further details.

5.3.3 The Bootstrap Estimator $\hat{e}^B$

Efron (1983) suggested several estimators of $e(\hat{r})$ based on the bootstrap. The principle is to use the bootstrap to estimate the bias of the resubstitution estimator, and then to subtract this from $\hat{e}^R$. The best of his estimators was the so called '.632

estimator' $\hat{e}^{.632}$ which turns out to be

$$\hat{e}^{.632}=.368\hat{e}^{R}+.632\epsilon$$

where ϵ is the observed error rate for the points in the training data which do not occur in each bootstrap replication. This is a very time consuming estimator to use in a simulation study, and so we follow Snapinn and Knoke (1985) in using an 'ideal' bootstrap estimator $\hat{e}^{B}$. This is simply the resubstitution estimator minus its true bias. This should provide an approximate upper bound to the performance of the true bootstrap, although it cannot be used in practice.

Efron's (1983) simulation results suggested that in some situations, the .632 estimator could in fact slightly outperform this 'upper bound', due to negative correlation between $\hat{e}^{R}$ and ϵ. This was an unusual occurence though, and we feel that $\hat{e}^{B}$ is still a useful guide to the performance of $\hat{e}^{.632}$.

It is easy to calculate $\hat{e}^{B}$ once all the simulations have been completed, since

$$\hat{e}^{B}=\hat{e}^{R}-(E(\hat{e}^{R})-E(e(\hat{r}))$$

The criterion used to compare estimators is UMSE defined as

$$UMSE=E(\hat{e}-e(\hat{r}))^{2}$$

hence

$$UMSE_{B}=UMSE_{R}-(E(\hat{e}^{R})-E(e(\hat{r})))^{2}$$

and $E(\hat{e}^{R})$ and $E(e(\hat{r}))$ can be estimated from the sample means of $\hat{e}^{R}$ and $e(\hat{r})$ over all of the simulations.

5.3.4 The Smoothed Resubstitution Estimator $\hat{e}^{S}$

Some smoothed resubstitution estimators were defined by Glick (1978) in the univariate case, and generalised by Snapinn and Knoke (1985).Their best estimator uses a normal smoothing

function g(.) where

$$g(x)=\Phi(-W(x)/bD)$$

where $D^2=(n_1+n_2-2)(\bar{X}_1-\bar{X}_2)^T S^{-1}(\bar{X}_1-\bar{X}_2)$, the estimated Mahalanobis distance between the populations, and b is the smoothing parameter. The estimator is now

$$\hat{e}^S=(1/n)\Sigma g(x_j)$$

Snapinn and Knoke (1985) consider several possibilities for b. The best, called by them the NS estimator, involves obtaining an approximate expression for $E(\hat{e}^S|\bar{X}_1,\bar{X}_2,S)$ as a function of b, and equating it with $E(\hat{e}^P|\bar{X}_1,\bar{X}_2,S)$, where $\hat{e}^P$ is a parametric estimator given in the next section. The value of b obtained by this method is

$$b=\left[\frac{(p+2)(n_1-1)+(n_2-1)}{(n_1(n_1+n_2-p-3))}\right]^{1/2}$$

where p is the dimension of the population distributions. If $n_1=n_2=n$ then

$$b=\left[\frac{(n-1)(p+3)}{n(2n-p-3)}\right]^{1/2}$$

Note that as b approaches zero, $\hat{e}^S$ approaches $\hat{e}^R$, and as b approaches ∞, $\hat{e}^S$ approaches 0.5.

5.3.5 The Parametric Estimator $\hat{e}^P$

The simplest parametric estimator of $e(\hat{r})$ is obtained by plugging estimates of μ_1 and Ω into equation 5.1. This was first suggested by Fisher (1936). This is known to be optimistically biased, and several variations have been proposed to correct this bias. We will consider the method of Lachenbruch and Mickey (1968), in line with Snapinn and Knoke (1984). The 'plug in' estimator is defined as

$$\hat{e}^{\text{plug in}}=\Phi(-D/2)$$

where D^2 is the estimated Mahalanobis distance between the populations. This is modified to be $\hat{e}^P$ where

$$\hat{e}^P=\Phi(-DS/2)$$

where $DS^2=(n_1+n_2-p-3)D^2/(n_1+n_2-2)$

Page (1985) notes that this is equivalent to plugging the unbiased estimator of Ω^{-1} (rather than Ω) into equation 5.1. She also suggests several other parametric estimators, all of which have broadly similar performance to $\hat{e}^P$.

5.3.6 The Interval Estimation Method $\hat{e}^I$

We propose a new method of error rate estimation based on interval estimation for the log odds ratio. Recall the 'uncertainty matrix' of chapter four. Here we are classifying observations from the training data as 'uncertain' if a 95% confidence interval for the log odds ratio contains zero. In other words we are saying that if the interval contains zero we do not have enough information on which to base a decision. Hence we define a new counting function h_I where

h_I=1 if a 95% confidence interval for $\theta(x)$ is wholly negative

=½ " " " " " " " contains zero

=0 " " " " " " " is wholly positive.

The estimate of the error rate is then

$$\hat{e}^I=(1/n)\Sigma h_I(x_j)$$

This is in fact another form of smoothed resubstitution, which will approach $\hat{e}^R$ if the groups are well separated, and approach ½ if they are identical. In this respect $\hat{e}^I$ is similar to $\hat{e}^S$, though here the degree of smoothing depends directly on the amount of uncertainty involved in classification of the samples, rather than only their sizes, dimensionality and separation.

5.4 A Simulation Study

5.4.1 Methods

in order to estimate the UMSE for each estimator, we performed a simulation study. Sample sizes of 10 and 25 were considered for dimensions of 1, 3 and 5. 5000 replications were performed in each case. Data were generated from populations pop_1 and pop_2 with multinormal distributions $N_p(\underline{0}_p, I)$ and $N_p((\Delta,0,\ldots,0)^T, I)$, where Δ varied from 0 to 3. Δ is of course equal to the square root of the Mahalanobis distance between the populations.

For each simulation the error rate $\hat{e}$ was estimated by each method, and its squared error $(\hat{e}-e(\hat{r}))^2$ calculated. This was then averaged over all simulations, to obtain an estimate of the UMSE for each value of Δ separately. The only exception to this procedure was the estimation of $UMSE_B$, which had to be performed after all the simulations were completed since the calculations require $UMSE_R$, $E(\hat{e}^R)$ and $E(e(\hat{r}))$, as described in section 5.3.3.

5.4.2 Results of Simulations

The results of the simulations are given in figures 5.1 to 5.6. The curves have been fitted by the GHOST (1985) graphics package, routine CURVEO, to the estimated UMSE at Δ=0, 0.5, 1, 1.5, 2, 2.5 and 3.

