U N I V E R S I T Y O F C A P E T O W N

DEPARTMENT OF MATHEMATICAL STATISTICS

THE APPLICABILITY OF DISCRIMINANT ANALYSIS TECHNIQUES ON THE MULTIVARIATE NORMAL AND NON-NORMAL DATA TYPES IN MARKETING RESEARCH

BY

PETRUS JACOBUS UYS VAN DEVENTER

A thesis prepared under the supervision of Professor C.G. Troskie in fulfilment of the requirement for the degree of Masters in Science in Mathematical Statistics.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to the following:

My Supervisor, Professor C.G. Troskie for his valuable guidance, assistance and enthusiasm.

Elmarie van Deventer for typing this thesis.

All members of staff and post graduate students of the Department of Mathematical Statistics who not only provided encouragement which was often very valuable, but never hesitated to lend a helping hand.

PREFACE

A. The purpose of the procedures described is to assign "objects" or "observations" in some optimum fashion to one of two or more populations. In routine banking a bank manager may wish to classify clients who wish to make loans as low or high credit risks on the basis of the elements of certain accounting statements. In such a case there are two definite distinct classes.

Another investigation may be initiated to determine whether buying habits are different with respect to the categories: urban, sub-urban and rural clients.

Note that in the first example the classes are determined before any sample of observations is investigated, i.e. the sample results do not influence the choice of groups. In the latter case one is trespassing on the terrain of cluster analysis.

In the first case we have two types of problems, namely that of devising a classification rule from samples of already classified objects and that of imposing the classification scheme on the objects. The term "discrimination" refers to the process of deriving classification rules from samples of classified objects and the term "classification" refers to applying the rules to knew objects of unknown class.

Although it is possible to convert raw data into more easily grasped forms like cartoon faces (Chernoff, 1973) this still represents the problem that any grouping or classification based on these diagrams is subjective.

This suggests the following reasons for developing formal statistical classification methods (see Hand, 1981):

(i) Formal methods are objective and can be repeated by other researchers.

(ii) One can assess the performance of the assignment rule.

(iii) One can formally measure the relative sizes of the classes.

(iv) One can determine how representative is a particular example of its class.

(v) One can investigate what aspects of the objects are important in producing a classification.

(vi) One can describe and test the differences between classes.

This agrees with four of the main objectives of discriminant analysis and classification, viz.

(i) Finding linear composites of the predictor variables that enable the analyst to separate the groups by maximising the among-groups variation relative to within-groups variation.

(ii) Establishing procedures for assigning new individuals whose profiles are known, but not the group identities, to one of the groups.

(iii) Testing whether significant differences exist between the mean predictor-variable profiles of the groups.

With large sample sizes it is not difficult to obtain a significant test ratio, even though the classification accuracy is poor.

(iv) One should eventually be guided by the classification accuracy.

To summarise: The problem of classification arises when an investigator wishes to classify an individual into one of several categories on the basis of some measurements. The problem may be considered as a problem of "statistical decision functions". Each observation occurs according to a given distribution. Now each distribution may be known or unknown. If it is known, the parameters may be known or unknown. In the latter case the parameters must be estimated from a sample from that population.

If there is no hint of which distribution is involved one may fall back on a non-parametric approach. It will be seen that non-parametric methods are sometimes as accurate as parametric methods.

B. Having discussed the basic "definition" of discriminant analysis one may well ask whether this procedure is useful for the type of data one often encounters in marketing research situations. The classical discriminant analysis was based on two groups and continuous data which are normally distributed. When the distributions involved, were not known, normal distributions were assumed, because of large enough sample sizes.

We know, however, that this is an idealistic way of looking at the real situation of small samples, discrete data, continuous and discrete data mixed, ordinal discrete data, nominal data and so forth which one so often, almost as a rule, encounters dealing with marketing research problems.

This thesis tries to give a summary of the classical approach to discriminant analysis as well as some other methods in order to see whether the techniques available are appropriate for handling all these different types of data as well as the typical small samples. As this field has been studied fairly well it is not easy to even mention all possible techniques (see section 5.4), but I think most of the basic methods have been touched.

Studying all these techniques it has become evident that much of the present day marketing research data can be analysed by means of the classical approach, even if the data are not normal continuous, but of an ordered nature (see section 3.10). If results are not satisfactory then one can use more data-specified techniques as described in chapter 3.

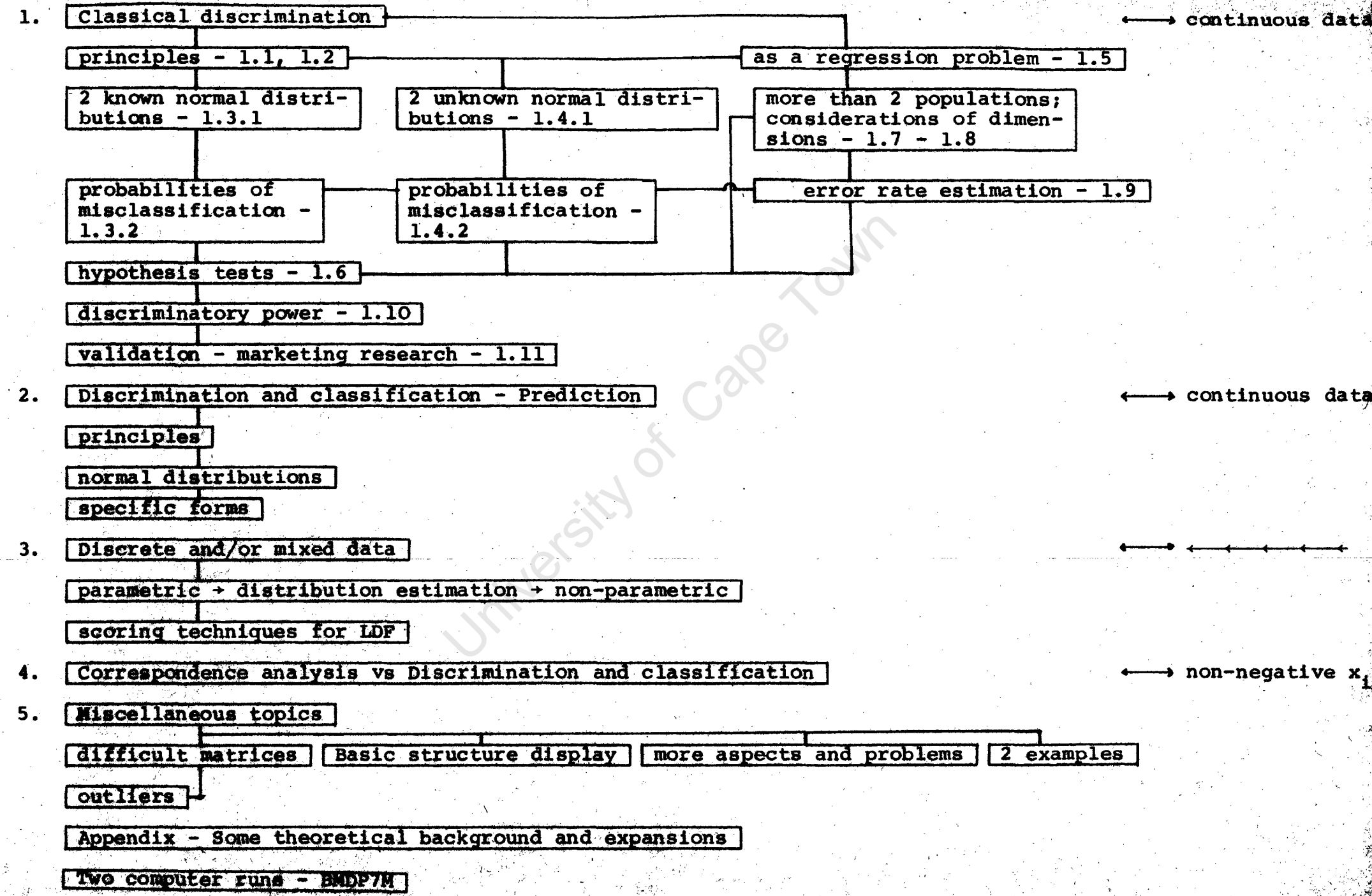
Chapter 4 has been included to show some of the relationships between discriminant analysis and correspondence analysis, where the latter is not such a well-known, but very useful and interesting procedure. In this chapter, i.e. chapter 4 and section 5.3 the

basic structure display of a data matrix was used to define the basic structures of the data matrices in correspondence analysis and discriminant analysis. Note that correspondence analysis is well suited for the positive nature of the data one often finds in marketing research.

In chapter 5 two of the more common problems in discriminant analysis (typical of multivariate analyses?) are briefly discussed and some other problems just mentioned. A very brief description is given of two typical problems in marketing research. In both cases the group sizes are fairly small and although almost no refinements, except for jackknifing, was added to the basic programs, the rates of correct classification were more than satisfactory - especially in example 2.

Finally some theoretical background and results are given in the Appendix.

C. Diagrammatically the text may be summarised as on the next page:



NOMENCLATURE

In general the following notation was used unless specified otherwise:

p	number of variables in $\underline{x}$
x_i	variable/element i of vector $\underline{x}$
$\underline{x}$	observation vector
Π_i	population i
π_i	probability associated with Π_i
$n(p,\underline{\mu},\Sigma)$	multivariate normal distribution with p variables in each vector observation
n_i, N_i	size of sample/population i
$W(p,q,\Sigma)$	The p-variate Wishart density with q degrees of freedom
Δ^2	Mahalanobis's distance squared
D^2	Mahalanobis's sample distance squared
A	the matrix of sums of squares and cross products
S	the sample variance-covariance matrix
$\Phi(\cdot)$	the cumulative normal distribution
E_{pq}=E:pxq	a pxq matrix with all elements equal to 1
$\text{diag}(a_1,\ldots,a_k)$	diagonal matrix with diagonal elements $a_1, a_2, \ldots, a_k$
tr(A)	trace of matrix A

Chapter I

CLASSICAL DISCRIMINATION

1.1 Principles leading to a solution

Initially we will discuss the case of two populations for ease of notation.

One of the principles which will help us to find a procedure is the minimisation of the cost of misclassification of an observation. There are other ways of tackling the problem as we shall see. For other short and easy reading references refer inter alia to Press (1972), Johnson and Wichern (1982) and P. Green (1978).

Let us define a rule that will classify among observations as follows (Anderson, 1958):

If an observation or individual is characterized by a certain set of values say $\underline{x}' = (x_1, x_2, x_3, \ldots, x_p)$, it will be classified as from Π_1; if it has other values it is classified as from Π_2 where Π_1 and Π_2 are the two populations. The classification thus depends on the vector of measurements $\underline{x}' = (x_1, x_2, x_3, \ldots, x_p)$. One can think in terms of a p-dimensional space which is divided into two distinct regions R_1 and R_2 referring to Π_1 and Π_2 respectively.

Let us now distinguish between the two kinds of misclassification errors with the corresponding costs of misclassification. The cost of classifying an individual as from Π_2 when it is in actual fact from Π_1 is $c(2/1)>0$ and the cost of classifying an individual from Π_1 when it is in fact from Π_2 is $c(1/2)>0$. It is of interest to note here that the units of cost is arbitrary, because eventually it is only the ratio of the two costs that is important. Although these costs may not be known exactly, the statistician may have a rough idea of their magnitude.

A good classification procedure is one which minimises the cost

1.2/...

of misclassification in some sense. Refer to table one for the appropriate costs.

		Decision	
		Π_1	Π_2
Population	Π_1	0	c(2/1)
	Π_2	c(1/2)	0

TABLE 1

1.2 The minimum cost procedure when a priori probabilities are known - A Bayes procedure.

Suppose the probability that one observation comes from Π_1 is q_1 and from Π_2 is q_2 where the density function of Π_1 is $f_1(\underline{x})$ and that of Π_2 is $f_2(\underline{x})$.

The probability of correctly classifying an observation from Π_1, i.e. region R_1 in p-space is

$$P(1/1,R) = \int_{R_1} f_1(\underline{x})d\underline{x} \qquad 1.2.1$$

where $d\underline{x} = dx_1, dx_2, dx_3, \ldots, dx_p$. The probability of correctly classifying an observation from population Π_2, i.e. region R_2 in p-space is

$$P(2/2,R) = \int_{R_2} f_2(\underline{x})dx \qquad 1.2.2$$

The probability of misclassifying an observation from Π_1 is

$$P(2/1,R) = \int_{R_2} f_1(\underline{x})dx \qquad 1.2.3$$

and the probability of misclassifying an observation from Π_2 is

$$P(1/2),R) = \int_{R_1} f_2(\underline{x})dx \qquad 1.2.4$$

The cost of procedures P(1/1,R) and P(2/2,R) is zero as can be seen from table one.

The probability of drawing an observation from Π_1 and misclassifying it is $q_1 P(2/1,R)$. The probability of drawing an observation from Π_2 and misclassifying it is $q_2 P(1/2,R)$.

The expected total cost which must be minimised is:

$$c(2/1)P(2/1,R)q_1 + c(1/2)P(1/2,R)q_2 \qquad 1.2.5$$

The minimisation of this average loss is determined by dividing the space into regions R_1 and R_2 in such a way that the expected loss is as small as possible.

Now 1.2.5 can be written as

$$c(2/1)q_1 \int_{R_2} f_1(\underline{x})d\underline{x} + c(1/2)q_2 \int_{R_1} f_2(\underline{x})d\underline{x} \qquad 1.2.6$$

Expression 1.2.6 for the expected loss can be minimised by minimising the probability of misclassification and this can be obtained by assigning to that population which has the higher "conditional cost" for each given vector $\underline{x}$. Therefore, if

$$\frac{c(2/1)q_1 f_1(\underline{x})}{c(2/1)q_1 f_1(\underline{x}) + c(1/2)q_2 f_2(\underline{x})} \geq \frac{c(1/2)q_2 f_2(\underline{x})}{c(2/1)q_1 f_1(\underline{x}) + c(1/2)q_2 f_2(\underline{x})} \qquad 1.2.7$$

we choose Π_1, otherwise we choose Π_2. The choice in case of equality is arbitrary.

Since minimisation is applied at each point, we minimise over the whole space.

The alternatives can be set out as follows:

Choose R_1 and R_2 according to

$$R_1 : c(2/1)q_1 f_1(\underline{x}) \geq c(1/2)q_2 f_2(\underline{x}) \quad \text{or} \quad \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{c(1/2)q_2}{c(2/1)q_1} \qquad 1.2.8$$

1.4/...

$$R_2 : c(2/1)q_1 f_1(\underline{x}) < c(1/2)q_2 f_2(x) \text{ or } \frac{f_1(\underline{x})}{f_2(\underline{x})} < \frac{c(1/2)q_2}{c(2/1)q_1} \quad 1.2.9$$

1.3.1 Classification into one of two known multivariate normal populations

Observe the two multivariate normal populations with equal covariance matrices, namely $\underline{x}_1 \sim n(p, \underline{\mu}_1, \Sigma)$ and $\underline{x}_2 \sim n(p, \underline{\mu}_2, \Sigma)$ where $\underline{x}_i' = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})$ and $\underline{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \ldots, \mu_{ip})$. Let us assume initially that $\Sigma_1 = \Sigma_2 = \Sigma$. Then it follows that $\frac{f_1(\underline{x})}{f_2(\underline{x})}$ which can also be seen as a likelihood ratio (Morrison, 1976), can be written as

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} = \lambda = \frac{(2\pi)^{-\frac{1}{2}p}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)\right]}{(2\pi)^{-\frac{1}{2}p}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\right]} \quad 1.3.1.1$$

$$= \exp\Big[-\tfrac{1}{2}\underline{x}'\Sigma^{-1}\underline{x} + \tfrac{1}{2}\underline{\mu}_1'\Sigma^{-1}\underline{x} + \tfrac{1}{2}\underline{x}'\Sigma^{-1}\underline{\mu}_1 - \tfrac{1}{2}\underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 + \tfrac{1}{2}\underline{x}'\Sigma^{-1}\underline{x} - \tfrac{1}{2}\underline{x}'\Sigma^{-1}\underline{\mu}_2 - \tfrac{1}{2}\underline{\mu}_2'\Sigma^{-1}\underline{x} + \tfrac{1}{2}\underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2\Big]$$

$$= \exp\left[(\underline{\mu}_1'-\underline{\mu}_2')\Sigma^{-1}\underline{x} - \tfrac{1}{2}(\underline{\mu}_1'-\underline{\mu}_2')\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2)\right] \quad 1.3.1.2$$

$$\therefore \log\lambda = (\underline{\mu}_1'-\underline{\mu}_2')\Sigma^{-1}\underline{x} - \tfrac{1}{2}(\underline{\mu}_1'-\underline{\mu}_2')\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2) \quad 1.3.1.3$$

We call $(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x}$ the discriminant function and $\log\lambda$ the classification function.

In 1.2.8 let $k = \frac{c(1/2)q_2}{c(2/1)q_1}$ where the costs of misclassification are equal and $q_1 = q_2$, then $k = 1$, i.e. $\log k = 0$.
Then the classification regions are:

R_1 when $\log\lambda \geq 0$, i.e. $(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) \geq 0$

R_2 when $\log\lambda < 0$, i.e. $(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2) < 0$

1.3.1.4

1.5/...

i.e.

$$R_1 : (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} > \tfrac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2)$$

$$R_2 : (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} < \tfrac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2) \qquad 1.3.1.5$$

1.3.2 The corresponding probabilities of misclassification

Let $U = (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} - \tfrac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2)$ where $\underline{x}\sim n(p,\underline{\mu}_1,\Sigma)$ of $\underline{x}$ is from Π_1, then

$$U \sim n\Big((\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{\mu}_1 - \tfrac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2),\ (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)\Big) \qquad 1.3.2.1$$

i.e.

$$U \sim n\Big(\tfrac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2),\ (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)\Big) \qquad 1.3.2.2$$

Let $\Delta_p^2 = (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$ where Δ_p is Mahalanobis's distance between Π_1 and Π_2, then

$$U \sim n(\tfrac{1}{2}\Delta_p^2,\ \Delta_p^2) \qquad 1.3.2.3$$

Similarly, if $\underline{x} \sim n(p,\ \underline{\mu}_2,\ \Sigma)$, then

$$U \sim n(-\tfrac{1}{2}\Delta_p^2,\ \Delta_p^2)$$

Taking into account the cost of misclassification and taking $\log k = c$, then

$$P(2/1) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\Pi}\Delta_p} \exp\left[-\tfrac{1}{2}(z-\tfrac{1}{2}\Delta_p^2)^2/\Delta_p^2\right] dz \qquad 1.3.2.5$$

i.e.

$$P(2/1) = \Phi\left[(c-\tfrac{1}{2}\Delta_p^2)/\Delta_p\right] \qquad 1.3.2.6$$

and

$$P(1/2) = 1 - \Phi\left[(c+\tfrac{1}{2}\Delta_p^2)/\Delta_p\right] \qquad 1.3.2.7$$

so that c can be solved and therefore the optimum ratio for

1.6/....

$\frac{q_2 c(1/2)}{q_1 c(2/1)}$ can be determined in a mini-max solution where

$$c(1/2)\ (1-\phi\left((c+\tfrac{1}{2}\Delta_p^2)/\Delta_p\right)) = c(2/1)\phi\left((c-\tfrac{1}{2}\Delta_p^2)/\Delta_p\right) \qquad 1.3.2.8$$

using a method of trial and error in the normal tables.

1.4.1 Classification into one of two multivariate normal populations with the parameters Σ, $\underline{\mu}_1$ and $\underline{\mu}_2$ unknown - a sampling procedure

When the parameters are unknown we would have liked to substitute for $\underline{\mu}_1$, $\underline{\mu}_2$ and Σ using the unbiased estimators. If one would substitute them into 1.3.1.3, 1.3.1.4 and further, there is no basis for assuming that the principle of optimality with respect to cost is still applicable (Anderson, 1958).

An appropriate method is Fisher's approach using the union intersection principle of Roy (see Morrison, 1976) to obtain the maximum distance between centroids.

Define the two multivariate normal populations as before with equal covariance matrices such that $\bar{\underline{x}}_1 \sim n\ (p,\ \underline{\mu}_1,\ \frac{1}{N_1}\Sigma)$ and $\bar{\underline{x}}_2 \sim n\ (p,\ \underline{\mu}_2,\ \frac{1}{N_2}\ \Sigma)$, then

$$\bar{\underline{x}}_1 - \bar{\underline{x}}_2 \sim n(p,\ \underline{\mu}_1-\underline{\mu}_2,\ \frac{N_1+N_2}{N_1N_2}\ \Sigma) \qquad 1.4.1.1$$

i.e.

$$\underline{a}\cdot(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \sim n(p,\ \underline{a}(\underline{\mu}_1-\underline{\mu}_2),\ \underline{a}'\frac{N_1+N_2}{N_1N_2}\ \Sigma\ \underline{a}) \qquad 1.4.1.2$$

which is a univariate normal distribution for any non-zero $\underline{a}$: pxl.

Let S be the unbaised estimator for Σ, then

1.7/...

$$t^2(\underline{a}) = \frac{\left[\underline{a}'(\bar{\underline{x}}_1-\bar{\underline{x}}_2)\right]^2 \cdot \frac{N_1N_2}{N_1+N_2}}{\underline{a}'S\underline{a}}$$

$$= \frac{\underline{a}'(\bar{\underline{x}}_1-\bar{\underline{x}}_2)(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'\underline{a}}{\underline{a}'S\underline{a}} \cdot \frac{N_1N_2}{N_1+N_2} \qquad 1.4.1.3$$

Now $(\bar{\underline{x}}_1-\bar{\underline{x}}_2)(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'$ is pxp and symmetrical and S is pxp and symmetrical so that the maximum for $t^2(\underline{a})$ exists where $\underline{a}$ is the eigenvector of $\frac{N_1N_2}{N_1+N_2} S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'$ which corresponds with the largest eigen root and this maximum of $t^2(\underline{a})$ equals the largest eigen root of $\frac{N_1N_2}{N_1+N_2} S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'$ and equals $(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)\frac{N_1+N_2}{N_1N_2} = T^2$ and the corresponding eigen-vector is $S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) = \underline{a}$. (See Graybill, 1969).

We now have an $\underline{a}$ such that $t^2(a)$ is maximised, i.e. the centroids of the two samples are thus determined that we have a "maximum" or "optimal" difference between them.

Our index is thus

$$\underline{a}'\underline{x} = \left[S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)\right]'\underline{x}$$

$$\text{i.e. } \underline{a}'\underline{x} = (\underline{x}_1-\underline{x}_2)'S^{-1}\underline{x} = y \qquad 1.4.1.4$$

The mean values of the indices of the two samples are now $\bar{y}_1 = (\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}\bar{\underline{x}}_1$ and $\bar{y}_2 = (\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}\bar{\underline{x}}_2$.

The midpoint of these mean values on a discriminant function scale is

$$\frac{\bar{y}_1+\bar{y}_2}{2} = \tfrac{1}{2}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1+\bar{\underline{x}}_2) \qquad 1.4.1.5$$

from which we have our classification rule:

Allocate the individual with response $\underline{x}$

to population 1 if $(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}\underline{x} > \frac{1}{2}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)$

to population 2 if $(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}\underline{x} < \frac{1}{2}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)$, 1.4.1.6

i.e. sample units are allocated to that group nearest to it with respect to the mean score.

$\frac{\bar{y}_1+\bar{y}_2}{2}$ is the value of a random variable and therefore the rule can be summarised in the classification function:

$$W = (\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}\bar{\underline{x}} - \tfrac{1}{2}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1+\bar{\underline{x}}_2) \qquad 1.4.1.7$$

where $\underline{x}$ is allocated to Π_1 if $W>0$ and to population 2 if $W<0$. If $\log k \neq 0$, this "optimal" rule for "maximum distance" would be changed by the subtraction of a constant.

It is important to note that prior to the application of any classification rule, one should determine first whether the group centroids differ significantly, i.e. whether any allocation will have any meaning. For this Hotelling's T^2 test measure can be applied:

$H_o : \mu_1 = \mu_2$ against the alternative hypothesis. Calculate

$T^2 = \frac{N_1N_2}{N_1+N_2}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)$ where $F = \frac{N_1+N_2-p-1}{(N_1+N_2-2)p} \cdot T^2$

is distributed as the F distribution with p and N_1+N_2-p-1 degrees of freedom if H_o is true.

1.4.2 Misclassification probabilities associated with 1.4.1

Estimation of misclassification probabilities is difficult in such a case.

If $\underline{x} \sim n(p, \underline{\mu}_1, \Sigma)$ then it follows that

$$\frac{(\underline{x}-\underline{\mu}_1)'\underline{a}}{\sqrt{\underline{a}'\Sigma\underline{a}}} \sim n(0,1) \qquad 1.4.2.1$$

if $\underline{a}$ is independent of $\underline{x}$, i.e.

$$P(2/1) = P(W<0)$$

$$= P\left[\frac{\underline{x}'\underline{a}-\underline{\mu}_1'\underline{a}}{\sqrt{\underline{a}'\Sigma\underline{a}}} < \frac{\frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'\underline{a}-\underline{\mu}_1'\underline{a}}{\sqrt{\underline{a}'\Sigma\underline{a}}}\right]$$

$$= \Phi\left[\frac{\frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'\underline{a}-\underline{\mu}_1'\underline{a}}{\sqrt{\underline{a}'\Sigma\underline{a}}}\right] \qquad 1.4.2.2$$

and similarly

$$P(1/2) = \Phi\left[\frac{\underline{\mu}_2'\underline{a} - \frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'\underline{a}}{\sqrt{\underline{a}'\Sigma\underline{a}}}\right] \qquad 1.4.2.3$$

but Σ, $\underline{\mu}_1$ and $\underline{\mu}_2$ are all unknown.

Various estimates of $P_1 = P(2/1)$ and $P_2 = P(1/2)$ were suggested in Lachenbruch and Mickey (February, 1968) and are shown in more detail in Kshirsagar (1972). We will have a look only at those estimates which are fairly useful and common, i.e. the OS, U, $\bar{U}$ and "apparent" methods as described below.

Lachenbruch et al. stated that if approximate normality can be assumed then the OS and $\bar{U}$ methods are good, but if $(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) = D^2$, the sample Mahalanobis distance squared, is small (say < 1,0) or the sample size is small relative to the number of parameters, the OS method should not be used. In such a case the $\bar{U}$ or U method should be used.

When normality is questionable and the sample size is small

1.10/...

relative to the number of variables the U method should be used.

(i) The OS method:

Estimate D^2 by $DS = (m-p-3)D^2 / (N_1+N_2-2)$ and use the asymptotic expansion for the distribution of W as given by Okamoto (1963) to estimate P_1 and P_2. Note that $m = N_1+N_2$.

(ii) The U method:

Morrison (1976) describes the U method as follows:

Calculate for each observation $\underline{x}_i$ from the k th sample.

$$W_i = \left\{\underline{x}_i - \tfrac{1}{2}\left[\bar{\underline{x}}_1+\bar{\underline{x}}_2 - \frac{1}{N_k-1}(\underline{x}_i-\bar{\underline{x}}_k)\right]\right\}' S_i^{-1}\left[\bar{\underline{x}}_1-\bar{\underline{x}}_2+\frac{(-1)^k}{N_k-1}(\underline{x}_i-\bar{\underline{x}}_k)\right]$$

1.4.2.4

where

$$S_i^{-1} = \frac{N_1+N_2-3}{N_1+N_2-2}\left(S^{-1} + \frac{C_k}{1-C_k(\underline{x}_i-\bar{\underline{x}}_k)'S^{-1}(\underline{x}_i-\bar{\underline{x}}_k)} S^{-1}(\underline{x}_i-\bar{\underline{x}}_k)(\underline{x}_i-\bar{\underline{x}}_k)'S^{-1}\right)$$

and $C_k = \frac{N_k}{(N_k-1)(N_1+N_2-2)}$. In this formulation N_1+N_2 statistics are calculated, but with the particular individual observation omitted in the calculation of the linear discriminant coefficients, means and covariance matrix.

As before if $W_i \geq 0$ allocate to population 1
and if $W_i < 0$ allocate to population 2.

The probabilities P_1 and P_2 are now estimated by the proportion of misclassified cases for each group.

Note that for computational purposes W_i can be expressed as

1.11/...

$$W_i = \frac{N-3}{N-2}\left(\underline{x}_i' S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) - \tfrac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)\right.$$

$$+ \frac{1}{1-C_k(\underline{x}_i-\bar{\underline{x}}_k)'S^{-1}(\underline{x}_i-\bar{\underline{x}}_k)}\cdot\left\{C_k\left[(\underline{x}_i-\bar{\underline{x}}_k)'S^{-1}(\underline{x}_i-\bar{\underline{x}}_k)\right]\left[(\underline{x}_i-\bar{\underline{x}}_k)'S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)\right]\right.$$

$$+ C_k\left[(\underline{x}_i-\bar{\underline{x}}_k)'S^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)\right]^2$$

$$\left.\left. + (-1)^k \frac{2N_k-1}{2(N_k-1)^2}\left[(\underline{x}_i-\bar{\underline{x}}_k)'S^{-1}(\underline{x}_i-\bar{\underline{x}}_k)\right]^2\right\}\right) \qquad 1.4.2.5$$

This technique is similar to the so-called "jackknife" approach as described in Section 1.11. This is one of the reasons why it is fairly robust with respect to normality.

For the derivation of these W_i's refer to Kshirsagar (1972) keeping in mind that Kshirsagar's S is equivalent to our A.

(iii) The $\bar{U}$ method:

Unlike the U method which is not influenced sharply by a deviation from normality, the $\bar{U}$ method depends explicitly on the assumption that the linear discriminant variate is normally distributed.

Compute

$$\bar{W}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} W_{ik}$$

where

$$s_k^2 = \frac{1}{N_k-1}\sum_{i=1}^{N_k} W_{ik}-\bar{W}_k \qquad k = 1, 2 \qquad 1.4.2.6$$

using the scores in 1.4.2.4 and 1.4.2.5, then the estimates will be

$$\hat{P}_1 = \left(\frac{-\bar{W}_1}{s_1}\right), \quad \hat{P}_2 = \left(\frac{\bar{W}_2}{s_2}\right) \qquad 1.4.2.7$$

The derivation of this can be found in Kshirsagar (1972).