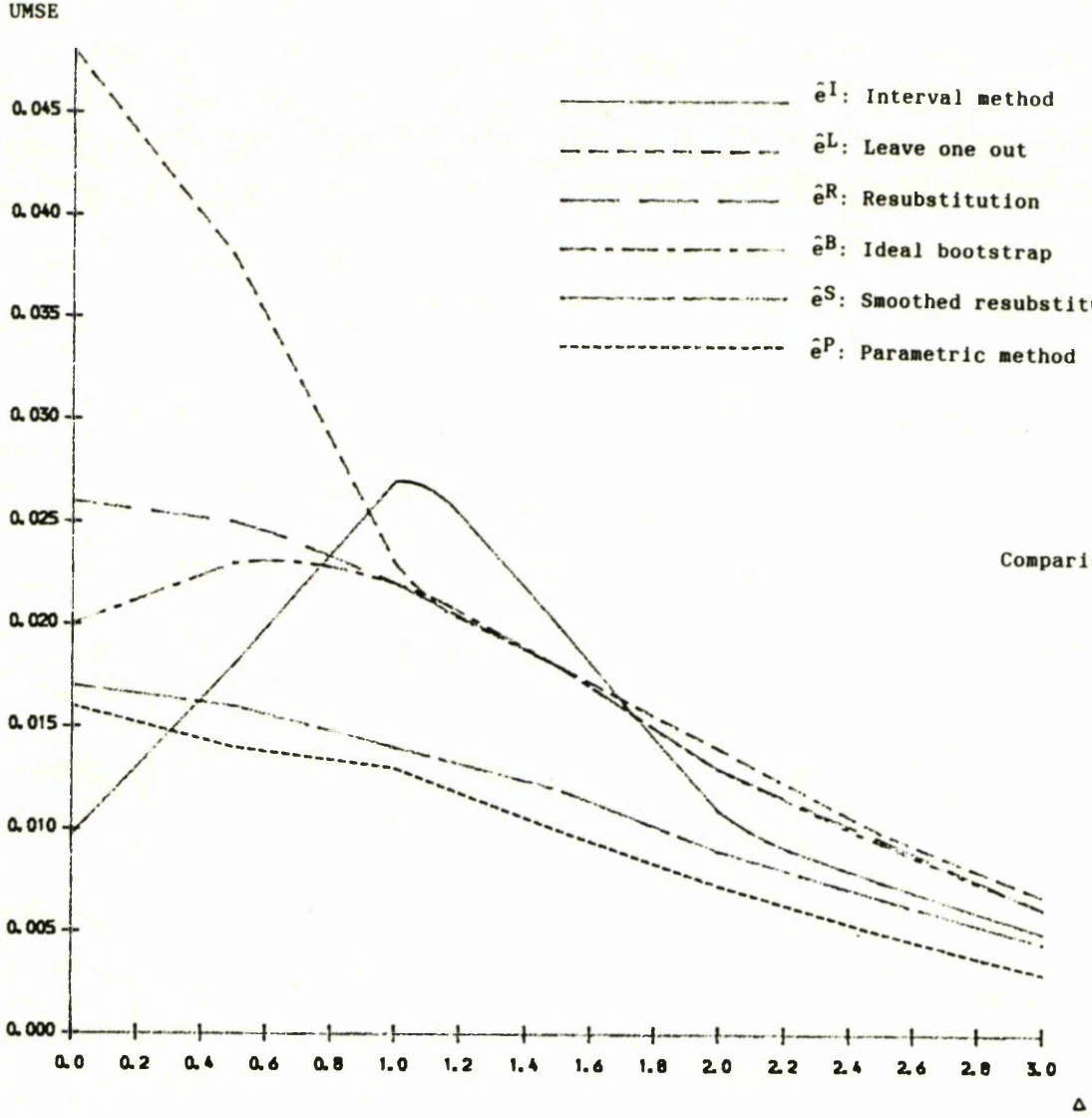

Figure 5.1
Comparison of error rate estimators
Sample size = 10
dimension = 1

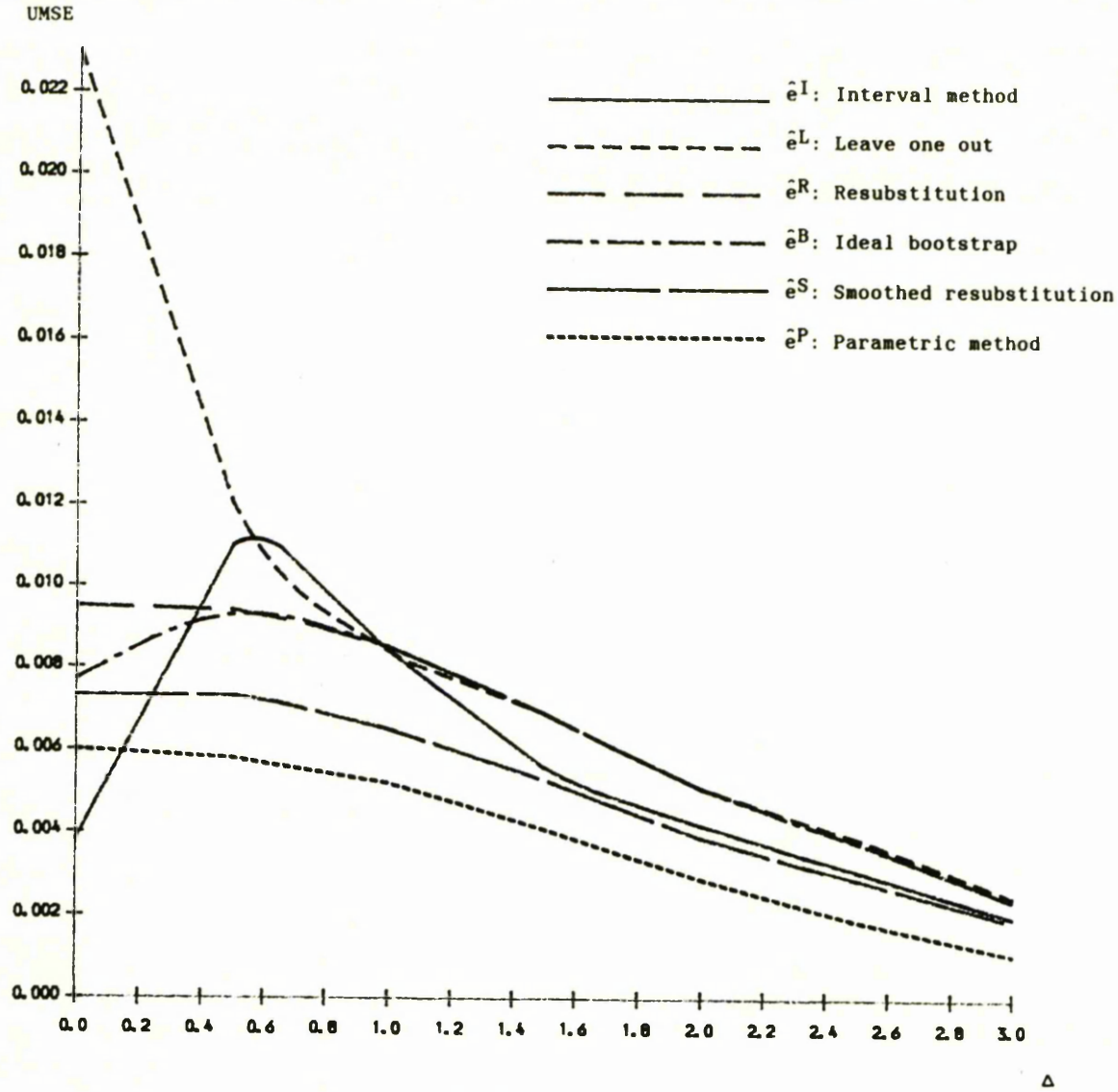

Figure 5.2
Comparison of error rate estimators
Sample size = 25
dimension = 1

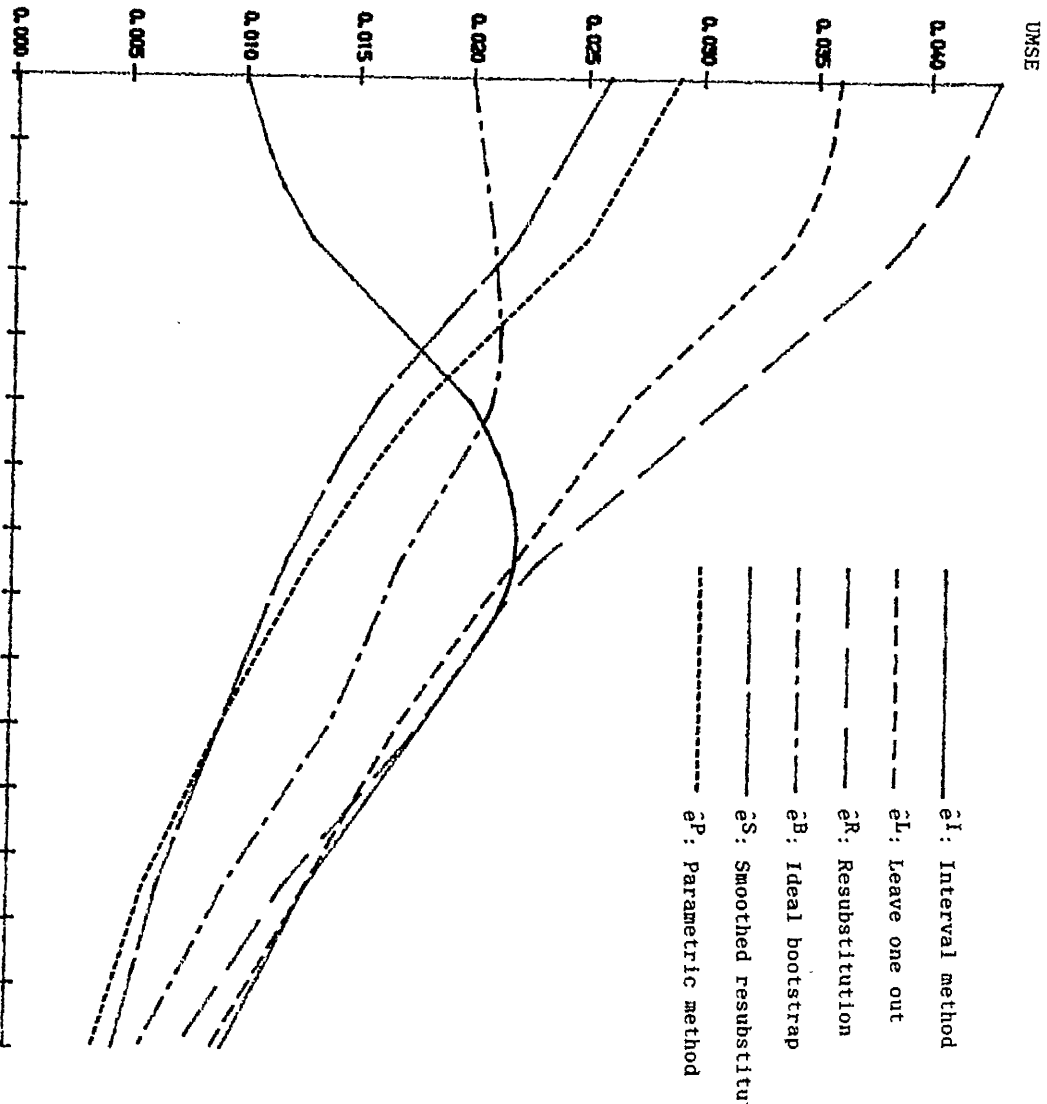

Figure 5.3
Comparison of error rate estimators
Sample size = 10
dimension = 3

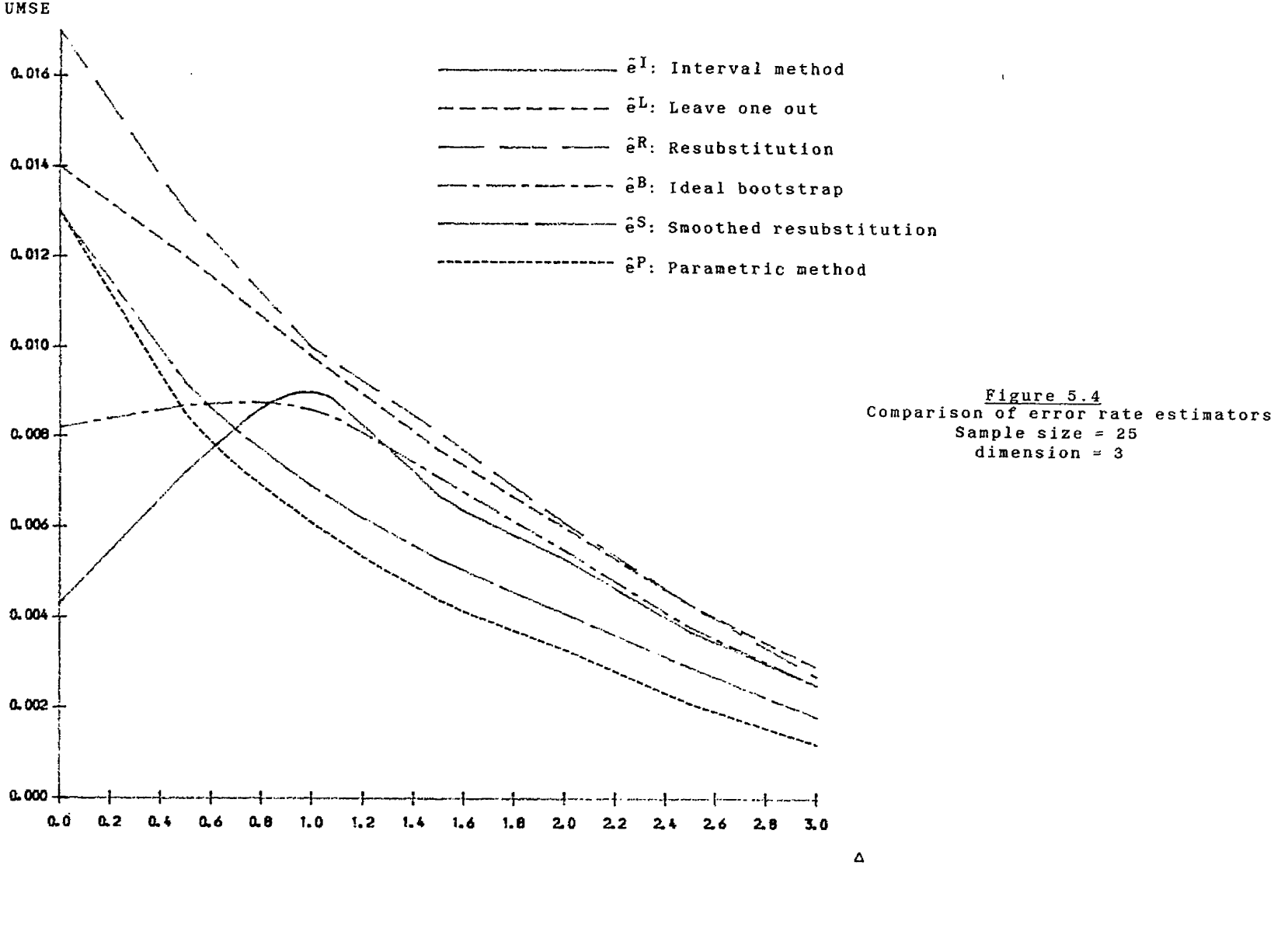

Figure 5.4
Comparison of error rate estimators
Sample size = 25
dimension = 3

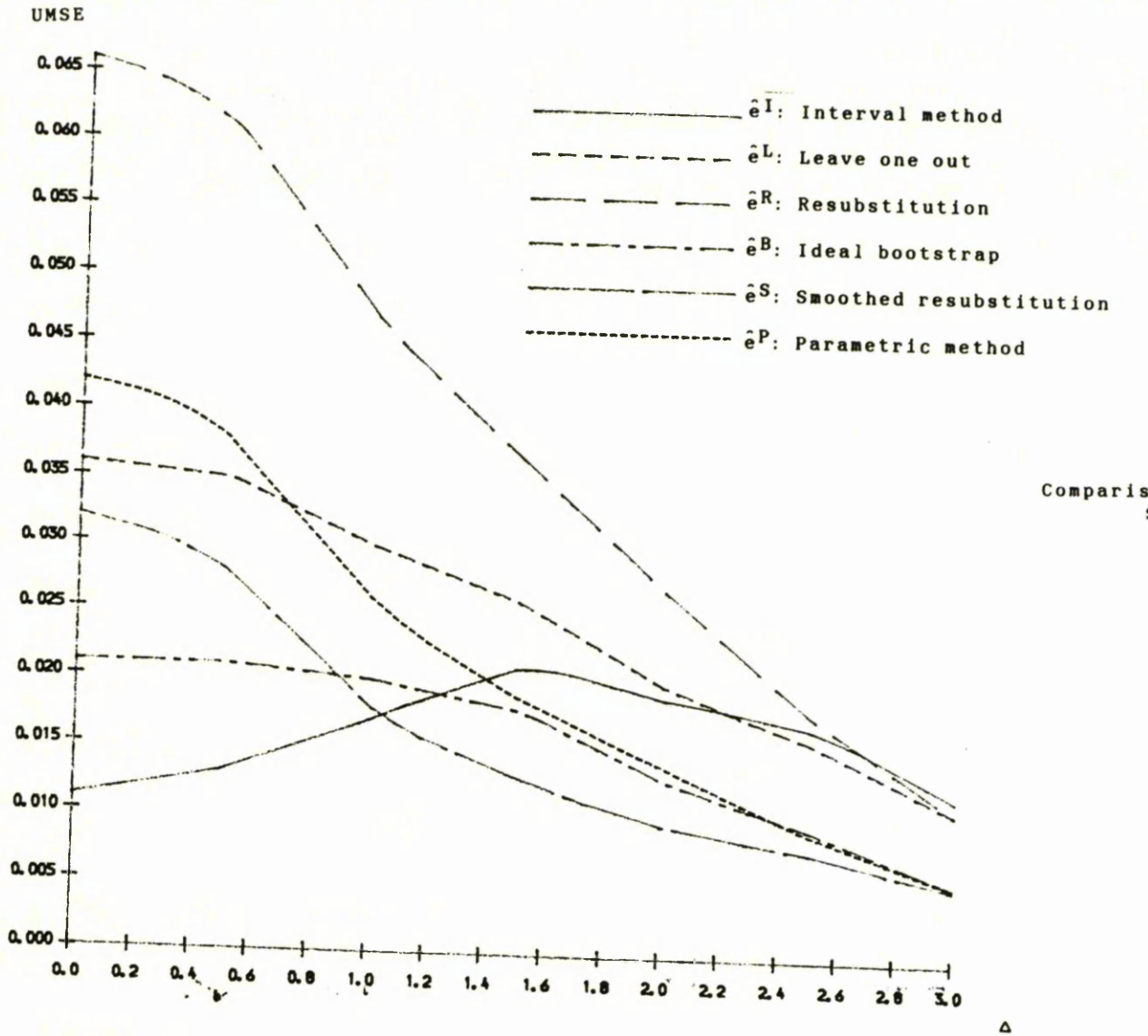

Figure 5.5
Comparison of error rate estimators
Sample size = 10
dimension = 5

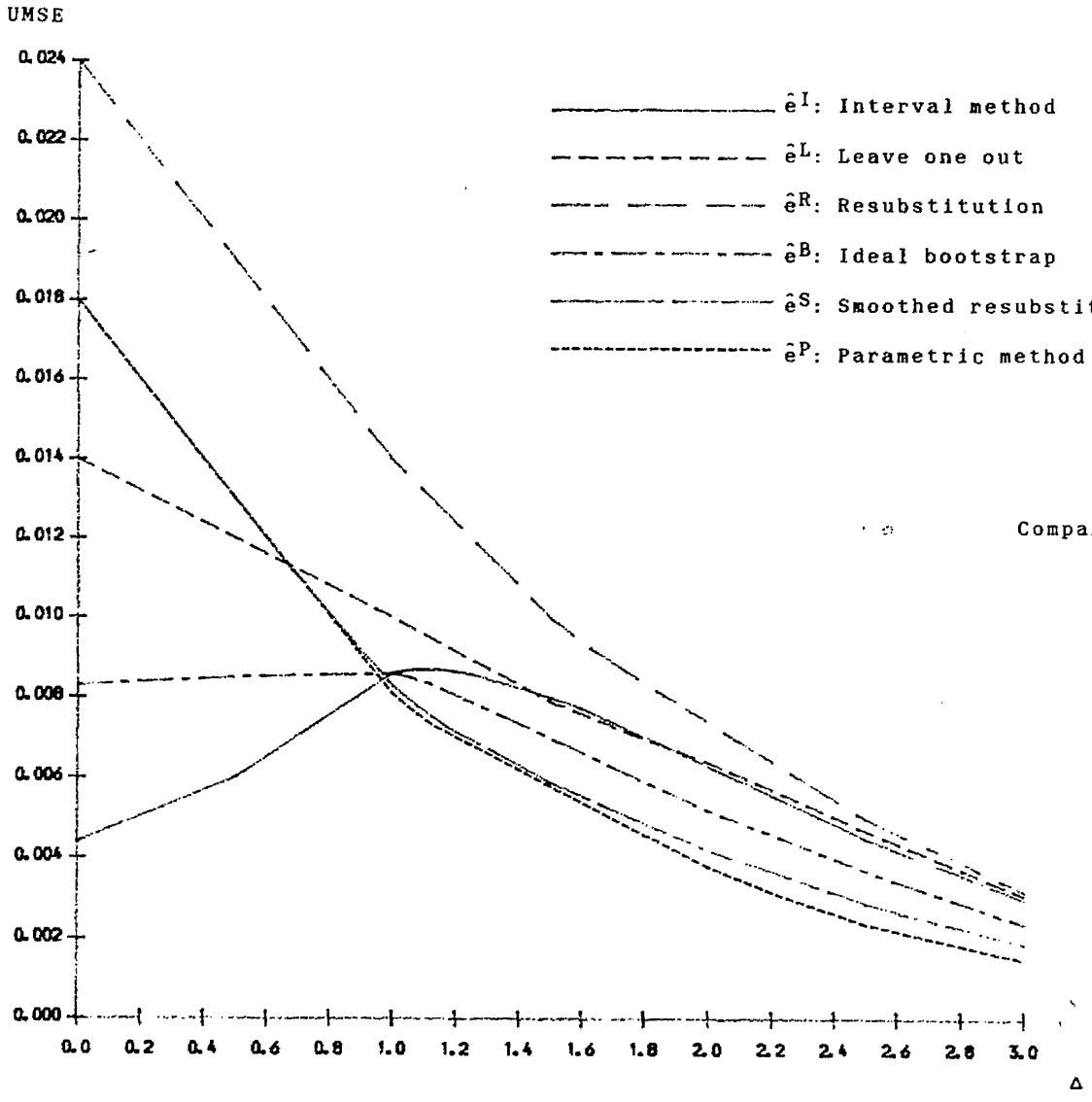

Figure 5.6
Comparison of error rate estimators
Sample size = 25
dimension = 5

5.4.3 Conclusions From Simulations

There are several points raised by the simulation study:-

1) Sample size affects all estimators approximately equally. All do worse with small sample sizes, but their relative performance is unchanged. (But see 4).

2) All estimators do well when the groups are well separated, ie when the true error rates are small.

3) $\hat{e}^I$ is good for large and small Δ, and not so good for moderate Δ.

4) The poorest estimator is $\hat{e}^R$ in almost all situations, the next worst being $\hat{e}^L$, and $\hat{e}^I$ for moderate values of Δ. The exception is for dimension p=1 (figures 5.1 and 5.2) where $\hat{e}^L$ is worse than $\hat{e}^R$.

5) The best estimators are $\hat{e}^P$ and $\hat{e}^S$, for all situations except small Δ, where $\hat{e}^I$ is best. $\hat{e}^P$ is better when the dimension is small, $\hat{e}^S$ is better when it is large.

6) The Mahalanobis distance between the populations at which $\hat{e}^P$ and $\hat{e}^S$ become better than $\hat{e}^I$ increases with the dimensionality, but sample size does not appear to have much effect.

7) The bootstrap estimator $\hat{e}^B$ performs reasonably well, but since this is only an upper bound on the performance of a true bootstrap estimator, this is perhaps rather disappointing.

5.4.4 Discussion of Simulation Results

It is not suprising that one of the best estimators is $\hat{e}^P$, since this is a parametric estimator, and in the simulations the populations were genuinely normally distributed. Equally, the smoothing parameter used in the construction of $\hat{e}^S$ was calculated to ensure that its expected value under assumptions of normality is equal to that of $\hat{e}^P$, and so it could be expected to have a

similar performance.

It is well known that $\hat{e}^R$ is optimistically biased, and this partly explains its poor performance. However, $\hat{e}^B$ is simply $\hat{e}^R$ with its bias subtracted off, and so the UMSE for $\hat{e}^B$ must be due to the variance of $\hat{e}^R$. Equally, $\hat{e}^L$ is known to be nearly unbiased, and so its large UMSE must be explained by a large variance, particularly for small Δ.

The performance of $\hat{e}^I$ is interesting. When the groups are well separated, the point estimates of the log odds ratio will have large magnitude, reflecting the degree of cetainty about their classification. Therefore the number of 95% confidence intervals for the log odds containing zero will be very small, and $\hat{e}^I$ will almost always be equal to $\hat{e}^R$. When Δ is small, most of the intervals will contain zero, reflecting large uncertainty about their true classification. Therefore $\hat{e}^I$ will always be very nearly equal to 0.5. This explains its good performance at each end of the graphs. In between, its relatively poor performance is probably due to oversmoothing, giving a pessimistic bias. It is possible that this could be rectified by using an interval confidence of something other than 95%, but a series of smaller simulations suggested that this was a sensible value to choose.

The observation that $\hat{e}^I$ is a good estimator for small values of Δ, and that $\hat{e}^P$ and $\hat{e}^S$ are good for large Δ suggests that some form of hybrid estimator could be successful. This is considered in the next section.

5.5 A Hybrid Error Rate Estimator

5.5.1 Introduction

Since $\hat{e}^I$ is a good estimator when Δ, the square root of the Mahalanobis distance between the populations, is small, and $\hat{e}^P$

and $\hat{e}^S$ are good when Δ is large, it is possible that a weighted average of $\hat{e}^I$ and one of the others could have the best features of both.

Let $hyb=\lambda\hat{e}^I+(1-\lambda)\hat{e}$

where $0\leqslant\lambda\leqslant1$, and $\hat{e}$ is either $\hat{e}^P$ or $\hat{e}^S$.

We want λ to be near 1 when Δ is zero, and to tend to 0 as Δ increases. We also want $\lambda=½$ where the two methods are equally good, and this point varies with dimension p, though apparently less so with sample size. The 'cross over' points are:-

p	'cross over'
1	$\Delta<0.5$
3	$0.5<\Delta<1.0$
5	$\Delta\approx1.0$

This suggests a weight of the form

$$\lambda=p/(p+5\Delta)$$

This gives $\lambda=1$ if $\Delta=0$, and $\lambda\to0$ as $\Delta\to\infty$. The 'cross overs', where $\lambda=½$ are

p	Δ
1	0.2
3	0.6
5	1.0

These are in the desired ranges, and so we define the two hybrid estimators

$$hyb1=(p/(p+5\hat{\Delta}))\hat{e}^I+(5\hat{\Delta}/(p+5\hat{\Delta}))\hat{e}^S$$

$$hyb2=(p/(p+5\hat{\Delta}))\hat{e}^I+(5\hat{\Delta}/(p+5\hat{\Delta}))\hat{e}^P$$

where $\hat{\Delta}$ is the estimator of Δ given by

$$\hat{\Delta}^2=(n_1+n_2-2)(\bar{X}_1-\bar{X}_2)^TS^{-1}(\bar{X}_1-\bar{X}_2)$$

Another method of determining λ would be to investigate the biases of $\hat{e}^I$, $\hat{e}^P$ and $\hat{e}^S$, with the idea of constructing an unbiased estimator with a moderate variance. It is perhaps a

weakness of our simulations that we did not directly estimate the bias of the estimators.

We now investigate the performance of these two estimators in a simulation study.

5.5.2 Simulations to test Hybrid Estimators

In order to test the two hybrid estimators, we performed similar simulations to those in section 5.4. The combinations of sample size and dimension were the same (ie n=10 and n=25 for dimensions 1, 3 and 5) but this time we used 200 replications to reduce the running time of the program. Since we are interested only in the relative performance of the methods the loss in precision should be of little importance. The methods tested in these simulations were $\hat{e}^I$, $\hat{e}^P$ and $\hat{e}^S$ (as before) and hyb1 and hyb2. The results are shown in figures 5.7 to 5.12.

5.5.3 Conclusions from Simulations

For small Δ, in all simulations, $\hat{e}^I$ is best, hyb1 and hyb2 are next best, with $\hat{e}^P$ and $\hat{e}^S$, being the poorest estimators. This would be expected from the definitions of hyb1 and hyb2. They will not of course be exactly equal to $\hat{e}^I$ for $\Delta=0$, since $\hat{\Delta}$ will always be greater than zero. For large Δ both the hybrids are better than $\hat{e}^I$, and virtually equal to their 'parents' $\hat{e}^S$ and $\hat{e}^P$ repectively. Again this is obvious from their definitions. For p=1 there is little to choose between the hybrids and their parents, though $\hat{e}^P$ and hyb2 are rather better than $\hat{e}^S$ and hyb1. As p increases the difference between $\hat{e}^P$ and $\hat{e}^S$ reverses, with $\hat{e}^S$ being rather better than $\hat{e}^P$ for p=5, n=10. The advantage of the hybrids over the other estimators becomes more noticable as p increases, with both hyb1 and hyb2 being better than their