1.12/....

(iv) The resubstitution method:

This method provides us with the estimated error rate which is commonly known as the "apparent" error rate. What the method obtains, is the ratio of misclassified observations where the classification function was determined on all N_1+N_2 elements and by resubstituting the observations by means of which these calculations had been made. This technique is often misleading and gives estimates of P_1 and P_2 that are too optimistic, i.e. too small, as the same observations are used to compute the classification function and also to evaluate its performance which obviously results in biased results.

Another not often used technique is to split the samples into two groups. Do all the calculations on the one group, i.e. use it as a training set. Assess the performance of the resulting function using the other group. This is not always practical as total sample sizes are often relatively small which disencourages the further decrease of the number of observations for calculation purposes in both cases.

Much work has been done on the estimation of error rates, with respect to already classified observations as well as new responses. Confidence intervals have been determined for the misclassification probabilities (see Lachenbruch, 1967). Sayre (1980) derived some actual and asymptotic error rates under certain conditions. A summary of some of the latest developments in this respect may be found in Dillon (1979).

1.5 Classification as a regression problem

Kshirsagar (1972) posed the classification also as a regression problem.

Let $\bar{\underline{x}} = \frac{1}{n_1} X E_{n_1 1}$, $\bar{\underline{y}} = \frac{1}{n_2} Y E_{n_2 1}$ where E_{pxq} is a pxq matrix with all elements equal to 1. Let X_{pxn_1} and Y_{pxn_2} be the sample observations from Π_1 and Π_2 respectively with $n(p,\underline{\mu}_i,\Sigma)$

populations. Let $\underline{x}$ be the observation to be classified. Then

$$A_x = X(I - \frac{1}{n_1} E_{n_1 n_1})X', \quad A_y = Y(I - \frac{1}{n_2} E_{n_2 n_2})Y', \quad A = A_x + A_y$$

$$S = \frac{1}{n_1+n_2-2} A, \quad \underline{d} = \bar{\underline{x}} - \bar{\underline{y}} ,$$

Using the exposition as given it follows that

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\, \underline{d} \sim n(p, \sqrt{\frac{n_1 n_2}{n_1+n_2}}\, \underline{\delta}, \Sigma)$$

where $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$. Note that $\sqrt{\frac{n_1 n_2}{n_1+n_2}}\underline{d}$ is independent of A where

$$A \sim W(p, n_1+n_2-2, \Sigma)$$

Therefore, $\underline{d}$ and S are unbiased estimates of $\underline{\delta}$ and Σ.

Then Fisher's discriminant function becomes

$$\underline{a}'\underline{x} = \underline{d}\, S^{-1}\, \underline{x} \qquad 1.5.1$$

and is known as the sample discriminant function.

Now, using the same notation, define the dummy variable ξ as

$$\xi = \lambda_1 \quad \text{if an individual belongs to } \Pi_1$$

$$= \lambda_2 \quad \text{if an individual belongs to } \Pi_2. \qquad 1.5.2$$

and we can write

$$E(\underline{x}) = \underline{\alpha} + \underline{\beta}\xi \qquad 1.5.3$$

where $\underline{x}$ is the vector of measurements for the individual, and

1.14/...

$$\underline{\alpha} = \frac{1}{\lambda_1-\lambda_2}(\lambda_1\underline{\mu}_2-\lambda_2\underline{\mu}_1),\ \underline{\beta} = \frac{1}{\lambda_1-\lambda_2}(\underline{\mu}_1-\underline{\mu}_2) \qquad 1.5.4$$

so that $E(\underline{x}) = \underline{\mu}_1$ when $\xi=\lambda_1$ and $E(\underline{x}) = \underline{\mu}_2$ when $\xi = \lambda_2$. From this it can be seen that 1.5.3 can be looked upon as the regression equation of $\underline{x}$ on ξ.

The problem however, is to predict ξ given $\underline{x}$ in order to be able to decide whether to assign the individual to Π_1 or Π_2, not the opposite. We must find the regression of ξ on $\underline{x}$.

Observe the following table 2 for the $N = n_1+n_2$ observations on $\underline{x}$

Variable	Observations on the n_1 individuals from Π_1	Observations on the n_2 individuals from Π_2
$\underline{x}$	X	Y
ξ	λ_1 (n_1 times)	λ_2 (n_2 times)

TABLE 2

The matrix of corrected sums of squares and sums of products of all the $n_1+n_2 = N$ observations on $\underline{x}$, $\underline{y}$ is

$$[X \vdots Y]\,[I - \frac{1}{N}E_{NN}]\,[X \vdots Y]'$$

$$= X(I - \frac{1}{n_1}E_{n_1n_2})X' + Y(I - \frac{1}{n_2}E_{n_2n_2})Y' + \frac{n_1n_2}{n_1+n_2}\underline{d}\underline{d}'$$

$$= A + \frac{n_1n_2}{n_1+n_2}\underline{d}\underline{d}' \qquad 1.5.5$$

1.15/...

The matrix of the corrected sums of products of $\underline{x}$ with ξ is

$$[X \vdots Y]\ (I - \tfrac{1}{N} E_{NN})[\lambda_1 E_{1n_1} \vdots \lambda_2 E_{1n_2}]'$$

$$= \frac{n_1 n_2}{n_1+n_2} (\lambda_1-\lambda_2)\underline{d} \qquad 1.5.6$$

The matrix of the sums of squares of observations of ξ is

$$\begin{bmatrix}\lambda_1 E_{n_1 1} \\ \lambda_2 E_{n_2 1}\end{bmatrix}' (I - \tfrac{1}{N} E_{NN}) \begin{bmatrix}\lambda_1 E_{1n_1} \vdots \lambda_2 E_{1n_2}\end{bmatrix}'$$

$$= \frac{n_1 n_2}{n_1+n_2} (\lambda_1 - \lambda_2)^2 \qquad 1.5.7$$

Let the regression of ξ on $\underline{x}$ now be

$$\text{constant} + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \qquad 1.5.8$$

where we use the method of least squares to minimise the sum of squares of deviations of ξ from its regression. Then $\underline{b}$ satisfies the normal equations

$$\frac{n_1 n_2}{n_1+n_2} (\lambda_1-\lambda_2)\underline{d} = (A + \frac{n_1 n_2}{n_1+n_2} \underline{d}\underline{d}')\underline{b} \qquad 1.5.9$$

using 1.5.5, 1.5.6 and 1.5.7. From this now follows that

$$\underline{b} = \frac{n_1 n_2}{n_1+n_2}(\lambda_1-\lambda_2)(A + \frac{n_1 n_2}{n_1+n_2} \underline{d}\underline{d}')^{-1} \underline{d} \qquad 1.5.10$$

Take now into account that[1)]

[1)] A_{pxp}; $|A| \neq 0$, $\underline{b}$: pxl, $c \neq 0$, then $(A+c\underline{b}\underline{b}')^{-1} = A^{-1} - \frac{c}{1+c\underline{b}'A^{-1}\underline{b}} A^{-1}\underline{b}\underline{b}A^{-1}$; Bartlett (1951).

1.16/...

$$(A + \frac{n_1 n_2}{n_1+n_2}\,\underline{d}\underline{d}')^{-1} = (I + \frac{n_1 n_2}{n_1+n_2}\,A^{-1}\,\underline{d}\underline{d}')^{-1}\,A^{-1} \qquad 1.5.11$$

$$= \left(I - \frac{n_1 n_2}{n_1+n_2} \cdot \frac{A^{-1}\,\underline{d}\underline{d}'}{1+\frac{n_1 n_2}{n_1+n_2}\,\underline{d}'A^{-1}\underline{d}}\right)^{-1} A^{-1}$$

1.5.12

Equation 1.5.10 now reduces to

$$\underline{b} = \frac{\frac{n_1 n_2}{n_1+n_2}\,(\lambda_1-\lambda_2)}{(1+\frac{n_1 n_2}{n_1+n_2} \cdot \frac{D_p^2}{n_1+n_2-2})}\,A^{-1}\,\underline{d} \qquad 1.5.13$$

where $D_p^2 = \underline{d}'S^{-1}\underline{d}$, so that $\underline{b}$ is proportional to $A^{-1}\underline{d}$ and thus $\underline{b}'\underline{x}$ to $\underline{a}'\underline{x}$, the sample discriminant function. Apart from a constant of proportionality the discriminant and regression functions are similar and will lead to the same classification procedure as before.

At this stage the following questions may well be asked:

(i) Are these b_i values valid?

(ii) How do we test before hand for the hypothesis $\underline{\mu}_1=\underline{\mu}_2$?

Thus let us take the discussion a little further, because if $\underline{\mu}_1 = \underline{\mu}_2$ there is not really any sense in trying to discriminate with respect to centroids. Further we would like to be able to apply a hypothesis test to each b_i.

The regression sum of squares of ξ on $\underline{x}$ is

$$SSR(\underline{x}) = \frac{n_1 n_2}{n_1+n_2}\,(\lambda_1-\lambda_2)\underline{b}'\underline{d} \qquad 1.5.14$$

1.17/...

$$= \frac{n_1 n_2}{n_1+n_2} (\lambda_1-\lambda_2)^2 \frac{\frac{n_1 n_2}{n_1+n_2} D_p^2}{n_1+n_2-2+\frac{n_1 n_2}{n_1+n_2} D_p^2} \qquad 1.5.15$$

because $((n_1+n_2-2)A^{-1}\underline{d})'\underline{d} = \underline{d}'S^{-1}\underline{d} = D_p^2$, so that we have the following ANOVA table.

Regression of ξ on $\underline{x}$

Source	d.f.	Sums of squares	F
SSR($\underline{x}$)	p	$\frac{n_1 n_2}{n_1+n_2} (\lambda_1-\lambda_2)^2 \frac{\frac{n_1 n_2}{n_1+n_2} D_p^2}{n_1+n_2-2+\frac{n_1 n_2}{n_1+n_2} D_p^2}$	$\frac{n_1+n_2-p-1}{p} \cdot \frac{\frac{n_1 n_2}{n_1+n_2} D_p^2}{n_1+n_2-2}$
Error S.S.	n_1+n_2-p-1	$\frac{n_1 n_2}{n_1+n_2} (\lambda_1-\lambda_2)^2 \frac{n_1+n_2-2}{n_1+n_2-2+\frac{n_1 n_2}{n_1+n_2} D_p^2}$	
Total	n_1+n_2-1	$\frac{n_1 n_2}{n_1+n_2} (\lambda_1-\lambda_2)^2$	

TABLE 3

There is no association between $\underline{x}$ and ξ if $\underline{\beta} = 0$, i.e. if $\underline{\mu}_1 = \underline{\mu}_2$.

Although the basic assumptions in a standard regression analysis are not satisfied - i.e. the dependent variable ξ must have a normal distribution and the independent variable $\underline{x}$ must be fixed - it can be proven that the hypothesis

$$H_o : \underline{\mu}_1 = \underline{\mu}_2$$

can be tested by

$$F = \frac{\text{Regression SS/d.f.}}{\text{Error SS/d.f.}} = \frac{n_1+n_2-p-1}{p} \cdot \frac{\frac{n_1n_2}{n_1+n_2} D_p^2}{n_1+n_2-2} \qquad 1.5.16$$

where F has the F distribution with p and n_1+n_2-p-1 degrees of freedom when $\underline{\mu}_1=\underline{\mu}_2$, i.e. when $\Delta_p^2=0$.

For further information regarding the distribution of D_p^2 refer Appendix B.

At this point it is interesting to note the relationship between Hotelling's T^2, Mahalanobis's D^2 and Fisher's R^2 where R is the multiple correlation coefficient between ξ and $\underline{x}$ (Lachenbruch, 1968).

$$R^2 = \frac{\text{regression S.S.}}{\text{total S.S.}}$$

$$= \frac{\frac{n_1n_2}{n_1+n_2} D_p^2 \;/\; (n_1+n_2-2)}{1+\frac{n_1n_2}{n_1+n_2} D_p^2 \;/\; (n_1+n_2-2)} \qquad 1.5.17$$

and from this follows directly that

$$R^2\left(1 + \frac{n_1n_2}{n_1+n_2} D_p^2 \;/\; (n_1+n_2-2)\right) = \frac{n_1n_2}{n_1+n_2} D_p^2 \;/\; (n_1+n_2-2) \qquad 1.5.18$$

$$= \frac{T_p^2}{n_1+n_2-2} \qquad 1.5.19$$

$$= \frac{R^2}{1-R^2} \qquad 1.5.20$$

This is handy because hypothesis tests on R^2 may measure the discriminating ability as determined by the regression method.

1.19/...

A further point of interest is the fact that λ_1 and λ_2 are never used in test 1.5.16 so that one can use more convenient values like $\lambda_1 = 1$ and $\lambda_2 = 0$ or 1, -1, etc. Fisher used

$\lambda_1 = \frac{n_2}{n_1+n_2}$ and $\lambda_2 = \frac{-n_1}{n_1+n_2}$ from which follows that $\lambda_1-\lambda_2 = 1$ and $\bar{\xi} = 0$. For a full discussion of this result refer Anderson (1958). Different choices of λ_1 and λ_2 yield different values of $\underline{b}$, but all such $\underline{b}$'s are proportional to each other and to $\underline{a}'\underline{x}$. This however doesn't matter in discriminant analysis, as we have to standardise the discriminant function by dividing by its standard deviation before using it (see Johnson and Wichern, 1982).

This difference between regression and discrimination is important; the regression coefficients are unique while the discriminant coefficients are not - only their ratios are unique.

In standard least square theory Var ($\underline{b}$) is equal to $(A + \frac{n_1n_2}{n_1+n_2}\,\underline{d}\underline{d}')^{-1}\sigma^2$ where σ^2 is the variance of ξ and is estimated from the analysis of variance table. This is not true here, because ξ is not normally distributed. If we however write

$$\hat{V}(\underline{b}) = (A + \frac{n_1n_2}{n_1+n_2}\,\underline{d}\underline{d}')^{-1}\hat{\sigma}^2$$

$$= (A + \frac{n_1n_2}{n_1+n_2}\,\underline{d}\underline{d}')^{-1}\,\frac{S.S.E.}{n_1+n_2-p-1}$$

$$= (A + \frac{n_1n_2}{n_1+n_2}\,\underline{d}\underline{d}')^{-1}\,\frac{\frac{n_1n_2}{n_1+n_2}(\lambda_1-\lambda_2)^2}{n_1+n_2-p-1}\cdot\frac{n_1+n_2-2}{n_1+n_2-2+\frac{n_1n_2}{n_1+n_2}D_p^2}$$

1.20/....

$$= (A + \frac{n_1 n_2}{n_1+n_2}\,\underline{d}\underline{d}') \frac{\frac{n_1 n_2}{n_1+n_2}(\lambda_1-\lambda_2)^2}{(n_1+n_2-2)-p+1} \cdot \frac{n_1+n_2-2}{n_1+n_2-2+\frac{n_1 n_2}{n_1+n_2} D_p^2}$$

1.5.21

then

$$\frac{b_i}{\{\hat{V}(b_i)\}^{\frac{1}{2}}} = t_{(n_1+n_2-2)-p+1} \qquad 1.5.22$$

to test the significance of b_i.

For the derivation of the covariance matrix of $\underline{b}$ see Kshirsagar (1972), Kendall (1982) or Appendix D.

1.6.1 Tests associated with discriminant analysis

1.6.1.1 Notation

Kshirsagar (1972) as well as Rao (1966) paid attention to the different tests which may be applied to a discriminant analysis. The notation as well as the tests are taken from these two references.

$D_p^2 = \underline{d}'S^{-1}\underline{d}$ is Mahalanobis's sample distance squared with distribution of D_p^2 given by

$$H_p' \left(\frac{n_1 n_2}{n_1+n_2} D_p^2 \,/\, n_1+n_2-2,\ \frac{n_1 n_2}{n_1+n_2}\Delta_p^2\right) d \left(\frac{n_1 n_2}{n_1+n_2} D_p^2\right) \qquad 1.6.1.1.1$$

with

$$E(D_p^2) = \frac{n_1+n_2-2}{(n_1+n_2-2)-(p-1)} \left(\Delta_p^2 + \frac{p(n_1+n_2)}{n_1 n_2}\right), \qquad 1.6.1.1.2$$

i.e. D_p^2 is not unbiased as an estimator for Δ_p^2, the

1.21/...

population Mahalanobis distance. See Appendix B for the distribution of D_p^2.

Therefore an unbiased estimator for Δ_p^2 is

$$\frac{(n_1+n_2-2) - (p-1)}{n_1+n_2-2} D_p^2 - \frac{(n_1+n_2)p}{n_1 n_2} \qquad 1.6.1.1.3$$

Partition $\underline{x}$, $\underline{\delta}$, Σ, $\underline{d}$, A, $\underline{a}$ as follows:

$$\underline{x} = \begin{bmatrix} \underline{x}_1 \\ \cdots \\ \underline{x}_2 \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix} \qquad \underline{\delta} = \begin{bmatrix} \underline{\delta}_1 \\ \cdots \\ \underline{\delta}_2 \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix}$$

$$\underline{a} = \Sigma^{-1}\underline{\delta} = \begin{bmatrix} \underline{a}_1 \\ \cdots \\ \underline{a}_2 \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix} \qquad \underline{d} = \begin{bmatrix} \underline{d}_1 \\ \cdots \\ \underline{d}_2 \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \cdots & \vdots & \cdots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix} \qquad A = \begin{bmatrix} A_{11} & \vdots & A_{12} \\ \cdots & \vdots & \cdots \\ A_{21} & \vdots & A_{22} \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix}$$

$$\begin{matrix} k & p-k \end{matrix} \qquad \begin{matrix} k & p-k \end{matrix} \qquad 1.6.1.1.4$$

with $\underline{d}$ and $\underline{\delta}$ as defined before. Further, let

$$\underset{(p-k)\,x\,k}{\beta} = \Sigma_{21}\Sigma_{11}^{-1} \qquad \underset{(p-k)\,x\,k}{B} = A_{21}A_{11}^{-1}$$

$$\underset{(p-k)\,x\,(p-k)}{\Sigma_{22.1}} = \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \qquad \underset{(p-k)\,x\,(p-k)}{A_{22.1}} = A_{22}-A_{21}A_{11}^{-1}A_{12}$$

1.22/...

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} + \beta\Sigma_{22.1}^{-1}\beta & \vdots & -\beta'\Sigma_{22.1}^{-1} \\ \cdots\cdots\cdots\cdots & \vdots & \cdots\cdots \\ -\Sigma_{22.1}\beta & \vdots & \Sigma_{22.1}^{-1} \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix}$$

$$\begin{matrix} k & & p-k \end{matrix}$$

and

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + BA_{22.1}^{-1}B & \vdots & -B'A_{22.1}^{-1} \\ \cdots\cdots\cdots\cdots & \vdots & \cdots\cdots \\ -A_{22.1}B & \vdots & A_{22.1}^{-1} \end{bmatrix} \begin{matrix} k \\ \\ p-k \end{matrix} \qquad 1.6.1.1.5$$

$$\begin{matrix} k & & p-k \end{matrix}$$

With respect to Δ_p^2 we can write

$$\Delta_p^2 = \underline{\delta}'\Sigma^{-1}\underline{\delta}$$

$$= \begin{bmatrix} \underline{\delta}_1 \\ \cdots \\ \underline{\delta}_2 \end{bmatrix}' \begin{bmatrix} \Sigma_{11}^{-1}+\beta'\Sigma_{22.1}^{-1}\beta & \vdots & -\beta'\Sigma_{22.1}^{-1} \\ \cdots\cdots\cdots\cdots & \vdots & \cdots\cdots \\ -\Sigma_{22.1}^{-1}\beta & \vdots & \Sigma_{22.1}^{-1} \end{bmatrix} \begin{bmatrix} \underline{\delta}_1 \\ \cdots \\ \underline{\delta}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \underline{\delta}_1'\left(\Sigma_{11}^{-1} + \beta'\Sigma_{22.1}^{-1}\beta\right) & -\underline{\delta}_2'\Sigma_{22.1}^{-1}\beta & \vdots & -\underline{\delta}_1'\beta'\Sigma_{22.1}^{-1} + \underline{\delta}_2'\Sigma_{22.1}^{-1} \end{bmatrix} \begin{bmatrix} \underline{\delta}_1 \\ \vdots \\ \underline{\delta}_2 \end{bmatrix}$$

$$= \underline{\delta}_1'\Sigma_{11}^{-1}\underline{\delta}_1 + \underline{\delta}_1'\beta\Sigma_{22.1}^{-1}\beta\underline{\delta}_1 - \underline{\delta}_1'\beta'\Sigma_{22.1}^{-1}\underline{\delta}_2 - \underline{\delta}_2'\Sigma_{22.1}\beta\underline{\delta}_1 + \underline{\delta}_2'\Sigma_{22.1}^{-1}\underline{\delta}_2$$

$$= \Delta_k^2 + (\underline{\delta}_2-\beta\underline{\delta}_1)'\,\Sigma_{22.1}^{-1}(\underline{\delta}_2-\beta\underline{\delta}_1) \qquad 1.6.1.1.6$$

Similarly

$$D_p^2 = D_k^2 + (\underline{d}_2-B\underline{d}_1)'\,S_{22.1}^{-1}(\underline{d}_2-B\underline{d}_1) \qquad 1.6.1.1.7$$

with $S_{22.1} = \frac{1}{n_1+n_2} A_{22.1}$, and Δ_k^2, D_k^2 are the true and "studentised" squared distances between Π_1 and Π_2 based on

1.23/....

the first k variables in $\underline{x}_1$ only. The increase in the distance between Π_1 and Π_2 due to variables $x_{k+1}, \ldots, x_p$ over the distance based on $x_1, \ldots, x_k$ is given by

$$\Delta_p^2 - \Delta_k^2 = (\underline{\delta}_2 - \beta\underline{\delta}_1)' \Sigma_{22.1}^{-1} (\underline{\delta}_2 - \beta\underline{\delta}_1) \qquad 1.6.1.1.8$$

with unbiased estimate $D_p^2 - D_k^2$ from 1.6.1.1.3 as

$$\frac{1}{n_1+n_2-2}\left[\{(n_1+n_2-2)-(p-1)\}D_p^2 - \{(n_1+n_2-2)-(k-1)\}D_k^2\right] - \frac{n_1+n_2}{n_1 n_2}(p-k)$$

$$1.6.1.1.9$$

1.6.2 Several hypotheses and an appropriate test measure

$$H_1 : a_{k+1} = \ldots = a_p = 0, \qquad 1.6.2.1$$

i.e. $\underline{a}_2 = \underline{0}$ where $\underline{a} = \Sigma^{-1}\underline{\delta}$ is the coefficient vector of the discriminant function $\underline{a}'\underline{x}$ and $\underline{a}$ is partitioned as in 1.6.1.1.4.

From 1.6.1.1.5 it follows that

$$\underline{a}_2 = -\Sigma_{22.1}^{-1}\beta\underline{\delta}_1 + \Sigma_{22.1}^{-1}\underline{\delta}_2$$

$$= \Sigma_{22.1}^{-1}(\underline{\delta}_2 - \beta\underline{\delta}_1) \qquad 1.6.2.2$$

so that H_1 is equivalent to

$$H_2 : \underline{\delta}_2 = \beta\underline{\delta}_1 \qquad 1.6.2.3$$

when $\underline{a}_2 = \underline{0}$, or equivalently $H_2 : E(\underline{x}_2/\underline{x}_1)$ is the same in both Π_1 and Π_2.

Further, from 1.6.1.1.6 H_1 and H_2 are both equivalent to

$$H_3 : \Delta_p^2 = \Delta_k^2 \qquad 1.6.2.4$$

1.24/...

i.e. $\Delta_p^2 - \Delta_k^2 = 0$ with the test measure

$$F = \frac{(n_1+n_2-2)-(p-1)}{p-k} \cdot \frac{\frac{n_1 n_2}{n_1+n_2}(D_p^2-D_k^2)}{(n_1+n_2-2)+\frac{n_1 n_2}{n_1+n_2} D_k^2} \qquad 1.6.2.5$$

where F has the F distribution with p-k and $(n_1+n_2-2)-(p-1)$ degrees of freedom if H_1, H_2 or H_3 is valid.

H_1, H_2 and H_3 all accepted means that the variables $x_{k+1}, \ldots, x_p$ do not have any discriminating ability once $x_1, \ldots, x_k$ have already been considered.

Consider also the following hypotheses:

H_4 : $E(\underline{x}_2)$ is the same in Π_1 and Π_2, given that $E(\underline{x}_1)$ is the same in Π_1 and Π_2, 1.6.2.6

i.e.

H_4 : $\underline{\delta}_2 = \underline{0}$, given $\underline{\delta}_1 = 0$, 1.6.2.7

or

H_5 : $\Delta_p^2 = 0$, given $\Delta_k^2 = 0$, 1.6.2.8

or

H_6 : $\Delta_{p-k}^2 = 0$, given $\Delta_k^2 = 0$ 1.6.2.9

where Δ_{p-k}^2 is based on $\underline{x}_2$, viz. $\Delta_{p-k}^2 = \underline{\delta}_2' \Sigma_{22}^{-1} \underline{\delta}_2$.

This can be summarised: If H_4, H_5 or H_6 is true, then from 1.6.1.1.6 we have that $\Delta_p^2 = \Delta_k^2$ and 1.6.2.5 is applicable again.

1.25/...

Note that if $\underline{x}_1$ has the same mean in both populations, $\underline{x}_1$ is known as the vector of concomittant variables or ancillary variables. They have no discriminating ability by themselves, but in the presence of other variables having discriminating ability, these ancillary variables, on account of their correlations with the main variables, may provide additional discrimination. In practice, however, when the correlations are unknown and have to be estimated from data, there may be loss of information, unless the correlations are high. (Rao, 1966).

A very special case worth pointing out is when k=p-1, in which case we wish to determine if a single specified variable has discriminating power. The F statistic will have the F distribution with 1 and n_1+n_2-p-1 degrees of freedom or equivalently the t^2 distribution with n_1+n_2-p-1 degrees of freedom. Significant values for 1.6.2.5 would lead to the conclusion that the measurement x_p is needed for discriminatory power. For further information see Berenson, Levine and Goldstein (1983).

1.7.1 Discrimination in the case of more than two populations

Let Π_i, $i = 1, \ldots, m$ be m populations with density functions $p_i(\underline{x})$ respectively. We want to partition our space of observations into m mutually exclusive and exhaustive regions R_i, $i = 1, \ldots, m$. Define the cost of misclassifying an observation from Π_i as coming from Π_j by $c(j/i)$. The probability of this misclassification is

$$P(j/i,R) = \int_{R_j} p_i(\underline{x})d\underline{x} \qquad 1.7.1.1$$

Then the conditional expected cost of misclassifying an $\underline{x}$ from Π_i into any one of Π_j, $j = 1, \ldots m$ but $j \neq i$, is

1.26/...

$$\sum_{\substack{j=1 \\ j\neq i}}^{m} c(j/i)\ P(j/i,R) \qquad 1.7.1.2$$

Let the a priori probabilities for Π_i be as before, i.e. q_i, i=1, ..., m. Then the total expected loss is

$$\sum_{i=1}^{m} q_i \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{m} c(j/i)\ P(j/i,R) \right\} \qquad 1.7.1.3$$

Our aim is to partition R into R_1, ..., R_m in order to make 1.7.1.3 a minimum.

The conditional probability of an observation $\underline{x}$ as coming from Π_i is

$$\frac{q_i p_i(\underline{x})}{\sum_{i=1}^{m} q_k p_k(\underline{x})} \qquad 1.7.1.4$$

So that if we classify the observation as from Π_j then the expected loss is

$$\sum_{\substack{j=1 \\ j\neq i}}^{m} \left[\frac{q_i p_i(\underline{x})}{\sum_{k=1}^{m} q_k p_k(\underline{x})} \cdot c(j/i) \right] \qquad 1.7.1.5$$

We minimise this expected loss if we minimise the numerator, i.e.

$$\sum_{\substack{i=1 \\ i\neq j}}^{m} q_i p_i(\underline{x})\ c(j/i), \qquad 1.7.1.6$$

1.27/...

i.e. we must calculate 1.7.1.6 for all j and select that j which gives a minimum - if two or more different indices give a minimum the choice is arbitrary. Assign $\underline{x}$ to R_j according to the "minimum" choice.

Follow this procedure for each $\underline{x}$ so that an observation is classified as from Π_j if it falls in R_j.

This argument can be summarised as follows:

If q_i is the a priori probability of drawing an observation from population Π_i with density $p_i(\underline{x})$, i=1, ..., m and if the cost of misclassifying an observation from Π_i as from Π_j is $c(j/i)$, then the regions of classification, R_1, ..., R_m that minimise the total expected cost of misclassification are defined by assigning $\underline{x}$ to R_k if

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i p_i(\underline{x})\, c(k/i) < \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i p_i(\underline{x})\, c(j/i), \quad j=1, \ldots, m; \quad j\neq k \qquad 1.7.1.7$$

If 1.7.1.7 holds for all j($j\neq k$) except for h indices and the inequality is replaced by equality for those indices, then this observation can be assigned to any of the (h+1) populations. If the probability of equality between right hand and left hand sides of 1.7.1.7 is zero for each k and j under Π_j for each i, then the minimising procedure is unique - except for sets of probability zero. Refer Anderson (1958) for further detail.

Let $c(j/i)=1$ for all i and j ($i\neq j$), then in R_k

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i p_i(\underline{x}) < \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i p_i(\underline{x}), \quad \forall\ j\neq k \qquad 1.7.1.8$$

1.28/...

This however, means that the term excluded from the left-hand side is bigger than the term excluded from the right-hand side, i.e.

$$q_j p_j(\underline{x}) < q_k p_k(\underline{x}), \qquad \forall j \neq k \qquad 1.7.1.9$$

or

$$\ln q_j p_j(\underline{x}) < \ln q_k p_k(\underline{x}), \qquad \forall j \neq k \qquad 1.7.1.10$$

In this form the observation $\underline{x}$ is in R_k if k is the index for which $q_i p_i(\underline{x})$ is a maximum; i.e. Π_k is the most probable population.

<u>If a priori probabilities are not available we define an expected loss on the condition that the observation comes from a given population</u>. The conditional expected loss if the population is from Π_i is

$$\sum_{\substack{j=1 \\ j \neq i}}^{m} c(j/i)\, P(j/i, R) \qquad 1.7.1.11$$

Let us apply this to a number of multivariate normal populations as described inter alia in Johnson and Wichern (1982).

Let

$$p_i(\underline{x}) = (2\pi)^{-\frac{1}{2}p} |\Sigma_i|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\underline{x}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)\right], \quad i=1, \ldots, m \qquad 1.7.1.12$$

with $c(i/i) = 0$, $c(k/i) = 1$, $k \neq i$, i.e. the misclassification costs are equal and a priori probability q_i.

Allocate $\underline{x}$ to Π_k when

$$\ln q_k p_k(\underline{x}) = \ln q_k - \tfrac{1}{2}p\ln(2\pi) - \tfrac{1}{2}\ln|\Sigma_k| - \tfrac{1}{2}(\underline{x}-\underline{\mu}_k)'\Sigma_k^{-1}(\underline{x}-\underline{\mu})$$

$$= \max_i \left[\ln q_i p_i(\underline{x})\right] \qquad 1.7.1.13$$

One can look at this classification rule from another angle by comparing the ${}_iD_p^2(\underline{x})$ two at a time. This leads to the following rule:

Allocate to Π_k if for all i=1, ..., m, k≠i

$${}_kD_p^2(\underline{x}) \geq {}_iD_p^2(\underline{x})$$

$$0 \leq {}_kD_p^2(\underline{x}) - {}_iD_p^2(\underline{x})$$

$$= -(\underline{\mu}_k-\underline{\mu}_i)'\Sigma^{-1}\underline{x} - \tfrac{1}{2}(\underline{\mu}_k-\underline{\mu}_i)'\Sigma^{-1}(\underline{\mu}_k-\underline{\mu}_i) + \ln(\frac{q_k}{q_i})$$

1.7.1.19

i.e.

$$(\underline{\mu}_k-\underline{\mu}_i)'\Sigma^{-1}\{\underline{x}-(\underline{\mu}_k+\underline{\mu}_i)\} \geq \ln(\frac{q_i}{q_k}) \quad \forall i=1, \ldots, m; \quad i\neq k$$

1.7.1.20

Write U_{ki} for the left-hand side of 1.7.1.20 so that U_{ki} defines region R_k.

If the costs of misclassification are equal, then $\ln(\frac{q_i}{q_k})$ will fall away. This then is the most common form of the discriminant function, because the a priori probabilities are often difficult to determine.

Note also that

$$U_{ki} = -U_{ik} \qquad 1.7.1.21$$

In the usual case where the covariance matrix is unknown - if a common one is assumed - and $\underline{\mu}_i$, i=1, ..., m are unknown, the sample statistics are used where $\hat{\underline{\mu}}_k = \bar{\underline{x}}_k$ and $\hat{\underline{\mu}}_i = \bar{\underline{x}}_i$ and

and $S = \dfrac{\sum_{i=1}^{m}(n_i-1)S_i}{\sum_{i=1}^{m} n_i - m}$. Then one can determine

1.31/...

$$u_{ki}(\underline{x}) = (\bar{\underline{x}}_k - \bar{\underline{x}}_i)' S^{-1} (\underline{x} - \tfrac{1}{2}(\bar{\underline{x}}_k + \bar{\underline{x}}_i)), \quad i=1, \ldots, m; \quad i \neq k \qquad 1.7.1.22$$

and allocate to Π_k if

$$u_{ki}(\underline{x}) \geq \ln\left(\frac{q_i}{q_k}\right), \qquad \forall_{i \neq k} \qquad 1.7.1.23$$

Here we have formed hyperplanes because $u_{ik}(\underline{x})$ is a linear combination of the components of $\underline{x}$ so that we find in fact the regions which are separated by these hyperplanes.

As in the two class case we have that the estimation of the population parameters may mean that the sample classification rules are not any longer necessarily optimal. Their performance can however still be estimated using Lachenbruch's hold-out procedure as described earlier.

If $n_{iM}^{(H)}$ is the number of misclassified hold-out observations in the i th group, i=1, ..., m, then an estimate of the expected actual error rate, E(AER), is provided by

$$\hat{E}(AER) = \frac{\sum_{i=1}^{m} n_{iM}^{(H)}}{\sum_{i=1}^{m} n_i} \qquad 1.7.1.24$$

where H refers to "hold-out" and M to "misclassified". For further detail refer to section 1.4.2.

We must emphasise, however, that rather strong assumptions of multivariate normality and equal covariances are involved. Before implementing a linear classification rule, these tentative assumptions should be checked in the order: Multivariate normality, then equality of covariances. If one or both of these assumptions are violated, improved classification is probably possible if data is first suitably transformed.

The quadratic rules are an alternative to classification with linear discriminant functions. The former are appropriate if normality appears to hold but the assumption of equal covariance matrices is seriously violated. The assumption of normality however, seems to be more critical for quadratic rules than linear rules. If doubt exists as to the appropriateness of a linear or quadratic rule, both rules can be constructed and their error rates examined using Lachenbruch's hold-out procedure.

1.7.2 The use of oneway MANOVA, regression analysis and canonical correlations to diminish dimensions in the case of a large number of variables as well as a large number of discriminant functions

Johnson and Wichern (1982) stated that the use of MANOVA, regression analysis and canonical correlations to diminish dimensions in discriminant analysis is the result of the need to obtain a reasonable representation of the populations that involves only a few linear combinations of the observations, such as $\underline{\ell}_1'\underline{x}$, $\underline{\ell}_2'\underline{x}$ etc.

Although the primary purpose of discriminant analysis is to separate populations, it can also be used to classify new observations. For the basic theory it is not necessary to assume that the k populations are multivariate normal, but in order to be able to apply the hypothesis tests for a possible decrease in dimensions we shall assume normality It is, however essential that we have pxp var.-covariance matrices of full rank, i.e. $\Sigma_1 = \ldots = \Sigma_k = \Sigma$.

We shall first of all define and derive the one way MANOVA notation and technique. The reason for this is that it supplies us with a relative easy check on whether the means are different so that a discriminant analysis is applicable. If so, we use this technique in combination with regression analysis and canonical correlation to discriminant between observations as well as to diminish the dimensions if possible.

1.7.3 Oneway MANOVA

Consider the k p-variate normal populations Π_α, $\alpha=1, \ldots, k$ with equal variance-covariance matrices (see Kshirsagar, 1972). Let $q = k-1$ and let

$$H_o : \underline{\mu}_1 = \ldots = \underline{\mu}_k \qquad 1.7.3.1$$

where the k independent samples are random of size n_α from population α, $\alpha = 1, \ldots, k$. The r th observation of the i th variable of the α th population is indicated by $x_{ir\alpha}$ $(i=1, \ldots, p;\ r=1, \ldots, n_\alpha;\ \alpha=1, \ldots, k)$.

X_α : pxn_α is the matrix of the n_α sample observations from Π_α and

$$X_{pxN} = \underset{\substack{\uparrow \\ n_1}}{X_1} \vdots \underset{\substack{\uparrow \\ n_2}}{X_2} \vdots \cdots\cdots \vdots \underset{\substack{\uparrow \\ n_k}}{X_k} \quad p \qquad 1.7.3.2$$

is the matrix of all $N = n_1 + \ldots + n_k$ observations. Let $n = N-1$ and as before

$$\underline{\bar{x}}_\alpha = \frac{1}{n_\alpha} X_\alpha E_{n_\alpha}, \qquad \alpha=1, \ldots, k \qquad 1.7.3.3$$

and

$$A_\alpha = X_\alpha \left(I - \frac{1}{n_\alpha} E_{n_\alpha n_\alpha}\right) X_\alpha' \qquad \alpha=1, \ldots, k \qquad 1.7.3.4$$

Further we have that

$$\underline{\bar{x}}_\alpha \sim n\left(p, \underline{\mu}_\alpha, \frac{1}{\sqrt{n_\alpha}} \Sigma\right) \qquad 1.7.3.5$$

is indepentently distributed from A_α where

$$A_\alpha \sim W(p, n_\alpha-1, \Sigma) \qquad \alpha=1, \ldots,k \qquad 1.7.3.6$$

All these distributions are independent, because the k samples are all independent. Then

1.34/...

$$A = \sum_{i=1}^{k} A_\alpha \sim W(p, n-q, \Sigma) \qquad 1.7.3.7$$

keeping in mind that $n-q = (N-1)-(k-1) = N-k = \sum_{\alpha=1}^{k} (n_\alpha - 1)$

is the total degrees of freedom.

Note that A is independently distributed from $\underline{\bar{x}}_\alpha$, $\alpha=1, \ldots, k$ and holds irrespective of H_o being true or not.

Define a new matrix of observations

$$Z = \left[\underline{z}_1 \vdots \ldots \vdots \underline{z}_q \vdots \underline{z}_k\right] = UG \text{ or } U = ZG \qquad 1.7.3.8$$

as G_{kxk} is orthogonal where

$$U = \left[\sqrt{n_1}\underline{\bar{x}}_1 \vdots \sqrt{n_2}\underline{\bar{x}}_2 \vdots \ldots \vdots \sqrt{n_k}\underline{\bar{x}}_k\right] \qquad 1.7.3.9$$

and

$$G' = \left[\underline{g}_1 \vdots \ldots \vdots \underline{g}_q \vdots \underline{h}\right]$$

$$= \underset{q}{\left[G_o' \vdots \underline{h}\right]}\ \underset{1}{k} \qquad 1.7.3.10$$

is any orthogonal matrix with

$$\underline{h}' = \left[\left(\frac{n_1}{N}\right)^{\frac{1}{2}}, \ldots, \left(\frac{n_k}{N}\right)^{\frac{1}{2}}\right] \qquad 1.7.3.11$$

When one takes the distribution in 1.7.3.3 and the form of U in 1.7.3.9, then it follows from the independence of $\underline{\bar{x}}_\alpha$, $\alpha=1, \ldots, k$ that U is distributed as follows, viz.

$$(2\pi)^{-\frac{1}{2}pk} |\Sigma|^{-\frac{1}{2}k} \exp\left[-\tfrac{1}{2}\mathrm{tr}\Sigma^{-1}(U-E(U))(U-E(U))'\right]dU \qquad 1.7.3.12$$

1.35/...

whereas the transfromation 1.7.3.8 has the Jacobian $|G|^{k}=1$ it follows that the transformed form of U as in 1.7.3.8 leads to

$$(2\pi)^{-\frac{1}{2}pk}|\Sigma|^{-\frac{1}{2}k}\exp\left\{-\tfrac{1}{2}\mathrm{tr}\Sigma^{-1}(Z-E(Z))(Z-E(Z))'\right\}dZ \quad {}^{1)} \qquad 1.7.3.13$$

<u>When H_o is true</u>, i.e. $\underline{\mu}_\alpha=\underline{\mu}$, $\alpha=1, \ldots, k$, <u>then</u>

$$\begin{aligned} E(Z) &= E(U)G' \\ &= \left[\sqrt{n_1}\,\underline{\mu} \vdots \sqrt{n_2}\,\underline{\mu} \vdots \ldots \vdots \sqrt{n_k}\,\underline{\mu}\right] G' \\ &= \underline{\mu}\cdot\left[\sqrt{n_1}, \sqrt{n_2}, \ldots, \sqrt{n_k}\right]G' \\ &= \underline{\mu}\cdot\sqrt{N}\,\underline{h}'G' \\ &= \underline{\mu}\cdot\sqrt{N}\left[0, 0, 0, \ldots, 0, 1\right] \\ &= \left[\underline{0} \vdots \underline{0} \vdots \ldots \vdots \underline{0} \vdots \sqrt{N}\,\underline{\mu}\right] \end{aligned} \qquad 1.7.3.14$$

as $\underline{h}'\underline{h}=1$ and $\underline{h}'\underline{q}_i=0$.

Note that this result is dependent on the condition H_o.

From this follows that $\underline{z}_\alpha$, $\alpha=1,\ldots, q(\alpha\neq k)$ are independently distributed as $n(p, \underline{0}, \Sigma)$ under H_o. Obviously $\underline{z}_k$ is also normally distributed but with non-zero mean.

At this stage we have A which has the $W(p, n-q, \Sigma)$ distribution as well as

$$B = \sum_{\alpha=1}^{q} \underline{z}_\alpha \underline{z}_\alpha' \qquad 1.7.3.15$$

1.36/...

1) $(ZG-E(ZG))(Z-E(ZG))'=(Z-E(Z))GG'(Z-E(Z))'=(Z-E(Z))(Z-E(Z))'$

Keeping H_0 and the distributions of $\underline{z}_\alpha$, $\alpha=1, \ldots, k$ in mind, as well as the fact that the $\underline{z}_\alpha$'s are functions of $\underline{x}_\alpha$, $\alpha=1, \ldots, k$ which are all independent of A, we know that A and B are independent with the distribution of B being

$$W(p, q, \Sigma) \qquad 1.7.3.16$$

which is a real Wishart distribution if $q \geq p$, but the pseudo-Wishart distribution if $q<p$.

From Appendix C it follows that

$$\Lambda = \Lambda(n, p, q) = \frac{|A|}{|A+B|} \qquad 1.7.3.17$$

has the Wilks $\Lambda(n, p, q)$ distribution whenever H_0 is true and can therefore be used as a test measure for H_0, which will be rejected at the $100\alpha\%$ level of significance whenever

$$-\{n-\tfrac{1}{2}(p+q+1)\}\ln\Lambda > W_\alpha(n, p, q)$$

$$= C_\alpha(p, q, M)\chi^2_{pq}(x) \quad , M = n-p-q+1, \qquad 1.7.3.18$$

as defined by Bartlett - the left-hand side of 1.7.3.18 being approximately distributed as a χ^2 distribution with pq degrees of freedom and a correction factor on the right-hand side obtained by Shatzoff(1966). The latter should be referred to whenever $W_\alpha(n, p, q) > \chi^2_{pq}(\alpha)$. For more detail refer to Kshirsagar.

Note that B is the "between groups" matrix of sums of squares and products while A is known as the "within groups" matrix of sums of squares and products, i.e.

$$B = \sum_{\alpha=1}^{q} \underline{z}_\alpha \underline{z}'_\alpha$$

$$= ZZ' - \underline{z}_k \underline{z}'_k$$

$$= UG'GU' - U\underline{hh}'U$$

1.37/...

$$= UU' - \left(\sum_{\alpha=1}^{k} \frac{n_\alpha}{\sqrt{N}} \bar{\underline{x}}_\alpha\right)\left(\sum_{\alpha=1}^{k} \frac{n_\alpha}{\sqrt{N}} \bar{\underline{x}}_\alpha\right)$$

$$= \sum_{\alpha=1}^{k} n_\alpha \bar{\underline{x}}_\alpha \bar{\underline{x}}_\alpha' - N\bar{\underline{\underline{x}}}\bar{\underline{\underline{x}}}'$$

$$= \sum_{\alpha=1}^{k} n_\alpha (\bar{\underline{x}}_\alpha - \bar{\underline{x}})(\bar{\underline{x}}_\alpha - \bar{\underline{x}})' \qquad 1.7.3.19$$

with

$$\bar{\underline{x}} = \sum_{\alpha=1}^{k} \frac{n_\alpha \bar{\underline{x}}_\alpha}{N} \qquad 1.7.3.20$$

From this we can now construct the multivariate analysis of variance table:

Source of variation	d.f.	Matrix of S.S. and S.P.
Between groups	q	B
Within groups	n-q	A
Total	n	A+B

TABLE 4

If H_o is not true we can carry on with the discriminant analysis - otherwise we may be wasting our time and effort.

Before we go further, it is good to have a look at the implications of the decision that H_o is not true, i.e. $E(\underline{z}_\alpha) \neq 0$, $\alpha = 1, \ldots, k$. Then

$$\sum_{\alpha=1}^{q} E(\underline{z}_\alpha)E(\underline{z}_\alpha') = \sum_{\alpha=1}^{k} n_\alpha (\underline{\mu}_\alpha - \bar{\underline{\mu}})(\underline{\mu}_\alpha - \bar{\underline{\mu}})' = \Delta, \qquad 1.7.3.21$$

1.38/...

but var $(\underset{\sim}{z}_\alpha) = \Sigma$ and from the distribution of $\underset{\sim}{z}_\alpha$, $\alpha=1, \ldots, k$ it follows that

$$\Omega = \Sigma^{-1}\Delta \qquad 1.7.3.22$$

is known as the noncentrality matrix which becomes the null matrix under H_o. Ω can thus be seen as a measure of departure from H_o.

From the distributions of A and B follow that

$$E(A) = (n-q)\Sigma \qquad 1.7.3.23$$

and

$$E(B) = q\Sigma + \Sigma\Omega \qquad 1.7.3.24$$

$$= q\Sigma + \Delta \qquad 1.7.3.25$$

(James, 1955). For more detail see Kshirsagar. See lemma in Appendix B.

Thus $\frac{A}{n-q}$ and $\frac{B}{q}$ provide indepent estimates of Σ when H_o is true. Wilks's Λ compares these two estimates and provides a test of H_o which, if rejected initiates the discriminant analysis of the data.

Use 1.7.3.24 and then an unbiased estimate of Ω according to the above-mentioned lemma is given by

$$\hat{\Omega} = (n-p-q-1)A^{-1}B - qI_p \qquad 1.7.3.26$$

We will see more about this in section 1.7.5.

1.7.4 The MANOVA problem as a regression problem

The one way MANOVA problem can be expressed as a regression

1.39/...

problem. Again we gain nothing in itself doing a regression analysis, but if H_o of 1.7.3 is rejected, then the basic theory gives us more insight into our group centroids.

Use the q dummy variables $y_1, \ldots, y_q$ where

$$y_\alpha = 1, \quad \underline{x} \text{ is from } \Pi_\alpha$$
$$= o \text{ otherwise}, \alpha=1, \ldots, q \qquad 1.7.4.1$$

From this we then have that

$$E(\underline{x}) = \underline{\mu}_k+(\underline{\mu}_1-\underline{\mu}_k)y_1 + \ldots + (\underline{\mu}_q-\underline{\mu}_k)y_q \qquad 1.7.4.2$$

so that

$$E(\underline{x}) = \underline{\mu}_k , \quad \alpha=1, \ldots, q$$
$$= \underline{\mu}_k \qquad 1.7.4.3$$

when $\underline{x}$ comes from Π_k since all the y's are zero in the latter case. Then we have further that

$$E(\underline{x}/\underline{y}) = \underline{\mu}_k+\beta\underline{y} \qquad 1.7.4.4$$

with

$$\beta = \left[\underline{\mu}_1-\underline{\mu}_k \vdots \cdots \vdots \underline{\mu}_q-\underline{\mu}_k\right] \qquad 1.7.4.5$$

Then $H_o : \underline{\mu}_1 = \ldots = \underline{\mu}_k$ is similar to $H_o : \beta=0$ so that the hypothesis test applicable in regression analysis may be applied in this case to see whether it is worthwhile carrying on with a discriminant analysis.

Corresponding to the matrix X_α of the n_α observations on $\underline{x}$ from Π_α, we have now the matrix $Y_\alpha : qxn_\alpha$ of "observations" on $\underline{y}$. The matrix of all N observations on $\underline{y}$ is then

1.40/...

$$Y = \left[Y_1 \vdots \ldots \vdots Y_q \right]_{qxN} \qquad 1.7.4.6$$

where $Y_\alpha = \begin{bmatrix} 0 \\ \ldots\ldots\ldots \\ 11 \ldots 11 \\ \ldots\ldots\ldots \\ 0 \end{bmatrix}$ with the 1's in the α th row.

Define now the matrices, with $J = I - N^{-1} E_{NN}$;

$$C_{xx} = XJX', \; C_{xy} = XJY', \; C_{yy} = YJY' \qquad 1.7.4.7$$

so that

$$C_{xx} = A+B, \; C_{xy} = \left[n_1(\bar{\underline{x}}_1 - \bar{\underline{x}}) \vdots \ldots \vdots n_q(\bar{\underline{x}}_q - \bar{\underline{x}}) \right] \qquad 1.7.4.8$$

$$C_{yy} = \text{diag}\,(n_1, \ldots, n_q) - \frac{1}{N}\left[n_1, \ldots, n_q \right]' \left[n_1, \ldots, n_q \right]$$

therefore

$$C_{yy}^{-1} = \text{diag}\,\left(\frac{1}{n_1}, \ldots, \frac{1}{n_q}\right) + \frac{1}{n_k} E_{qq}$$

$$C_{xy}\, C_{yy}^{-1}\, C_{yx} = B, \; C_{xx.y} = A$$

$$= C_{xx} - C_{xy}\, C_{yy}^{-1}\, C_{yx} \qquad 1.7.4.9$$

which are found in the regression analysis of $\underline{x}$ on $\underline{y}$ and is exactly the same as table 4 in section 1.7.3.

Source	d.f.	S.S. and S.P. matrix
Regression of $\underline{x}$ on $\underline{y}$	q	$B = C_{xy}\, C_{yy}^{-1}\, C_{yx} = \hat{\beta}\, C_{yy}\, \hat{\beta}'$
Error	n-q	$A = C_{xx.y} = C_{xx} - C_{xy}\, C_{yy}^{-1}\, C_{yx}$
Total	n	$A + B = C_{xx}$

TABLE 5

$$s = \text{rank}\ (\Omega) = \text{rank}\ (\Delta) = (\beta) \qquad 1.7.5.3$$

But β is a pxq matrix so that $s<p,q$ whichever is the smaller.

If one can determine the rank of one of the above-mentioned matrices, part of our problem is solved. In practise, however, these matrices are unknown and s must be estimated from samples from which $\hat{\Omega}$, $\hat{\Delta}$ or $\hat{\beta}$ may be determined.

We know, however, from 1.7.3.26 that $\hat{\Omega} = (n-q-p-1)A^{-1}B-qI_p$. Even if the number of non-zero roots is s, more than s roots will (may) be non-zero because it is an estimation only. Assuming n to be large enough one may accept that the first s roots will be significant, so that $\hat{s}$ equals the number of significant roots of $(n-q-p-1)A^{-1}B-qI_p$.

If λ_i are the roots, they will satisfy

$$|(n-q-p-1)A^{-1}B-qI_p-\lambda I_p| = 0 \qquad 1.7.5.4$$

or when $r^2 = \frac{\lambda+q}{n-p-1+\lambda}$, then

$$|(n-q-p-1)A^{-1}B - qI_p - \lambda I_p| = 0$$

$$|(n-q-p-1)\ B - qA - \lambda A| = 0$$

$$|(n-p-1+\lambda)\ B-(q+\lambda)\ B - (q+\lambda)\ A| = 0$$

$$\left| - \frac{q+\lambda}{n-p-1+\lambda}\ (B+A) + B \right| = 0$$

$$|\ - r^2\ (B+A) + B| = 0 \qquad 1.7.5.5$$

or

$$|\ - r^2\ C_{xx} + C_{xy}\ C_{yy}^{-1}\ C_{yx}| = 0 \qquad 1.7.5.6$$

1.43/...

so that the roots r_i^2 , i=1, ..., f; f = min (p, q) are the squares of the sample canonial correlations between the random variables $\underline{x}$ and the dummy variables $\underline{y}$ defined in 1.7.4.1.

The tests of significance for determining the number of significant roots in the case of canonical correlations may be used as it is briefly described:

For each value of s prepare the table 6 , starting with s=1 and use the χ^2 test as shown sequentially until a non-significant result is achieved.

Source	d.f.	χ^2
lst s roots	pq-(p-s)(q-s)	$- m \ln \prod_{i=1}^{s} (1-r_i^2)$
Remaining roots	(p-s)(q-s)	$- m \ln \prod_{i=s+1}^{k} (1-r_i^2)$
Total	pq	$- m \ln \Lambda(n,p,g)$

TABLE 6

The dimensionality of the group means is then inferred from the number of significant roots of 1.7.5.5 using the criterian

$$\chi^2 = -(n - \tfrac{1}{2}(p+q+1)) \ln \prod_{i=s+1}^{f} (1-r_i^2) \qquad 1.7.5.7$$

where $\prod_{i=s+1}^{f} (1-r_i^2) = \Lambda_i(n,p,q)$ for testing the hypothesis that the dimensionality of the space spanned by $\underline{\mu}_\alpha$, α=1, ..., k is s. See also section 1.8 for a detailed description of Wilks's Stepwise procedure.

1.44/...

We can look at this measure as the total "distance" measured by $\sum_{i=1}^{s} \delta_i^2 = tr(\Sigma^{-1}\Delta) = tr(\Omega)$ or alternative measures in order to use the χ^2- criterion. Note that if s=p we have Pillai's V or Hotelling's generalised T^2.

1.7.6 How to determine the number of significant discriminant functions

In the case of two groups Fisher's discriminant function was obtained by maximising the ratio of the "between groups S.S." to the "within groups S.S." in the analysis of variance of an arbitrary linear function $\underline{\ell}'\underline{x}$. Here we are going to do something similar.

Assume k p-variate normal populations Π_α, $\alpha=1, \ldots, k$ with means $\underline{\mu}_\alpha$ and equal variance-covariance matrix Σ. From table 6 in section 1.7.3 it follows that for $\underline{x}$ and a linear combination $\underline{\ell}'\underline{x}$, we want to maximise $\underline{\ell}'B\underline{\ell}/\underline{\ell}'A\underline{\ell}$. We must therefore find the $\underline{\ell}$ which will maximise $\underline{\ell}'B\underline{\ell}/\underline{\ell}'A\underline{\ell}$ or for that matter

$$\underline{\ell}'B\underline{\ell}/\underline{\ell}'(A+B)\underline{\ell} \qquad 1.7.6.1$$

and as in the previous section by differentiation this "optimum" $\underline{\ell}$ satisfies the equation

$$\left[- r^2(A+B) + B \right]\underline{\ell} = \underline{0} \qquad 1.7.6.2$$

where r^2 is the canonical correlation, i.e.

$$|- r^2(A+B) + B| = 0$$

From 1.7.5.5 and further we know 1.7.6.3 has $f = \min(p,q)$ roots $r_1^2 > r_2^2 > \ldots r_f^2$.

1.45/...

Corresponding to each root r_i^2 there will be a vector $\underline{\ell}_i$ satisfying 1.7.6.2, so that we have f discriminant functions $\underline{\ell}_i'\underline{x}$, i=1, ..., f which are also the sample canonical variables of the X-space. Since r_1^2 is the largest root of 1.7.6.3 the corresponding function $\underline{\ell}_1'\underline{x}$ provides the maximum separation of the group means and is useful as a discriminant function. Similarly $\underline{\ell}_2'\underline{x}$ provides separation in a different direction, and so on. The question is now whether all f discriminant functions are really important.

We have to measure the discriminating ability of these discriminant functions.

In section 1.5 it was shown that the performance of the discriminant function in the case of two groups could be measured by Mahalanobis's D^2 or using the distribution of R^2, the square of the multiple correlation coefficient between $\underline{x}$ and ξ, the single dummy variable.

Now R^2 is replaced by r_i^2, i=1, ..., f, where the single ξ is replaced by a dummy variable $\underline{y}$. Therefore it is natural to expect r_1^2 to measure the discriminating ability of $\underline{\ell}_1'\underline{x}$, r_2^2 of $\underline{\ell}_2'\underline{x}$ etc. If only r_1^2, r_2^2, ..., r_s^2 are significantly large and the rest are insignificant we employ only $\underline{\ell}_i'\underline{x}$, i=1, ..., s, for discrimination. This procedure has been described in full in section 1.7.4 and 1.7.5.

The result is that the adequate number of discriminant functions is the same as the dimensionality of the space spanned by the k group means.

1.7.7 Discrimination in the case of a large number of populations and a large number of variables - a summary

We can use the preceding results in the following way.

Use the s discriminant functions $\underline{\ell}_i'\underline{x}$ as determined to classify a new observation $\underline{x}_o$ in one of the k groups.

Form

$$u_{o1} = \underline{\ell}_1'\underline{x}_o; \quad u_{o2} = \underline{\ell}_2'\underline{x}_o, \ \ldots, \ u_{os} = \underline{\ell}_s'\underline{x}_o \qquad 1.7.7.1$$

where u_{oi}, i=1, ..., s are known as the new co-ordinates of the point $\underline{x}_o$. Immediately the dimensionality of the problem is reduced.

In a similar way the co-ordinates of the estimated mean $\bar{\underline{x}}_\alpha$ of Π_α will be

$$u_{\alpha 1} = \underline{\ell}_1'\bar{\underline{x}}_\alpha, \quad u_{\alpha 2} = \underline{\ell}_2'\bar{\underline{x}}_\alpha, \ \ldots, \ u_{\alpha s} = \underline{\ell}_s'\bar{\underline{x}}_\alpha \qquad 1.7.7.2$$

Therefore, with respect to Π_α the distance squared between $\underline{x}_o$ and the mean $\bar{\underline{x}}_\alpha$ of Π_α based on the new co-ordinate system is

$$\Delta^2_{\alpha o} = \sum_{i=1}^{s} (u_{\alpha i} - u_{oi})^2 \qquad 1.7.7.3$$

and this will be so for all $\alpha = 1, \ldots, k$.

Finally the observation $\underline{x}_o$ should be assigned to Π_a if the point $\underline{x}_o$ is nearer to the mean of Π_a than the mean of any other Π_α, i.e. assign $\underline{x}_o$ to Π_a when

$$\Delta^2_{ao} = \min_{i=1}^{k} (\Delta^2_{io}) \qquad 1.7.7.4$$

1.8 Wilks's Stepwise Procedure - Decreasing the number of variables considered for discrimination purposes

We have seen in sections 1.7.5 - 1.7.7 how the dimensions and so the number of discriminant functions as well as the

1.47/...

the number of transformed variables could be decreased. In this process all the original p-variables were still being used.