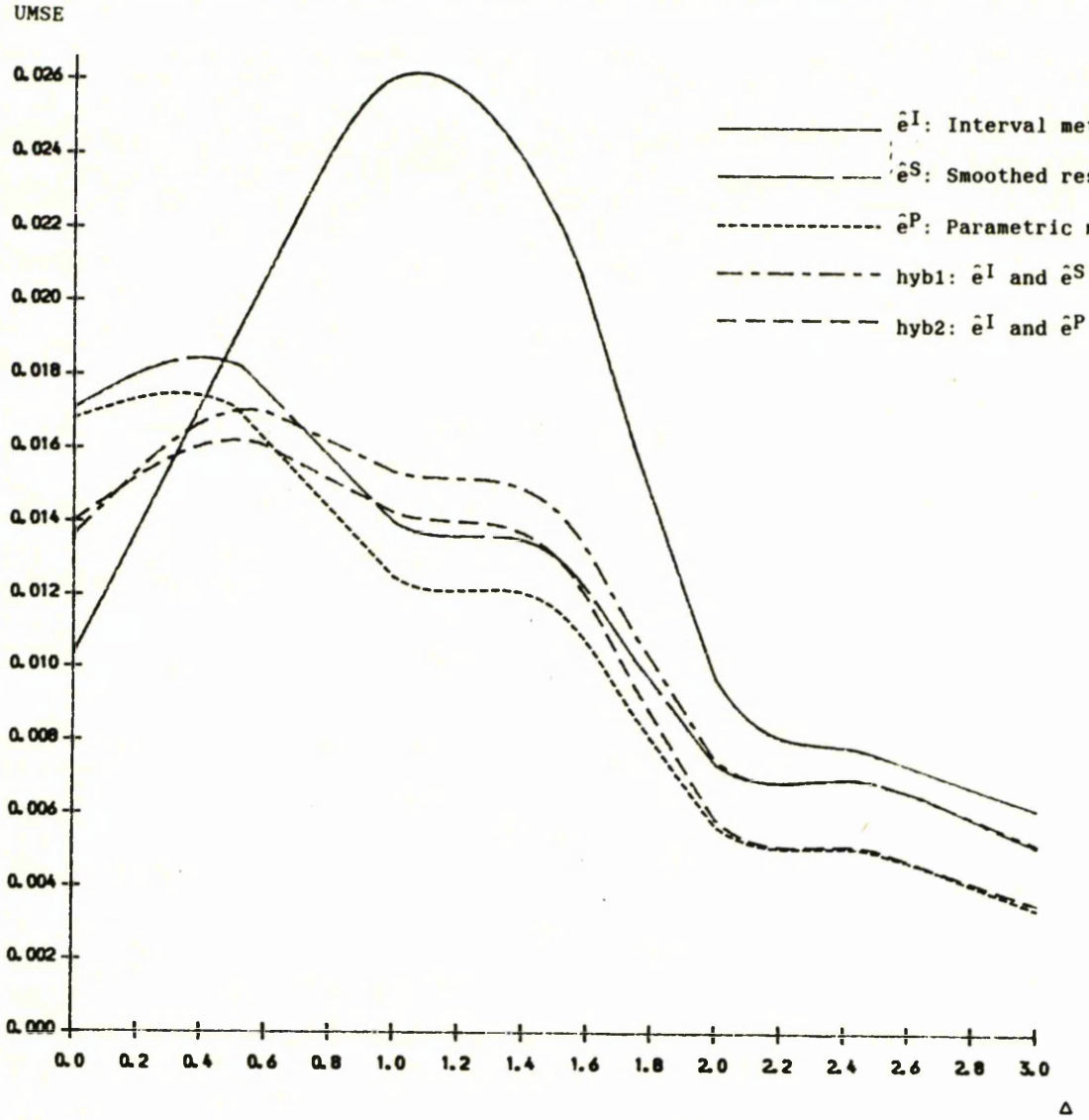

Figure 5.7
Comparison of error rate estimators
Sample size = 10
dimension = 1

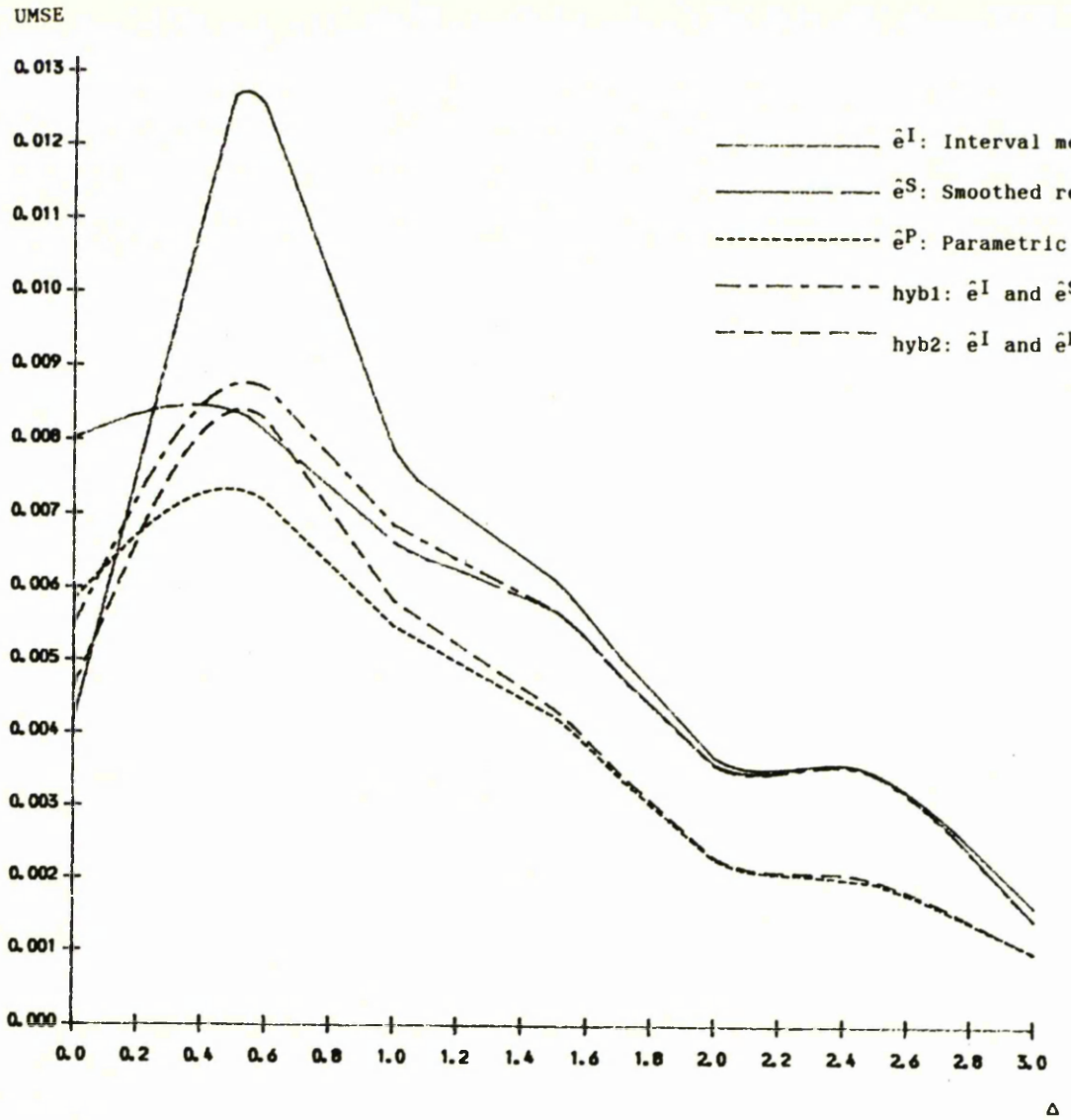

Figure 5.8
Comparison of error rate estimators
Sample size = 25
dimension = 1

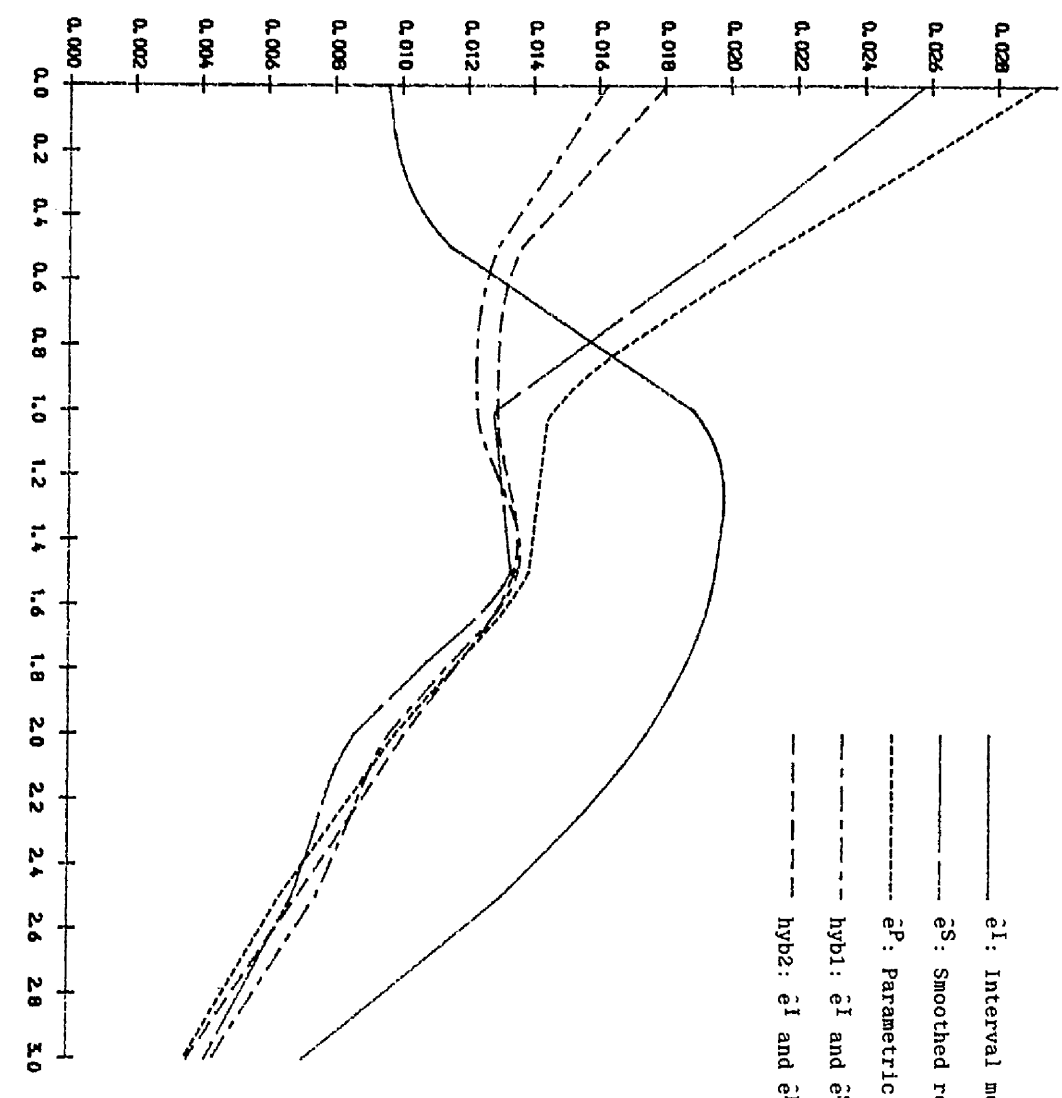

Figure 5.9
Comparison of error rate estimators
Sample size = 10
dimension = 3

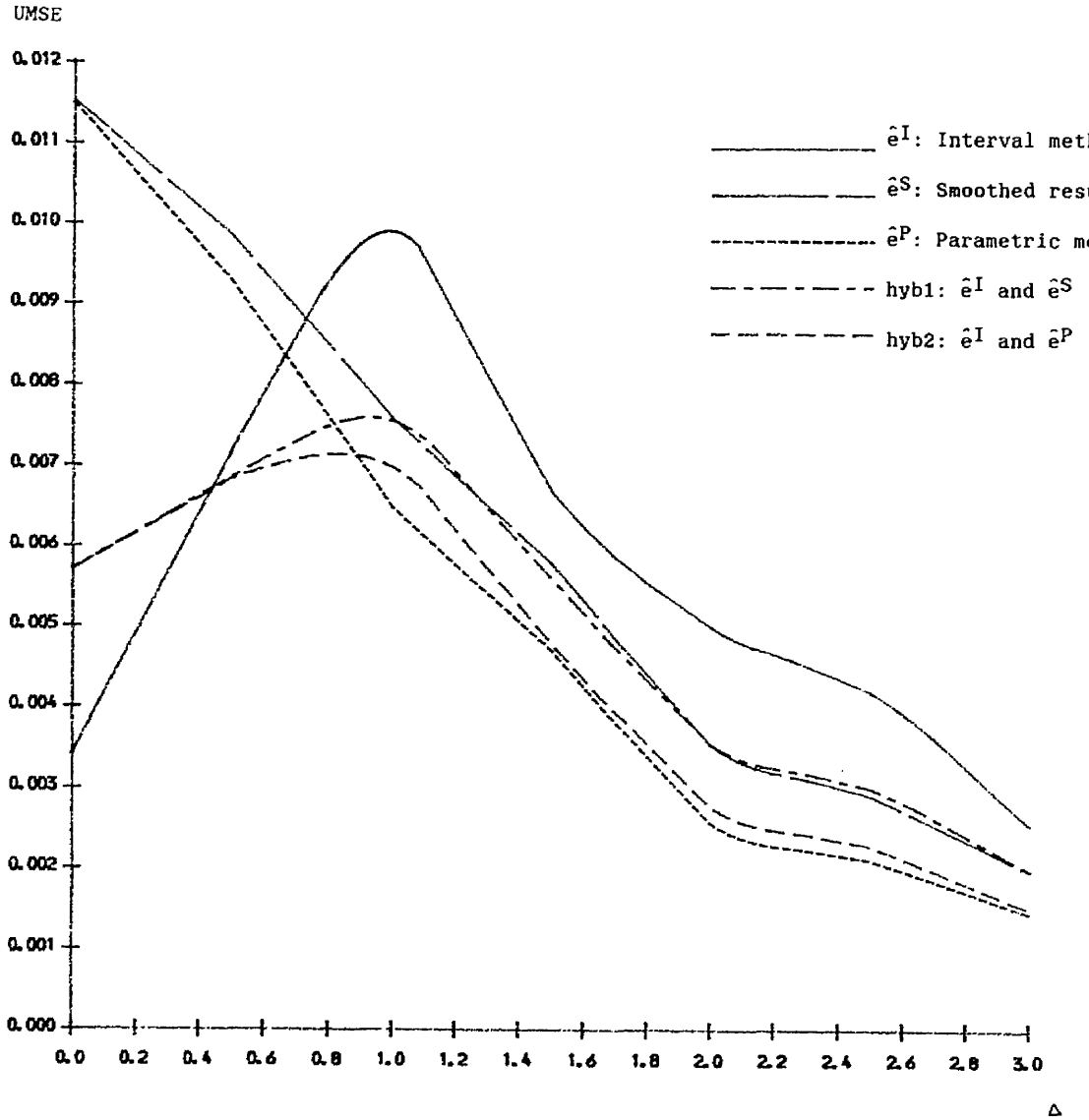

Figure 5.10
Comparison of error rate estimators
Sample size = 25
dimension = 3

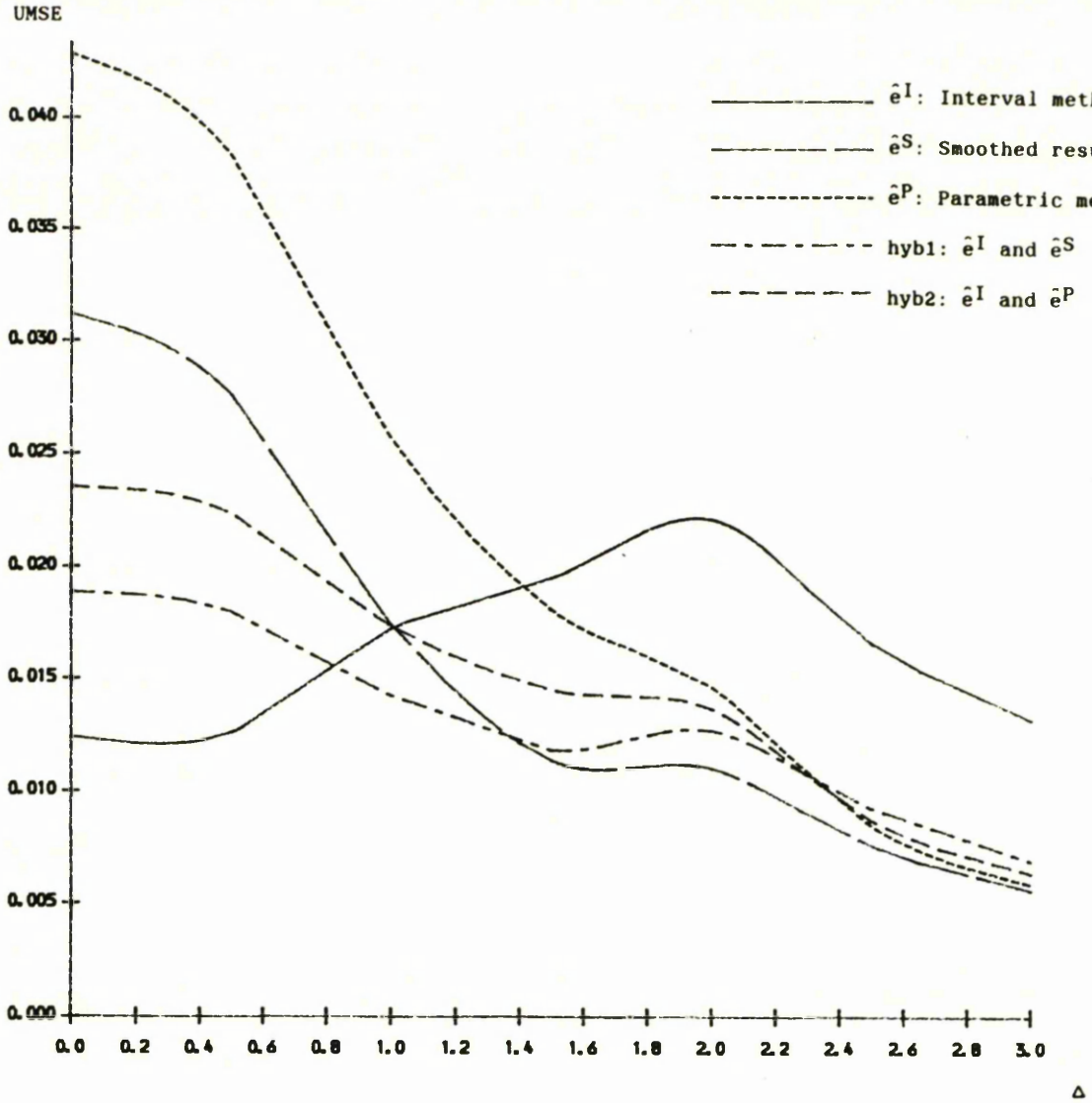

Figure 5.11
Comparison of error rate estimators
Sample size = 10
dimension = 5

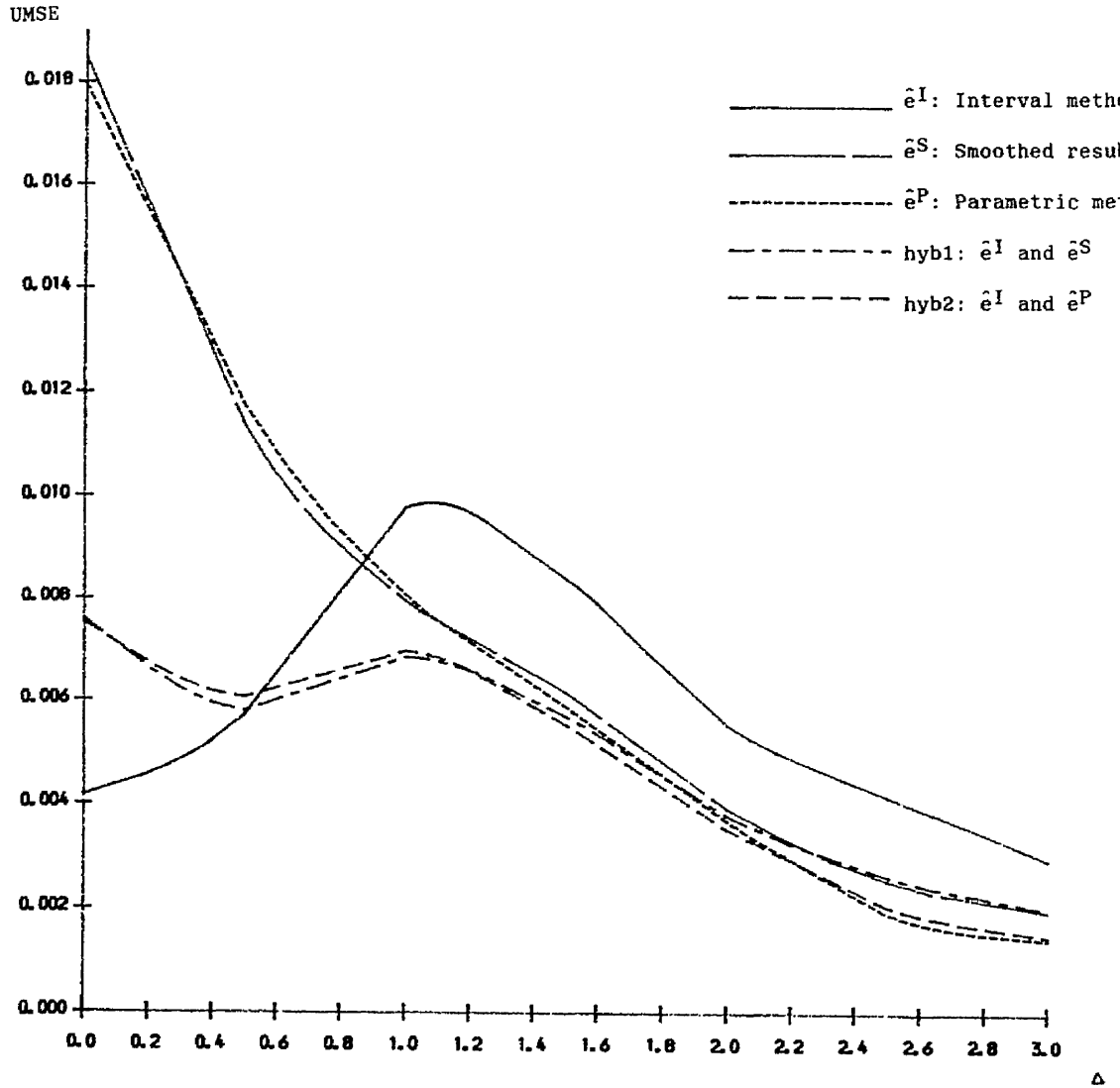

Figure 5.12
Comparison of error rate estimators
Sample size = 25
dimension = 5

parents over a large range of Δ, and at no point being much worse.

5.5.4 Discussion

For p=1, the best error rate estimator would appear to be $\hat{e}^P$, but as the dimension increases the hybrid estimators become the best, with a substantial gain at small values of Δ, and very little loss for larger Δ. There is little to choose between hyb1 and hyb2. The number of parameters in Ω increases with the square of the dimension, and so estimators which require accurate estimates of Ω, such as $\hat{e}^P$ and $\hat{e}^S$ could be expected to perform relatively poorly at high dimensions when compared with the 'partially non parametric' hybrid estimators. It would appear however that even allowing for this loss in accuracy in estimating Ω, the parametric estimators out-perform the totally nonparametric leave one out estimator, at least in the situation where the populations are genuinely normally distributed. Since normality is assumed in constructing most of the other estimators, for example it is assumed when constructing the interval estimates used in $\hat{e}^I$, we now go on to investigate robustness to non-normality.

5.6 Robustness To Non-Normality

5.6.1 Introduction

The performances of the estimators $\hat{e}^I$, $\hat{e}^L$, $\hat{e}^S$, $\hat{e}^P$, hyb1 and hyb2 were investigated when the distributions of the two populations were non-normal, though still with equal covariances. The training data were generated from mixtures of two normal distributions, using results from Johnson and Kotz (1970). The main results are as follows:-

Let $f(x)=\lambda N(\mu_1,\sigma_1^2)+(1-\lambda)N(\mu_2,\sigma_2^2)$

where $0\leqslant\lambda\leqslant1$

Then $E(X)=\lambda\mu_1+(1-\lambda)\mu_2$

$E(X^2)=\lambda(\mu_1^2+\sigma_1^2)+(1-\lambda)(\mu_2+\sigma_2^2)$

$E(X^3)=\lambda(\mu_1^3+3\mu_1\sigma_1^2)+(1-\lambda)(\mu_2^3+3\mu_2\sigma_2^2)$

$E(X^4)=\lambda(\mu_1^4+6\mu_1^2\sigma_1^2+3\sigma_1^4)+(1-\lambda)(\mu_2^4+6\mu_2^2\sigma_2^2+3\sigma_2^4)$

If we set $\lambda=½$, $\mu_1=a$, $\mu_2=-a$, then X will have mean zero. Let $\sigma_1^2=b$, $\sigma_2^2=c$.

We require Var(X)=1, so

$E(X^2)=½(a^2+b)+½(a^2+c)=1$

ie $2a^2+b+c=1$

Let the skewness of the mixture be β, and its kurtosis be γ. Then

$\beta=m_3/(m_2\sqrt{m_2})$

$\gamma=m_4/(m_2^2)$

where $m_k=E(X-E(X))^k$

We are interested in two situations:-

<u>Case one</u> $\beta=0$, increasing γ.

Setting a=0 guarantees $\beta=0$. Hence, since Var(X)=1 and E(X)=1,

$\gamma=E(X^4)=(3/2)(b^2+c^2)$

where b+c=2

<u>Case two</u> $\gamma=3$, varying β.

$\gamma=E(X^4)=a^4+3a^2(b+c)+(3/2)(b^2+c^2)=3$

and $2a^2+b+c=2$

We can substitute a into the equation for γ, and then find b in terms of c. It turns out that

$b=5c-4\pm2\sqrt{6(c-1)}$

(where b must be positive). Therefore by varying c and calculating b and a, we can ensure var(X)=1 and $\beta=0$, and obtain varying γ.

5.6.2 The Simulations

We set n=10 and p=3 as a representative sample size and dimension, and change the skewness and kurtosis of each variable to the same degree. We consider the following situations:-

Case one $\beta=0$, $\gamma=3.12$, $\gamma=4.08$, $\gamma=5.43$

Case two $\gamma=3$, $\beta=.85$, $\beta=.46$, $\beta=-.46$, $\beta=-.85$

These represent quite large deviations from Normality.Each simulation consisted of 200 replications. The results are shown in figures 5.13 to 5.19.

5.6.3 Discussion of Simulations

The results of these simulations are very similar to those in the previous sections. It is perhaps suprising that even when the population distributions are very non-normal, the relative performance of the estimators is unchanged. The non parametric estimator $\hat{e}^L$ has nearly the same UMSE, and the other estimators do rather worse than in the normal case, but not sufficiently badly to be worse than $\hat{e}^L$, or to change their order. This is explained by table 5.1. 5.1a shows the optimal error rates for Normal populations, for different values of Δ. 5.1b gives the optimal error rates for non-normal data. It can be seen that the largest difference between corresponding error rates is only 0.14, in the situation where $\beta=0.85$, $\gamma=3$, $\Delta=0.5$. This is equivalent to an increase of at most .02 in UMSE, and the more representative difference of .05 equals an increase in UMSE of no more than .0025, small when compared with the magnitude of the UMSE for most of the estimators. Of course, as the sample size increases the UMSE of all the estimators will decrease, and the difference between the optimal probabilities in table 5.1a and 5.2b will become more important.

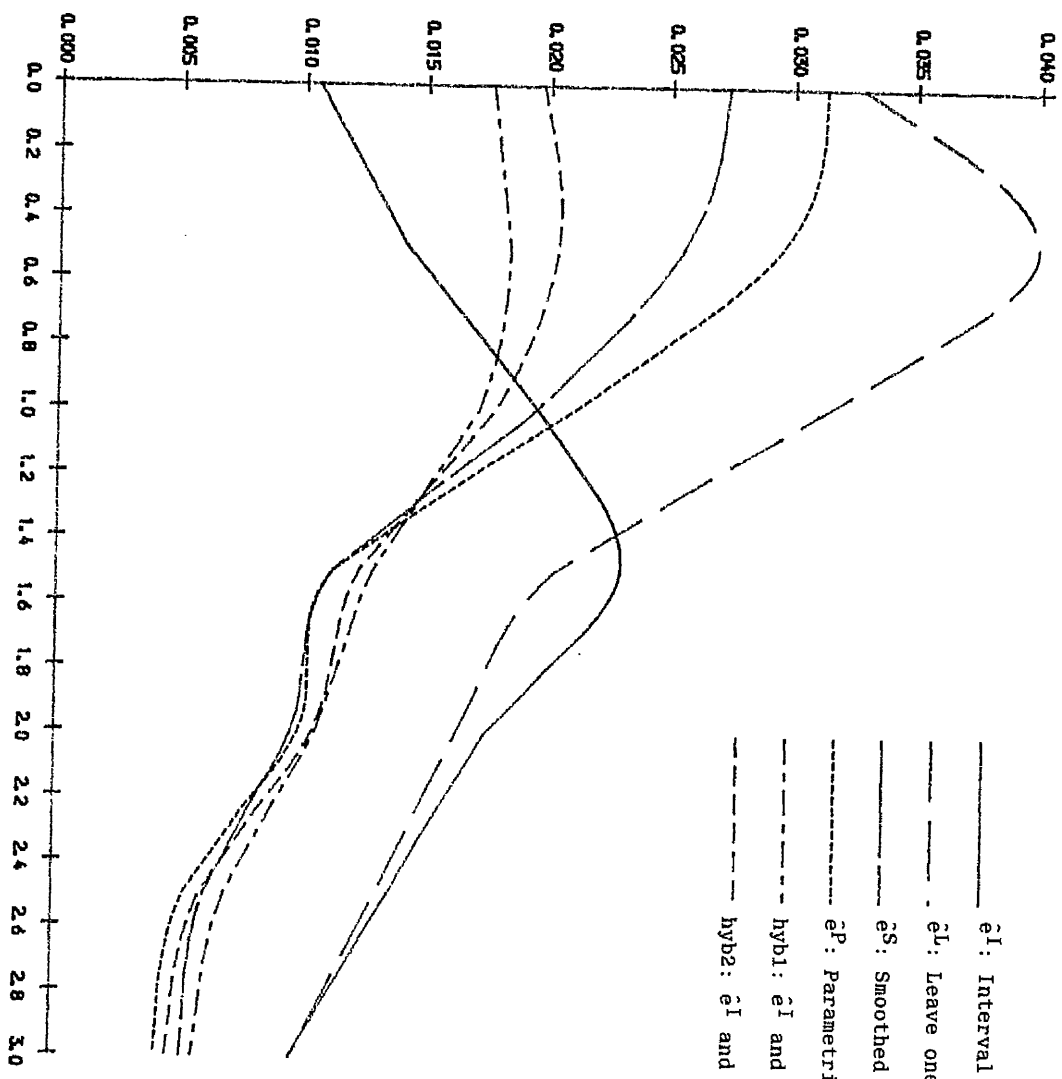

Figure 5.13
Comparison of error rate estimators
Non normal data
$\gamma = 3.12$
$\beta = 0$

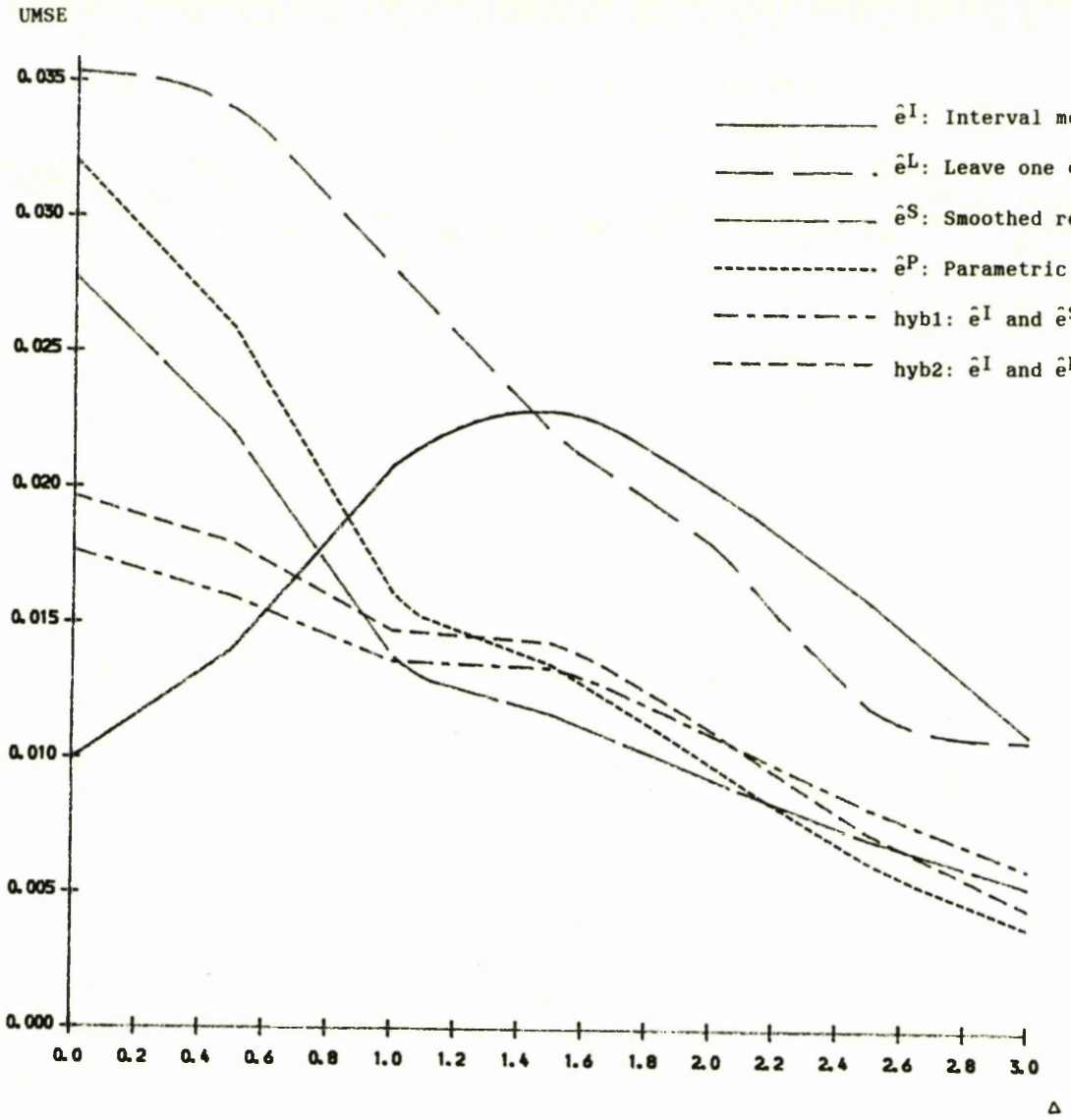

Figure 5.14
Comparison of error rate estimators
Non normal data
$\gamma = 4.08$
$\beta = 0$

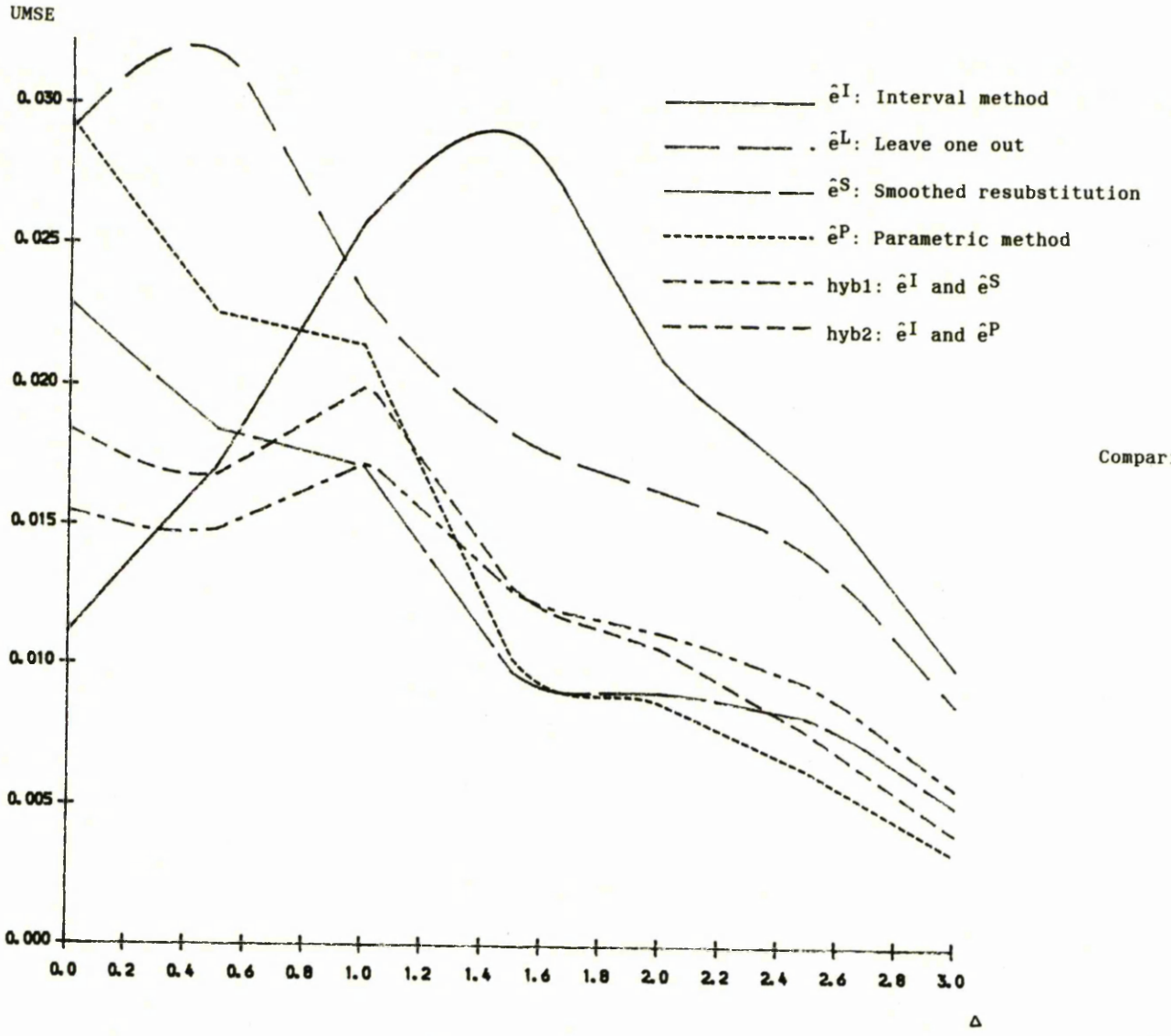

Figure 5.15
Comparison of error rate estimators
Non normal data
γ = 5.43
β = 0

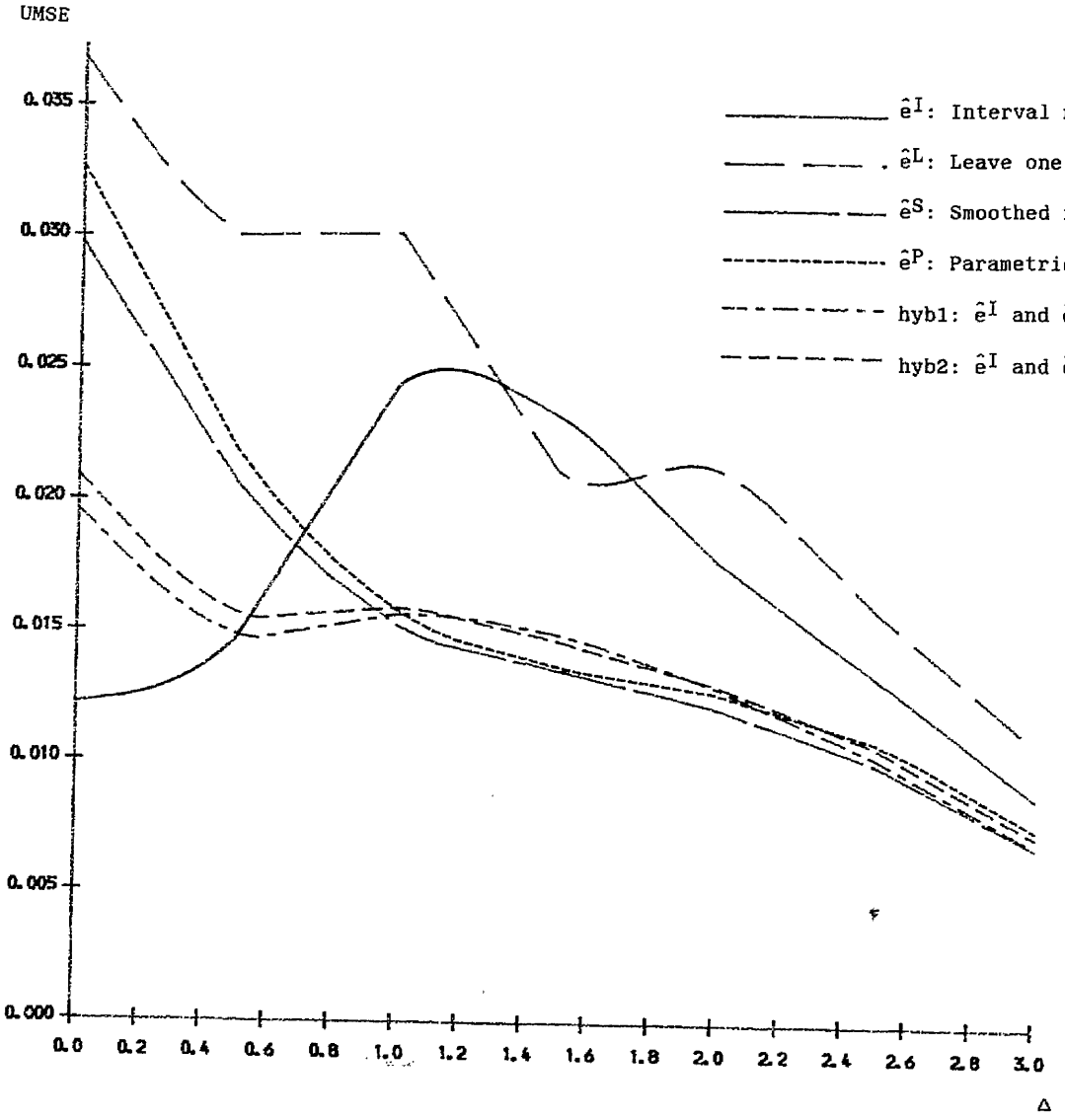

Figure 5.16
Comparison of error rate estimators
Non normal data
$\gamma = 3$
$\beta = 0.85$