If one can determine the approximate dimensionality as described in those sections one might like to know which s of the original p-variables are significant in the analysis. Various methods have been proposed based on different approaches.

McKay and Campbell (1982, I) compared some of the techniques under the headings:

(a) Canonical variate analysis approach,

(b) Stepwise F-methods and

(c) All subsets procedures.

Their conclusion was that although (a) gives useful information the recommended method would be (c) with specific technique the use of simultaneous test procedures by means of MANCOVA likelihood ratio statistics. This is,however, a computationally cumbersome method when the number of variables are large. If this is so they recommended the use of a combination forward-backward selection procedure. There are, however, grave disadvantages, e.g. the significance levels are unknown in the case of individual tests.

Having established the approximate dimensionality of the centroids however, one may use this knowledge as an aid in the determination of a stopping rule.

Farmer and Freund (1975) compared 4 backward elimination techniques in which at each step the variable is deleted which

1.48/...

(a) has the largest R^2 value computed from the rows of the error matrix

(b) has the smallest absolute correlation with the "best" linear discriminant function

(c) is similar to (b), but the two "best" discriminant functions are considered

(d) makes the smallest change in Wilks's Λ.

They concluded that procedure (d) based on the decomposition of Wilks's Λ was by far the most effective and popular - and is incidentally the method used by the BMDP-7M package.

Rencher and Larson (1980) investigated Wilks's Λ using Monte Carlo methods with the emphasis on the downward bias present and gave the following brief description of the technique:

Let A(1,2, ...,p) and B(1,2, ...,p) be the within and between group matrices based on n observations from a oneway MANOVA with g groups and p variables x_1, ..., x_p. Then

$$\Lambda(1, \ldots, p) = \frac{|A(1, \ldots, p)|}{|A(1, \ldots, p) + B(1, \ldots, p)|} \qquad 1.8.1$$

is Wilks's Λ-statistic with parameters n, p, g-1. If one adds a variable x_{p+1}, to $\underline{x}$, one can form the multiplicative increment

$$\Lambda(p+1) = \frac{\Lambda(1, \ldots, p, p+1)}{\Lambda(1, \ldots, p)} \qquad 1.8.2$$

which is called a partial Wilks's Λ-statistic which corresponds to

1.49/...

$$F = \frac{n-g-p}{g-1} \cdot \frac{1-\Lambda(p+1)}{\Lambda(p+1)} \qquad 1.8.3$$

provided that x_{p+1} is an arbitrary variable rather than one which maximises F. In such a case F has the F-distribution. (See Appendix C on Wilks's Λ.)

In the stepwise procedure 1.8.3 is used as criterion to decide which variable must be entered and which one is to be removed. At each stage the variable with the largest F- to enter is added to the set of variables if its F-value is larger than a specified threshold, F_{in} (say). After this step all variables are re-examined and the one with the smallest F- to remove is deleted if the F-value is less than a threshold, F_{out} say.

Obviously the condition"x_{p+1} is an arbitrary variable" is not satisfied, with the result that 1.8.3 does not have the F-distribution, because in such a case Wilks's Λ is downward biased. Hawkins (1976) suggested that the level of significance, α will approximately be attained if the level

$\frac{\alpha}{(p-k)}$ is used; α being the required level of significance, p-k the number of variables available for inclusion and k the number of variables already included.

The bias in Λ may lead to problems as described in Rencher and Larson:

(a) Inclusion of too many variables in the subset and probably unstable subsets.

(b) Selection of an entirely spurious subset. If predetermined significant levels are used the correct classification rates may be good even if none of the available variables are good discriminators. This is especially so if the number of variables is large and the sample sizes are small.

1.50/...

They found large reductions in both the average and lower percentage points of Λ in the cases where the number of variables was more than the degrees of freedom of the error matrix, i.e. $p>g-1$. That is the reason why the one way MANOVA must first be carried out to determine the dimensionalities (see also Jennrich, 1977). It is thus important to note that if $p>g-1$ than the stepwise procedure is not recommended, unless the one way MANOVA has been carried out first.

When $p<g-1$ however, it was found that Wilks's Λ gave the "best" subset if compared with the other three stepwise procedures.

1.9 More on the Estimation of Error Rates in the case of several populations

In section 1.4.2 it was pointed out that Lachenbruch's hold-out U-technique can be used to estimate the error rate, especially so as the resubstitution or apparent estimated error rates and the "split into two sets" or split sample methods show a substantial downward bias and reduce the effective sample size respectively.

It is difficult to select a "best" estimating method as there is no "best" estimator for several reasons. One is that most estimators are sensitive to application - specific factors such as sample size and violation of distributional assumptions.

The U-method mentioned above is desirable as only a small bias is introduced, because the classification rule is determined using (n-1) rather than n observations. Only the observation to be classified is being held out and the remaining observations are used as the training sample

for the determination of the classification rule. Then this observation is classified and it is noted whether it is correctly classified or not. This procedure is repeated for all observations. Formules for calculation purposes are given in section 1.4.2.

Glick (1978), Moore, Whitsell and Lundgrebe (1976) developed an error rate called the posterior probability estimated error rate. Hora and Wilcox (1982) found that by amalgamating the U and posterior techniques they could obtain low variance, low biased estimators of the error rate.

If an observation is assigned to the population having the largest posterior probability, one has a set of optimal discriminant functions. Let p_j, $j = 1, \ldots, J$ be the prior probability of an observation belonging to the j th population and $f(\underline{x}/j)$ the probability density function of the observation $\underline{x}$ in population j. The unconditional density of $\underline{x}$ is

$$f(\underline{x}) = \sum_{j=1}^{J} p_j \, f(\underline{x}/j) \qquad 1.9.1$$

and the posterior probability of the j th population is

$$P(j/\underline{x}) = \frac{p_j \, f(\underline{x}/j)}{f(\underline{x})} \qquad 1.9.2$$

Determine the set of vectors $\underline{x}$ which are such that $P(1/\underline{x})$ is the largest of all the J posterior probabilities. Call the set R_1. Determine similar sets which are such that $P(j/\underline{x})$ are the largest among the J posterior probabilities and call these sets of vectors R_j, $j = 2, \ldots, J$.

1.52/...

The error rate for observations belonging to the j th population (j = 1, ..., J), is then

$$e_j = 1 - \int_{R_j} f(\underline{x}/j)\, d\underline{x} \qquad 1.9.3$$

where

$$f(\underline{x}/j) = \frac{P(j/\underline{x}).f(\underline{x})}{p_j} \qquad 1.9.4$$

from equation 1.9.2. Thus 1.9.3 becomes

$$e_j = 1 - \int_{R_j} \frac{P(j/\underline{x}).f(\underline{x})}{p_j}\, d\underline{x} \qquad 1.9.5$$

Now $\int_{R_j} P(j/\underline{x}) f(\underline{x}) d\underline{x}$ is the expected value, i.e. $\bar{y}_{.j}$

of the random variable

$$y_{ij} = P(j/\underline{x}_i) \quad \text{if } \underline{x}_i \epsilon R_j \qquad 1.9.6$$

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

for i = 1, ..., N = sample size, taken with respect to $f(\underline{x})$ in 1.9.1, i.e. $E_{P(j/\underline{x})}(f(\underline{x}))$.

The posterior probability estimator of e_j can thus be determined as

1.53/...

$$\hat{e}_j = 1 - \frac{\bar{y}_{.j}}{p_j}$$

$$= 1 - \frac{1}{p_j N} \sum_{i=1}^{N} y_{ij} \qquad 1.9.7$$

from which follows that $E(\hat{e}_j) = e_j$ so that $\hat{e}_j$ is an unbiased estimator of the error rate for the optimal discriminant function.

To estimate the "overall error rate" it is just natural to use the a priori population probabilities as weights, thus

$$\hat{e} = \sum_{j=1}^{J} p_j \hat{e}_j \qquad 1.9.8$$

If one uses the definition of R_j as in the paragraph after 1.9.2, then from 1.9.6 and 1.9.7 it follows that

$$\hat{e} = 1 - \frac{1}{N} \sum_{i=1}^{N} \max\{P(1/\underline{x}_i), P(2/\underline{x}_i), \ldots, P(J/\underline{x}_i)\} \qquad 1.9.9$$

Therefore the overall error rate estimator is calculated from the average of the maximum posterior probabilities for each observation. Note that when an estimated discriminant function is used the unbiasedness of the posterior error rate estimator does not necessarily hold. Glick (1978) however, noted that when an estimate is used then the posterior error rate estimator will still be a good estimator of the error rate of the sample discriminant function.

A unique feature of the posterior type of error rate estimators is that the calculation of the estimator is not

1.54/...

dependent on classified observations. The estimator can thus be found even if classification may become known in the future only. This is so because the estimator is calculated by using each observation's largest posterior probability without regard to whether this probability is associated with the correct classification.

Hora et al. integrated the U-estimator and the posterior probability estimators. They calculated the posterior probability of an observation belonging to each population without using that observation to estimate the unknown population parameters.

They compared the apparent, U, posterior and posterior/U estimators using simulation in the Monte Carlo technique with normal densities. In the comparison they used the ratio

$$\frac{\text{Mean square error of the error rate estimator}}{\text{Best mean square error in that row}},$$

i.e. they normalised the results per simulated sample using the "best" one in a row as the standard unit.

They found that only in the case of a large number of variables and a small population separation in terms of the Mahalanobis distance was the estimator of the posterior/U method inferior to the others with respect to the estimation of the overall error rate. In general however, the posterior/U method was by far superior to the other techniques.

In a comparison of the techniques with respect to availability, accuracy etc. only one real problem was pointed out, and that was that the specifications of a family of probability functions is required. This is not a severe limitation as any appropriate probability function may be used.

The BMDP-7M discriminant analysis programs (Dixon and Brown, 1977) supply an estimate of the posterior probability of group membership so that the posterior probability error rate estimator can be determined easily as the arithmetic mean of these values - the U-method is available when the stepwise program is used. This package also supplies the posterior probabilities with the single observation withheld, so that the posterior/U method requires only the calculation of a simple numerical average of the posterior probability error rate estimates as they are supplied by the printout.

Hora et al. recommended the posterior/U method if the assumption of normality is not severely violated, otherwise the U method or apparent method must be used, because of the bias which may make the first method inferior if normality can not be assumed. Note also that unequal covariance matrices may have the same effect (Hawkins, 1981).

1.10 A test for discriminatory power

Having established a discrimination procedure it is of interest to know the discriminating ability of the discrimination function(s). If the sample N is large enough, where $N = \sum_{i=1}^{k} n_i$, the following approximate method may be applied to determine whether the discrimination procedure does better than chance (Press, 1972).

Define the confusion matrix C with elements n_{ij} being the number of observations from population Π_i misclassified into Π_j. Assume k populations and n_i observation per population Π_i, i = 1, ..., k, then the matrix is

1.56/...

True Π_i's	predicted Π_j's: Π_1	Π_2	Π_k
Π_1	n_{11}	n_{12}	n_{1k}
Π_2	n_{21}	.....	n_{2k}
..	.....		
Π_k	n_{k1}	.....	n_{kk}

$= C_{kxk}$ 1.10.1

Note that the diagonal elements denote the number of correct classifications.

Use n for the total number of correct classifications and $\bar{n}$ for the number of misclassifications, e for the number of expected correct classifications and $\bar{e}$ for the expected number of misclassifications if classification is made at random where the probability of a succesful random classification is $\frac{1}{k}$, i.e.

$$n = \sum_{i=1}^{k} n_{ii}\ ,\quad \bar{n} = N - n,\quad e = \frac{N}{k}\ ,\quad \bar{e} = N - \frac{N}{k} \qquad 1.10.2$$

Use the chi-square test, i.e.

$$Q = \frac{(n-e)^2}{e} + \frac{(\bar{n}-\bar{e})^2}{\bar{e}} \qquad 1.10.3$$

$$= \frac{(n-\frac{N}{k})^2}{\frac{N}{k}} + \frac{\left[(N-n)-(N-\frac{N}{k})\right]^2}{N-\frac{N}{k}}$$

$$= \frac{N - nk}{N(k-1)} \qquad 1.10.4$$

using 1.10.2 to find 1.10.4. The zero hypothesis is H_0: Correct classification is random; against H_a : Correct classification is better than pure chance. Under H_0 being

1.57/...

correct we find that Q is approximately distributed as the χ^2-statistic with 1 degree of freedom.

1.11 Validation of the technique of discriminant analysis in Market Research with special reference to the estimation of error rates in small sample discriminant analysis

Crash and Perrault (1977) asked whether the results of sample based discriminant analysis are valid with respect to the broader population of interest. They wanted to know whether the classification potential is as high as sample estimates indicate, whether the true population profiles, i.e. characteristics of groups which are dominant in terms of discrimination, are what they appear to be from the sample results and whether the underlying sample-based dimensions are generalisable to the population. They concentrated on small sample results as the marketing researchers are often forced to use small samples.

Glick showed that despite the good properties of the U method which is relatively robust to the assumption of normality and yields almost unbiased estimates of misclassification probabilities, the low magnitude of bias reduction is overwhelmed by the large standard deviation of the estimate. See Dillon (1979) for more detail as well as confidence limits for the probabilities of misclassification. The researcher wants to have confidence in his results - and in the case of small samples it is so much more difficult to find estimates with small confidence intervals. (See Glick, 1978).

Another reason why further investigation is necessary, is because of the results experienced by researchers with res-

pect to performance by the different techniques. The most common methods, like the OS, $\bar{U}$ etc. are not as robust to the assumption of normality as is the U-method and in the case of small samples one is quite often not sure of the distribution(s) involved.

Crash et al. discussed several methods of validation. Most of these methods are based on the bias in the error rates of classification as we have seen. In section 1.4.2 we also saw that when the sample size is small and the number of variables is small then neither the apparent or resubstitution method, nor the U-method is applicable.

At a basic level the validity of discriminant function analysis results resides in the stability of the coefficients of the discriminant functions derived. Research has been on its way to find validation methods which use all available information in the sample data to determine the stability of parameter estimates while allowing unbiased estimation of error rates. The U-method does not allow the determination of the stability of the coefficients. If the jackknife method is combined to the U-method however, one achieves two goals in one, viz. the error rates are estimated with stability of the coefficients.

Observe a sample of N sampling units and partition this sample into k subsets of M sampling units each. Compute the discriminant function based on all N sampling units. Hold out one subset and calculate the discriminant function based on the remaining k-1 subsets. Repeat this latter process for all k subsets, i.e. withholding subset i from the k subsets in the complete sample. Call the discriminant function based on the complete sample $f(\theta_o')$ and the other discriminant functions $f(\theta_i')$, i = 1, ..., k where $f(\theta_i')$ refers to the discriminant function calculated with subset i withheld.

Calculate the coefficient of each variable according to the jackknife method, i.e. for $\hat{a}_{it}$ the estimated coefficient of variable t in $f(\Theta_i')$, i = 0, ..., k calculate the pseudo-values

$$\hat{a}'_{it} = k\hat{a}'_{ot} - (k-1)\hat{a}'_{it} \qquad 1.11.1$$

and then

$$\hat{a}_{it} = \frac{\sum_{i=1}^{k} \hat{a}'_{it}}{k} = k\hat{a}'_{ot} - (k-1)\hat{a}'_{it} \qquad 1.11.2$$

When the researcher deals with small samples he may regard each observation as a subset. In any case, apply the U-method to estimate the rate of misclassification and simultaneously the jackknife method to estimate stable coefficients. Confidence intervals can be constructed using student's t with k-1 degrees of freedom. (See Tukey , 1958 and Mosteller and Tukey, 1968).

Another approach may be followed. When a subset is being held out - make a complete jackknife analysis of the remaining (k-1) subsets resulting in k equations and classify the results in the hold-out set according to all k (i.e. k-1 pseudo and 1 complete) discriminant functions. Do this to all k subsets, this results in k^2 cross validations and thus also yields a good measure of the performance of the variables. The pseudo-values also can be used to determine confidence intervals for the coefficients.

The bias in the jackknife estimate in the case of linear estimators is less than the bias in the original sample estimate and frequently approaches zero (see Green, 1964 and Crash et al.).

1.60/...

The jackknife method therefore provides a basis on which the strength of classification, not just the number classified, can be evaluated. It also provides us with some confidence intervals and is excellently applicable to small sample analysis.

At the moment however, "canned" computer software using the complete jackknife method is not yet available. Large samples result in quite a number of computer runs so that the time, effort and cost must be weighed against stability of coefficients. With large samples, however there is not much to gain from the other more common techniques of determining the discriminant functions.

Chapter 2

PREDICTIVE DISCRIMINATION

2.1 Introduction

Sometimes it is of interest to an investigator to assess in some way the relative odds or probability that a new multivariate observation $\underline{z}$ belongs to one of k multivariate normal populations Π_i, i = 1, ..., k. This problem has been thoroughly investigated by Geisser (1964) and briefly summarised by Kendall and Stuart (1982) and Fatti, Hawkins and Raath (1982).

A criticism against the linear discriminant function of Fisher is that it takes no account of the relative sizes of the training samples from the different populations. Another criticism against the LDF is that the assumptions validating the procedure do not necessarily validate the procedure when estimators must be used for parameters (Anderson, 1958).

One way of looking at these problems to overcome them, is to look at the predictive distribution for a new vector of observations $\underline{z}$, given the sample means and covariances for each population. There are several ways to find these solutions, for example a direct odds method (Kendall et al.), a likelihood ratio introduced by Anderson (1958), and a formulation based on a Bayesian framework considered by Geisser and Dunsmore (1966).

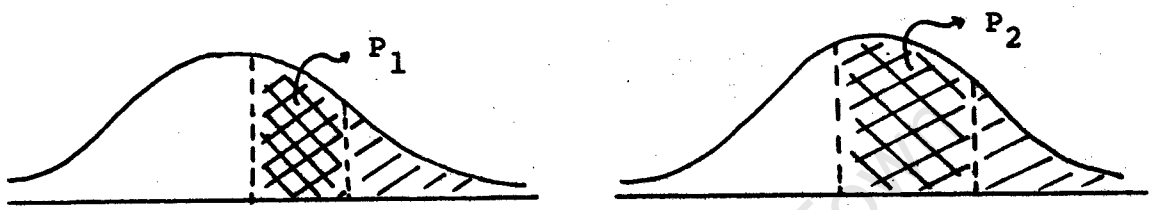
In the first case let i = 1, 2. Consider the Studentized variates

$$T_i^2 = (\underline{z} - \bar{\underline{x}}_i)' S^{-1} (\underline{z} - \bar{\underline{x}}_i) \qquad 2.1.1$$

where S is the pooled sample covariance matrix based on

$m = n_1+n_2-2$ degrees of freedom. Given that $\underline{z}$ comes from population i, $n_i T_i^2/(n_i+1)$ has Hotellings's T^2 distribution with m-1 degrees of freedom. The procedure is now to allocate the new individual to the more probable population; i.e. allocate $\underline{z}$ to population 1 when

$$P\ (T_1^2 > \text{observed}/H_1) > P\ (T_2^2 > \text{observed}/H_2) \qquad 2.1.2$$



T_{1obs} T_{1crit} relatively > T_{2obs} T_{2crit}

$P_1 < P_2$ ⇒ decide on 2

Fig. 2.1

Now both variables have T^2 distributions with m-1 degrees of freedom so that the decision rule becomes the discriminant

$$x^* = \frac{n_1 T_1^2}{n_1+1} - \frac{n_2 T_2^2}{n_2+1} \qquad 2.1.3$$

where $\underline{z}$ is allocated to population 1 if $x^* < 0$.

When the population covariance matrices are different the separate S_i must be used in 2.1.1 which then has a T^2 distribution with n_i-1 degrees of freedom. Note that in this case the one-sample form of T^2 is used. To find the appro-

2.2/...

priate values for T^2 use the F tables together with the correct multiplication factor. As the probabilities in 2.1.2 are conditional upon $\underline{z}$ coming from population i, the rule may be extended to cover different prior probabilities for H_1, H_2 etc.; for example

$$\pi_1 \, P\,(T_1^2 > \text{observed}/H_1) > \pi_2 \, P\,(T_2^2 > \text{observed}/H_2) \qquad 2.1.4$$

Anderson considered a likelihood ratio test for comparing the hypothesis that the new individual comes from population i, (i=1, 2) based upon the likelihood for all n_1+n_2+1 observations. That is, assign to population 1 if

$$\pi_1 \sup \left\{ \prod_{i=1}^{n_1+1} f_i(\underline{x}_{1i}) \cdot \prod_{i=1}^{n_2} f_2(\underline{x}_{2i}) \right\} > \pi_2 \sup \left\{ \prod_{i=1}^{n_1} f_1(\underline{x}_{1i}) \cdot \prod_{i=1}^{n_2+1} f_2(\underline{x}_{2i}) \right\} \qquad 2.1.5$$

where in each case the supremum is evaluated over the parameter space. When the two populations are $n(p, \underline{u}_i, \Sigma)$ we find that the likelihood ratio statistic yields the posterior odds ratio

$$\lambda = \frac{\pi_1}{\pi_2} \left[\left\{ 1 + \frac{n_2 T_2^2}{m(n_2+1)} \right\} / \left\{ 1 + \frac{n_1 T_1^2}{m(n_1+1)} \right\} \right]^{\frac{1}{2}(n_1+n_2+1)} \qquad 2.1.6$$

where population 1 is preferred when $\lambda > 1$. When $\pi_1 = \pi_2$ we find 2.1.3. This rule differs from Geisser's results by a constant only.

2.3/...

2.2 The Bayesian approach to predictive discrimination

Geisser (1966) summarised the procedure for the general case applicable to any continuous distribution and any number of parameters.

Assume k populations Π_i, $i = 1, \ldots, k$ each specified by a continuous density $f(\cdot/\theta_i, \psi_i)$, θ_i and ψ_i being the sets of distinct unknown parameters of Π_i. Let X_i be the matrix set of data obtained on Π_i based on N_i independent vector observations. Let $\underline{z}$ be the new vector observations to be assigned which has a prior probability q_i of belonging to Π_i where

$$\sum_{i=1}^{k} q_i = 1.$$

Let $\theta = \bigcup_{i=1}^{k} \theta_i$, $\psi = \bigcup_{i=1}^{k} \psi_i$ then $g(\theta/\psi)$ is the joint prior density of θ for known ψ and $L(X_i/\theta_i, \psi_i)$ the likelihood of the sample obtained from Π_i, with the joint likelihood obtained on Π_i, $i = 1, \ldots, k$ given by

$$L(X/\theta, \psi) = \prod_{i=1}^{k} L(X_i/\theta_i, \psi_i) \qquad 2.2.1$$

where X is the set of all the data samples $X_1, \ldots, X_k$. The posterior density, if it exists, will be

$$P(\theta/X, \psi) \propto L(X/\theta, \psi) g(\theta/\psi) \qquad 2.2.2$$

From this the predictive density of $\underline{z}$, under the hypothesis that it was obtained from Π_i, is

$$f(\underline{z}/X, \psi, \Pi_i) = \int f(\underline{z}/\theta_i, \psi_i, \Pi_i) P(\theta/\psi)\, d\theta \qquad 2.2.3$$

or

2.4/...

$$f(\underline{z}/X, \psi, \Pi_i) = \int f(\underline{z}/\theta_i, \psi_i, \Pi_i)P(\theta_i/X, \psi)d\theta_i \qquad 2.2.4$$

where

$$P(\theta_i/X, \psi) = \int P(\theta/X, \psi)\, d\theta_i^c \qquad 2.2.5$$

using θ_i^c as the complement of θ_i, i.e. $\theta_i \cup \theta_i^c = \theta$

Then the posterior probability that $\underline{z}$ belongs to Π_i can be calculated, i.e.

$$P(\underline{z}\varepsilon\Pi_i/X, \psi, q) \propto q_i f(\underline{z}/X, \psi, \Pi_i) \qquad 2.2.6$$

For classification purposes we may choose to assign $\underline{z}$ to that Π_i for which 2.2.6 is a maximum. Sets of regions R_i, $i = 1, \ldots, k$ can be constructed for the observation space of $\underline{z}$ where R_i is the <u>set</u> of regions for which $u_i(\underline{z}) = q_i\ f(\underline{z}/X, \psi, \Pi_i)$ is a maximum and use these as classification regions for future observations.

Classification errors could be determined as follows:

$$P(\Pi_i/\Pi_i) = q_i \int_{R_i} f(\underline{z}/X, \psi, \Pi_i)d\underline{z} \qquad 2.2.7$$

$$P(\Pi_j/\Pi_i) = q_i \int_{R_j} f(\underline{z}/X, \psi, \Pi_i)d\underline{z}\ , \quad i\neq j \qquad 2.2.8$$

$$P(\Pi_i^c/\Pi_i) = q_i\left(1 - \int_{R_i} f(\underline{z}/X, \psi, \Pi_i)d\underline{z}\right) \qquad 2.2.9$$

Hence the predictive probability of <u>a</u> misclassification is

2.5/...

$$\sum_{i=1}^{k} P(\Pi_i^c/\Pi_i) = 1 - \sum_{i=1}^{k} P(\Pi_i/\Pi_i) \qquad 2.2.10$$

Geisser also investigated the joint predictive density in the classification of n jointly independent observations $\underline{z}_1, \ldots, \underline{z}_n$ and found that

$$P(\underline{z}_1 \epsilon \Pi_{i_1}, \ldots, \underline{z}_n \epsilon \Pi_{i_n}/X, \psi, q)$$

$$\propto \left(\prod_{j=1}^{n} q_{i_j}\right) f(\underline{z}_1, \ldots, \underline{z}_n/X, \psi, \Pi_{i_1}, \ldots, \Pi_{i_n}) \qquad 2.2.11$$

with the second factor on the right hand side being equal to

$$\int\left\{\int P(\theta/\psi, X)\, d\left\{\bigcup_{j=1}^{n} \theta_{i_j}\right\}^c\right\} \prod_{j=1}^{n} f(\underline{z}_j/\theta_{i_j}, \psi_{i_j}, \Pi_{i_j})\, d \bigcup_{j=1}^{n} \theta_{i_j} \qquad 2.2.12$$

With 2.2.11 in mind the procedure in 2.2.6 may be called a marginal assignment.

2.3 Application of the Bayesian approach to predictive discriminant analysis in the case of multivariate normal distributions

It is important to note that in classification applications our interest focuses primarily on a statement concerning the relative likelihood or probability that an observation belongs to one or another of the populations as a basis for assignment,

2.6/...

and not the more Bayesian application of making a probability statement about the locality of parameter(s).

Let $\underline{x}_{ij}$, $j = 1, \ldots, N_i$, $i = 1, \ldots, k$ be the set of pxl vectors coming from the k populations with

$$\bar{\underline{x}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \underline{x}_{ij}$$

$$A_i = \sum_{j=1}^{N_i} (\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)' = (N_i - 1)S_i$$

$$A = \sum_{i=1}^{k} A_i = (N-k)S, \quad N = \sum_{i=1}^{k} N_i \qquad 2.3.1$$

Assume now that an observation $\underline{z}' = (z_1, \ldots, z_p)$ is observed with known prior probability π_i of belonging to Π_i which is $n(p, \underline{\mu}_i, \Sigma_i)$ distributed with $\underline{\mu}_i' = (\mu_{1i}, \ldots, \mu_{pi})$, $\Sigma_i = [\sigma]_{uti}$, $i = 1, \ldots, k$; $u, t = 1, \ldots, p$.

When $\underline{\mu}_i$ and Σ_i are both known the posterior density function is identical to the discriminant function in the classical case for known parameters. The two other situations most commonly found in practise are the following:

(a) Σ_i, $\underline{\mu}_i$ all unknown, and

(b) $\Sigma_i = \Sigma\ \forall_i$, $\underline{\mu}_i$ all unknown

Press (1972) gives a brief but complete derivation of the results.

2.7/...

Let $\underline{z}$ be the new observation. The posterior density of the parameters in the i th population, given the observed data, is

$$P(\underline{\mu}_i, \Sigma_i^{-1} | \bar{\underline{x}}_i, A_i, \Pi_i) \propto P(\bar{\underline{x}}_i, A_i | \underline{\mu}_i, \Sigma_i^{-1}, \Pi_i) P(\underline{\mu}_i, \Sigma_i^{-1}) \qquad 2.3.2$$

i.e. it is proportional to the product between the likelihood and the prior densities. But $\bar{\underline{x}}_i$ and A_i are independent and

$$\bar{\underline{x}}_i \sim n(p, \underline{\mu}_i, \frac{1}{N_i}\Sigma_i) \qquad 2.3.3$$

$$A_i^{-1} \sim W(p, N_i - 1, \Sigma_i) \qquad 2.3.4$$

so that

$$P(\bar{\underline{x}}_i, A_i | \underline{\mu}_i, \Sigma_i^{-1}, \Pi_i) \propto |A_i|^{\frac{1}{2}(N_i - p - 2)} |\Sigma_i^{-1}|^{\frac{1}{2}N_i} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\Sigma_i^{-1}\left[A + N_i(\bar{\underline{x}}_i - \underline{\mu}_i)(\bar{\underline{x}}_i - \underline{\mu}_i)'\right]\right\} \qquad 2.3.5$$

Assume that no prior information is available, so we use the diffuse prior density

$$P(\underline{\mu}_i, \Sigma_i^{-1}) \propto |\Sigma_i^{-1}|^{-\frac{1}{2}(p+1)} \qquad 2.3.6$$

Therefore the posterior density of the parameters as in 2.3.2 becomes

2.8/...

$$P(\underline{\mu}_i, \Sigma_i^{-1}/\underline{\bar{x}}_i, A_i, \Pi_i) \propto |\Sigma_i^{-1}|^{\frac{1}{2}(N_i-p-1)} \cdot \exp\left\{-\tfrac{1}{2}\mathrm{tr}\Sigma_i^{-1}\left[A_i + N_i(\underline{\bar{x}}_i-\underline{\mu}_i)(\underline{\bar{x}}_i-\underline{\mu}_i)'\right]\right\} \quad 2.3.7$$

The predictive density of $\underline{z}$ for classification into population i is

$$P(\underline{z}/\underline{\bar{x}}_i, A_i, \Pi_i) = \iint P(\underline{z}|\underline{\mu}_i, \Sigma_i^{-1}, \Pi_i) P(\underline{\mu}_i, \Sigma^{-1}/\underline{\bar{x}}_i, A_i, \Pi_i) d\underline{\mu}_i d\Sigma_i^{-1} \quad 2.3.8$$

Substitute 2.3.7 and the sampling density of $\underline{z} \sim n(p, \underline{\mu}_i, \Sigma_i)$ into 2.3.8, then

$$P(\underline{z}/\underline{\bar{x}}_i, A_i, \Pi_i) \propto \iint |\Sigma_i^{-1}|^{\frac{1}{2}(N_i-p-1)} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\Sigma_i^{-1}\left[A_i+N_i(\underline{\bar{x}}_i-\underline{\mu}_i)(\underline{\bar{x}}_i-\underline{\mu}_i)'\right]\right\}$$

$$\cdot \frac{1}{(2\pi)^{\frac{1}{2}p}} |\Sigma_i^{-1}|^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\underline{z}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{z}-\underline{\mu}_i)\right\} d\underline{\mu}_j d\Sigma^{-1}$$

$$\propto \iint |\Sigma_i^{-1}|^{\frac{1}{2}(N_i-p)} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\Sigma_i^{-1}\left[A_i+N_i(\underline{\bar{x}}_i-\underline{\mu}_i)(\underline{\bar{x}}_i-\underline{\mu}_i)' + (\underline{z}-\underline{\mu}_i)(\underline{z}-\underline{\mu}_i)'\right]\right\} d\underline{\mu}_j d\Sigma^{-1} \quad 2.3.9$$

$$\propto \int \frac{d\underline{\mu}_i}{|A_i+N_i(\underline{\bar{x}}_i-\underline{\mu}_i)(\underline{\bar{x}}_i-\underline{\mu}_i)' + (\underline{z}-\underline{\mu}_i)(\underline{z}-\underline{\mu}_i)'|^{\frac{1}{2}(N_i+1)}} \quad 2.3.10$$

2.9/...

$$\propto \int \frac{d\underline{\mu}_i}{|B_i + (\underline{\mu}_i - \bar{\underline{\mu}}_i)(\underline{\mu}_i - \bar{\underline{\mu}}_i)'|^{\frac{1}{2}(N_i+1)}} \qquad 2.3.11$$

where

$$\bar{\underline{\mu}}_i = \frac{N_i \bar{\underline{x}}_i + \underline{z}}{N_i+1} \quad \text{and} \quad B_i = \frac{A_i}{N_i+1} + \frac{N_i}{N_i+1} \bar{\underline{x}}_i \bar{\underline{x}}_i' + \frac{\underline{z}\underline{z}'}{N_i+1} - \frac{(N_i \bar{\underline{x}}_i + \underline{z})(N_i \bar{\underline{x}}_i + \underline{z})'}{(N_i+1)^2}$$

Then 2.3.11 becomes

$$P(\underline{z}/\bar{\underline{x}}_i, A_i, \Pi_i) \propto \int \frac{d\underline{\mu}_i}{|B_i|^{\frac{1}{2}(N_i+1)} |I + B_i^{-1}(\underline{\mu}_i - \bar{\underline{\mu}}_i)(\underline{\mu}_i - \bar{\underline{\mu}}_i)'|^{\frac{1}{2}(N_i+1)}} \qquad 2.3.12$$

which gives us

$$P(\underline{z}/\bar{\underline{x}}_i, A_i, \Pi_i) \propto \frac{1}{|B_i|^{\frac{1}{2}(N_i+1)}} \int \frac{d\underline{\mu}_i}{[1 + (\underline{\mu}_j - \bar{\underline{\mu}}_j)' B_i^{-1} (\underline{\mu}_i - \bar{\underline{\mu}}_i)]^{\frac{1}{2}(N_i+1)}} \qquad 2.3.13$$

(from $|I_p + AB| = |I_p + \underline{a}\underline{b}'| = 1 + \underline{b}'\underline{a}$ when A_{pxn} and B_{nxp} are such that $n = 1$) so that the integrand is proportional to the multivariate student t-density. Therefore

$$P(\underline{z}/\bar{\underline{x}}_i, A_i, \Pi_i) \propto \frac{|B_i|^{\frac{1}{2}}}{|B_i|^{\frac{1}{2}(N_i+1)}} = |B_i|^{-\frac{1}{2}N_i} \qquad 2.3.14$$

$$\propto |A_i + \frac{N_i}{N_i+1} (\underline{z} - \bar{\underline{x}}_j)(\underline{z} - \bar{\underline{x}}_i)'|^{-\frac{1}{2}N_i} \qquad 2.3.15$$

2.10/...

To find now the predictive probability density for classifying the new $\underline{z}$ into Π_i, we must find

$$P(\Pi_i/\text{data}) = P(\underline{z}/\bar{\underline{x}}_i, A_i, \Pi_i)q_i \qquad 2.3.16$$

Note that 2.3.15 can be written in exact form by taking the multivariate t-density $\underline{t}$: pxl with N_i-1 degrees of freedom as

$$g(\underline{t}) = \frac{k|\Sigma_i|^{-\frac{1}{2}}}{\left[N_i+(\underline{t}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{t}-\underline{\mu}_i)\right]^{\frac{1}{2}(N_i+p)}} \qquad -\infty<t_j<\infty,\ N_i>0 \qquad 2.3.17$$

where $k = \dfrac{N_i^{\frac{1}{2}N_i}\Gamma\left(\frac{1}{2}(N_i+p)\right)}{\Pi^{\frac{1}{2}p}\Gamma(\frac{1}{2}N_i)}$

From this follows that

$$P(\Pi_i/\underline{z},\{\underline{x}_{ij}\}) \propto q_i\left[\frac{N_i}{N_i+1}\right]^{\frac{1}{2}p}\frac{\Gamma(\frac{1}{2}N_i)}{\Gamma[\frac{1}{2}(N_i-p)]|S_i|^{\frac{1}{2}}}\cdot\left[1+\frac{N_i(\underline{z}-\bar{\underline{x}}_i)'A_i^{-1}(\underline{z}-\bar{\underline{x}}_i)}{N_i+1}\right]^{\frac{1}{2}N_i} \qquad 2.3.18$$

When the Σ_i are assumed to be equal the multivariate t density has N-k degrees of freedom and

$$P(\Pi_i/\underline{z},\{\underline{x}_{ij}\}) \propto q_i\left[\frac{N_i}{N_i+1}\right]^{\frac{1}{2}p}\left[1+\frac{N_i(\underline{z}-\bar{\underline{x}}_i)'A^{-1}(\underline{z}-\bar{\underline{x}}_i)}{N_i+1}\right]^{-\frac{1}{2}(N-k+1)} \qquad 2.3.19$$

2.11/...

It can be shown that 2.3.18 and 2.3.19 both tend to the classical case as $N_i \to \infty$.

Overall the differences between the predictive and classical approaches are slight in terms of the allocation rule applied, but they lead to very different estimates of the probability of misclassification for the new individual, that is the predictive approach yields more reliable estimators.

Aitchison, Habbeman and Kay (1977) made the statement that if the statistician wished to have some measure of confidence in the reality of the plausibilities that he reports for the type, he would be well advised on theoretical grounds to use the predictive approach.

McLachlan (1979) also found in his asymptotic approach, which does not rely on Monte Carlo methods that the atypicality indices favour the predictive method. He found that for equal probabilities the predictive method generally gives less extreme estimates of the posterior probabilities.

Raath and Hawkins (1980) discussed the problem of the choice between the heteroscedastic or homoscedastic models. In the derivation of 2.3.18 it was assumed a priori that the Σ_i are mutually independent with diffuse prior distributions. In practise these Σ_i are not so different as may be implied since it would be expected that measurements of the same characteristics in different populations would give rise to similar if not identical covariance matrices.

Marks and Dunn (1974) showed that in such a case Anderson's estimative technique in the heteroscedastic case gives worse classifications than the same technique in the homoscedastic form if the samples are of "moderate" size. Raath and Hawkins stated that it seems reasonable that this will also hold for the

predictive discrimination approach. This is quite useful, because it takes care of the condition that $n_i > p$ in order that S_i be non-singular.

2.4 Alternative forms and simplifications in preditive discriminant analysis assuming normal distributions

Although one often has no knowledge of what the covariance matrices are, one often has reason to believe that they are related. Let the $\underline{x}_{ij}$ be defined as in the last paragraph, then

$$\bar{\underline{x}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \underline{x}_{ij} \sim n(p, \underline{\mu}_i, \frac{\Sigma i}{N_i}) \qquad 2.4.1$$

and

$$A_i = \sum_{j=1}^{N_i} (\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)' \sim W(p, N_i - 1, \Sigma_i) \qquad 2.4.2$$

Let $N_i - 1 = n_i$ and let $\underline{\mu}_i$ have a diffuse prior and Σ_i be distributed a priori as a $W^{-1}(p, \upsilon, \Gamma)$ distribution where Γ has a diffuse prior. Then Raath and Hawkins showed that if $\underline{z}$ comes from a population Π_r with prior probability q_r, then $\underline{z}_r \sim n(p, \mu_r, \Sigma_r)$ and

$$f(\underline{z}, \Gamma/\text{data}) \propto |\Gamma|^m \prod_{i=1}^{k} \frac{|A_i|^{\frac{1}{2}(n_i - p - 1)}}{|\Gamma + B_i|^{\frac{1}{2}(n_i^* + \upsilon - p - 1}} \qquad 2.4.3$$

and from this follows that

$$f(\underline{z}/\text{data}) \propto \int_{\Gamma} |\Gamma|^m \prod_{i=1}^{k} \frac{|A_i|^{\frac{1}{2}(n_i - p - 1)}}{|\Gamma + B_i|^{\frac{1}{2}(n_i^* + \upsilon - p - 1)}} \, d\Gamma \qquad 2.4.4$$

2.13/...

where $n_i^* = \begin{cases} n_i & i \neq r \\ n_i+1 & i=r \end{cases}$

$m = \frac{1}{2}[k(\upsilon-p-1)-p-1]$

$$B_i = \begin{cases} A_i & i \neq r \\ A_i + \frac{N_i}{N_i+1}(\underline{z}-\bar{\underline{x}}_i)(\underline{z}-\bar{\underline{x}}_i)' & i=r \end{cases} \qquad 2.4.5.$$

Now, 2.4.4 may be evaluated numerically in special cases. Γ may be known, "plugged" in or diagonal but unknown. We will have a look at each of these alternatives - presenting the results only.

In the first instance let 2.4.6

$$f(\underline{z}/\text{data}, \underline{z}_r) \propto \prod_{i=1}^{k} \frac{|A_i|^{\frac{1}{2}(n_i-p-1)}}{|\Gamma+B_i|^{\frac{1}{2}(n_i^*-\upsilon-p-1)}} \qquad 2.4.7$$

so that

$$P(\Pi_r/\underline{z}, \text{data}) \propto q_r\, f(\underline{z}/\text{data}, \Pi_r) \qquad 2.4.8$$

Further it follows that for a general known Γ

$$f(\underline{z}/\text{data}, \Pi_r) = \pi^{-\frac{1}{2}}\left(\frac{N_r}{N_r+1}\right)^{\frac{1}{2}p} \frac{\Gamma_p(\frac{1}{2}(n_r+\upsilon-p))}{\Gamma_p(\frac{1}{2}(n_r+\upsilon-p-1))} \cdot \frac{|\Gamma+A_r|^{\frac{1}{2}(n_r+\upsilon-p-1)}}{|\Gamma+A_r+\frac{N_r}{N_r+1}(\underline{z}-\bar{\underline{x}}_r)(\underline{z}-\bar{\underline{x}}_r)'|^{\frac{1}{2}(n_r+\upsilon-p)}} \qquad 2.4.9$$

2.14/....

where $\Gamma_p(a) = \pi^{\frac{1}{4}p(p-1)} \prod_{i=1}^{p} \Gamma(a-\frac{1}{2}(i-1))$

If Γ is not known it can be estimated using A_i in the sense that a pooled estimator of Γ will be

$$\hat{\Gamma} = \frac{\sum_{i=1}^{k} w_i G_i}{\sum_{i=1}^{k} w_i}$$

$$= \frac{\sum_{i=1}^{k} w_i \frac{\upsilon-2p-2}{N_i-1} A_i}{\sum_{i=1}^{k} w_i} \qquad 2.4.10$$

with the condition that the weights w_i are such that $\Sigma w_i \neq 0$, e.g. w_i could be equal to $(1-\frac{1}{N_i})$. Note that G_i is an unbiased estimator of Γ since

$$E(A_i/\Sigma_i) = (N_i-1)\Sigma_i \qquad 2.4.11$$

and

$$E(\Sigma_i) = \Gamma/(\upsilon-2p-2) \qquad 2.4.12$$

provided that $\upsilon > 2p+2$. From this follows that

$$E(A_i/\Gamma) = \frac{N_i-1}{(\upsilon-2p-2)} \Gamma \qquad 2.4.13$$

so that

$$G_i = \frac{\upsilon-2p-2}{N_i-1} A_i \qquad 2.4.14$$

2.15/...

Only in case k is small and N_i small will the error be large and must one resort to more exact theory.

In the case of a diagonal Γ, though unknown, Γ can be estimated using an iteration model. If

$$f(\Gamma/\text{data}) \propto |\Gamma|^{m} \prod_{i=1}^{k} |\Gamma+A_i|^{-\frac{1}{2}(n_i+\upsilon-p-1)} \qquad 2.4.15$$

take the logarithm and differentiate with respect to Γ; then

$$m\Gamma^{-1} = \sum_{i=1}^{k} \tfrac{1}{2}(n_i+\upsilon-p-1)[(\Gamma+A_i)^{-1}]_d \qquad 2.4.16$$

where $[(\Gamma+A_i)^{-1}]_d$ is the diagonal of the inverse matrix $(\Gamma+A_i)^{-1}$. Use 2.4.10 now as an estimate for Γ in 2.4.3 so that with the constant of proportionality being calculated

$$f(\underline{z}/\text{data}) = \pi^{-\frac{1}{2}p} \left(\frac{N_r}{N_r+1}\right)^{\frac{1}{2}p} \frac{\Gamma_p(\frac{1}{2}(n_r+\upsilon-p))}{\Gamma_p(\frac{1}{2}(n_r+\upsilon-p-1))} \cdot$$

$$\cdot \frac{|\hat{\Gamma}+A_r|^{\frac{1}{2}(n_r+\upsilon-p-1)}}{|\hat{\Gamma}+A_r+\frac{N_r}{N_r+1}(\underline{z}-\bar{\underline{x}}_r)(\underline{z}-\bar{\underline{x}}_r)'|^{\frac{1}{2}(n_r+\upsilon-p)}} \qquad 2.4.17$$

so that $P(\Pi_i/\underline{z}, \text{data}) \propto q_i\, f(\underline{z}/\text{data}, \Pi_i)$.

Raath and Hawkins also discussed transforms which may be necessary to diagonalise the matrix if it is not so in the

2.16/...

first place. That will however only be possible if the Σ_i are believed to be approximately proportional to an intra-class correlation matrix. In such a case a preliminary Helmert transformation of the data will reduce Σ_i to an approximately diagonal form.

They found that the full predictive discrimination model obtained by integrating Γ out is unworkable unless p=1 and k=2.

Plugging in an estimator of Γ gives results that seem quite useful.

They concluded that the model of a general Γ may be unreliable if there are only a few small populations. When the assumption of a diagonal Γ is true however, fewer parameters have to be estimated with an increase in accuracy.

A note on the choice of υ is given as this reflects the strength of one's a priori conviction in the similarities of the different Σ_i.

Consider σ_{jji} - the variance of the jth component in the ith population. The assumption that $\Sigma_i \sim W^{-1}(p, \upsilon, \Gamma)$ implies a priori that

$$\sigma_{jji}/\sigma_{jjk} \sim F(\upsilon-2p, \upsilon-2p) \qquad 2.4.18$$

an F-distribution. Suppose now that one is 95% sure that the ratio $\sigma_{jji}/\sigma_{jjk}$ for an arbitrary i, j and k will lie in the range $(\frac{1}{4}, 4)$, then tables of the F-distribution show

2.17/...

that $\upsilon-2p = 9$ or $\upsilon = 2p + 9$. By specifying intervals of suitable width and suitable confidence, υ may also be determined in this way.

The general assumptions of discriminant analysis are either that all groups have identical covariance matrices, or that these matrices are completely general. The true situation is usually intermediate. While not exactly equal, the covariance matrices are similar. Such similarity may be modelled by introducing hyperparameters Γ and υ supposing the Σ_i to be centred on Γ.

Although the efficiency of the classification rule obtained by the estimative method doesn't differ much from the rule obtained by the predictive method, especially if the N_i are not small and/or markedly unequal, the posterior probabilities may be quite different.

Aitchison, Habbema and Kay (1977) considered two p-dimensional multivariate normal populations and data sets that are small in relation to the parameter dimension. They found in their study that the predictive method with unequal covariance matrices is slightly better than the same method with equal covariance matrices, which in turn is better than the estimative method with equal covariance matrices (LDF). The estimative method with unequal covariance matrices (QD) came off badly.

McLachlan (1979) questioned Aitchison et al's (1977) theoretical motivation for the results they obtained on the grounds that their n_1 and n_2 were very small relatively to p. He investigated the relative performances of the different methods and found that the predictive method's results are less extreme asymptotically for the posterior probabilities in the case of equal prior probabilities as well as unequal prior probabilities

2.18/...

except when the predictive and estimative rules are on opposite sides of 0,5 such that the predictive probability is further away from 0,5 than the estimative probability. He further found that the predictive method is usually less biased than the estimative method.

In a mechanical classification procedure however, it doesn't matter much which of the predictive or estimative method is used.

Chapter 3

DISCRIMINATION OF DISCRETE AND OR MIXED DATA

3.1 Introduction

It has been found by several statisticians that the estimative procedures of Fisher and Anderson are robust for deviations from the normality assumption (Lachenbruch and Goldstein, 1979). Fisher's approach in particular is not dependent on normality, but as in the case of unequal covariance matrices skew distributions will lead to unequal misclassification probabilities. It is interesting to note that in the case of unequal covariance matrices the quadratic procedures tend to be more unstable for deviations from normality than the linear discriminant procedures. By deviation from normality one refers inter alia to continuous distributions (probably multivariate) which are not distributed normally, discrete data (probably multivariate) or mixed data where one may have a vector for which some variables are discrete and some are continuous of a normal type or otherwise.

Discrete data may be of a type where some variables may take on more than two values and some are dichotomous. Some of these variables may represent ordered measurements and some nominal.

In the case of ordered discrete variables relatively satisfying results may be possible using the linear discriminant function (LDF) - i.e. treating the data as an approximation of a multivariate normal distribution. If the ordered variables form part of a mixed vector where the other variables are continuous, one may do the same or treat the mixed variables also as discrete variables by collapsing each continuous variable as a discrete variable by grouping the observations of each continuous variable, i.e. by using dummy variables. In the latter case a purely discrete technique may be used (Cochran and Hopkins, 1961).

If however, the discrete variables are of the nominal type it is obvious that one may not use a continuous data method like the LDF. In such a case one must use a pure discrete technique.

There are various other approaches to discriminant analysis. If one does not want to make any specific assumptions about the distributions associated with the groups one may assume that the posterior probabilities of each of the groups, given a particular observation, has a logistic form. The location model as well as the multinomial procedures are based on specific distributional assumptions, although in the latter case they are very broad. Non-parametric methods do exist like the kernel method where a density function is estimated before discrimination takes place. Other non-parametric methods are the rank transformation procedure and the nearest neighbourhood procedure (Hills, 1967).

It is noteworthy that where some methods are applicable to binary (dichotomous) variables only, a variable representing more than one category can be replaced by more than one dummy variable of the dichotomous type.

For example let variable x have categories 1, 2, 3 and 4, then any representation of a category can be represented by dummy variables y_i as follows:

x	1	2	3	4
y_1	0	1	1	1
y_2	0	0	1	1
y_3	0	0	0	1

that is, category 3 in x is represented by $\underline{y} = (1, 1, 0)$.

3.2/...

Theoretically the fact that the method can be applied to binary variables only is therefore no restriction. Practically however, it may cause problems regarding the cell frequencies.

A brief review of some of these techniques follows - with the exception of the kernel function method which is described in more detail.

The LDF temptation however, will always be with us and this chapter is concluded with a section on scoring techniques on several types of data so that it can be used in an ordinary LDF analysis.

3.2 The multinomial model

Any combination of qualitative items may be regarded as a classification variable $\underline{x}$ with a discrete sample space. N individuals may be sampled from a mixed population or two independent samples of size n_1 and n_2 are given from populations Π_1 and Π_2 with prior probabilities usually specified.

In the first case denote the number of individuals from Π_i with $\underline{X} = \underline{x} = (x_1, \ldots, x_p)$ as the binomial random variable $N_i(\underline{x})$ with expected value $N\delta_i f_i(\underline{x})$, i=1, 2. Note that in $\underline{x} = (x_1, \ldots, x_p)$ each x_i, i=1, ..., p assumes at most a finite number, s_i, i=1, ..., p of distinct values; i.e. there are $\prod_{i=1}^{p} s_i$ possible types of measurements with underlying density within each group being multinomial. In each group each of these $s = \prod_{i=1}^{p} s_i$ points has a probability attached to it, i.e. group Π_j is characterised by the multinomial probability $p(\underline{\theta}) = p(\theta_{j1}, \ldots, \theta_{js})$. (Berenson, Levine and Goldstein, 1983).

3.3/...

The total number of individuals is $N = n_1+n_2 = \sum_{\underline{x}} N_1(\underline{x}) + \sum_{\underline{x}} N_2(\underline{x})$

Intuitive estimates for prior probabilities are given by $\hat{\delta}_i = \frac{n_i}{N}$ and the non-parametric estimates of the class conditional densities or state probabilities by $\hat{f}_i(\underline{x}) = \frac{n_i(\underline{x})}{n_i}$,

$i=1, 2.$ (Goldstein and Dillon, 1978).

The resulting discriminant scores are then given by

$$\hat{g}_i(x) = \hat{\delta}_i \hat{f}_i(\underline{x}) = \frac{n_i}{N} \cdot \frac{n_i(\underline{x})}{n_i} = \frac{n_i(\underline{x})}{N}, \quad i=1, 2. \qquad 3.2.1$$

With independent random samples, prior probabilities are usually specified and not estimated. The random variables $n_i(\underline{x})$, $i=1, 2$ are still defined as the number of individuals from Π_i with $\underline{X} = \underline{x}$.

Now, $E[n_i(\underline{x})] = n_i \, f_i^*(\underline{x})$, where $f_i^*(\underline{x})$ is the density at $\underline{x}$ or the state probability defined by $\underline{x}$ from Π_i as opposed to $f_i(\underline{x})$ as defined above. Now, $f_i^*(\underline{x})$ can be estimated by $n_i(\underline{x})/n_i$. If δ^* is specified a positive prior probability associated with Π_i, then the discriminant score becomes

$$\hat{g}_i(\underline{x}) = \hat{\delta}^* \, \hat{f}_i^*(\underline{x}) = \hat{\delta}^* \cdot \frac{n_i(\underline{x})}{n_i} \qquad 3.2.2$$

If the sample space is partitioned by $\hat{D} = \langle \hat{D}_1, \hat{D}_2 \rangle$ then from 3.2.1 and 3.2.2 classify $\underline{x}$ as from $\hat{D}_1$ if $\hat{g}_1(\underline{x}) \geq \hat{g}_2(\underline{x})$ and to Π_2 if $\hat{g}_1(\underline{x}) < \hat{g}_2(\underline{x})$.

Thus in the case of a mixed sample the classification rule becomes: Assign to $\hat{D}_1$ if

$$n_1(\underline{x}) \geq n_2(\underline{x}) \qquad 3.2.3$$

and to $\hat{D}_2$ otherwise.

3.4/...

In the case of independent samples the rule becomes:
$\underline{x} \in \hat{D}_1$ if

$$\delta^* \, n_1(\underline{x})/n_1 \geq (1-\delta^*) n_2(\underline{x})/n_2 \qquad 3.2.4$$

and $\underline{x} \in \hat{D}_2$ if

$$\delta^* \, n_1(\underline{x})/n_1 < (1-\delta^*) n_2(\underline{x})/n_2 \qquad 3.2.5$$

Note that this procedure can easily be extended to more than two groups.

A big problem with this discrimination method is the fact that a large number of observations relative to the number of variables is required if sufficient data in each state are to be available for the estimation of state probabilities. Another objection to the full multinomial model, as is also known, is that a zero state in Π_1 may mean something different to a zero state in Π_2. Moore (1973) as well as Goldstein and Rabinowitz (1975) came to the conclusion that this method is applicable only when heteroscedasticity exists between the groups and when large training samples are available.

3.3 A variation on the full multinomial model

Hills (1967) suggested the nearest neighbourhood rule as a kind of solution to the problem of state sparseness as set out above. It should be noted that Hills's method fixes a volume and determines the proportion of points falling in that volume, so that it is really a discrete variable kernel function method (see section 3.6) with a kernel which is uniform within a certain range and zero outside it (Hand, 1981).

The technique introduced by Hills was summarised by Goldstein and Dillon (1978).

3.5/...

Let all data be dichotomies or use dichotomised dummy variables - see the previous section, i.e. all variables are of the (o;1) type.

Use a sample-based likelihood ratio procedure for classifying a particular response vector $\underline{x}$. All responses differing from $\underline{x}$ in no more than r components are incorporated into the rule. For a given response vector $\underline{x}$, let

$$T_j = \{\underline{y}_j / (\underline{x}-\underline{y}_j)'(\underline{x}-\underline{y}_j) \leq r\} \qquad 3.3.1$$

T_j is now the set of responses $\{\underline{y}_j\}$ such that each of the elements differs in no more than r components from $\underline{x}$, i.e. we have an r-level nearest neighbourhood rule. Assume further that the samples i=1, 2 are independent with size n_1 and n_2 from Π_i, i=1, 2. Then the discrimination rule is

$$\underline{x} \in \Pi_1 \quad \text{if} \quad \delta^* \sum_{T_j} \frac{n_1(\underline{y}_j)}{n_1} > (1-\delta^*) \sum_{T_j} \frac{n_2(\underline{y}_2)}{n_2}$$

$$\underline{x} \in \Pi_2 \quad \text{if} \quad \delta^* \sum_{T_j} \frac{n_1(\underline{y}_j)}{n_1} < (1-\delta^*) \sum_{T_j} \frac{n_2(\underline{y}_j)}{n_2} \qquad 3.3.2$$

and randomly allocate if equality holds. Note that δ^* is used as defined in the previous section.

Hills briefly discussed the approach in the case of more than 2 populations. He found that this method is less subject to sampling variability, but it is not necessarily consistent.

Hand discussed the problem related to the uncertainty of

3.6/...

acceptance or rejection for the calculation purposes in 3.3.1 when an observation is on the "cell limit."

3.4 Logistic discriminant analysis

A series of popular approaches to binary discrimination problems have been based on logarithmic transformations. A simple one is obtained when we assume that each class-conditional probability function can be written as

$$P(\underline{x}/\Pi_i) = \exp(\alpha_i' \underline{X}) \qquad 3.4.1$$

i.e. the logarithms of the class-conditional functions are linear. Sometimes $\underline{X} = \underline{x}$, but often $\underline{X}' = (1, \underline{x}')$ to allow for a constant term. Hand (1982) gave an excellent brief discussion of the former and Cox (1970) discussed the technique from a regression point of view.

More generally however, we can assume only that the difference between the logarithms of the class-conditional probability functions is linear, making no assumption about the linearity of the class-conditional probability functions themselves. Note this model does imply the first one, but that the first does not imply the latter. This more general assumption is equivalent to assuming that the coefficients of all terms not explicitly included in $\underline{x}$ are identical in the two populations.

The motivation for our approach was summarised by Day and Kerridge (1967) with some theoretical background.

Suppose we have vectors $\underline{x}$ where the components are not normally distributed and where the observations are not independent. If in addition to this the sample vectors are not close to the mode of the distribution of $\underline{x}$ in any of the

3.7/...

populations then the logistic model supplies part of the solution to the problem. It is of interest to note that the logistic distribution shows very desirable properties in the sense that it is unimodal in the univariate sense and that the distribution has no heavy tails.

Anderson (1972) adopted this approach, except that he assumed that

$$\ln P(\Pi_i/\underline{x}) - \ln P(\Pi_j/\underline{x}) \qquad 3.4.2$$

is linear rather than

$$\ln P(\underline{x}/\Pi_i) - \ln P(\underline{x}/\Pi_j) \qquad 3.4.3$$

The difference in approach involves only a constant term.

According to this more general approach now

$$\ln (P(\Pi_i/\underline{x})/P(\Pi_j/\underline{x})) = \underline{\alpha}_i\underline{x}\ , \quad i=1, \ldots, k-1 \qquad 3.4.4$$

This model is exact for $\underline{x}' = (1, \underline{X}')$ and X_i independent binary. It is also exact if the class-conditional probability density functions are multivariate normal with identical variance-covariance matrices.

In fact a number of distributions used in discriminant analysis have posterior probabilities with logistic form, e.g.

(a) as mentioned the multivariate normal distribution with equal dispersion matrices and

(b) also mentioned the multivariate dichotomous or (0;1) variables and further

3.8/...

(c) the multivariate Bernoulli distribution where the variables follow a log-linear model with equal second and higher order effects in all k groups and

(d) a combination of the distributions mentioned in (a), (b) and (c).

Note that the posterior probabilities are not based on distributions for which parameters must be estimated. The parameters in the posterior probabilities are estimated directly, i.e. in stead of

$$P(\Pi_i/\underline{x}) \propto \pi_i \; f_i(\underline{x}) \qquad i = 1, \ldots, k \qquad 3.4.5$$

where $\underline{x}$ is classified in the group with the highest posterior probability, $f_i(\underline{x})$ being the density function corresponding to Π_i which we have estimated up to now, let

$$\ln \frac{P(\Pi_i/\underline{x})}{P(\Pi_k/\underline{x})} = \ln \frac{p_{i\underline{x}}}{p_{k\underline{x}}} = \underline{\alpha}_i \underline{x} \qquad i=1, \ldots, k-1$$

i.e. the difference between the logs is linear, where $\underline{\alpha}_i' = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{ip})$ and $\underline{x}' = (x_o, x_1, \ldots, x_p)$ with $x_o = 1$.