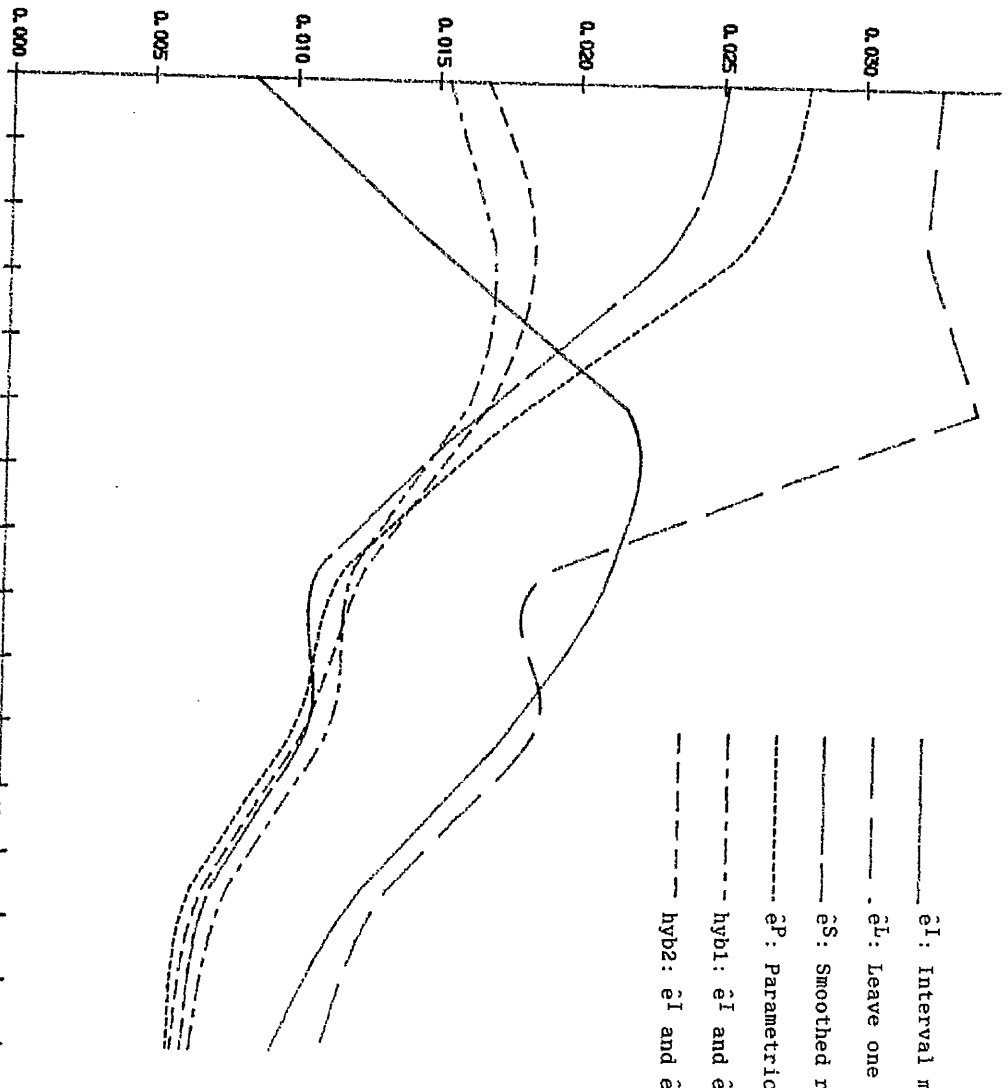

Figure 5.17
Comparison of error rate estimators
Non normal data
$\gamma = 3$
$B = 0.46$

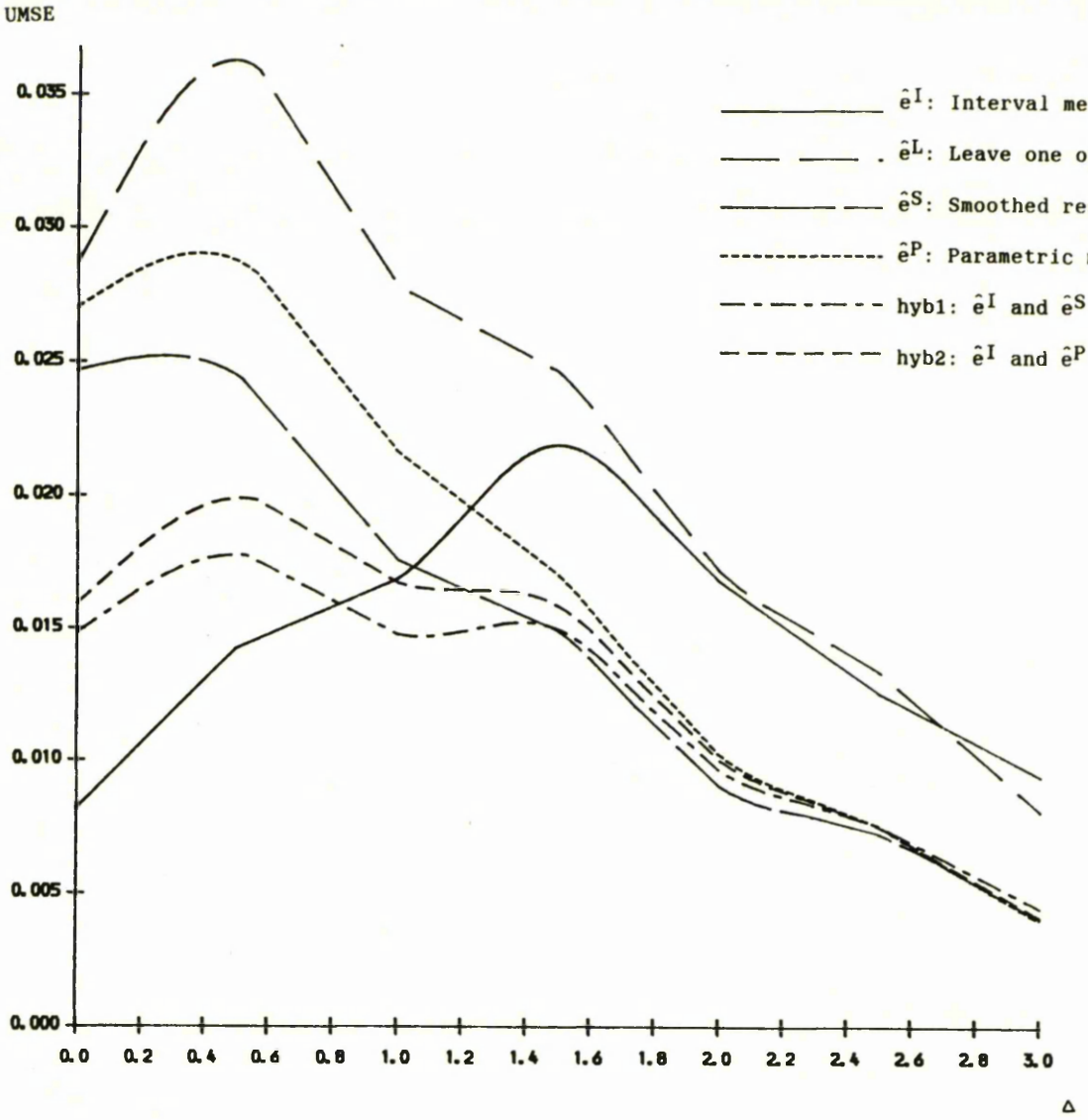

Figure 5.18
Comparison of error rate estimators
Non normal data
$\gamma = 3$
$\beta = -0.46$

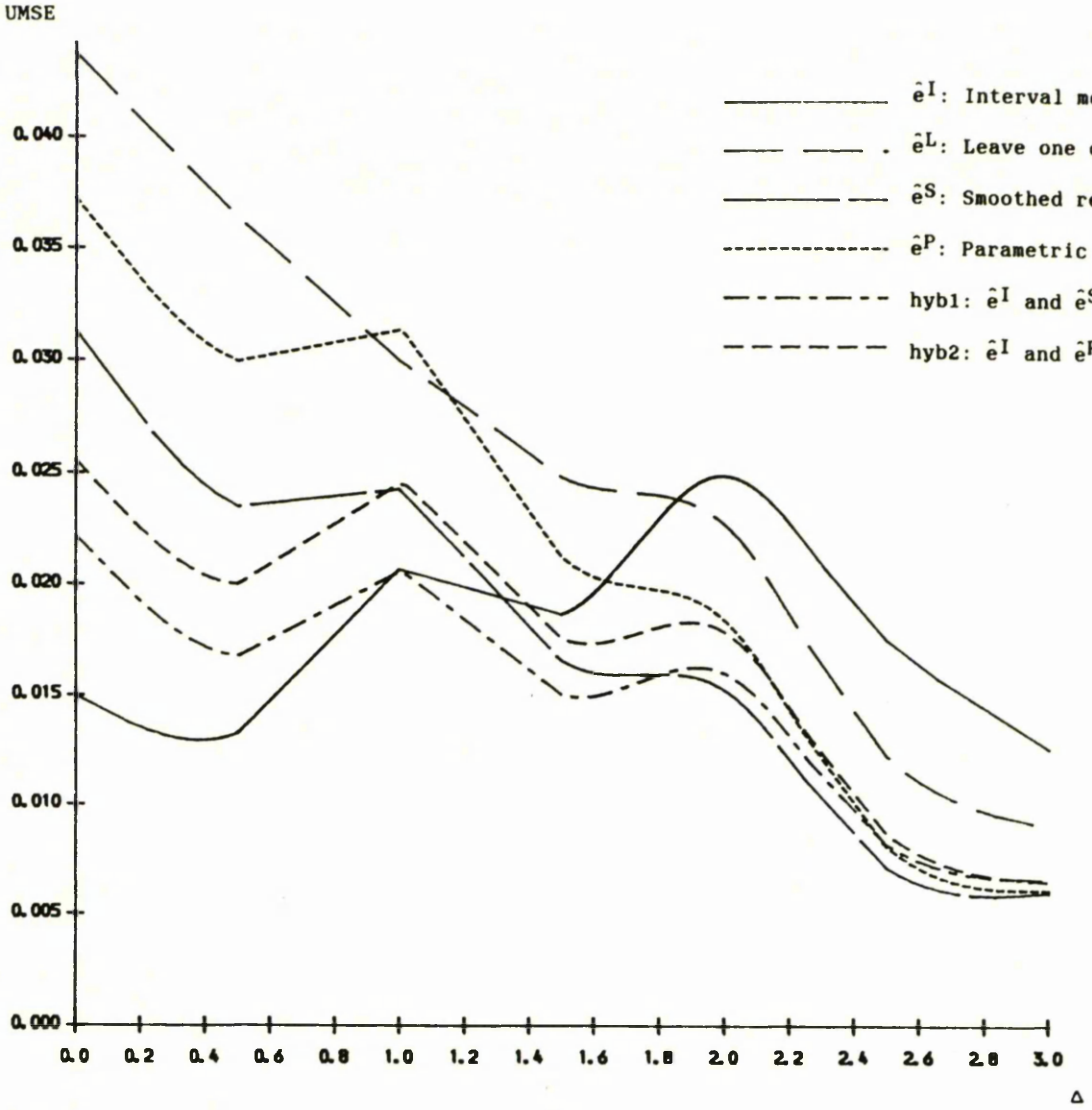

Figure 5.19
Comparison of error rate estimators
Non normal data
$\gamma = 3$
$\beta = -0.85$

5.7 Robustness to Unequal Covariances

As a final test of the estimators we consider the case where the assumption of equal covariance is not valid. Consider the situation where the populations are multinormal $N(0,k_1I)$ and $N(\delta,k_2I)$, where $\delta=(\Delta,0,\ldots,0)^T$ and $k_1+k_2=2$. Then the linear discriminant given the true parameters (but pooling the covariance matrices) will assume a common identity covariance, and the true error rate will be $p(X>\frac{1}{2}\Delta|X\sim N(0,k_1I))$. These probabilities are give in table 5.2 for $k_1=2/3$ (ie $k_1/k_2=\frac{1}{4}$) and $k_1=4/3$ (ie $k_1/k_2=2$). An error rate estimator based on the assumption of equal covariances, such as the parametric method, willl attempt to estimate $p(X>\frac{1}{2}\Delta|X\sim N(0,I))$, and these probabilities are given in table 5.1a. It can be seen that the difference between these probabilities and those in table 5.2 are small, giving a maximum additional UMSE of only .0025. Hence the relative performance of the estimators should be unchanged when the covariances differ by a factor of at least two, for small sample sizes.

5.8 Overall Conclusions

There are several important conclusions to be drawn from this chapter. Firstly, the very commonly used leave one out error rate estimator is very inaccurate for small samples, especially when the true error rate is large. This was noted by Glick (1978) and others, who drew attention to its large variance. An UMSE of .04 is equivalent to an expected error of 20%. There are other, better estimators available, particularly the parametric estimator $\hat{e}^P$ and the smoothed resubstitution estimator $\hat{e}^S$. These are good even when the distributional assumptions on which they are based do not hold, at least for small samples. The estimators

Table 5.1

5.1a Optimal error rates for Normal populations with equal covariances, Mahalanobis distance Δ apart.

Δ	0	.5	1	1.5	2	2.5	3
error rate	.50	.40	.31	.23	.16	.11	.06

5.1b Optimal error rates for non Normal populations with equal covariances.

Case one $\beta=0$

	Δ						
γ	0	.5	1	1.5	2	2.5	3
3.12	.50	.40	.30	.22	.15	.10	.07
4.08	.50	.38	.27	.20	.13	.09	.07
5.43	.50	.32	.21	.15	.11	.09	.07

Case two $\gamma=3$

	Δ						
β	0	.5	1	1.5	2	2.5	3
.85	.39	.34	.29	.24	.19	.14	.10
.46	.45	.36	.29	.23	.17	.12	.08
-.46	.55	.44	.34	.24	.15	.08	.04
-.85	.61	.54	.42	.25	.10	.03	.01

Table 5.2

Error rates for unequal covariance situations

ie $p(X>\frac{1}{2}\Delta | X\sim N(0,k_1))$

	Δ						
k_1	0	.5	1	1.5	2	2.5	3
2/3	.50	.38	.27	.18	.11	.06	.03
4/3	.50	.41	.33	.26	.19	.14	.10

which did best in our simulation studies were the two hybrids hyb1 and hyb2, both of which performed well under all the sampling situations considered.

It should be stressed that we only considered small sample situations. In large sample situations both $\hat{e}^R$ and $\hat{e}^L$ will have smaller UMSEs, and in non-normal or unequal covariance situations they should eventually outperform estimators based on the assumptions of normality or equal covariance. However, if only a small training data set is available, we would recommend the use of either of the estimators hyb1 and hyb2.

CHAPTER SIX

Error Rate Estimation When There Are More Than Two Groups

6.1 Introduction

Very little work has been done on error rate estimation in the multi-group situation. Perhaps the only paper published in this area is Chernick, Murthy and Nealy (1985) who were interested in comparing various bootstrap estimators. In this chapter we explain the difficulties involved and give some suggestions as to how the problem can be addressed. We propose several estimators and a small simulation study gives some idea of how well they perform. First of all we describe the discriminant analysis methodology with multiple groups, and explain some of the problems.

6.2 Multiple Group Discriminant Analysis

Let x be an observation from one of k groups with distributions $N_p(\mu_i, I)$, i=1,...,k. The linear discriminant is based on the set of $\alpha_i(.)$ where

$$\alpha_i(x)=(x-\mu_i)^T\Omega^{-1}(x-\mu_i) \quad i=1,\ldots,k$$

and x is assigned to the population for which $\alpha_i(x)$ is minimised, ie the 'closest' population. This is of course equivalent to the two group case if k=2, where we are interested in a function of the form $\theta(x)=(\alpha_2(x)-\alpha_1(x))$, and x is assigned to population 1 if $\theta(x)>0$. The functions α_i partition the sample space into k regions, one for each group, and x is assigned to the group into whose region it falls.

One of the major problems in error rate estimation with more than two groups is calculating the true conditional error rate. We are interested in the probability that an observation chosen at random from population 1 will fall in the correct region of

the sample space to be classified as group 1. This involves integrating a multinormal density function over a p-dimensional space defined by up to k-1 hyperplanes, and can only be done numerically. There is no simple formula as there is in the two group case. This is particularly a problem in a simulation study as the conditional error rate changes with each new training data set. The approach of Chernick, Murthy and Nealy (1985) was to estimate the conditional error rate by a small simulation within each main simulation. This is computationaly expensive, and it is not clear how many repetitions in each small simulation are neccessary in order to get sufficiently accurate estimates. A simplification would be to replace the conditional error rate by the optimal error rate, since this is dependant only on the known population parameters and so only needs to be calculated once. This is the approach we take in the simulation study described later.

6.3 Some Possible Error Rate Estimators

It is not easy to generalise all of the estimators considered in the previous chapter. The parametric method has no obvious equivalent since no formula exists for the true error rate. It is difficult to generalise the smoothed resubstitution method since there are different $\theta(x)$'s for each pair of groups and they are not independant. A rather unsatisfactory possibility is discussed later. The resubstitution, leave one out and interval methods all do generalise though, and are now discussed.

6.3.1 Resubstitution $\hat{e}^R$

This is directly analogous to the method described in the previous chapter. We define a counting function $h_R(.)$ where

$h_R(x)=0$ if $\hat{\alpha}_1(x)<\hat{\alpha}_j(x)$ for all $j=2,\ldots,k$

$=1$ otherwise

where $\hat{\alpha}_i(x)=(n_1+n_2-p-3)(x-\bar{X}_i)^T S^{-1}(x-\bar{X}_i)-1/n_i$

The estimate of the error rate is then

$$\hat{e}^R=(1/n)\Sigma h_R(x_i)$$

where x_i, $i=1,\ldots,n$ are the training data from group 1.

6.3.2 Leave One Out $\hat{e}^L$

This again is directly analogous to the two group estimator. We redefine $\hat{\alpha}_i(x)$ by omitting x from the training data, and so produce the estimator $\hat{e}^L$ based on the new counting function $h_J(x)$, derived from $h_R(x)$ in the obvious way.

6.3.3 The Interval Method $\hat{e}^I$

In the two group case we defined a form of smoothed resubstitution where x was assigned to group 1 if a 95% confidence interval for $\theta(x)$ was wholly positive, to group 2 if it was wholly negative, and both groups were given a weight of ½ if the interval contained zero. An equivalent here is to find the set of groups that are 'equally likely' and give them equal weights, with the other groups getting zero weight. The method is as follows:-

1) Define $\theta_{ij}(x)=\alpha_j(x)-\alpha_i(x)$ $i,j=1,\ldots,k$, $i\neq j$.

2) Find the 'most likely' group, ie the one for which $\hat{\alpha}(x)$ is minimised (this is the group to which resubstitution would assign x).Say this is group g.

3) Construct 95% confidence intervals for $\theta_{gj}(x)$ and count the number of these intervals that contain zero. Say there are m of them.

4) Define $h_I(.)$ as follows:-

$h_I(x)=1$ if $g\neq 1$ and the interval for θ_{g1} does not contain zero.

$=1-1/(m-1)$ otherwise.

5) The estimate of the error rate is $\hat{e}^I$ where

$\hat{e}^I=(1/n)\Sigma h_I(x_i)$

The rationale behind this estimator is that since a 95% confidence interval for $\theta_{ij}(x)$ contains zero, there is little evidence on which to base a choice between groups i and j. Therefore x should in some sense be assigned equally to the two groups. Similarly, x should be assigned equally to all groups indistinguishable from the most likely group.