Then

$$P(\Pi_i/\underline{x}) = p_{i\underline{x}} = \exp(\underline{\alpha}_i' \underline{x}) p_{k\underline{x}} \qquad i=1, \ldots, k-1 \qquad 3.4.6$$

But

$$\sum_{i=1}^{k-1} p_{i\underline{x}} + p_{k\underline{x}} = 1$$

3.9/...

therefore

$$\sum_{i=1}^{k-1} \exp(\underline{\alpha}_i'\underline{x})p_{k\underline{x}} + p_{k\underline{x}} = 1$$

i.e.

$$P(\Pi_k/\underline{x}) = p_{k\underline{x}} = (1 + \sum_{i=1}^{k-1} \exp(\underline{\alpha}_i'\underline{x}))^{-1} \qquad 3.4.7$$

In this more general case some of the x_i may be continuous and some even polychotomous (Fatti, 1982).

Anderson used the maximum likelihood method to obtain a set of equations which can be solved for the $\underline{\alpha}_i$ applying an iterative Newton-Raphson method to find these solutions.

Let $P(\underline{x}/\Pi_s) = f_s(\underline{x})$ be the likelihood of $\underline{x}$ given Π_s and denote the number of points from class or population s in cell $\underline{x}$ by $n_{s\underline{x}}$. Let $n_{\underline{x}} = \sum_{i=1}^{k-1} n_{i\underline{x}}$ and $A = [\underline{\alpha}_1, \underline{\alpha}_2, \ldots, \underline{\alpha}_{k-1}]$. Further denote the mixture density of $\underline{x}$ by $\phi_{\underline{x}}$, and π_i the probability associated with population Π_i, then

$$L = \prod_{s=1}^{k} \prod_{\underline{x}} \{L(\underline{x}/\Pi_s)\}^{n_{s\underline{x}}} \qquad 3.4.8$$

but $L(\underline{x}/\Pi_s) = P(\underline{x}/\Pi_s)$

$$= \frac{P(\underline{x} \cap \Pi_s)}{P(\Pi_s)}$$

$$= \frac{P(\Pi_s/\underline{x}) \cdot P(\underline{x})}{P(\Pi_s)}$$

3.10/...

$$= \frac{p_{s\underline{x}} \cdot \phi_{\underline{x}}}{\pi_s} \qquad 3.4.9$$

from which follows that

$$L = \prod_{s=1}^{k} \prod_{\underline{x}} \left(\frac{p_{s\underline{x}} \cdot \phi_{\underline{x}}}{\pi_s} \right)^{n_{s\underline{x}}}$$

$$= \prod_{s=1}^{k} \prod_{\underline{x}} \left(\frac{p_{s\underline{x}} \cdot p_{\underline{x}}}{\pi_s} \right)^{n_{s\underline{x}}} \qquad 3.4.10$$

$$\log L = \sum_{s=1}^{k} \log \prod_{\underline{x}} \left(\frac{p_{s\underline{x}} \cdot p_{\underline{x}}}{\pi_s} \right)^{n_{s\underline{x}}}$$

$$= \sum_{s=1}^{k} \sum_{\underline{x}} n_{s\underline{x}} \{ \log p_{s\underline{x}} p_{\underline{x}} - \log \pi_s \}$$

$$= \sum_{s=1}^{k} \sum_{\underline{x}} n_{s\underline{x}} \{ \log p_{s\underline{x}} p_{\underline{x}} \} - \sum_{s=1}^{k} \sum_{\underline{x}} n_{s\underline{x}} \log \pi_s \qquad 3.4.11$$

To find the first partial derivative w.r.t. α_{sj}, define $f_{ij} = \frac{\partial \log L}{\partial \alpha_{ij}}$, then, keeping in mind that the second term on the right hand side as well as $p_{\underline{x}}$ do not contain an α_{sj}, let

$$\log L_1 = \sum_{s=1}^{k} \sum_{\underline{x}} n_{s\underline{x}} \log p_{s\underline{x}}$$

$$= \sum_{s=1}^{k} \sum_{\underline{x}} n_{s\underline{x}} \log e^{\underline{\alpha}_s' \underline{x}} p_{k\underline{x}}$$

3.11/...

$$= \sum_{\underline{x}} n_{1\underline{x}} \underline{\alpha}_1' \underline{x} + \sum_{\underline{x}} n_{2\underline{x}} \underline{\alpha}_2' \underline{x} + \ldots + \sum_{\underline{x}} n_{s\underline{x}} \underline{\alpha}_s' \underline{x} + \ldots + \sum_{\underline{x}} n_{k-1,\underline{x}} \underline{\alpha}_{k-1}' \underline{x}$$

$$+ \sum_{\underline{x}} n_{1\underline{x}} \log p_{k\underline{x}} + \ldots + \sum_{\underline{x}} n_{s\underline{x}} \log p_{k\underline{x}} + \ldots + \sum_{\underline{x}} n_{k-1,\underline{x}} \log p_{k\underline{x}} +$$

$$\sum_{\underline{x}} n_{k,\underline{x}} \log p_{k\underline{x}}$$

$$\frac{\partial \log L_1}{\partial \alpha_{sj}} = \sum_{\underline{x}} n_{s\underline{x}} x_j + \sum_{\underline{x}} \sum_{j=1}^{k} \left\{ n_{j\underline{x}} \cdot \frac{\partial \log p_{k\underline{x}}}{\partial \alpha_{sj}} \right\}$$

$$= \sum_{\underline{x}} n_{s\underline{x}} x_j + \sum_{\underline{x}} \left\{ n_{\underline{x}} \cdot \left(1 + \sum_{i=1}^{k-1} e^{\underline{\alpha}_i' \underline{x}}\right) \cdot \frac{-x_j \, e^{\underline{\alpha}_s' \underline{x}}}{\left(1 + \sum_{i=1}^{k-1} e^{\underline{\alpha}_i' \underline{x}}\right)^2} \right\}$$

$$= \sum_{\underline{x}} \{ n_{s\underline{x}} x_j - n_{\underline{x}} \, p_{s\underline{x}} \, x_j \}$$

So that eventually

$$f_{sj} = \frac{\partial \log L}{\partial \alpha_{sj}} = \sum_{\underline{x}} \{ n_{s\underline{x}} - n_{\underline{x}} \, p_{s\underline{x}} \} x_j \qquad 3.4.12$$

after an amount of manipulation.

Let f_{sj} be equal to zero and solve iteratively for the α_{ij} using the Newton-Raphson procedure as described in Anderson.

Let f_{sj} be as in 3.4.12 for j = 0, 1, ..., p then

3.12/...

$$\frac{\partial^2 \log L}{\partial \alpha_{t\ell} \partial \alpha_{sj}} = \frac{\partial f_{sj}}{\partial \alpha_{t\ell}} = \sum_{\underline{x}} n_{\underline{x}} p_{s\underline{x}} p_{t\underline{x}} x_j x_\ell \quad , \quad s \neq t \qquad 3.4.13$$

and

$$\frac{\partial^2 \log L}{\partial \alpha_{s\ell} \partial \alpha_{sj}} = \frac{\partial f_{sj}}{\partial \alpha_{s\ell}} = -\sum n_{\underline{x}} p_{s\underline{x}} (1-p_{s\underline{x}}) x_j x_\ell \quad , \quad s=t \qquad 3.4.14$$

Let

$$F_\alpha = \left[\frac{\partial f_{sj}}{\partial \alpha_{t\ell}}\right]_{(k-1)(p+1)\times(k-1)(p+1)} = \left[F_{sj,t\ell}\right] \qquad 3.4.15$$

i.e. s, t = 1, ..., k-1 and j, ℓ = 0, ...,p with the ordered pair form for (s,j) and (t,ℓ) of (1,0), (1,1), ..., (1,p), (2,0), (2,1), ..., (k-1, p) and $\underline{f}_\alpha$ the column vector with (k-1)(p+1) rows, i.e. $\underline{f}_\alpha = \{f_{ij}\}$ with the same ordering. Let $\underline{f}_a$ and F_a denote the values of $\underline{f}_\alpha$ and F_α when $\underline{\alpha}_1$, ..., $\underline{\alpha}_{k-1}$ take on the values indicated by the column vector $\underline{a}$ where $\underline{a}' = (\underline{\alpha}_1', \ldots, \underline{\alpha}_{k-1}')$. Use as starting values for the $\underline{a}_i$ as estimator of $\underline{a}$ the vector $\underline{o}$ and call it $\underline{a}_o$. The next value $\underline{a}_1$ will be

$$\underline{a}_1 = \underline{a}_o - F_{a_o}^{-1} \underline{f}_{a_o}$$

$$\underline{a}_2 = \underline{a}_1 - F_{a_1}^{-1} \underline{f}_{a_1} \qquad 3.4.16$$

As F^{-1} changes very slowly for changes in $\underline{f}$, one does not have to recalculate F^{-1} for each $\underline{a}_i$, but only for say, every tenth repetition. Having calculated the $\hat{\underline{\alpha}}_i$, the logistic preference for any population given any $\underline{x}$ can be determined, i.e. allocate

3.13/...

to that population for which the index gives a maximum for $\underline{\alpha}_i'\underline{x}$, i=1, ..., k.

For a more detailed derivation see also Day and Kerridge (1967). Anderson (1972) showed that this estimation is viable irrespective of the case of mixed populations or already separated data, although the maximum likelihood estimators are not unique when there is a complete separation of groups in the training sample, i.e. when a linear discriminant function can be found that classifies all N sample points correctly. In this case, however any solution to equation 3.4.12 will tend to give good discrimination although the α_{ij} may not be consistent.

For further development regarding log-linear models refer the next section.

As a last remark it is worthwhile looking at the conclusions of Press and Wilson (1978) where they motivated their preference of the logistic model above the estimative technique. Their studies show that whenever a single qualitative variable is present the logistic approach gives more acceptable results, although they stress the fact that differences will be small except in some specific cases.

3.5.1 The log-linear model and the discriminant problem

An alternative to the full multinomial model which suffers from the probability of state sparseness and an extension to the logistic model of the previous section wassuggested and described by Berenson, Levine and Goldstein (1983), Goldstein and Dillon (1978), Haberman (1974), Hand (1981) etc. Haberman discussed the use of this technique in the specific case of ordered categories. The model uses the goodness of fit statistics as a tool of model building, provides a solution to the state sparseness problem and incorporates information with respect to orderings of the variables. In addition to this the model takes care of association factors among variables.

Before discussing the application of the model itself, which is straightforward, because the model gives us a probability for $\underline{x}$ which together with the group probability, gives us the final procedure and solution, we give a brief description of the basis of the log-linear model and then its layout and extension to the logit form.

We will see that the sufficient statistics for log-linear models are easy to obtain and yield the expected cell frequencies without having to determine the underlying parameters.

3.5.2 The log-linear model, calculation of some statistics and an extension to the logit model

These models represent the probability distribution by an expansion (see Hand, 1981).

$$P(\underline{x}) = \exp(\alpha_o+\alpha_1(x_1)+\alpha_2(x_2)+\alpha_3(x_3)+ \ldots +\alpha_p(x_p)$$
$$+\alpha_{12}(x_1,x_2)+\alpha_{13}(x_1,x_3)+ \ldots +\alpha_{1p}(x_1,x_p)$$
$$+\alpha_{1,2,\ldots,p}(x_1,x_2,\ldots,x_p)) \qquad 3.5.2.1$$

where the $\alpha_{i,\ldots,j}$ for respective i, j are defined as

$$\alpha_o = \sum_{\underline{x}} \ln \hat{P}(\underline{x}) / \prod_{i=1}^{p} g_i$$

where g_i is the number of categories for the i th variable.

$$\alpha_1(a) = \sum_{\substack{\underline{x} \\ x_1=a}} \ln \hat{P}(\underline{x}) / \prod_{i=2}^{p} g_i - \alpha_o$$

.....

$$\alpha_p(a) = \sum_{\substack{\underline{x} \\ x_p=a}} \ln \hat{P}(\underline{x}) / \prod_{i=1}^{p-1} g_i - \alpha_o$$

3.15/...

$$\alpha_{12}(a,b) = \sum_{\substack{\underline{x} \\ x_1=a, x_2=b}} \ln\hat{P}(\underline{x}) / \prod_{i=3}^{p} g_i - \alpha_2 - \alpha_1 - \alpha_o$$

.....

$$\alpha_{1\ldots p}(a, \ldots, c) = \sum_{\substack{\underline{x} \\ \underline{x}=(a,\ldots,c)}} \ln\hat{P}(\underline{x}) - \ldots - \alpha_o \qquad 3.5.2.2$$

$\hat{P}(\underline{x})$ is the expected value of $P(\underline{x})$ under the model assumptions.

We will view the $\alpha_{i,\ldots,j}$'s as main effects and interactive terms analogous to those of the analysis of variance. In terms of this it is easy to see that for p=2 and both variables binary, α_o becomes

$$\alpha_o = \frac{\ln\hat{P}((0,0)) + \ln\hat{P}((0,1)) + \ln\hat{P}((1,0)) + \ln\hat{P}((1,1))}{4}$$

and α_1 becomes

$$\alpha_1 = \frac{\ln\hat{P}((0,0)) + \ln\hat{P}((0,1))}{2} - \alpha_o$$

$$= \frac{\ln\hat{P}((0,0)) + \ln\hat{P}((0,1)) - \ln\hat{P}((1,0)) - \ln\hat{P}((1,1))}{4}$$

3.5.2.3

With higher order interactions we will see that it is possible to economise on the number on interaction terms.

To describe the modelling technique a two-way table, i.e. a two variable model is assumed. It will be appreciated that this model can be expanded very easily to any realistic number of variables. A brief but useful summary can be found in Berenson, Levine and Goldstein (1983).

Let the variables A and B with I and J levels respectively form

3.16/...

an IxJ contingency table. Let x_{ij} be the observed frequency and m_{ij} be the expected frequency associated with cell (i,j) if a sample of n observations was distributed across the table. Using the more common notation of Bishop et al. let

$$m_{ij} = e^{\mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{AB(ij)}} \qquad 3.5.2.4$$

so that

$$\ln m_{ij} = \mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{AB(ij)} \qquad 3.5.2.5$$

with constraints

$$\sum_i \mu_{A(i)} = \sum_j \mu_{B(j)} = \sum_i \mu_{AB(ij)} = \sum_j \mu_{AB(ij)} = 0 \qquad 3.5.2.6$$

Bishop, Fienberg and Holland (1975) showed that it is immaterial whether the cell probabilities or the expected counts in each cell are used as the mean terms, i.e. the μ - terms are just transformations of cross product ratios which are invariant under row and column multiplications, e.g. if I=J=2, then

$$\begin{aligned}\mu_{AB(11)} &= \tfrac{1}{4} \ln\left(\frac{m_{11}}{m_{12}}\Big/\frac{m_{21}}{m_{22}}\right) \\ &= \tfrac{1}{4} \ln\left(\frac{m_{11}\, m_{22}}{m_{12}\, m_{21}} \times \frac{\frac{1}{N^2}}{\frac{1}{N^2}}\right) \\ &= \tfrac{1}{4} \ln \frac{p_{11}p_{22}}{p_{12}p_{21}} \qquad 3.5.2.7\end{aligned}$$

As above the μ terms can now be defined as

3.17/...

$$\mu = \sum_{ij} \frac{\ln m_{ij}}{IJ} \qquad 3.5.2.8$$

$$\mu_{A(i)} = \sum_{j} \frac{\ln m_{ij}}{J} - \mu \qquad 3.5.2.9$$

$$\mu_{B(j)} = \sum_{i} \frac{\ln m_{ij}}{I} - \mu \qquad 3.5.2.10$$

$$\mu_{AB(ij)} = \ln m_{ij} - \mu_{A(i)} - \mu_{B(j)} + \mu \qquad 3.5.2.11$$

so that there are (I-1)(J-1) independent parameters needed to explain all possible association in the IxJ table. Note that calculation of $\mu_{AB(ij)}$ necessitates the calculation of $\mu_{A(i)}$ and $\mu_{B(j)}$ which require the value of μ. This order of calculation will always be necessary as higher order μ-terms measure the deviations between lower order μ-terms. This requires the use of the hierarchy principle, i.e.

(a) for any two sets of indexes $\{\theta\}$ and $\{\theta'\}$ such that $\{\theta\} \subset \{\theta'\}$, $\mu_{\{\theta\}} = 0$ implies that $\mu_{\{\theta'\}} = 0$.

(b) for any set of indices $\{\theta''\}$, the fact that $\mu_{\{\theta''\}} \neq 0$ implies that all lower order relatives of $\mu_{\{\theta''\}}$ are non zero.

If all μ-terms are present in a model, the model is said to be saturated, otherwise it is said to be unsaturated. As one always tries to use the smallest possible number of parameters without changing the goodness of fit seriously, focus will be on the unsaturated models. If for example $\mu_{AB(ij)} = 0 \ \forall i,j$ in a model with three variables A, B and C with levels I,J and K then A and B are independent conditional on C. This model could also be described as an [AC][BC] model, i.e. leaving out that μ-term which is absent.

3.18/...

In deciding on a model there are three basic objectives, namely parsimony with respect to the number of parameters, goodness of fit with respect to deviance of cell estimates from cell observations and interpretability.

To estimate the theoretical cell frequencies the most advocated technique is the method of maximum likelihood. The advantages are that the cell estimates will not be influenced by the fact whether the table was constructed using a Poisson sampling scheme, a product multinomial sampling scheme or a full multinomial sampling scheme on the condition that in the case of product multinomial sampling the μ-terms corresponding to fixed margins are included in the fitted model (Bishop et al).

The theoretical cell frequencies are easy to compute and are often simple functions of marginal totals. The estimates have favourable large sample properties and where closed form expressions for the maximum likelihood estimates are not available the method of iterative proportional fitting works well (see Bishop et al. or Fienberg (1980) for rules for detecting existence of direct (closed form) estimates and the application thereof.)

Bishop et al. discussed the G^2-statistic which is used to assess the goodness of fit in chapter 14 in detail with special reference to its asymptotic behaviour. The G^2-statistic is preferred to the well known Pearson statistic:

$$\chi^2 = \sum_i \frac{(x_i - m_i)^2}{m_i} \qquad 3.5.2.12$$

χ^2 is asymptotically chi-square distributed with degrees of freedom equal to the number of cells in the table minus the number of independent fitted parameters defined by the model.

3.19/...

The G^2-statistic is based on the likelihood ratio and possesses a useful partitioning property which can be applied in model building - a property not shared by the χ^2-statistic.

$$G^2 = 2\sum_i x_i \ln\frac{m_i}{x_i} \qquad 3.5.2.13$$

with degrees of freedom equal to the number of cells in the table minus the number of independent fitted parameters defined by the model. As an example assume the A, B and C variable model with IxJxK cells. Assume the model [AB][C], i.e.

$$\ln m_{ijk} = \mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{C(k)}+\mu_{AB(ij)} \qquad 3.5.2.14$$

The degrees of freedom = IJK(number of cells) - 1(μ) - (I-1) (number of independent parameters with respect to variable A) - (J-1)(with respect to variable B) - (K-1)(variable C)-(I-1)(J-1) (with respect to the association parameters which are independent)= (K-1)(IJ-1).

In order to partition the model, define

$$L(\underline{x}) = \sum_i x_i \ln x_i \qquad 3.5.2.15$$

$$L(\hat{\underline{m}}) = \sum_i x_i \ln \hat{m}_i \qquad 3.5.2.16$$

then G^2 can be found as

$$G^2 = -2(L(\hat{\underline{m}}) - L(\underline{x})) \qquad 3.5.2.17$$

Suppose two models are denoted by [1] and [2] and that all the μ-terms in [2] are contained in model [1], i.e. [2] is

3.20/...

nested within [1]. Let the parameter estimates for model [1] be denoted by $\left(\hat{\underline{m}}^{[1]}\right)$. Then the G^2-statistic for model [2] can be written as

$$G^2([2]) = -2\left(L(\hat{\underline{m}}^{[2]}) - L(\underline{x})\right)$$

$$= 2\left(L(\hat{\underline{m}}^{[1]}) - L(\hat{\underline{m}}^{[2]})\right) + 2\left(L(\underline{x}) - L(\hat{\underline{m}}^{[1]})\right)$$

$$= G^2([2]/[1]) + G^2([1]), \qquad 3.5.2.18$$

i.e. $G^2([2]) > G^2([1])$ by $G^2([2]/[1])$, the conditional likelihood ratio statistic for model [2] given by model [1]. Now $G^2([2])$, $G^2([1])$ and $G^2([2]/[1])$ are asymptotically distributed chi-square with respective degrees of freedom ν_2, ν_1 and $\nu_2-\nu_1$. This then gives us a way to determine whether a given μ-term must be added or ignored from the model. For a full description of this stepwise method refer Bishop et al.

A natural extension of the model is being used in discriminant analysis, namely the logit models in which interest centers upon the relationship of a set of explanatory variables on at least one responsive variable, i.e. the relationships between the explanatory variables as such are not as important as the nature of the effects on the response variable(s). The technique can be illustrated best by means of an example as in Berenson et al.

Assume a one response variable with two levels so that the log odds table can be constructed after a suitable log-linear model has been found. Say the response variable is C, where C has two levels. Suppose the best parsimonious model is given by

3.21/....

[AB][AC][BC], i.e. $\mu_{ABC(ijk)} = 0$, then

$$\ln m_{ijk} = \mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{C(k)}+\mu_{AB(ij)}+\mu_{BC(jk)}+\mu_{AC(ik)} \qquad 3.5.2.19$$

The logit model is then defined by

$$\text{logit}_{ij} = \ln\left(\frac{m_{ij1}}{m_{ij2}}\right) = 2(\mu_{C(1)}+\mu_{AC(i1)}+\mu_{BC(j1)}) \qquad 3.5.2.20$$

because

$$\mu_{C(1)} = -\mu_{C(2)} \; ; \quad \mu_{AC(i1)} = -\mu_{AC(i2)} \; ; \quad \mu_{BC(j1)} = -\mu_{BC(j2)} \qquad 3.5.2.21$$

Therefore it follows that

$$\text{logit}_{ij} = \ln\left(\frac{m_{ij1}}{m_{ij2}}\right) = w + w_i^{A\bar{C}} + w_j^{B\bar{C}} \qquad 3.5.2.22$$

with $w = 2\mu_{C(1)}$, $w_{(i)}^{A\bar{C}} = 2\mu_{AC(i1)}$ and $w_{(j)}^{B\bar{C}} = 2\mu_{BC(j1)}$ where the bar over C indicates the variable which the odds refer to.

Note that the estimated frequencies may be compared directly.

The odds can be converted to a proportion by using the transformation suggested by Berkson (1944) as

$$\hat{p}_{ij} = \frac{e^{\text{logit}_{ij}}}{1+e^{\text{logit}_{ij}}} = \frac{m_{ij1}/m_{ij2}}{1+m_{ij1}/m_{ij2}} \qquad 3.5.2.23$$

which will estimate the ratio of odds of level 1 of variable C

3.22/...

against level 2 of variable C. The discrimination rule therefore is:

Allocate to population 1, i.e. level 1 of C if $\hat{p}_{ij} \geq 0,5$

Allocate to population 2, i.e. level 2 of C if $\hat{p}_{ij} < 0,5$

with an arbitrary allocation in the case of equality.

It will be appreciated that a large amount of work and time is saved by these models when closed form estimators may be used and even when the iterative procedure must be followed, because according to Birch's theorem the μ-terms then do not have to be calculated (see Berenson et al. and Bishop et al.).

3.5.3 The application of the log-linear and logit models in discriminant analysis - a summary

The application is illustrated by means of an example as given in Berenson et al.

Assume a 2xJxK contingency table where the first variable, x_1 identifies the group, i.e. assign to population Π_1 if $x_1=1$ and to Π_2 when $x_1=2$. For the three variable model, x_i, i=1,2,3 the saturated model for the theoretical frequencies is

$$\ln m_{ijk} = \mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{C(k)}$$
$$+\mu_{AB(ij)}+\mu_{AC(ik)}+\mu_{BC(jk)}+\mu_{ABC(ijk)} \qquad 3.5.3.1$$

Assuming equal prior probabilities and a sample based partition $D = (D_1, D_2)$ against the optimal partition $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2)$ to define a sample based logit rule:

Assign a response characterised by $\underline{x} = (x_2, x_3) = (j,k)$ as follows

3.23/...

$$\ln\left(\frac{m_{1jk}}{m_{2jk}}\right) > 0 \quad \text{then } \underline{x}\in\Pi_1$$
$$< 0 \quad \text{then } \underline{x}\in\Pi_2 \qquad 3.5.3.2$$
$$= 0 \quad \text{randomly}$$

Note that in terms of a log-linear representation, the full multinomial rule is equivalent to using a completely saturated design.

Discrimination may also be made on ground of the w values in the model as in the previous section.

The procedure as described above should be followed after a suitable, more parsimonious model has been constructed, thereby diminishing the number of parameters considerably.

If there are more than two populations to be considered, say p populations then variable x_1 has p levels, i.e. $\underline{x}_1 = (x_1)$ and $\underline{x}_2' = (x_2, \ldots, x_k)$, where every x_i, $i = 2, \ldots, k$ can take on s_i different categorical values and vector $\underline{x}_2'$ can be classified as belonging to Π_j when $\ln\left(\frac{m_{j2\ldots k}}{m_{\cdot j2\ldots k}}\right)$ is a maximum for all $s = 1, \ldots, p;\ s\neq j;\ j=1, \ldots, p,$

$$\ln\frac{m_{j2\ldots k}}{m_{\cdot 2\ldots k}} = \max_{\substack{s=1,\ldots,p\\ i=1,\ldots,p\\ s\neq i}}\left\{\ln\left(\frac{m_{i2\ldots k}}{m_{s2\ldots k}}\right)\right\}. \qquad 3.5.3.3$$

3.6 Discrimination using a distribution free procedure based on the estimation of probability distributions using the kernel function method

3.6.1 Introduction

Welch (1939) showed that if we want to classify an object that

3.24/...

comes from one of two populations with associated densities f_1 and f_2 then a classification rule should be based upon the likelihood ratio f_1/f_2. Under the assumption of sampling from normal populations with equal covariance matrices or even different covariance matrices the procedures are fairly easy. Studies have shown that these procedures perform reasonably well for some non-normal populations, but there is interest in approaching the problem in a less constrained way.

The question is whether it is possible to estimate the likelihood ratio f_1/f_2 without estimating f_1 and f_2. It is true however, that much more is known about density estimation than likelihood ratio estimation. The most direct way to proceed therefore is to estimate the individual densities f_1 and f_2 and then the likelihood ratio f_1/f_2.

Wahba (1975) dealt with the optimal convergence properties of various classes of nonparametric density estimates and the advantages of these estimates in the classification problem. Lachenbruch and Goldstein (1979) summarised these results:

Let f_1 and f_2 be the two underlying continuous densities from subpopulations Π_1 and Π_2, and p_1 and p_2 are the two prior probabilities. If $D = (D_1, D_2)$ is a partition of the sample space X such that $x \in D_j$ implies that x belongs to Π_j for any $x \in X$, then the probability of correct classification for the partition of rule D is given by

$$r(D) = \sum_{j=1}^{2} \int_{D_j} p_j f_j(x)\,dx \qquad 3.6.1.1$$

If $\mathcal{D}$ is the set of all possible classification rules then D^* achieves the optimal probability of correct classification if

$$r(D^*) = \sup_{D \in \mathcal{D}} r(D) = r^* \qquad 3.6.1.2$$

where according to Welch D^* is defined by

3.25/...

$$D_j^* = \{x/j \text{ is the smallest integer such that } p_j f_j(x) = \max_{j=1,2}(p_j f_j(x))\} \qquad 3.6.1.3$$

If a random sample of size n is available from the mixture of Π_1 and Π_2, say Π and it is determined that n_j of the observations are from Π_j we estimate p_j by $\hat{p}_j = {}^{n_j}/_{n}$ (refer 3.2.1).

Denote the partition in 3.6.1.3 by $\hat{D}_j$ and an estimate of $f_j(x)$ by $\hat{f}_j(x)$ at the point x, then with

$$\hat{r}(D) = \sum_{j=1}^{2} \int_{D_j} \hat{p}_j f_j(x)dx \qquad 3.6.1.3$$

it can be shown that

$$\hat{r}(\hat{D}) = \sup_{D \in \mathcal{D}} \hat{r}(D) \qquad 3.6.1.4$$

Glick (1972) showed that if the density estimates are consistent and provided that $\int \sum_j \hat{p}_j \hat{f}_j(x)dx$ converges to unity or is bounded by a finite constant then $\sup |\hat{r}(D) - r(D)| \to 0$, $r(\hat{D}) \to r^*$ and $\hat{r}(\hat{D}) \to r^*$, which are very desirable error-rate convergence properties. Note that no restrictive conditions relating to the underlying family of distributions generating the data are assumed.

There are many density estimators, inter alia the kernel estimates, the orthogonal series estimates, the Pearsonian system, the Gram-Charlier approach and the maximum likelihood procedure proposed by Fisher, references of which may be found in Wegman (1972).

This section gives a summary of the properties of the kernel

3.26/...

functions as well as the application of these estimates in discriminant analysis. Given the a priori class probabilities and cost constants, if not equal, the usual procedure is followed to find the classification rule using the kernel function method to find the a priori class conditional densities.

3.6.2 The kernel function

Parzen (1962) showed in his paper how one may construct a family of estimates of f(x) and the mode such that the estimates are consistent and asymptotically normal. The problem of estimating the mode of a probability density function is comparable to the problem of maximum likelihood estimation of a parameter.

According to Everitt and Hand (1981) the estimation of a probability density function may be looked upon as one way of approximating a mixture of distributions by a single density by increasing the number of parameters. The introduction of more parameters would be expected to yield a better approximation than a mixture distribution. The number of parameters is increased to the number of sample points in the mixture. In application to discriminant analysis where one may have two or more distinct classes, each population will be associated with a unique density.

Parzen uses a smooting parameters λ(or 2λ) which is dependent on n, i.e. $\lambda = \lambda(n)$, where n is the number of sample observations under discussion. Then $f_n(x)$ can be estimated by

$$f_n(x) = [F_n(x+\lambda) - F_n(x-\lambda)]/(2\lambda) \qquad 3.6.2.1$$

As n increases, λ will decrease.

3.27/...

This gives rise to the estimation of a probability using a mixture of densities so that when $\underline{x}_{i1}, \ldots, \underline{x}_{in_i}$ form a sample from a multivariate population Π_i, then the distribution of $\underline{y}$ in population Π_i can be estimated by

$$P(\underline{y}/\Pi_i, D) = P(\underline{y}/D_i, \lambda_i)$$

$$= \frac{1}{n_i} \sum_{j=1}^{n_i} K(\underline{y}/\underline{x}_{ij}, \lambda_i) \qquad 3.6.2.8$$

where D_i is the subset of D which contains data from population Π_i.

Parzen used the following example to illustrate the method in the case of a univariate distribution. (See also Van der Merwe and Kotze, 1980).

Assume n_i sample elements from population Π_i, i=1, ..., k. The sample elements from Π_i are the vectors $\underline{x}_{i1}, \ldots, \underline{x}_{in_i}$. We shall use for our example only the 1st element of each vector, i.e. $x_{i11}, x_{i21}, \ldots, x_{in_i1}$. Let the number of 1st elements which fall in the interval $(y-\lambda, y+\lambda)$ be n.

We estimate $P(\underline{y}/\Pi_i, D)$ as follows

$$P(\underline{y}/\Pi_i, D) = \frac{\text{number in interval}}{\text{total number}} \cdot \frac{1}{\text{length of interval}}$$

$$= \frac{n}{n_i \cdot 2\lambda}$$

$$= \frac{1}{n_i \lambda} \cdot \sum_{j=1}^{n_i} K\left(\frac{y-x_{ij1}}{\lambda}\right) \qquad 3.6.2.9$$

where

3.29/...

$$K(t) = \tfrac{1}{2} \text{ for } |t| \leqslant 1 \text{ i.e. } |y-x_{ijl}| \leqslant \lambda$$

$$= 0 \text{ for } |t| > 1 \text{ i.e. } |y-x_{ijl}| > \lambda \qquad 3.6.2.10$$

Here we can see why the kernel function is also known as the window function, because of the "window" $[y-\lambda;\ y+\lambda]$.

Therefore each observation has a contribution to make to the $\hat{P}(\underline{y}/\Pi_i, D)$. This contribution is determined by K(t) as well as the value of λ where λ determines the interval size.

Meisel (1972) gave the ideal characteristics of the kernel function as follows:

(i) The mode of the function must be at $\underline{x}_{ij}=\underline{y}$.

(ii) The contribution of $\underline{x}_{ij}$ through the kernel function must decrease as the distance between $\underline{x}_{ij}$ and $\underline{y}$ increases, i.e. the contribution must be approximately zero when $\underline{x}_{ij}$ is far from $\underline{y}$.

(iii) Equivalent differences between $\underline{x}_{im}$, $\underline{x}_{in}$, $m \neq n$ respectively and $\underline{y}$ must result in equal contributions throught the kernel function.

Murthy (1966) showed that p one-dimensional kernel functions can be expressed as a single p-dimensional kernel function and vice versa so that

$$K(\underline{x}) = K(x_1, \ldots, x_p)$$

$$= K_1(x_1) \cdot K_2(x_2) \ldots K(x_p) \qquad 3.6.2.11$$

Parzen showed that the kernel function must satisfy the following conditions in order to give estimates which are stable and asymptotically normal.

3.30/...

From 3.6.2.10

$$K(\underline{y}/\underline{x}_{ij}, \lambda) = \prod_{t=1}^{k} K_t(y_t/x_{ijt}, \lambda)$$

$$= \prod_{t=1}^{k} \lambda^{z_t}(1-\lambda)^{1-z_t} \qquad 3.6.3.4$$

where $z_t = 0$ when $x_{ijt} \neq y_t$

$= 1$ when $x_{ijt} = y_t$ and $\frac{1}{2} \leqslant y \leqslant 1$ so that when each binary variable has its own smoothing parameter λ_t equation 3.4.3.3 transforms to

$$K(\underline{y}/\underline{x}_{ij}, \underline{\lambda}) = \prod_{t=1}^{k} \lambda_t^{z_t}(1-\lambda_t)^{1-z_t} \qquad 3.6.3.5$$

where z_t is as in 3.6.3.4.

Note in 3.6.3.3 that when $\lambda=1$ then the density is estimated by means of the relative frequency, i.e.

$$K(\underline{y}/\underline{x}_{ij}, 1) = 1 \qquad \underline{y} = \underline{x}_{ij}$$
$$= 0 \qquad \underline{y} \neq \underline{x}_{ij} \qquad 3.6.3.6$$

and when $\lambda = \frac{1}{2}$ the density is estimated by the uniform distribution, i.e.

$$K(\underline{y}/\underline{x}_{ij}, \tfrac{1}{2}) = (\tfrac{1}{2})^k, \quad \underline{y} \text{ any vector} \qquad 3.6.3.7$$

(c) (i) When the data is multivariate categorical with more than 2 categories, Aitchison and Aitken distinguished between nominally scaled and ordinally scaled data. They suggest the following kernel function for the former (see also Van der Merwe and Kotze, 1980).

3.33/...

$$K(\underline{y}/\underline{x}_{ij}, \lambda) = \prod_{t=1}^{k} K_t(y_t/x_{ijt}, \lambda)$$

$$= \prod_{t=1}^{k} \lambda^{z_t} \left(\frac{1-\lambda}{c_t-1}\right)^{1-z_t} \qquad 3.6.3.8$$

where $z_t = 0$ when $x_{ijt} \neq y_t$

$= 1$ when $x_{ijt} = y_t$ and c_t is the number of categories, i.e. 0, 1, ..., c_t-1 in which y_t is partitioned.

(ii) For categorical data which are ordinally scaled, the nearness of y_t to x_{ijt} is taken into account by means of a weighting system. The following kernels are suggested:

When $c_t = 4$: $K_t(y_t/x_{ijt}, \lambda)$ for ordinal variables with 3 categories:

	$y_t = 0$	$= 1$	$= 2$
$x_{ijt} = 0$	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$
$= 1$	$\frac{1}{2}(1-\lambda^2)$	λ^2	$\frac{1}{2}(1-\lambda^2)$
$= 2$	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2

When $c_t = 4$: $K_t(y_t/x_{ijt}, \lambda)$ for ordinal variables with 4 categories:

	$y_t = 0$	$= 1$	$= 2$	$= 3$
$x_{ijt} = 0$	λ^3	$3\lambda^2(1-\lambda)$	$3\lambda^2(1-\lambda)$	$1-\lambda^3$
$= 1$	$\frac{\lambda(1-\lambda^3)}{1+\lambda}$	λ^3	$\frac{\lambda(1-\lambda^3)}{1+\lambda}$	$\frac{(1-\lambda)(1-\lambda^3)}{1+\lambda}$
$= 2$	$\frac{(1-\lambda)(1-\lambda^3)}{1+\lambda}$	$\frac{\lambda(1-\lambda^3)}{1+\lambda}$	λ^3	$\frac{\lambda(1-\lambda^3)}{1+\lambda}$
$= 3$	$(1-\lambda)^3$	$3\lambda(1-\lambda)^2$	$3\lambda^2(1-\lambda)$	λ^3

3.34/...

(d) The kernel function method makes it possible to handle the problem of discrimination when there is a population with mixed binary and continuous features. Kotze and Van der Merwe expanded on the reference given by Aitchison and Aitken by giving an example.

Let each observation vector consist of k measurements of which k_1 are of the binary type and k_2 measurements are continuous variables so that $k_1+k_2 = k$, i.e. $\underline{x}_{ij} = (x_{ij1}, x_{ij2}, \cdots, x_{ijk_1}, x_{ijk_1+1}, x_{ijk_1+2}, \cdots, x_{ijk})$.

Let $\underline{y} = (\underline{u}, \underline{v})$ be a typical vector from the space $B^{k_1} \times R^{k_2}$ then

$$\hat{P}(\underline{u}, \underline{v}/\Pi_i, D) = \hat{P}(\underline{u}, \underline{v}/D_i, \lambda_i, \delta_i)$$

$$= \frac{1}{n_i} \sum_{j=1}^{n_i} K(\underline{u}/\underline{x}_{ij}, \lambda).L(\underline{v}/\underline{x}_{ij}, \delta) \qquad 3.6.3.9$$

Where K is the cubical binomial density function with smoothing parameter λ and L the spherical normal distribution with smoothing parameter δ. Note that K as well as L may be the results of products of "variable-type" kernel functions themselves. Note further that factorisation in kernel functions does not imply independency - e.g. 3.6.3.9 does not imply independency between the binary and continuous variables.

In a similar way other types of variables can be included and taken care of by expanding on the product characteristic of the kernel function.

There are several techniques available for estimation of the smoothing parameter(s).

Aitchison and Aitken proposed a jackknife likelihood method to determine λ (or λ_t, δ etc.), i.e. maximise $V(\lambda_i/D_i)$ using λ_i as variable in

3.35/...

$$V(\lambda_i/D_i) = \prod_{j=1}^{n_i} P(\underline{x}_{ij}/D_i, \lambda_i) \qquad 3.6.3.10$$

They found however, in the case of the cubical binomial kernel function that λ_i turns out to be $\lambda_i = 1$, while $\lambda_i = 0$ for all n_i in the case of the multivariate normal kernel function, which gives rise to a "point" parameter. Habbema, Hermans and Van der Broek (1974) proposed a variation of the jackknife likelihood technique which Van der Merwe and Kotze showed as follows; The proposed function is

$$W(\lambda_i/D_i) = \prod_{j=1}^{n_i} P(\underline{x}_{ij}/D_i - \underline{x}_{ij}, \lambda_i) \qquad 3.6.3.11$$

$$= \prod_{j=1}^{n_i} \left(\frac{1}{n_i - 1} \sum_{r=1}^{n_i} \frac{1}{\lambda_i^k} K\left(\frac{\underline{x}_{ij} - \underline{x}_{ir}}{\lambda_i}\right) \right) \qquad 3.6.3.12$$

where W must be maximised. $D_i - \underline{x}_{ij}$ denotes the set D_i with the single vector $\underline{x}_{ij}$ excluded.

Aitchison and Aitken showed that $\hat{\lambda}_i$ converges in probability to 1 and that $P(\underline{y}/D_i, \hat{\lambda}_i)$ converges in probability to $P(\underline{y}/\Pi_i, D_i)$ as $n \to \infty$.

For more examples of estimation methods for the smoothing parameter refer to Van Ness (1979).

It is of interest to note that it was the general consensus of Van Ness and Simpson (1976) and Van Ness (1979) that the choice of the kernel function itself is not so crucial, but that the choice of the smoothing parameter is critical. The latter must be adapted to the covariances. Van Ness used several values for λ and chose that one which gave the best classification rate. This technique however is heavy on computer time. They found that the kernel function method compares very well

3.36/...

with all other discrimination techniques, especially in the case of a small number of observations on many variables, or where there are variables with large variances, e.g. discrete variables with many categories. It is further noteworthy that new data with new combinations of the variable may be classified on grounds of the existing data bank.

3.7 The location model

Krzanowski (1975) proposed a model for the classification of mixed data where the observation vectors contain binary as well as continuous variables.

Let the observation vector be $\underline{w}' = (\underline{x}', \underline{y}')$ where $\underline{x}' = (x_1, \ldots, x_q)$ is the vector of q binary variables and $\underline{y}' = (y_1, \ldots, y_p)$ is the vector of p continuous variables. Express the q binary variables as a multinomial $\underline{z}' = (z_1, \ldots, z_k)$ where $k = 2^q$ so that each distinct pattern of $\underline{x}$ defines a multinomial cell uniquely. An observation $(x_1, \ldots, x_q)$ will be an element in cell $c = 1 + \sum_{i=1}^{q} x_i 2^{i-1}$

Assume a further k cells and that $\underline{y}$ is multivariate normally distributed with equal dispersion matrix Σ in all cells and mean $\underline{\mu}^{(m)}$ in cell m, where $m = 1, \ldots, k$. Let the probability of obtaining an observation in cell m be p_m (see also Olkin and Tate, 1975). In the case of a two group discrimination problem the model is generalised to multivariate means $\underline{\mu}_i^{(m)}$ and probabilities p_{im} for $i = 1, 2$ with common Σ for all cells of both populations. From this we can summarise that

$$(\underline{y}/z_m = 1,\ z_j = 0,\ j \neq m) \sim n(p, \underline{\mu}_i^{(m)}, \Sigma) \qquad 3.7.1$$

is the conditional distribution of $\underline{y}$ given $\underline{x}$.

The optimum allocation rule, given all population parameters can now be derived.

3.3.7/...

Let $P_i(\underline{w})$ be the probability density of $\underline{w}$ in Π_i, i = 1,2. If costs and a priori probabilities are assumed to be equal, then $\underline{w}$ is assigned to Π_1 if $P_1(\underline{w})/P_2(\underline{w}) > 1$ and otherwise to Π_2. But

$$P_i(\underline{w}) = P_i(\underline{x}, \underline{y}) = P_i(\underline{x}) \cdot P_i(\underline{y}/\underline{x}) \qquad i=1,2 \qquad 3.7.2$$

Suppose that the observation $\underline{x}$ falls in cell m, then it follows that

$$P_i(\underline{x}, \underline{y}) = p_{im} \cdot P_i(\underline{y}/z_m), \qquad 3.7.3$$

i.e. $p_{im} = P_i(\underline{x})$ for $\underline{x}$ in cell m and z_m the corresponding cell in the $\underline{z}$ vector.

From 3.7.1 allocate to Π_1 if

$$(\underline{\mu}_1^{(m)} - \underline{\mu}_2^{(m)})\Sigma^{-1}\left\{\underline{y} - \tfrac{1}{2}(\underline{\mu}_1^{(m)} + \underline{\mu}_2^{(m)})\right\} \geq \ln(p_{2m}/p_{1m}) \qquad 3.7.4$$

and otherwise to Π_2 with $m = 1 + \sum_{i=1}^{q} x_i 2^{i-1}$. It follows that there is a discriminant function for each of the multinomial cells where the discrete components of the model determine the cut off point in each case.

In order to determine the misclassification probabilities P(1/2) and P(2/1) the Mahalanobis distance squared,

$$D_m^2 = (\underline{\mu}_1^{(m)} - \underline{\mu}_2^{(m)})'\Sigma^{-1}(\underline{\mu}_1^{(m_1)} - \underline{\mu}_2^{(m_2)}) \qquad 3.7.5$$

is applied, i.e. the distance squared between Π_1 and Π_2 conditional on the observation falling in multinomial cell m. If the overall probability of misclassification from Π_i is taken as the sum of the probabilities of misclassification for each

3.38/...

multinomial cell of Π_i weighted by the probability of its occurence, we find that

$$P(2/1) = \sum_{m=1}^{k} p_{1m} \Phi\left[(\ln \frac{p_{2m}}{p_{1m}} - \tfrac{1}{2}D_m^2)/D_m \right] \qquad 3.7.6$$

and

$$P(1/2) = \sum_{m=1}^{k} p_{2m} \Phi\left[(\ln \frac{p_{1m}}{p_{2m}} - \tfrac{1}{2}D_m^2)/D_m \right] \qquad 3.7.7$$

where $\Phi(x)$ is the cumulative standard normal distribution function.

When the parameters are not known, they must be estimated from the initial samples of sizes n_1 and n_2 respectively.

Let n_{1m} and n_{2m} denote the number of observations falling in cell m of the q-way contingency table containing $k = 2^q$ cells each for Π_1 and Π_2. Let $\underline{Y}_{ji}^{(m)}$ denote the vector of continuous variables associated with the j th observation in cell m of the sample from Π_i. Then, if

$$\bar{\underline{Y}}_i^{(m)} = \frac{1}{n_{im}} \sum_{j=1}^{n_{im}} \underline{Y}_{ji}^{(m)} \qquad 3.7.8$$

the maximum likelihood estimates of the population parameters p_{im}, $\underline{\mu}_i^{(m)}$ and Σ are given by

$$\hat{p}_{im} = \frac{n_{im}}{n_i} \ , \ \hat{\underline{\mu}}_i^{(m)} = \bar{\underline{Y}}_i^{(m)}$$

$$V = \frac{1}{n_1+n_2-2k} \sum_{i=1}^{2} \sum_{m=1}^{k} \sum_{j=1}^{n_{im}} (\underline{Y}_{ji}^{(m)} - \bar{\underline{Y}}_i^{(m)})(\underline{Y}_{ji}^{(m)} - \bar{\underline{Y}}_i^{(m)})' \qquad 3.7.9$$

3.39/...

This procedure however, may lead to a problem regarding the size of n_{im} which may in some cases be very small if n_1 and n_2 are not large relative to k, the number of cells and this may give rise to poor parameter estimates. For this Krzanowski gave some approximations of which one is a second order log-linear model as an approximation for the p_{im}. For further detail refer to Goldstein and Dillon (1978).

Krzanowski found that the location model gives better results than Fisher's LDF when there is evidence of interactions between binary variables and between populations. The usefulness of the location model seems to diminish as the number of binary variables is more than approximately 7 in random sample cases of $n_1=n_2=50$. He has however, not done any extensive research regarding the advantages of the location model.

3.8 Gatekeeping analysis for completely nominal data and some other approaches

This simple hierarchical thresholding analysis is often easy to apply with relatively good results (Montgomery, 1975 and Kendall, 1963).

(i) Search for a variable and for a value of that variable which will enable us to reach a classification decision for all or a part of our sample while making very few errors. In effect we seek a variable and a value of that variable above or below, if the data is ordinal in nature, which there is little or no overlap in the sample distributions from the two (or more) populations. This search may be made on the basis of prior logic and/or heuristic methods.

(ii) Remove from the data base those observations which we are ready to classify.

3.40/...

(iii) Return to step 1 with the remainder of the data. Repeat until sample sizes become very small, or no variable can be found which will achieve the objectives, or you are satisfied with the classification success.

One advantage of this approach is that it may be used on relatively small samples. Unfortunately it cannot be applied directly to continuous data, unless categorised. Further it is not possible to find theoretical distributions for the misclassification rates. Refer also Kendall (1963) and Montgomery for examples as well as Kendall, Stuart and Ord (1981) for further comment and examples.

Further approaches for fully nominal data may be found in Gitlow (1979). He discussed the dummy variable multiple discriminant function (MDF) and the multivariate nominal scaled analysis (MNA). He found that MNA gave better results than MDF in the data he used in his analysis. These two techniques, which are relatively new are easy to apply and deserve more attention in the future.

The MDF is often used when the dependent as well as the independent variables are all nominally scaled. The model is

$$y_{kl} = b_{lo} + \sum_{i=1}^{p} \sum_{j=1}^{c_i-1} b_{ijl} x_{ijk} , \quad l=1, \ldots, G \qquad 3.8.1$$

where i, j and l refer to variable, category and group respectively.

G equals the number of groups.
b_{lo} is the constant term for the lth group.
y_{kl} indicates the group membership of observation k, i.e. if observation k is a member of group 2 ($l=2$) then $y_{kl}=2$.
p is the number of original independent variables.
c_{i-1} is the number of dummy variables needed to represent the c_i categories of the jth original independent variable.

3.41/...

$b_{ij\ell}$ is the MDF coefficent for the dummy variable formed from the j th category of the i th original independent variable for group ℓ.

x_{ijk} = 1 if observation k shows the typical characteristics of the j th category of the i th variable

= 0 otherwise

In the case of more than two groups where the dependent variable is of a nominal nature the MNA procedure is appropriate if the independent variables are nominally scaled and an additive model is suitable to represent the data set. The model is:

$$y_{k\ell} = \bar{y}_{\ell} + \sum_{i=1}^{p} \sum_{j=1}^{c_i} a_{ij\ell} x_{ijk} \qquad \ell=1, \ldots, G \qquad 3.8.2$$

where as before i, j and ℓ refer to variable, category and group respectively.

G equals the number of groups.

$y_{k\ell}$=1 if observation k is a member of group ℓ

=0 otherwise.

$\bar{y}_{\ell}$ is the percentage of observations in group ℓ.

p is the number of independent variables.

$a_{ij\ell}$ is the MNA coefficient which shows the percentage deviation from the overall percentage of observations in group ℓ caused by the j th category of the i th independent variable - see Andrews and Messinger (1973) for the derivation of $a_{ij\ell}$.

x_{ijk}=1 if observations k shows the characteristic represented by category j of the i th independent variable.

=0 otherwise.

MDF has a distribution theoretical basis whereas MNA has no theoretical basis. MNA on the other hand does not need homogeneity

3.42/...

of variance to perform well and the results are easier to interpret. It is known that MDF is inappropriate for many data structures, e.g. if group covariance matrices are not equal, whereas it is not known for what data structures MNA may be inappropriate.

Gitlow found in his study that for the data he used the MNA profile performed much better than the dummy variable MDF profile using a hold out discriminatory power test.

The MNA profile showed less confusion among misclassified observations than the dummy MDF profile. Press's $\chi^2(Q)$ statistic using hold out samples showed that the MNA's predictive ability is far superior to that of the dummy variable MDF's ability.

In general it was found that the MNA profile is more representative of the true profile than is the case with the dummy variable MDF profile.

As MNA is a relatively new and untested procedure Gitlow advocated more attention for this method which may give very satisfactory results when discriminating in nominally scaled data sets.

3.9 Discriminant Analysis based on ranks - A technique which gives one some control over the rate of misclassification

Randles, Broffitt, Ramburg and Hogg (1978) and Broffitt, Randles and Hogg (1976) described a "distribution-free" partial discriminant analysis technique which has a very desirable property, namely the ability to keep the probabilities of misclassification in the case of two sample discrimination to more or less an equal degree. In addition, the region of uncertainty - or "not classification" can be diminished and specific ratios of P_1 to P_2 can be determined.

Let $\underline{x}_1, \ldots, \underline{x}_{n_1}$ and $\underline{y}_1, \ldots, \underline{y}_{n_2}$ be independent random samples from two p-variate populations Π_1 and Π_2 respectively. Another random observation say $\underline{z}$ will generally be classified by means of Fisher's LDF say

$$D_L(\underline{z}) = \left[\underline{z} - \tfrac{1}{2}(\bar{\underline{x}}+\bar{\underline{y}})\right]' S^{-1}(\bar{\underline{x}}-\bar{\underline{y}}) \qquad 3.9.1$$

with decision rule: Classify

$$\underline{z}\in\Pi_1 \quad \text{when} \quad D_L(\underline{z}) \geqslant c$$
$$\underline{z}\in\Pi_2 \quad \text{when} \quad D_L(\underline{z}) < c \qquad 3.9.2$$

so that the probabilities of misclassification are

$$P_1 = P(D_L(\underline{z}) < c/\underline{z}\in\Pi_1)$$

$$P_2 = P(D_L(\underline{z}) \geqslant c/\underline{z}\in\Pi_2) \qquad 3.9.3$$

A large overlap between Π_1 and Π_2 and/or a $\underline{z}$ near to c may cause uncertainty whether to classify or not, and if - where to classify.

The object is to minimise this uncertainty and this can be obtained by fixing P_1 and P_2 within certain limits. With respect to the latter it should be noted that Fisher's LDF and QDF do not result in a minimum of the average $\frac{(P_1+P_2)}{2}$ or in a more or less equal P_i, i=1,2, because of deviations from normality and differences in variances.

This method will enable us not only to fix $P_1 \doteq P_2$, but even to fix desirable ratios if we would like to, e.g. $P_1 \doteq k.P_2$.

We define A_i as the occurence or event which favours $\underline{z}\in\Pi_i$ and

3.44/...

define $\underline{x}_i$, i=1, ..., n, and $\underline{y}_i$, i=1, ..., n_2 as before. Assume that $\underline{z} \in \Pi_1$ so that the two samples are $\underline{x}_1$, ..., $\underline{x}_{n_1}$, $\underline{x}_{n_1+1}$ and $\underline{y}_1$, ..., $\underline{y}_{n_2}$ where $\underline{x}_{n+1} = \underline{z}$. Determine a discriminant function $D_{\underline{x}}(\cdot)$ based on these two final samples in such a way that $D_{\underline{x}}(\cdot)$ will be positive if the observation is from Π_1 and negative if otherwise. Any function which is symmetrical with respect to $\underline{x}_i$ as well as $\underline{y}_i$ is acceptable, e.g. Fisher's linear discriminant function or quadratic discriminant function as well as others may be used.

The symmetrical discriminant function therefore must satisfy the following condition; let

$$D_{\underline{x}}(\cdot) = D_{\underline{x}}(\cdot/\underline{x}_1, \ldots, \underline{x}_{n_1}, \underline{x}_{n_1+1}; \underline{y}_1, \ldots, \underline{y}_{n_2}) \qquad 3.9.4$$

Let $i_1, \ldots, i_{n_1+1}$ and $j_1, \ldots, j_{n_2}$ be arbitrary permutations of the integers 1, ..., n_1+1 and 1, ..., n_2, then $D_{\underline{x}}(\cdot)$ must satisfy

$$D_{\underline{x}}(\cdot/\underline{x}_{i_1}, \ldots, \underline{x}_{i_{n_1+1}}; \underline{y}_{j_1}, \ldots, \underline{y}_{j_{n_2}})$$

$$= D_{\underline{x}}(\cdot/\underline{x}_1, \ldots, \underline{x}_{n_1+1}; \underline{y}_1, \ldots, \underline{y}_{n_2}) \qquad 3.9.5$$

Let $R_{\underline{x}}(\underline{z}) = R_{\underline{x}}(\underline{x}_{n_1+1})$ be the rank of $D_{\underline{x}}(\underline{z})$ among $D_{\underline{x}}(\underline{x}_1)$, ..., $D_{\underline{x}}(\underline{x}_{n_1})$, $D_{\underline{x}}(\underline{x}_{n_1+1})$, ranking from smallest to largest. As $R_{\underline{x}}(\underline{z})$ becomes bigger it seems more likely that $\underline{z}$ may be an "$\underline{x}$" rather than a "$\underline{y}$" in terms of its $D_{\underline{x}}(\cdot)$ value. Now if $\underline{z} \in \Pi_1$ then $R_{\underline{x}}(\underline{x}_i)$, i=1, ..., n_1+1 and $R_{\underline{x}}(\underline{z})$ in particular have the uniform distribution over the integers 1, ..., n_1+1 as long as the distribution of $D_{\underline{x}}(\underline{z})$ is continuous. This result is however independent of the distributions of Π_1 or Π_2. For a proof of or

3.45/...

rather, a heuristic argument for this statement, see Randles et al.

Let $\alpha_1 = \frac{a_1}{n_1+1}$, a_1 being an integer and define event A_1 such that

$$A_1 : R_{\underline{x}}(\underline{z}) > a_1 \qquad 3.9.6$$

i.e. classify $\underline{z}$ as an observation from Π_1.

In cases where Π_1 and Π_2 overlap substantially

$$\begin{aligned} P_1 &= P(\text{classify in } \Pi_2 / \underline{z} \in \Pi_1) \\ &= P(\bar{A}_1 / \underline{z} \in \Pi_1) \\ &= P(R_{\underline{x}}(\underline{z}) \leq a_1) \\ &= \frac{a_1}{n_1+1} \qquad 3.9.7 \\ &= \alpha_1 \end{aligned}$$

To specify A_2 assume that $\underline{z}$ is from Π_2, so that the sample is $\underline{y}_1, \ldots, \underline{y}_{n_2}, \underline{y}_{n_2+1}$ where $\underline{y}_{n_2+1} = \underline{z}$. Compute $D_{\underline{y}}(\cdot)$ using samples of sizes n_1 and n_2+1 and determine $R_{\underline{y}}(\underline{z}) = R_{\underline{y}}(\underline{y}_{n_2+1})$ as the rank of $-D_{\underline{y}}(\underline{y}_{n_2+1})$ among the n_2+1 values $-D_{\underline{y}}(\underline{y}_1)$, $-D_{\underline{y}}(\underline{y}_2)$, ..., $-D_{\underline{y}}(\underline{y}_{n_2+1})$. Let $\alpha_2 = \frac{a_2}{n_2+1}$ with a_2 an integer so that A_2 can be defined as

$$A_2 : R_{\underline{y}}(\underline{z}) > a_2 \qquad 3.9.8$$

i.e. classify $\underline{z}$ as an observation from Π_2, and again if Π_1 and Π_2 overlap substantially, then

3.46/...

$$P_2 = P(\text{Classify in } \Pi_1 / \underline{z} \in \Pi_2)$$

$$= P(\bar{A}_2 / \underline{z} \in \Pi_2)$$

$$= P(R_{\underline{y}}(\underline{z}) \leq a_2) \qquad 3.9.9$$

$$= \frac{a_2}{n_2+1}$$

$$= \alpha_2$$

If $R_{\underline{x}}(\underline{z}) \leq a_1$ and $R_{\underline{y}}(\underline{z}) > a_2$ do not classify or use another approach.

Note that $D_{\underline{x}}(\cdot)$ and $D_{\underline{y}}(\cdot)$ are different, but in the case of n_1, n_2 large the difference may become insignificantly small. To ignore the difference however, will spoil the distribution free nature of the approach, although we will still have $P_1 \approx \alpha_1$ and $P_2 \approx \alpha_2$.

If there is not a substantial overlap a forced procedure may be followed, i.e. if

$$P_{\underline{x}}(\underline{z}) > P_{\underline{y}}(\underline{z}) \quad \text{classify } \underline{z} \text{ as from } \Pi_1$$

$$P_{\underline{x}}(\underline{z}) < P_{\underline{y}}(\underline{z}) \quad \text{classify } \underline{z} \text{ as from } \Pi_2 \qquad 3.9.10$$

where $P_{\underline{x}}(z) = \frac{R_{\underline{x}}(\underline{z})}{n_1+1}$ and $P_{\underline{y}}(\underline{z}) = \frac{R_{\underline{y}}(\underline{z})}{n_2+1}$

In this case the level of misclassification probabilities is not controlled, but Broffitt et al. stated that preliminary investigations indicated that they are approximately equal. In the case of equality use another non-ranking technique.