6.3.4 Smoothed Resubstitution $\hat{e}^S$

In the two group situation Snapinn and Knoke (1985) suggested the smoothed estimator $(1/n)\Sigma g(x_i)$ where

$g(x_i)=\Phi(-\hat{\theta}(x_i)/bD)$

(see previous chapter). It is not possible to use this in the multiple group situation since there are values of $\hat{\theta}(x)$ for each pair of groups, and they are not independant. One possible generalization of this would be to use $\hat{\theta}_{1j}(x)$, where $\hat{\alpha}_j(x)$ is the smallest value of $\hat{\alpha}_i(x)$, other than $\hat{\alpha}_1(x)$ if this is smaller. That is we are only considering the true group and the 'next most likely' group. This is not entirely satisfactory, but should provide some degree of smoothing to reduce the bias of $\hat{e}^R$. It is now neccessary to choose the value of the smoothing constant b. We use almost the same value as that chosen by Snappin and Knoke (1985), except that our estimate of Ω is now based on k groups rather than only 2, and so we change n_1+n_2-2 in the equation for b given in the previous chapter to $\Sigma(n_i-1)$. This gives the equation

$$b^2=(\Sigma(n_i-1)+(n_1-1)(p+1))/(n_1(\Sigma(n_i-1)-(p+1)))$$

if the training data consist of samples of size n_i, i=1,...,k, or if $n_1=\ldots=n_k=n$,

$$b^2=(n-1)(k+p+1)/(n(k(n-1)-p-1)).$$

We now define $h_S(.)$ as

$$h_S(x)=\Phi(-\hat{\theta}_{1j}(x)/bD)$$

Here D is the Mahalanobis distance between groups 1 and j. The estimate of the error rate is now

$$\hat{e}^S=(1/n)\Sigma h_S(x_i)$$

6.3.5 The Hybrid Estimator $\hat{e}^H$

As before, it is possible that some combination of $\hat{e}^S$ and $\hat{e}^I$ will have good properties. It is not sensible simply to take some weighted average as before, since the value of Δ can change with each point of the training data (ie the 'next most likely' group will not always be the same). Therefore we take the weighted average for each point separately, ie define the counting function

$$h_H(x)=(p/(5\hat{\Delta}+p))h_I(x)+(5\hat{\Delta}/(p+5\hat{\Delta}))h_S(x)$$

where h_I and h_S are the counting functions defined for $\hat{e}^I$ and $\hat{e}^S$ respectivly and $\hat{\Delta}$ is the estimate of the Mahalanobis distance between group 1 and the 'next most likely' group j given by

$$\hat{\Delta}^2=(\Sigma n_k-k)(\bar{X}_i-\bar{X}_j)^T S^{-1}(\bar{X}_i-\bar{X}_j)$$

The estimate of the error rate is now

$$\hat{e}^H(x)=(1/n)\Sigma h_H(x)$$

6.4 A Simulation Study

We performed a simulation study in order to get some idea of the relative performance of the estimators suggested above. We restricted our attention to the two dimensional case p=2, and

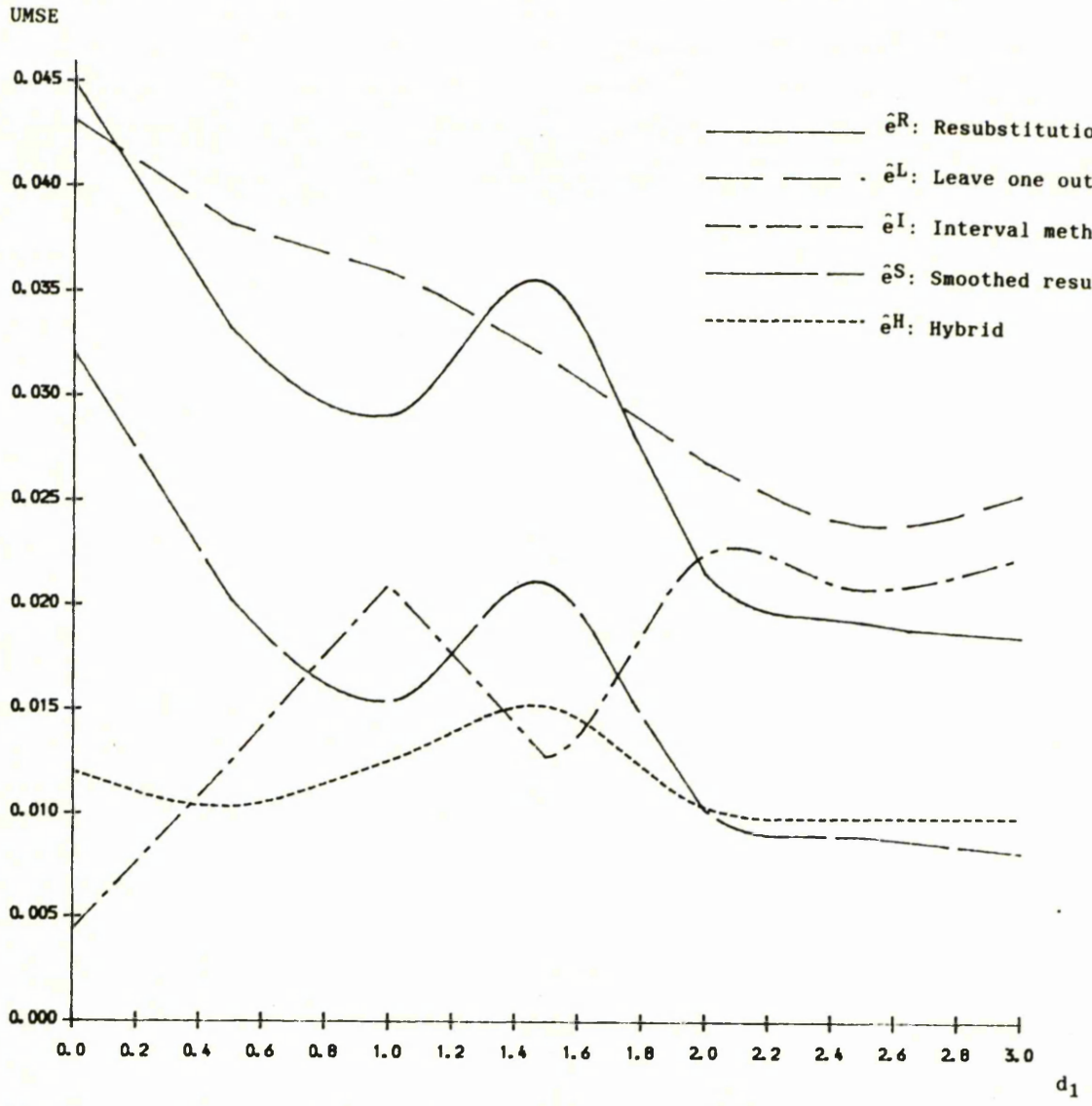

Figure 6.1
Comparison of error rate estimators
Four groups
$d_2=1$

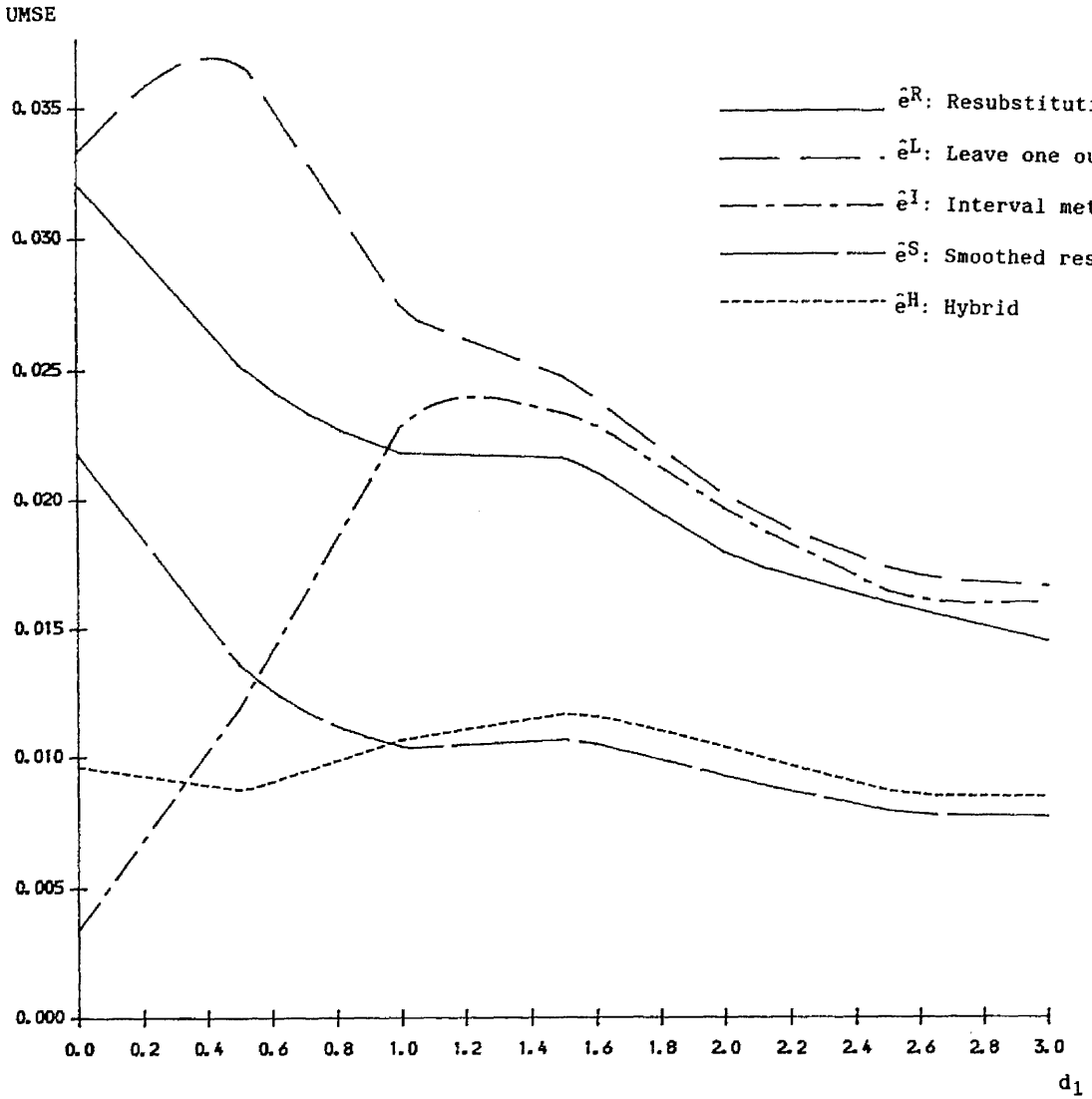

Figure 6.2
Comparison of error rate estimators
Four groups
$d_2=2$

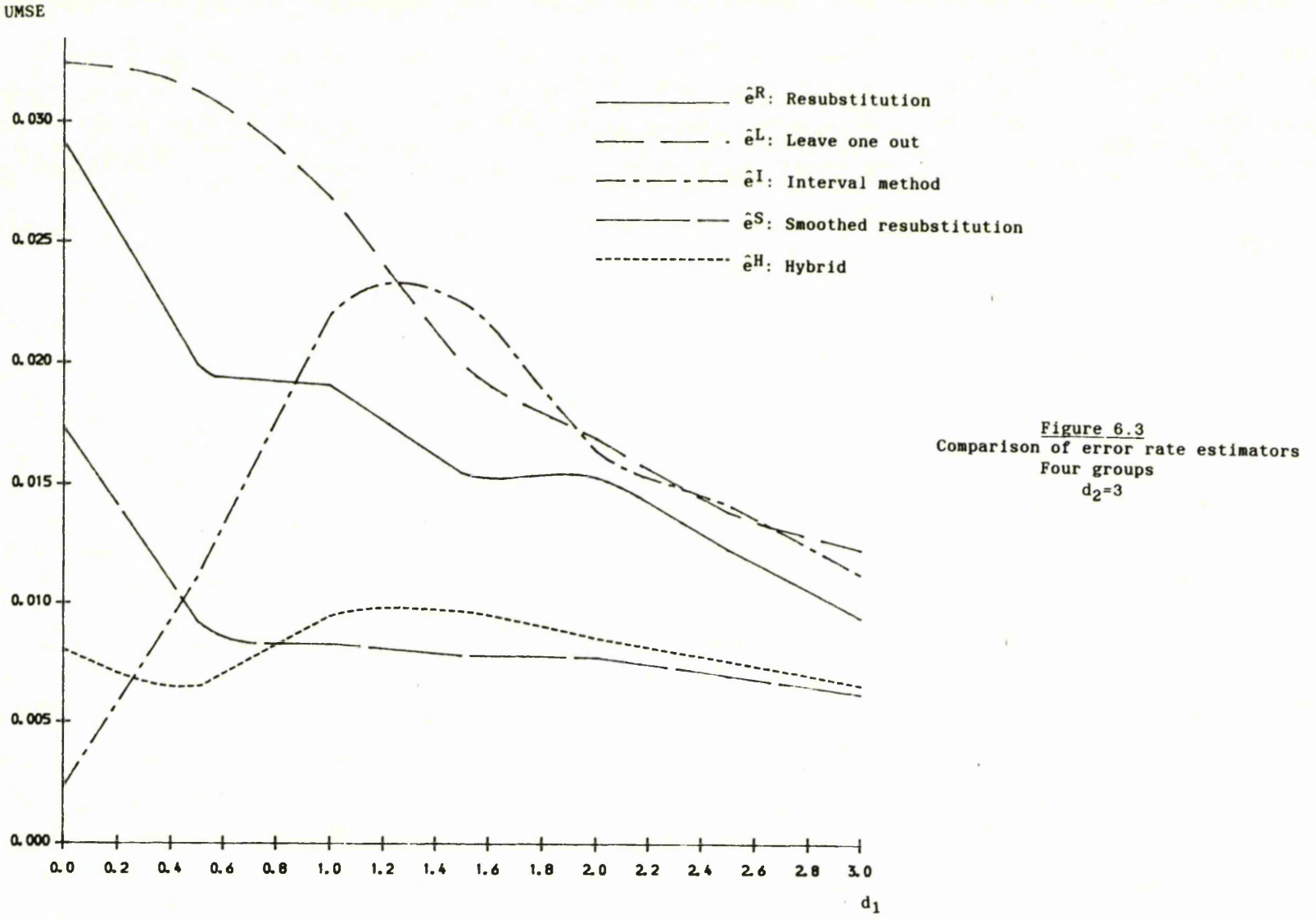

Figure 6.3
Comparison of error rate estimators
Four groups
$d_2=3$

only considered equal sample sizes of n=10. We chose k=4, with the populations being distributed $N(\mu_i,I)$, $i=1,\ldots,4$, where $\mu_1=(0,0)$, $\mu_2=(d_1,0)$, $\mu_3=(0,d_2)$, $\mu_4=(d_1,d_2)$. d_1 was fixed at 1, 2 and 3 while d_2 varied from 0 to 3 as in the previous chapter. This arrangement of the μ_i allows the true optimal error rate to be calculated very easily:-

We are interested in the probability p_{corr} that an observation x from group 1 is closer to μ_1 than to any of the other group means. Defining $|a-b|$ as $[(a-b)^T(a-b)]^{\frac{1}{2}}$, then because of the rectangular arrangement of the means,and the identity covariances, $p(|x-\mu_1|<|x-\mu_2|)$ is independant of $p(|x-\mu_1|<|x-\mu_3|)$. Also, if $|x-\mu_1|<|x-\mu_2|$ and $|x-\mu_1|<|x-\mu_3|$ then $p(|x-\mu_1|<|x-\mu_4|)=1$. Hence $p_{corr}=p(|x-\mu_1|<|x-\mu_2|)xp(|x-\mu_1|<|x-\mu_3)$. Also $p(|x-\mu_1|<|x-\mu_i|) = p(Y<\frac{1}{2}d_i)$ where $Y\sim N(0,1)$, and so it is easy to find the optimal error rate $e_{opt}=1-p_{corr}$.

The results of the simulations are given in figures 6.1 to 6.3.

6.5 Conclusions From Simulations

Although this was a very small simulation study, some interesting points have arisen. Firstly, the resubstitution estimator $\hat{e}^R$ is almost always better than the leave one out estimator $\hat{e}^L$, though neither was particularly good. This confirms the findings of Glick (1978) and others who noted that the increased variance of $\hat{e}^L$ over $\hat{e}^R$ was a serious problem in the two group case. It is perhaps not surprising that this should be even more of a problem when there are four groups.

The interval method $\hat{e}^I$ does fairly well. As would be expected in the light of the two group results it is best when there is a large overlap between group 1 and at least one other group, ie in

figure 6.1, and for low values of d_1 in figures 6.2 and 6.3. When d_2 is large (figure 6.3), ie when groups 3 and 4 are well separated from groups 1 and 2, the performance of $\hat{e}^I$ is very similar to that in the two group case, being good for small and large values of d_1 and not so good in between.

The smoothed resubstitution method $\hat{e}^S$ is very good, being much better than $\hat{e}^R$ or $\hat{e}^L$, and only performing poorly when there is a large overlap between groups. This is perhaps suprising in view of its rather arbitrary nature, and its performance may be affected by the relative positions of the population means. Further simulations would be neccessary to test this. The hybrid estimator is clearly best, being much the same as $\hat{e}^S$ except for situations of large overlap, where it gains from the influence of $\hat{e}^I$.

These results are all similar to those in the two group case, and though such a small simulation cannot be regarded as conclusive, they are very promising, indicating that, at least in the case of normal data, the hybrid estimator is very reliable. We now illustrate the use of these estimators when applied to a real data set.

6.6 Cushings Syndrome

Cushings syndrome (Aitchison and Dunsmore (1975)) is a form of hypertension which occurs in four forms - adenoma (type a), hyperplasia (type b), ectopic carcinoma (type c) and carcinoma (type d). In order to distinguish between the forms, data is available on 50 patients of known type - 8 of type a, 27 of type b, 5 of type c and 10 of type d. The data consist of measurements of the excretion rates of 14 steroid metabolites, and are given in appendix five. The variables are :-

1) Tetra hydro cortisol

2) Allo-tetra hydra cortisol

3) Tetra hydro cortisone

4) Reichsteins compound U

5) Cortisol

6) Cortisone

7) Tetra hydro-11-desoxycortisol

8) Tetra hydro corti costeron

9) Allo-tetra hydro corti costeron

10) Tetra hydro-11-dehydro corti costeron

11) Corticosteron

12) 11-dehydro corti-costeron

13) Pregnantriol

14) pregnentriol

For the purposes of this example we will assume equal covariances. Since the sample sizes are small there is no evidence that this is unreasonable. The aim is to produce classification matrices to assess the performance of the linear discriminant rule when different subsets of the variables are used in its construction. Two subsets have been chosen for illustrative purposes. They are variables 5,6,7 and 13 (subset 1) - chosen as having good discriminant power, and variables 9,10,11 and 12 (subset 2), which are not so good.

It is straightforward to extend the error rate estimators to produce classification matrices. The methods are as follows:-

Resubstitution and Leave one out:- For each data point x give weight 1 to the 'most likely' type, zero to each other type.

Interval Method:- For each x give weight equally to the 'most likely' type and all other 'equally likely' types (with weights summing to one for each x).

Smoothed Resubstitution:- For each x give the true type weight $\Phi(\hat{\theta}_{ij}(x)/b_i\hat{\Delta})$, where i is the true type and j the 'next most likely' type, and $\hat{\Delta}$ is the estimated Mahalanobis distance between types i and j. Since the sample sizes are different, the smoothing parameter b_i will depend on the true type. The 'next most likely' type is given weight $1-\Phi(\hat{\theta}_{ij}(x)/b_i\hat{\Delta})$, and the other types are given zero weight

Hybrid Method:- For each x give each type the appropriate weighted average of the interval and smoothed resubstitution weights.

The classification matrices for each method are then obtained by summing the weights over all x, and here we have divided by the sample sizes for each type to obtain proportions rather than total numbers assigned to each type.

The matrices for each method are given in table 6.1 (for subset 1) and table 6.2 (for subset 2), along with the estimated correct classification probability for an observation drawn at random from one of the types, with equal prior probabilities. This is simply the trace of the matrix, divided by the number of types. It is clear that there are large differences between the methods. For each subset resubstitution is the most optimistic and the interval method the most pessimistic, with the difference being .83 to .66 for subset 1 and .64 to .42 for subset 2. If we look at type c, which has the smallest sample size, and so could be expected to be the hardest to correctly classify, estimates of the correct classification probability range from .43 to .80 for subset 1, and from 0.00 to .60 for subset 2. For the largest group (type b), estimates of the correct classification probability range from .76 to .85 (subset 1) and from .42 to .67 (subset 2). As would be expected from the simulation results, the

Table 6.1

Classification Matrices for Cushings Syndrome - Subset 1

6.1a Resubstitution

		classified type			
		1	2	3	4
	1	.75	.13	.13	.00
true	2	.15	.85	.00	.00
type	3	.00	.00	.80	.20
	4	.10	.00	.00	.90

6.1b Leave-One-Out

		classified type			
		1	2	3	4
	1	.75	.13	.13	.00
true	2	.18	.81	.00	.00
type	3	.00	.00	.60	.40
	4	.10	.00	.00	.90

6.1c Interval Method

		classified type			
		1	2	3	4
	1	.71	.15	.15	.00
true	2	.19	.78	.04	.00
type	3	.23	.07	.43	.27
	4	.08	.05	.13	.73

Table 6.1 continued

6.1d Smoothed Resubstitution

		classified type			
		1	2	3	4
	1	.70	.18	.13	.00
true	2	.22	.76	.02	.00
type	3	.00	.05	.70	.26
	4	.10	.00	.17	.73

6.1e Hybrid Method

		classified type			
		1	2	3	4
	1	.68	.21	.11	.00
true	2	.22	.76	.02	.00
type	3	.10	.05	.60	.26
	4	.10	.00	.17	.73

Table 6.2

Classification Matrices for Cushings Syndrome - subset 2

6.2a Resubstitution

		classified type			
		1	2	3	4
	1	.50	.38	.00	.13
true	2	.11	.67	.22	.00
type	3	.00	.20	.60	.20
	4	.00	.10	.10	.80

6.2b Leave-One-Out

		classified type			
		1	2	3	4
	1	.50	.38	.00	.13
true	2	.15	.59	.22	.04
type	3	.20	.40	.00	.40
	4	.00	.10	.10	.80

6.2c Interval Method

		classified type			
		1	2	3	4
	1	.34	.28	.22	.16
true	2	.28	.42	.22	.08
type	3	.17	.27	.37	.20
	4	.06	.16	.24	.54

6.2d Smoothed Resubstitution

		classified type			
		1	2	3	4
	1	.60	.31	.00	.09
true	2	.20	.55	.22	.03
type	3	.00	.24	.55	.22
	4	.00	.11	.28	.62

6.2e Hybrid Method

		classified type			
		1	2	3	4
	1	.46	.33	.11	.10
true	2	.22	.53	.22	.04
type	3	.06	.27	.41	.25
	4	.01	.11	.27	.61

difference between methods is greatest for small sample sizes (type c), and large overlap (subset 2), but even in the best situation (subset 1, type b), there is a considerable difference. Although we obviously cannot say which is the correct estimate in this case, it is clear that there is a need for reliable error rate estimators.

CHAPTER SEVEN

Conclusions and Further Work

In chapter one we introduced the discriminant analysis problem and, motivated by an example, showed the neccessity for interval estimation of the log odds-ratio. The methods described in chapter two are different attempts to construct these interval estimates using various approximations. We also briefly discuss approaches to non-normal data. Chapter three describes a simulation study to compare these methods. We go on to describe ways of assessing a discriminant rule in chapter four, going into details of error rate estimation in chapters five and six.

The most serious limitation of our work is its dependence on the populations being normally distributed. All of the interval estimation techniques investigated by simulation in chapter three are specifically designed for normal populations. While the performances of Rigby's method, and Davis' for linear discrimination, were very encouraging, the restriction to normality excludes a wide variety of distributions common in medical problems. For example it is not clear how to incorporate discrete or categorical variables.

The logistic regression and profile likelihood approaches to interval estimation offer some hope of progress here, but no work has been done to determine for what sample sizes it is feasible to construct the asymptotic interval estimates suggested.The simulation results for the profile likelihood applied to normal data suggest that sample sizes of thirty at least would be neccessary before confidence levels approaching the required values could be reached. There is much scope for further work here. Incorporating some form of Bartlett correction is one possible approach, and the work of Krzanowski (1975), may also be

of interest.

The restriction to normal populations is also a limitation to the usefulness of the error rate estimation techniques presented in chapter five. The best estimator, the hybrid of the interval method and smoothed resubstitution, requires normality in the construction of interval estimates although in this context it is probably fairly robust. Probably all that is required here is some measure of the uncertainty in the estimation of the log odds ratio, and the exact confidence of the interval is unimportant. This is suggested by the fact that this method was very good for high dimensional data and small sample sizes, conditions under which the empirical confidence was seen not to be very close to the nominal 95% level. For this reason there is hope that a similar technique, using perhaps intervals obtained by logistic regression, could be as good for non normal data. It should be remembered however, that the smoothing constant for the smoothed resubstitution technique was determined to give the estimator the same expectation as a parametric estimator. This may not be so good for other distributions.

The very concept of the interval and hybrid methods was rather ad hoc, and could probably be improved upon, though the idea of smoothing by uncertainty in estimation is, we believe, a good one. It is difficult to find a theoretical justification for the precise methods proposed, their greatest strengths being that they appear to work. Since we have approximations to the first four moments of the distribution of the log odds ratio, it should be possible to incorporate some or all of this information into an error rate estimator, rather than using the rather crude technique of merely determining whether or not an interval of arbitrary confidence contains zero. In constructing the hybrid

estimator it would be of interest to examine the biases of the constituent estimators, with the aim of either finding a better weight, or a more satisfactory justification for the value of λ chosen.

The same limitations apply to the work in chapter six on error rate estimation with more than two groups. Here however there is also the need for further work on the normal case. We replaced the conditional error rate by the optimal error rate, and while this is reasonable given the inaccuracy of the estimators, it would be more satisfactory to include an estimate of the conditional rate in the simulations. This could be done by simulation within each simulation, the problem being to determine the number of repetitions required in each simulation. It is also neccessary to investigate different arrangements of the populations. The rectangular arrangement we used was very convenient for the calculation of the optimal error rates, but may not give a truly representative picture. This convenience would not apply of course if the conditional error rate was being estimated.

We should also investigate further the performance of bootstrap estimators. Our results with the 'ideal bootstrap' of chapter five, and the results of Chernick, Murthy and Nealy (1985) with three groups, suggested that at least with normal data, its performance does not match that of the best of the estimators we have proposed. Improvements to the bootstrap technique are constantly being found though, and given its very wide applicability it certainly merits further work.

APPENDIX ONE

Conn's Syndrome Data

		Variable							
Patient	Type	1	2	3	4	5	6	7	8
1	1	40	140.6	2.3	30.3	4.6	121.0	192	107
2	1	37	143.0	3.1	27.1	4.5	15.0	230	150
3	1	34	140.0	3.0	27.0	0.7	19.5	200	130
4	1	48	146.0	2.8	33.0	3.3	30.0	213	125
5	1	41	138.7	3.6	24.1	4.9	20.1	163	106
6	1	22	143.7	3.1	28.0	4.2	33.0	190	130
7	1	27	137.3	2.5	29.6	5.4	52.1	220	140
8	1	18	141.0	2.5	30.0	2.5	50.2	210	135
9	1	53	143.8	2.4	32.2	1.5	68.9	160	105
10	1	54	144.6	2.9	29.5	3.0	144.7	213	135
11	1	50	139.5	2.3	26.0	2.6	31.2	205	125
12	1	44	144.0	2.2	33.7	3.9	65.1	263	133
13	1	44	145.0	2.7	33.0	4.1	38.0	203	115
14	1	66	140.2	3.1	29.1	4.7	43.1	195	115
15	1	39	144.7	2.9	27.4	0.9	65.1	180	120
16	1	46	139.0	3.1	31.4	2.8	192.7	228	133
17	1	48	144.8	1.9	33.5	3.8	103.5	205	132
18	1	38	145.7	3.7	27.4	2.8	42.6	203	117
19	1	60	144.0	2.2	33.0	3.2	92.0	220	120
20	1	44	143.5	2.7	27.5	3.6	74.5	210	114
21	2	46	140.3	4.3	23.4	6.4	27.0	270	160
22	2	35	141.0	3.2	25.0	8.8	26.3	210	130
23	2	50	141.2	3.6	25.8	4.1	20.9	181	113
24	2	41	142.0	3.0	22.0	4.7	20.4	260	160

Patient	type	variable							
		1	2	3	4	5	6	7	8
25	2	57	143.5	4.2	27.8	4.3	23.7	185	125
26	2	57	139.7	3.4	28.0	5.2	46.0	240	130
27	2	48	141.1	3.6	25.0	2.5	37.3	197	120
28	2	60	141.0	3.8	26.0	6.5	23.4	211	118
29	2	52	140.5	3.3	27.0	4.2	24.0	168	104
30	2	49	140.0	3.6	26.0	6.3	39.8	220	120
31	2	49	140.0	4.4	25.6	5.1	47.0	190	125
32	unknown	49	142.6	2.3	36.0	6.2	35.7	192	125
33	"	35	145.8	2.8	28.0	3.8	24.0	250	140
34	"	64	143.0	3.3	27.6	2.9	35.0	210	130
35	"	56	142.0	3.7	29.0	3.5	33.0	223	125

APPENDIX TWO

Generation of Random Numbers

Sufficient statistics for $\hat{\theta}(x)$ and the estimated interval are $\bar{x}_1$, $\bar{x}_2$, the sample means, and S_1 and S_2, the sample sums of squares and cross products matrices. Their distributions are

$$x_i \sim N(\mu_i, 1/n_i D_i)$$

where $D_i=\text{diag}(d_1,\ldots,d_p)$, and

$$S_i \sim W_p(n_i, D_i)$$

The $\bar{x}_i$ were generated using the Nag (1984) routine G05DDF to generate univariate normal random variables, since the elements of $\bar{x}_i$ are independent. In order to generate the Wishart matrix S_i it is helpful to make use of the Bartlett decomposition (Kendal and Stuart (1966)). First generate matrix B, where

$$b_{ii} \sim \chi^2(n-i+1),\ i=1,\ldots,p$$

$$b_{ij} \sim N(0,1),\ i>j,\ i=2,\ldots,p,\ j=1,\ldots,p-1$$

$$b_{ij}=0 \text{ otherwise}$$

and the elements of B are independent.

Then $A=B^TB \sim W_p(n,I)$, and $S=D^{1/2}AD^{1/2} \sim W_p(n,D)$. The Chi-squared random variables were generated using Nag (1984) routine G05DHF.

APPENDIX THREE

An Algorithm for the Profile Likelihood

We have the graph of the profile likelihood in the form

$$\hat{\theta}_\lambda = f_1(\lambda)$$

$$p(\hat{\theta}_\lambda) = f_2(\lambda)$$

where we know $\lambda \in \Lambda$, and at any point λ is equal to the gradient of the graph. We require to find $\hat{\theta}_\lambda$ such that

$$p(\hat{\theta}_\lambda) - p(\hat{\theta}) = -\tfrac{1}{2}\chi^2(1;0.95) \qquad (1)$$

where $\hat{\theta}$ corresponds to $\lambda=0$.

It is not possible to obtain explicitly f_1^{-1} or f_2^{-1}. Therefore the solution must be found numerically. One possibility would be to run through all possible values of λ, which is easily done graphically, but would be very time consuming in a large simulation. However, the nature of the problem allows for an algorithm which greatly increases the speed of the simulation.

It is not neccessary to find an exact solution to equation (1). We are only interested in whether or not the true log odds θ_T is contained in the interval, ie we score a 'hit' if

$$p(\theta_T) > h$$

where $h = p(\hat{\theta}) - \tfrac{1}{2}\chi^2(1;0.95)$.

Critchley, Ford and Rijal (1987) show that the graph is convex. An equivalent problem is therefore to find whether the point x in figure A2.1 is above or below the curve.

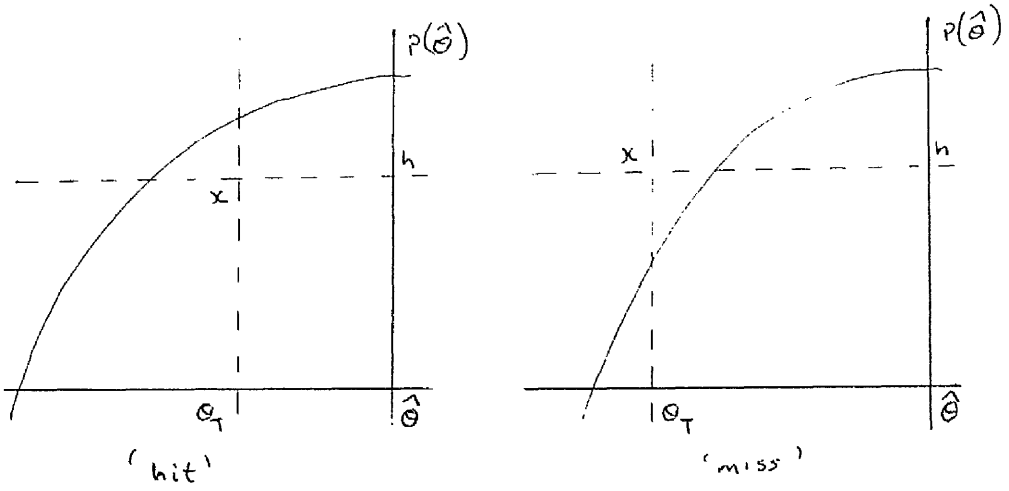

Fig. A 2.1

Without loss of generality let $\theta_T<\hat{\theta}$. Known points on the graph are $x=(\theta_T,h)$ and $y=(\hat{\theta},p(\hat{\theta}))$, and the asymptotic gradient as $\theta\to-\infty$ is known.

1) Let g_r=gradient of line xy, g_ℓ=asymptotic gradient. If $g_\ell<g_r$ stop. 'Hit' (see fig A2.2).

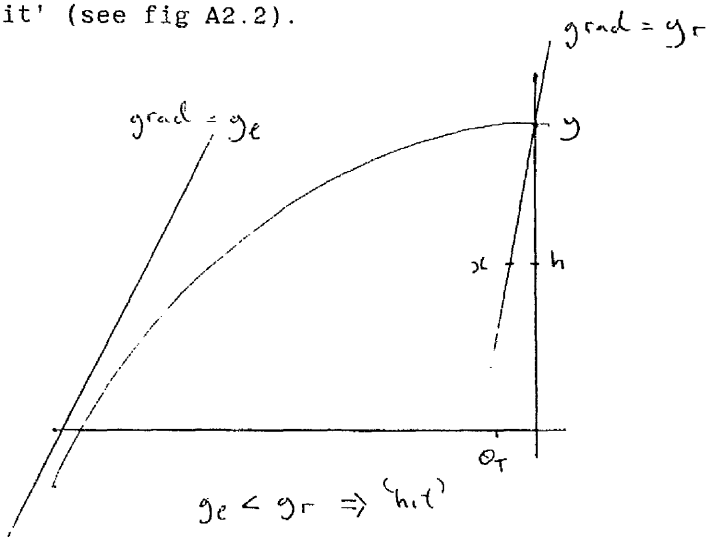

fig A2.2

2) let $\lambda=\frac{1}{2}(g_\ell+g_r)$, and let y_λ be the point $(\theta_\lambda,p(\theta_\lambda))$ corresponding to this λ (ie with gradient λ), and let g be the gradient of the line xy_λ. There are four possibilities:-

a) $\theta_T>\theta_\lambda$, $g>\lambda$ $\Rightarrow$ stop, 'miss' (fig A2.3)

b) $\theta_T<\theta_\lambda$, $g<\lambda$ $\Rightarrow$ stop, 'miss' (fig A2.4)

c) $\theta_T>\theta_\lambda$, $g<\lambda$ $\Rightarrow$ $g_\ell=g$, go to 3

d) $\theta_T<\theta_\lambda$, $g>\lambda$ $\Rightarrow$ $g_r=g$, go to 3

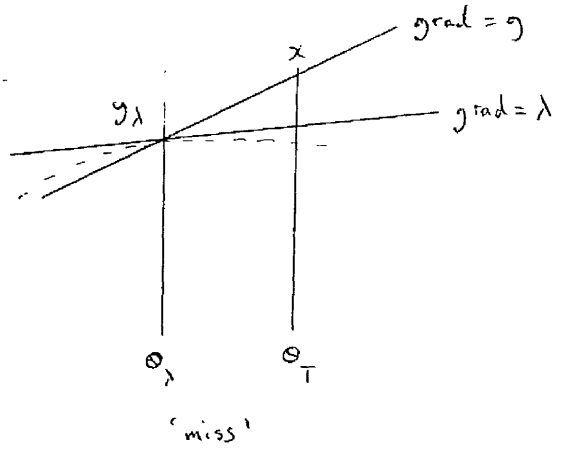

fig A 2.3

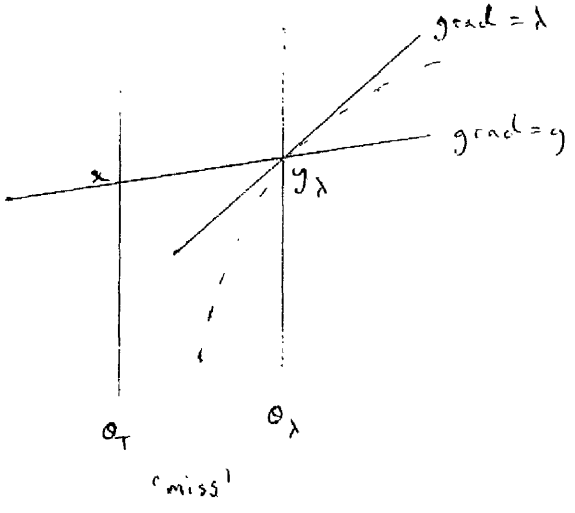

fig A 2.4

3) If $g_\ell < g_r$, stop 'hit' (fig A2.5).

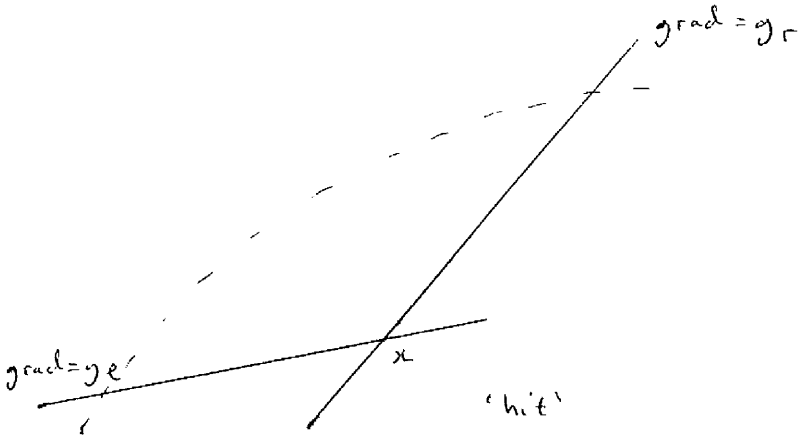

fig A 2.5

If $g_\ell > g_r$ go to 2.

APPENDIX FOUR

The Leave-One-Out Method

This form of error-rate estimation by cross-validation was first suggested by Lachenbruch (1967). The idea is to reduce the bias of the resubstitution estimator by calculating the discriminant rule for each training observation, omitting this observation from the calculations. In general this technique involves a large increase in computation but, for the situations we are interested in, Lachenbruch (1975, p36) gives formulae which greatly reduce this extra work. They are reproduced below.

1) Linear discrimination:

The resubstitution estimate of the linear discriminant is $\theta_1(x)$, where

$$\theta_1(x)=\tfrac{1}{2}(\alpha_2(x)-\alpha_1(x))$$

here $\alpha_i(x)=(x-\bar{X}_i)'S^{-1}(x-\bar{X}_i)$, $i=1,2$.

The leave-one-out equivalent of $\theta_1(x)$ is $\theta_1'(x)$, where

$$\theta_1'(x)=\tfrac{1}{2}\left[\frac{v-1}{v}\alpha_2(x)+\frac{C_1(v-1)(\alpha_{12}(x))^2}{1-(C_1/v)\alpha_1(x)}-C_1{}^2\left[\frac{v-1}{v}\alpha_1(x)+\frac{C_1(v-1)(\alpha_1(x))^2}{v^2(1-(C_1/v)\alpha_1(x))}\right]\right]$$

if x is from population 1, and

$$\theta_1'(x)=\tfrac{1}{2}\left[C_2{}^2\left[\frac{v-1}{v}\alpha_2(x)+C_2\frac{(v-1)\alpha_2{}^2(x)}{v^2(1-(C_2/v)\alpha_2(x)}\right]-\frac{v-1}{v}\alpha_1(x)-C_2\frac{(v-1)\alpha_2{}^2(x)}{v^2(1-(C_2/v)\alpha_2(x))}\right]$$

if x is from population 2. Here

$$\alpha_{12}(x)=(x-\bar{X}_1)'S^{-1}(x-\bar{X}_2)$$

$$v=n_1+n_2-2$$

$$C_i=n_i/(n_i-1),\ i=1,2.$$

For quadratic discrimination, the resubstitution estimator of the discriminant function is $\theta_q(x)$ where

$$\theta_q(x)=\tfrac{1}{2}(\alpha_2(x)-\alpha_1(x))+\tfrac{1}{2}\ln(|S_2|/|S_1|)$$

where $\alpha_i(x)=(x-\bar{X}_i)'S_i^{-1}(x-\bar{X}_i)$

The leave-one-out equivalent is $\theta_q'(x)$ where

$$\theta_q'(x)=\theta(x)-\tfrac{1}{2}\left[\frac{\alpha_1(x)+\alpha_1^2(x)}{n_1-1-\alpha_1(x)} + \ln\left[1-\frac{\alpha_1(x)}{n_1-1}\right]+p\ln\frac{n_1}{n_1-1}\right]$$

if x is from population 1, and

$$\theta_q'(x)=\theta_q(x)+\tfrac{1}{2}\left[\frac{\alpha_2(x)+\alpha_2^2(x)}{n_2-1-\alpha_2(x)} +\ln\left[1-\frac{\alpha_2(X)}{n_2-1}\right]-p\ln\frac{n_2}{n_2-1}\right]$$

if x is from population 2.

Here p is the dimension of the observation vector, x.

To obtain an estimate of the error rate, each observation x in the training data is classified as type 1 or type 2 according to whether $\theta_1'(x)$ (or $\theta_q'(x)$) is positive or negative. The estimate is then simply the proportion of observations incorrectly classified. This technique is almost unbiased, but suffers from an increase in variance when compared with the resubstitution estimator. As an example of the problems involved with this technique, consider the situation where the sample means are very close together. Then $\alpha_1(x)\approx\alpha_2(x)$ and so $\theta_1(x)\approx 0$ and an error rate of 50% would be expected. However, if the

observation (from group 1 say) is removed to calculate the leave-one-out discriminant function, the mean of the remainder of sample 1 will move slightly further away from x, but the mean of sample 2 will remain where it was. Hence $\alpha_1'(x)>\alpha_2'(x)$, and so $\theta_1'(x)<0$ and x is misclassified. This will happen for every observation in the training data, giving an error rate estimate of 100%. This type of problem is particularly serious if sample sizes are small or there are more than two groups.

APPENDIX FIVE

Cushings Syndrome Data ($\log_{10}$(raw data) +5)

patient	variable						
	1	2	3	4	5	6	7
a1	5.58	3.48	5.49	4.00	4.28	3.78	5.01
a2	5.34	4.60	5.48	2.70	4.34	4.04	5.05
a3	5.20	4.00	5.41	3.60	4.60	4.30	4.83
a4	5.58	4.20	5.28	4.38	4.38	4.00	4.56
a5	5.81	4.51	5.58	4.20	4.60	4.30	4.51
a6	5.34	2.70	5.61	4.04	4.62	4.15	5.64
a7	5.46	4.00	5.28	3.30	4.48	3.70	4.75
a8	6.11	5.15	6.02	4.20	4.94	4.60	4.95
b1	5.95	4.90	6.19	4.00	4.61	4.58	4.60
b2	5.51	2.70	5.85	3.60	4.00	4.26	4.11
b3	5.62	4.60	5.71	2.70	4.41	4.18	3.85
b4	6.15	5.51	6.20	3.60	4.53	4.52	4.30
b5	6.30	4.60	6.13	2.70	5.05	4.78	4.30
b6	5.51	4.30	5.62	2.70	4.18	4.15	3.70
b7	5.65	4.48	5.74	2.70	4.30	4.20	4.08
b8	6.11	4.85	6.11	3.60	4.41	4.26	3.30
b9	5.84	4.70	5.76	3.90	4.75	4.51	4.20
b10	5.58	4.90	5.92	3.90	4.41	4.30	3.70
b11	5.51	4.30	5.58	3.60	4.20	4.11	3.60
b12	6.11	4.70	5.92	4.08	4.75	4.52	5.38
b13	5.76	5.11	5.59	3.90	4.56	3.90	4.45
b14	5.81	5.20	5.81	3.90	4.41	4.34	3.90
b15	6.19	4.48	5.92	3.78	4.68	4.56	4.90
b16	6.11	5.20	5.89	3.70	4.90	4.59	3.90
b17	5.81	4.70	5.81	4.08	4.60	4.38	4.86

Cushings Data Continued

b18	5.89	5.26	5.69	3.95	4.34	4.00	4.00
b19	5.63	5.15	5.89	4.20	4.38	4.75	4.11
b20	6.01	4.30	5.89	4.08	4.75	4.62	4.45
b21	5.68	4.78	5.59	2.70	4.60	4.20	4.20
b22	5.95	5.20	5.89	4.08	4.94	4.68	4.98
b21	5.68	4.78	5.59	2.70	4.60	4.20	4.20
b22	5.95	5.20	5.89	4.08	4.94	4.68	4.98
b23	6.01	5.00	5.96	3.78	4.64	4.41	4.75
b24	5.76	5.00	5.89	3.78	4.38	4.20	4.51
b25	5.98	5.52	5.58	3.78	4.43	4.28	4.20
b26	5.65	4.20	5.65	3.90	4.26	4.26	4.26
b27	6.08	4.90	5.58	3.95	4.78	4.54	4.36
c1	6.19	5.18	6.01	4.08	4.90	4.60	5.49
c2	6.19	4.60	5.96	2.70	5.16	4.20	6.15
c3	5.98	4.30	5.98	3.90	5.11	3.90	5.36
c4	6.79	4.78	6.73	3.90	5.58	4.76	5.20
c5	5.99	4.60	6.20	4.48	4.94	4.62	5.60
d1	6.45	4.60	6.10	2.70	5.30	4.30	5.20
d2	6.41	4.78	6.19	2.70	5.11	4.66	5.41
d3	6.01	5.08	5.59	2.70	5.30	3.85	5.43
d4	6.19	3.70	5.95	2.70	5.20	4.56	4.78
d5	6.41	3.48	6.31	2.70	5.11	4.08	5.26
d6	6.79	5.58	6.49	5.83	6.13	4.81	5.72
d7	6.49	4.70	6.37	2.70	5.71	4.82	6.05
d8	6.67	4.30	6.41	2.70	6.01	4.70	5.83
d9	6.61	5.00	6.21	2.70	6.01	4.60	6.41
d10	6.19	4.70	5.51	4.20	5.58	4.34	5.20

Cushings Data Continued

patient	variable						
	8	9	10	11	12	13	14
a1	2.70	4.72	4.08	2.70	2.70	6.07	6.19
a2	4.51	4.63	4.36	2.70	2.70	5.11	5.08
a3	3.30	3.30	3.60	3.60	4.00	4.00	5.71
a4	3.78	3.60	3.60	3.78	3.48	4.00	2.70
a5	4.98	4.60	4.68	4.20	3.60	3.60	3.60
a6	2.70	2.70	4.00	3.60	3.30	5.04	4.48
a7	3.70	3.48	3.30	3.00	2.70	4.60	2.70
a8	4.20	4.20	4.08	3.78	2.70	4.70	4.00
b1	4.70	4.53	4.57	3.60	3.30	5.56	3.78
b2	3.90	3.78	3.85	3.30	2.70	4.95	2.70
b3	4.20	4.18	4.04	3.60	3.00	4.78	4.90
b4	4.48	4.52	4.38	3.60	2.70	4.90	4.90
b5	5.03	4.46	4.58	3.78	2.70	5.20	5.30
b6	3.90	3.48	3.70	2.70	2.70	4.70	2.70
b7	4.51	4.41	4.08	3.30	2.70	4.78	3.60
b8	4.93	4.69	4.45	4.00	2.70	4.90	4.53
b9	4.30	4.00	3.90	2.70	2.70	4.60	2.70
b10	4.41	4.38	4.45	3.60	3.30	4.60	2.70
b11	4.11	4.18	4.00	3.78	2.70	4.00	4.20
b12	4.86	4.51	4.51	3.78	3.30	5.20	4.15
b13	4.00	4.26	3.60	2.70	3.60	4.78	2.70
b14	4.78	4.81	4.28	3.70	3.70	5.20	4.00
b15	4.98	4.45	4.18	3.60	2.70	5.00	4.60
b16	5.06	5.02	4.51	3.85	2.70	5.20	4.48
b17	4.78	4.30	4.30	3.48	3.00	4.60	4.00
b18	4.41	4.51	4.11	3.60	3.30	4.85	4.00

Cushings Data Continued

b19	4.11	4.45	4.36	2.70	3.30	4.90	4.30
b20	4.81	4.34	4.51	3.48	2.70	5.15	4.85
b21	2.70	4.00	4.11	3.30	2.70	4.00	3.85
b22	4.81	4.57	4.30	3.30	2.70	5.08	4.30
b23	4.66	4.58	4.00	3.48	3.30	4.78	4.00
b24	4.38	4.34	4.34	3.78	2.70	4.60	4.00
b25	4.89	4.89	4.20	2.70	2.70	4.60	4.30
b26	4.38	4.04	4.18	2.70	2.70	4.30	2.70
b27	4.85	4.54	4.15	3.00	2.70	4.30	4.00
c1	4.54	4.48	4.34	3.60	2.70	5.81	5.52
c2	4.34	4.11	4.43	4.30	4.51	5.90	6.03
c3	5.16	5.26	4.63	4.20	3.90	5.49	5.23
c4	4.68	4.04	3.48	3.85	2.70	5.40	5.00
c5	5.11	5.33	4.30	2.70	2.70	5.88	6.03
d1	5.25	4.78	4.60	4.30	4.00	5.26	4.20
d2	4.97	4.51	4.48	4.48	3.48	4.70	4.20
d3	4.00	4.04	4.00	4.08	3.00	4.48	2.70
d4	5.13	4.51	4.34	3.48	2.70	4.30	4.20
d5	5.58	4.90	4.95	4.38	2.70	4.70	4.00
d6	6.06	5.51	5.41	5.28	4.20	4.78	4.60
d7	5.59	4.86	4.60	4.38	3.60	4.90	3.90
d8	5.41	4.72	4.90	4.60	4.20	5.51	4.30
d9	5.51	5.18	5.28	4.51	4.32	5.65	4.08
d10	4.60	4.30	4.30	4.34	3.85	5.00	2.70

REFERENCES

Abramowitz,M. and Stegun,I.A. (1965) Handbook of Mathematical Functions. New York: Dover Publications

Aitchison,J. and Dunsmore,I.R. (1975) Statistical Prediction Analysis. Cambridge University 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. (1975) Principles, practice and performance in decision making in clinical medicine. In Proceedings of the 1973 NATO conference on the role and effectiveness of theories of decision in practice (D.J.White and K.C.Brown eds.) London: Hodder and Stoughton

Ambergen,A.W. and Schaafsma,W. (1984) Interval estimates for posterior probabilities, applications to Border Cave. In Multivariate Statistical Methods in Physical Anthropology (G.N.van Vark and W.W.Howells eds.) J.D.Reidel publishing company.

Amemiya,T. and Powell,J.L. (1983) A comparison of the logit model and normal discriminant analysis when the independent variables are binary. In Studies in Econometrics, Time Series and Multivariate Statistics (S.Karlin, T.Amemiya and L.A.Goodman eds) Academic Press

Anderson,T.W. (1958) An Introduction to Multivariate Statistical Analysis New York: Wiley

Anderson,J.A. (1979) Multivariate logistic compounds. Biometrika,66,17-26.

Anderson,J.A. (1982) Logistic discrimination. In Handbook of Statistics, Vol 2 (P.R.Krishnaiah ed) New York: North Holland.

Bernado,J.M. (1976) Psi (digamma) function, Algorithm AS103. Appl.Statist.,25,315-317.

Chang,W-C. (1987) A graph for two training samples in a discriminant analysis. Appl. Statist.,36,82-91.

Chernick,M.R., Murthy,V.K. and Nealy,C.D. (1985) Application of bootstrap and other resampling techniques: evaluation of classifier performance. Pattern Recognition Letters,3,167-178

Critchley,F. and Ford,I. (1984) On the covariance of two non-central F random variables and the variance ofthe estimated linear discriminant function. Biometrika,71,637,638.

Critchley,F. and Ford,I. (1985) Interval estimation in discrimination: the multivariate normal equal covariance case. Biometrika,72,109-116.

Critchley,F., Ford,I. and Rijal,O. (1987) Uncertainty in discrimination. In Proceedings of the Conference DIANA II (F.Zitec ed.),83-106. Mathematical Institute of the Czechoslovak Academy of Sciences, Prague.

Critchley,F., Ford,I. and Rijal,O. (1988) Interval estimation based on the profile likelihood: Strong Lagrangian theory with applications to discrimination. Biometrika,75,21-28.

Davis,C.S. and Stephens,M.A. (1983) Approximate percentage points using Pearson curves. Algorithm AS192. Appl.Statist,32,321-327.

Davis,A.W. (1987) Moments of linear discriminant functions, and an asymptotic confidence interval for the log odds ratio. Biometrika,74,829-840.

Day,N.E. and Kerridge,D.F. (1967) A general maximum likelihood discriminant. Biometrics,23,313-323.

Efron,B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. JASA,70,892-898.

Efron,B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. JASA,78,316-331.

Elderton,W.P. and Johnson,N.L. (1969) Systems of Frequency Curves London: Cambridge University Press.

Fisher,R.A. (1936) The use of multiple measurement in taxonomic problems. Ann. Eugen.,8,376-386.

Foley,D.H. and Sammon,J.W. (1975) An optimal set of discrimination vectors. IEEE Trans. Computers,C-24,281-289.

Geisser,S. (1964) Posterior odds for multivariate normal classifications. J.R.Statist.Soc.B,26,69-76.

Geisser,S. (1977) Discrimination, allocatory and separatory, linear aspects. In Classification and Clustering (J.Van Ryzin ed.) New York: Academic Press.

GHOST (1985) GHOST-80 computer graphics package, release 7. UKAEA, Culham Laboratory, Abingdon.

Glick,N. (1978) Additive estimators for probabilities of correct classification. Pattern Recognition,10,211-222.

Habbema,J.D.F. and Hermans,J. (1974) Cases of doubt in allocation problems. Biometrika,61,313-324.

Hand,D.J. (1986) Recent advances in error rate estimation. Pattern Recognition Letters,4,335-346.

Hilden,J., Habbema,J.D.F. and Bjerregard,B. (1978) The measurement of performance in probabalistic diagnosis II. Meth.Inform.Med,17,227-237.

Hilden,J., Habbema,J.D.F. and Bjerregard,B. (1978) The measurement of performance in probabalistic diagnosis III Meth.Infrom.Med,17,238-246.

Johnson,N.L. and Kotz,S. (1970) Distributions in Statistics, Continuous Univariate Distributions, vol 1. Boston: Houghton Mifflin Company.

Kalbfleisch,J.G. (1979) Probability and Statistical Inference II. New York: Springer-Verlag.

Kalbfleisch,J.G. and Sprott,D.A. (1970) Application of likelihood methods to models involving large numbers of parameters. J.R.Statist.Soc.B,32,175-208.

Kay,R. and Little,S. (1987) Transformations of the explanatory variables in the logistic regression model for binary data. Biometrika,74,495-501.

Kendall,M.G. and Stuart,A. (1966) The Advanced Theory of Statistics,III. London: Griffin.

Krzanowski,W.J. (1975) Discrimination and classification using both binary and continuous variables. JASA,70,782-790.

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. (1975) Discriminant Analysis Hafner Press.

Lachenbruch,P.A. and Mickey,M.R. (1968) Estimation of error rates in discriminant analysis. Technometrics,10,1-10.

McLachlan,G.J. (1974) Estimation of the errors of misclassification on the criterion of asymptotic mean square error. Technometrics,16,255-260.

McLachlan,G.J. (1977) The bias of small sample posterior probabilities. Biom.J.,19,421-426.

McLachlan,G.J. (1979) A comparison of the estimative and predictive methods of estimating posterior probabilities. Comm.Statist.Theor.Meth.A,8,919-929.

McLachlan,G.J. (1980) The efficiency of Efron's bootstrap approach applied to error rate estimation in discriminant analysis. J.Statist.Comp.Simul,11,273-279.

Nag (1984) FORTRAN library manual. Numerical Algorithms Group, Oxford.

Moran,M.A. and Murphy,B.J. (1979) A closer look at two alternative methods of statistical discrimination. Appl.Statist.,28,223-232.

Page,J.T. (1985) Error rate estimation in discriminant analysis. Technometrics,27,189-198.

Peers,H.W. and Iqbal,M. (1985) Asymptotic expansions for confidence limits in the presence of nuisance parameters with applications. J.R.Statist.Soc.B,47,547-554.

Rigby,R.A. (1982) A credibility interval for the probability that a new observation belongs to one of two multivariate normal populations J.R.Statist.Soc.B,44,212-220.

Sammon,J.W..Jr. (1970) An optimal discriminant plane. IEEE Trans.Computers,C-19,826-829.

Schaafsma,W. (1982) Selecting Variables in discriminant analysis for improving on classical proceedures. In Handbook of Statistics, Vol. 2 (P.R.Krishnaiah and L.N.Kanal eds.) Amsterdam: North Holland.

Schaafsma,W. (1984) Some aspects of discriminant analysis. Report, Department of Mathematics, University of Groningen Netherlands.

Schaafsma,W. and Van Vark,G.N. (1979) Classification and discrimination problems with applications II. Statist.Neerlandica,33,91-126.

Sedransk,N. and Okamoto,M. (1971) Estimation of the probabilities of misclassification for a linear discriminant function in the univariate normal case. Annals of the Institute of Statistical Mathematics,23,419-435.

Shapiro,A.R. (1977) The evaluation of clinical predictions, a method and initial application. N.Engl.J.Med,296,1509-1514.

Snapinn,S.M. and Knoke,J.D. (1984) classification error rate estimators evaluated by unconditional mean squared error. Technometrics,26,371-378.

Snapinn,S.M. and Knoke,J.D. (1985) An evaluation of smoothed classification error-rate estimators. Technometrics,27,199-206.

Sorum,M.J. (1971) Estimating the conditional probability of misclassification. Technometrics,13,333-343.

Sorum,M.J. (1972) Estimating the expected and the optimal probabilities of misclassification Technometrics,14,935-943.

Sorum,M.J. (1973) Estimating the expected probability of misclassification for a rule based on the linear discriminant function: univariate normal case. Technometrics,15,329-339.

Toussaint,G.T. (1974) Bibliography on estimation of misclassification. IEEE Transactions on Information Theory,IT-20,472-479.

Van der Sluis,D.M. and Schaafsma,W. (1984) POSCON-a decision support system in diagnosis and prognosis based on a statistical approach. In Compstat 1984 (T.Havrenek et al eds.) Vienna: Physica-Verlag.

Additional References

Mosteller and Wallace (1963) Inference in an authorship problem. JASA, 58, 275-309

Mitchell and Payne (1971). A conservative confidence interval for a likelihood ratio. JASA, 66, 861-866.