3.47/...

The probabilities in 3.9.10 are valid on the condition that $D_{\underline{x}}(\underline{z})$, $D_{\underline{y}}(\underline{z})$ have a continuous distribution for Π_1 when $\underline{z}\in\Pi_1$ and for Π_2 when $\underline{z}\in\Pi_2$. Note that $D_{\underline{x}}(\cdot)$ and $D_{\underline{y}}(\cdot)$ do not even have to be of the same form. The choice of the forms of the discriminant functions should be made after having examined the data if possible.

Randles et al. showed heuristically that as n_1, n_2 increase the uniform distribution of $P_{\underline{x}}(\underline{z})$ over $(\frac{1}{n_1+1}, \frac{2}{n_1+1}, \ldots, 1)$ converges to the asymptotic distribution of $P_{\underline{x}}(\underline{z})$, i.e. the uniform distribution over [0, 1]. A similar statement applies to $P_{\underline{y}}(\underline{z})$.

They further show that $P_{\underline{x}}(\underline{z})$ and $P_{\underline{y}}(\underline{z})$ are inversely related as $\underline{z}$ changes.

Monte Carlo studies by Broffitt et all. and Randles et al. showed that not only is it possible to control α_1 and α_2, but it is also possible to keep $(P_1+P_2)/2$, i.e. the average misclassification probability comparatively small. The results which are given in these two articles are quite impressive.[1)]

3.10 Scoring discrete data for application of the LDF analysis

As noted earlier it is possible to apply the LDF to discrete data if the variables are of an ordered type. The problem however, is to decide on the method of scoring. Dillon and Western (1982) found that if scoring is done carefully and the levels of the variables increases, a simple dummy variable coding of the raw

1) (a) Note, when determining the functions $D_{\underline{x}}(\cdot)$ and $D_{\underline{y}}(\cdot)$ initially, $\underline{z}$ may be omitted completely, thereby saving computer time.
(b) Refer also the R-adapt method for possible quadratic forms in Randles et al.

frequency data followed by a LDF analysis does as well as the discrete discriminant procedure.

Dillon and Western found that reversals in ratios among groups lead to poor results when the LDF is applied, because of the non-monotomic behaviour of the likelihood ratio. The scoring which one applies therefore, must dampen the effects of reversals. A further requirement is that the numerical scores must be applied in such a way that the resulting mean difference vectors are ordinal and finally the distance between the respective groups must be as large as possible.

The scoring methods discussed are as follows:

(i) K-level coding, called the K-method: The categories are from 1 to K if the variable has K levels.

(ii) Dummy variable coding, called the Dummy-code method: For each variable having K categories construct K-1 dummy variables.

(iii) The Info-Gain method: Let $P^{i}_{j(k)}$ = P [respondent belongs to group i, i = 1,2, in the kth category of variable j]. The scoring is done as follows

$$s_{j(k)} = \ln(P^{(1)}_{j(k)}/P^{(2)}_{j(k)}) \qquad 3.10.1$$

This scoring method gives the kth category of a variable a value equal to the amount of information it furnishes for discrimination in favour of group 1. (See also Gokhale and Kulback, 1978).

(iv) The Diff-PP method based on differences in posterior probabilities: Let $P^{(1/j(k))}$ and $P^{(2/j(k))}$ denote the respective posterior probabilities in each group for the kth category of the jth variable, then

$$s_{j(k)} = P^{(1/j(k))} - P^{(2/j(k))}$$

3.49/...

$$= \frac{P_{j(k)}^{(1)} - P^{(2)}_{(j(k)}}{P_{j(k)}^{(1)} + P_{j(k)}^{(2)}} \qquad 3.10.2$$

for the kth category of the jth variable.

(v) The Diff-DS method based on differences in discriminant scores. If q_i, i = 1,2, are the prior group probabilities and j indicates a particular response pattern, i.e., j = (x_1 = k; x_2 = ℓ ; ...; x_p = m) and $qP^{(1)}_{k\ell\ldots m}$; $q_2P^{(2)}_{k\ell\ldots m}$ are the discriminant scores for response patterns (kℓ...m) where the P's are the respective cell probabilities, assign the score

$$S_j = q_1P^{(1)}_{k\ell\ldots m} - q_2P^{(2)}_{k\ell\ldots m} \qquad 3.10.3$$

to pattern j. Here a combination of data is summarised in one score.

Bahadur's method was used for simulating multivariate frequency data as described in Bahadur (1961). The number of variables as well as levels were changed during the study.

Dillon and Western come to the following conclusions:

(i) In the case of binary predictors use discrete procedures unless the group means are very different and the variables are definitely uncorrelated. The latter point is of the utmost importance.

(ii) Procedures (ii), (iii) and (iv) performed on approximately the same level, but with (iii) and (iv) slightly better than (ii). Procedure (iii) however is so easy to apply that it is recommended unless the number of variables and/or levels become large.

(iii) Procedures (i) and (v) should be ignored.

(iv) The full multinomial rule which was also used in this study performed poorly, because of sparse cells.

3.50/...

It is worthwhile noting that although the K-level coding method did poorly in this study it deserves further attention for the case where the number of categories are 10 or more as this may occur frequently in marketing applications.

Chapter 4

CORRESPONDENCE ANALYSIS AND DISCRIMINANT ANALYSIS

4.1 Introduction

Correspondence analysis is a fairly recent technique where it is equivalent to a number of techniques derived in different contexts since the mid 1930's. (See Geenacre, 1981 and Greenacre 1984). Some of the techniques involved are reciprocal averaging (Hill, 1973), simultaneous linear regression (Hirschfeld, 1935) dual scaling (Nishisato, 1980), a special case of canonical correlation analysis etc.

The analysis is primarily a technique for displaying the rows and columns of a two-way contingency table as points in corresponding low-dimensional vector spaces. These spaces may be superimposed for a joint display. Note that the analysis is basically suitable for nonnegative data.

Whereas correspondence analysis may be described as a special case of canonical correlation analysis where the latter is the study of linear relationships between two sets of variables, say Y and Z which tries to maximise the correlations between the linear combinations of Y and Z we may show that discriminant analysis is a special case of this problem.

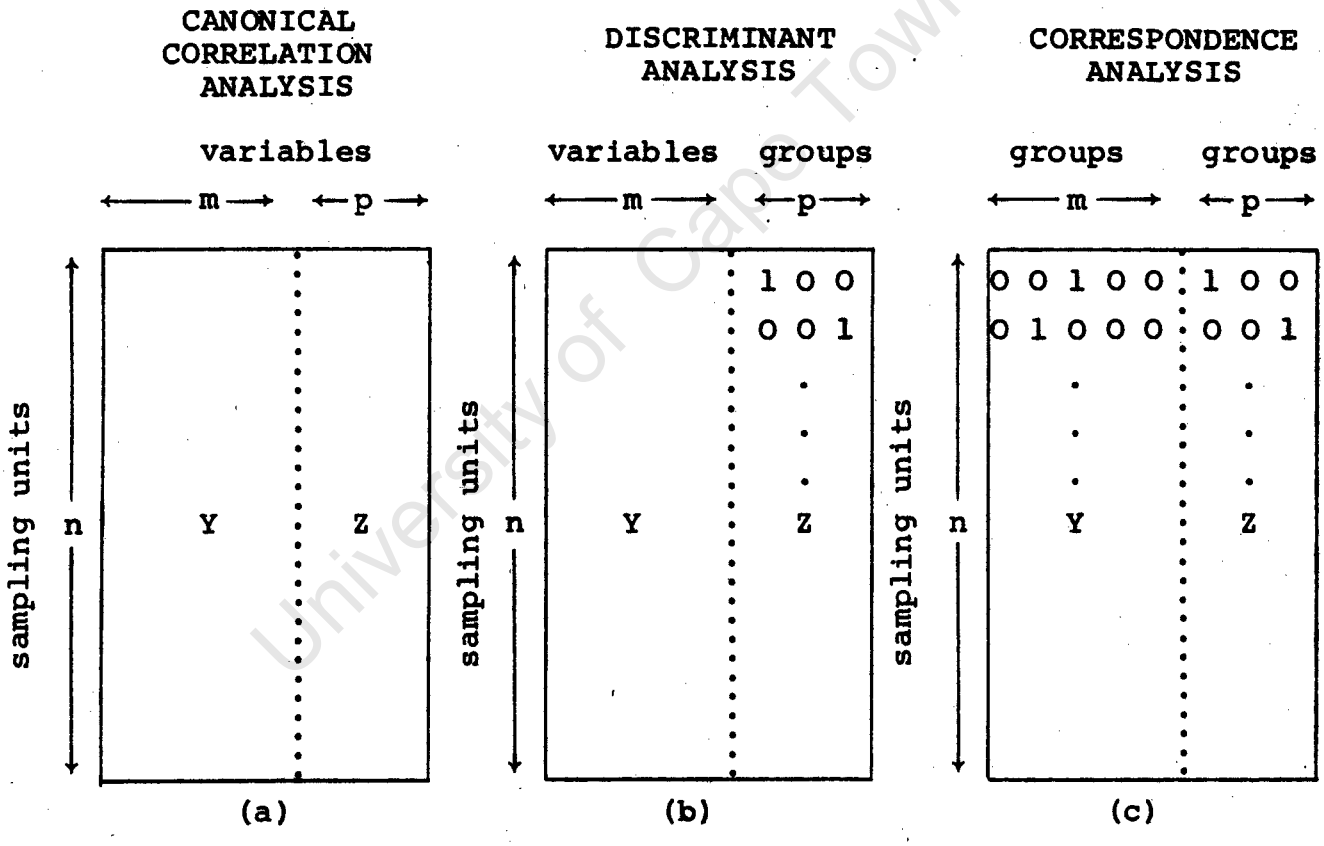
For the theoretical background refer to the appendix section G. An intuitive description can be found in Greenacre (1981).

Discriminant analysis consists of the investigation into the variables and patterns of observations to see how groups in a given partition of say I subjects can be characterised and separated, with the possibility of some more unclassified objects which must be classified. Thus, we have a set of variables say Z, which is a logical matrix indicating the groups to which the sampling units belong. The rows of Z contain zeros apart

4.1/...

from a 1 in the appropriate column to indicate the group membership with the result that Z forms a single qualitative variable defining a partition of the sampling units into groups. The canonical correlations would measure the ability of the variables in Y to linearly discriminate between these groups.

When Y is a logical matrix as well, then correspondence analysis investigates the dependence of two partitions of the same sampling units and the process may be called double discriminant analysis. Diagrammatically it can be shown as follows:



Discriminant analysis (b) and correspondence analysis (c) as special cases of canonical correlation analysis (a). In each case the objective of the analysis can be described as maximising the correlation of linear combinations $\underset{\sim}{Y}\underset{\sim}{u}$ and $\underset{\sim}{Z}\underset{\sim}{v}$ of the 2 sets of columns (Greenacre, 1981).

Figure 4.1.1

4.2/...

The analogy with canonical correlation analysis gives a restrictive view of the problem. The geometric approach of the French school (see for example, Benzécri, 1973) on the other hand gives a much broader view of correspondence analysis.

4.2 Geometric definition of correspondence analysis

According to Greenacre (1981) correspondence analysis can be described by defining firstly a cloud of points in a multidimensional vector space, secondly the metric structure on this space and thirdly the fit of this cloud to a variable low dimensional subspace onto which the points are projected for display and interpretation. Keep also in mind that there are two problems involved: The display of a cloud of points representing the rows of a contingency table and similarly, the columns.

Divide the rows by the respective row totals and determine the average relative row profile from the column totals. This average row profile is the weighted average of the row profiles where each row profile is weighted proportionally to the respective row sum in the original data matrix; this can also be seen as a centre of gravity. The row masses are defined as the row totals divided by the grand total.

The metric in the space of profiles is a general Euclidian metric, i.e. each squared difference in coordinates divided by the respective element of the average row profile may be used. In the latter case it's known as the chi-squared metric. If the row profiles are denoted by $\underline{a}_i'$ and the diagonal matrix by $D_{\underline{c}}$ where $\underline{c}'$ is the average row profile, then this metric measures between $\underline{a}_i'$ and $\underline{a}_j'$ by

$$d^2(\underline{a}_i , \underline{a}_j) = (\underline{a}_i - \underline{a}_j)' \, D_{\underline{c}}^{-1} \, (\underline{a}_i - \underline{a}_j) \qquad 4.2.1$$

4.3/...

The problem now is to find the p-dimensional subspace which is closest to all the points. The measure of closeness is defined as the weighted sum of squared distances from the points to the subspace, where the weights are once again the row masses.

If r_i denotes the mass of the ith point and z_i the distance of this point to the subspace, then we must find the subspace that minimises $\sum_i r_i z_i^2$ - refer the sketch.

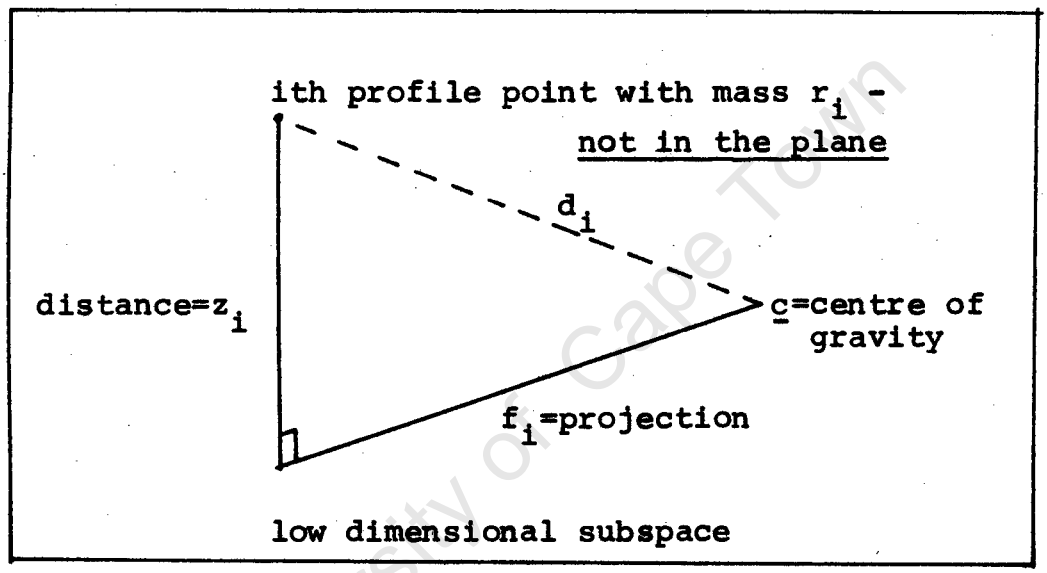


Figure 4.2.1

Because the triangle formed by the centre of gravity, the point and its projection has a fixed-length hypotenuse d_i, the minimum of $\sum_i r_i z_i^2$ correspondes to the maximum of $\sum_i r_i f_i^2$ where f_i is the distance of the ith point's projection from the centre of gravity.

The line along which $\Sigma r_i f_i^2$, i.e. the moment of inertia is maximised is called the principal axis of inertia, i.e. in m-dimensional space

4.4/...

the subspace will have m-1 orthogonal principal axes of inertia and moments of inertia. The total inertia of the cloud of points is then $\Sigma r_i d_i^2$, i.e. the chi-squared statistic devided by the total of the matrix. Geometrically, inertia may be thought of as weighted dispersion.

The principal axes solution is contained in an eigen-equation, where the eigen values are $\lambda_1 \geq \ldots \geq \lambda_{m-1} > 0$ are the moments of inertia with eigen vectors $\underline{u}_1, \ldots, \underline{u}_{m-1}$ being the principal axes of inertia. The axes form an orthonormal basis in the space of row profiles, using the previously defined metric D_c^{-1}, i.e. $\underline{u}_i' D_c^{-1} \underline{u}_j = \delta_{ij}$. The quality of the display is gauged by the moments of inertia expressed as percentages τ of the total inertia; i.e.

$$\tau = 100 \times \lambda_\alpha / \sum_{k=1}^{m-1} \lambda_k \qquad (\alpha=1, \ldots, m-1) \qquad 4.2.2$$

The display will now be an approximate representation of the points in a space of reduced dimension.

Each point's contribution to the moment of inertia of the axes can be examined closer by looking at the terms of $\Sigma r_i f_i^2$. Usually these are expressed as percentages. The angles between point vectors and the principal axes will also give more insight. The sum of the squared cosines of angles of a point over the complete set of orthogonal principal axes equals 1. If therefore, $\cos^2 \theta$ is high, θ must be low and the vector can be said to lie in the direction of the axis, or correlate with the axis.

The examination of the contribution of the profile points to each axis as well as the angles between the point vectors and the axes aid a great deal to the interpretation of the basic graphical display.

The above-mentioned argument with respect to row profiles applies

4.5/...

in a symmetrical fashion to the column profiles, i.e. a cloud of column profiles with masses in n-dimensional space as well as a generalised Euclidian matrix $D_{\underline{r}}^{-1}$, $\underline{r}$ being the average column profile or the vector of row masses previously defined, are defined. The subspaces of closest fit are then obtained by identifying the principal axes of inertia of the cloud of column points.

It is the duality in the last paragraph where this technique obtains its name from, i.e. correspondence analysis. There is even a correspondence of the positions of the row and column points on the axes.

If $\underline{f}$ contains the coordinates of a row profile point along the first principal axis with moment of inertia λ and $\underline{b}$ contains the coordinates of a column profile, then the coordinates of the latter point with respect to the first principal axis in its space is $\underline{b}'\underline{f}/\sqrt{\lambda}$. Now, the elements of $\underline{b}$ add to 1 so that $\underline{b}'\underline{f}$ is the weighted average (centre of gravity) of the coordinates of the row profiles.

Let A and B denote the matrices of row and column profiles respectively. Let $\underline{f}$ be the vector which contains the coordinates of a point with respect to the first principal axis, with moment of interia λ. The following transition formulae can be proved (see Appendix A and E) for any column vector $\underline{g}$.

$$A'B'\underline{f} = \lambda\underline{f} \qquad 4.2.3$$

$$B'A'\underline{g} = \lambda\underline{g} \qquad 4.2.4$$

The solutions for $\underline{f}$ and $\underline{g}$ simultaneously are equivalent to the solutions of a basic structure problem (see Appendix A).

In summary the following have been mentioned:

The inertias and their decompositions along the principal axes are identical in the two clouds of points. In the respective subspaces row points are attracted to regions of column points for which the row profiles are large, and vice versa. Accordingly

4.6/...

we can merge the displays into one and represent the row and column points on the same principal axes.

Note that the positions of all the row profiles collectively determine the position of each column profile point and vice versa. As a result of this individual row profiles and column profiles can not be compared on a distance basis on the simultaneous display.

The axes themselves must be interpreted in terms of the percentages of inertias along successive axes in the cases of two-way contingency data and rank order data.

4.3 Correspondence analysis and discriminant analysis

Previously correspondence analysis was described as a form of double discriminant analysis. Let us extend the Y matrix to represent a number of qualitative variables Y_1, Y_2, ..., Y_s and Z as the matrix of final classification, where Y_i is an n x m_i matrix, m_i being the dimension of variable i so that Y is an n x m matrix where $m = m_1 + \ldots + m_s$. The correspondence between the classification Z and the variables in Y is condensed into the contingency table Z'Y where Z'Y represents the totals of the rows of Y for each of the p groups, i.e. Z'Y is a p x m matrix. (Greenacre, 1981).

The correspondence analysis of this matrix reveals now graphically the correspondence between the groups and the set of predictor variables. In order to read this display more easily it is useful to link up the points representing the categories of each variable as shown in the sketch below in matrix form

4.7/...

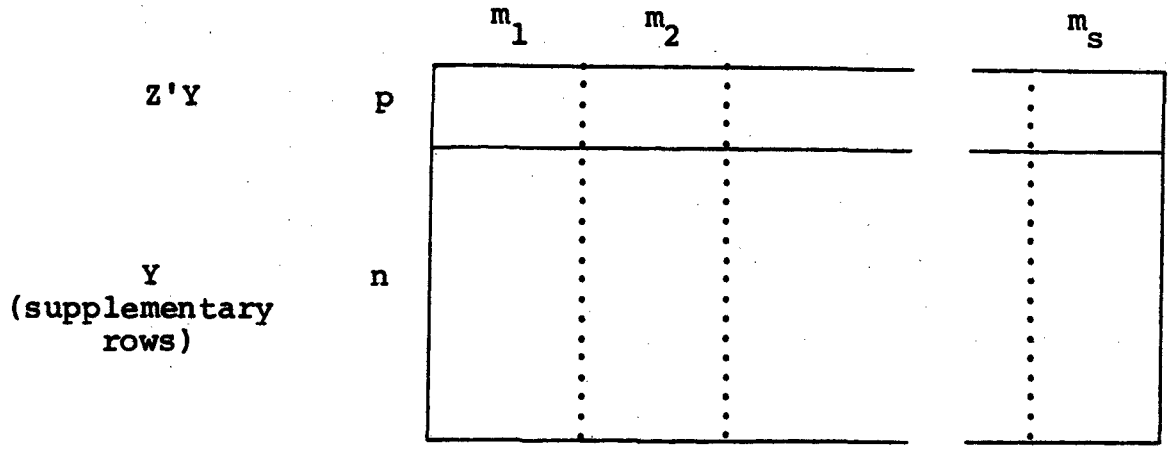


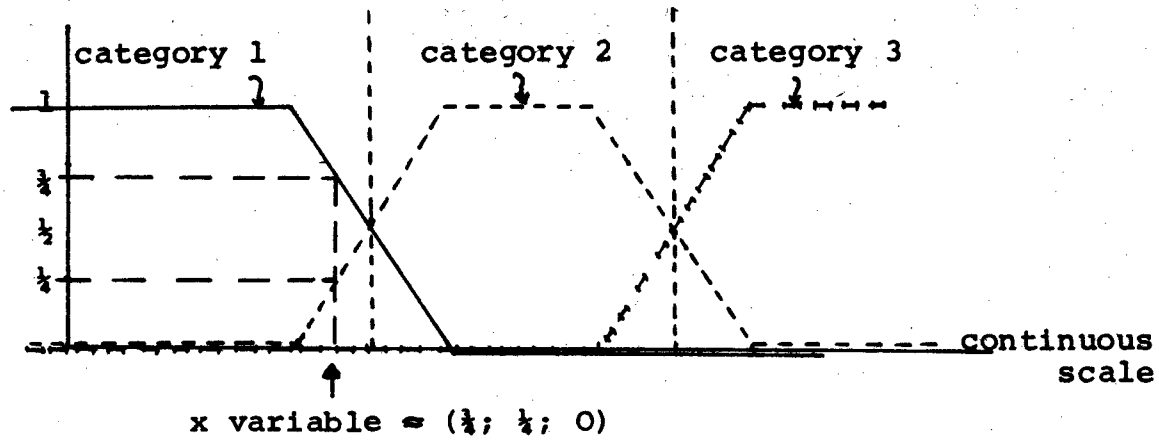
Fig. 4.3.1

This now permits the construction of a classification rule for new observations. Given a new observation, form the logical vector of categorised data, then determine its position with respect to the principal axes, using the transition formula. Plot this point in the discriminant space and identify the neighbouring points which determine its classification.

The number and radius of nearest, neighbouring points are also of interest and may influence the classification. For further detail see Greenacre (1981).

The use of correspondence analysis to determine a discriminant space can also be classified as a non-parametric method applicable to continuous as well as discrete data, although there will obviously be some loss of information involved in replacing a value on a continuous scale by its corresponding category. To attenuate for this latter problem a number of solutions are available, one being the coding scheme where e.g. a continuous variable must be recorded as one of 3 categories. To show that the variable in the sketch is almost in category 2, use a generalised qualitative variable (¼, ¾, 0).

4.8/...



Chapter 5

MORE ASPECTS WORTHY OF CONSIDERATION

5.1 Ill-conditioned data matrices and/or dependent variables

5.1.1 Introduction

In practice one will often find that measurements are highly correlated. This will obviously play a large role in the results obtained from an analysis, especially in the linear discriminant or quadratic discriminant function techniques of Fisher, for the covariance matrices will tend to be ill-conditioned.

Looking at the LDF $(\bar{x}_1-\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2)$ it is obvious that the inverted var-cov. matrix must be of such a nature that it is not ill-conditioned in any way and certainly not singular. Even if the observations are measured accurately, the minimum mean error squared of coefficients in the discriminant function can not be guaranteed, because of rounding off errors, intercorrelation of the independent variables which may give rise to ill-conditioned situations as well as a too large a number of variables which may lead to "dummy" intercorrelation.

As sample sizes are often small, biasing could probably be used whenever classification problems arise. Biasing most greatly affects the variables corresponding to the smallest eigenvalues of Σ. However, Σ is usually unknown and the effectiveness of biasing will have to be determined from the samples.

Before one starts off on biasing or other techniques to correct for intercorrelation etc, one must first try to determine whether there really is a problem. Several authors (see Forsythe and Maler, 1967; Marshall and Olkin, 1968; Vinoid, 1976) proposed the use of the "condition number" to measure the instability of a matrix when solving for a system of linear equations (Troskie, 1980). This is usually defined as

$$C_Q = Q(A).Q(A^{-1}) \qquad 5.1.1.1$$

5.1/...

where Q is usually taken as the norm. In the linear model the condition number for A (i.e. X'X) is

$$C_Q = C(X'X) = \frac{\lambda_1}{\lambda_p} \qquad 5.1.1.2$$

where $\lambda_1 > \ldots > \lambda_p$, the characteristic roots of X'X.

The condition number is a better measure of the nearness to singularity than the determinant of A. According to Belsey, Kuh and Welsch (1980) a condition number of more than 10 indicates weak dependencies. A condition number of 15-30 has an associated correlation coefficent of 0,9 and a condition number of 100 and more shows serious problems in the solutions. According to Troskie condition numbers shouldn't be much larger than 10p.

5.1.2 The effect of biasing

The application of bias to discriminant analysis was discussed inter alia by DiPillo (1976). His discussion was based on a reduction in variance and the effect of that on the probability of misclassification. The following will be based on the same frame work but instead of a reduction on the variance of the classification rule by biasing S directly, a bias is introduced on the correlation matrix.

Let $\underline{x}' = (x_1, \ldots, x_p)$ be an observation from one of two p-variate normal populations, say $\Pi_i \sim n(\underline{\mu}_i, p, \Sigma)$, i=1,2. Assume that $\Sigma_1 = \Sigma_2 = \Sigma$. The pooled sample var-cov. matrix is S, and R is the sample correlation matrix, i.e.

$$S = D^{\frac{1}{2}} R D^{\frac{1}{2}} \qquad 5.1.2.1$$

and the equation of interest is

$$S^* = D^{\frac{1}{2}} (R+kI) D^{\frac{1}{2}} \qquad 5.1.2.2$$

5.2/...

This way the elements of the main diagonal are not only increased directly, but the ratio between main-diagonal and off-diagonal elements are increased in the matrix of intra-correlations.

DiPillo called the common LDF the sample-based minimum χ^2-rule and modified S by replacing S by S+kI in

$$Q_j(\underline{x}) = (\underline{x}-\bar{\underline{x}}_j)'S^{-1}(\underline{x}-\bar{\underline{x}}_j), \quad j = 1,2 \qquad 5.1.2.3$$

with rule: If $\min_j[Q_j(\underline{x})] = Q_i(\underline{x})$, classify $\underline{x}$ into Π_i.

From 2.1

$$Q_j(\underline{x}) = (\underline{x}-\bar{\underline{x}}_j)'(D^{\frac{1}{2}}RD^{\frac{1}{2}})^{-1}(\underline{x}-\bar{\underline{x}}_j) \qquad 5.1.2.4$$

We will replace R now by R+kI, i.e. S by $D^{*\frac{1}{2}}(R+kI)D^{\frac{1}{2}}$,

$$\begin{aligned} Q_j^*(\underline{x}) &= (\underline{x}-\bar{\underline{x}}_j)'(D^{\frac{1}{2}}(R+kI)D^{\frac{1}{2}})^{-1}(\underline{x}-\bar{\underline{x}}_j) \\ &= (\underline{x}-\bar{\underline{x}}_j)'(D^{\frac{1}{2}}RD^{\frac{1}{2}} + kD)^{-1}(\underline{x}-\bar{\underline{x}}_j) \\ &= (\underline{x}-\bar{\underline{x}}_j)'(S+kD)^{-1}(\underline{x}-\bar{\underline{x}}_j) \qquad 5.1.2.5 \end{aligned}$$

Let $G(\underline{x}) = Q_1(\underline{x}) - Q_2(\underline{x})$, and if $G(\underline{x}) \geq 0$ classify in population 1, otherwise in population 2. Define $G^*(\underline{x})$ in a similar way, i.e.

$$G(\underline{x}) = (\bar{\underline{x}}_2-\bar{\underline{x}}_1)'S^{-1}\underline{x} - \tfrac{1}{2}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'S^{-1}(\bar{\underline{x}}_2+\bar{\underline{x}}_1) \qquad 5.1.2.6$$

and

$$G^*(\underline{x}) = (\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-1}x - \tfrac{1}{2}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-1}(\bar{\underline{x}}_2+\bar{\underline{x}}_1)' \qquad 5.1.2.7$$

Given that $\underline{x}$ is from Π_1 it can be shown that $D^*(\underline{x})$ is normally distributed with

$$5.1.2.8$$

$$E[G^*(\underline{x})/\bar{\underline{x}}_1,\bar{\underline{x}}_2,S,\Pi_1] = (\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-1}\underline{\mu}_1 - \tfrac{1}{2}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)(S+kD)^{-1}(\bar{\underline{x}}_2+\bar{\underline{x}}_1)$$

5.1.2.8

and

$$\text{var}[G^*(\underline{x})/\bar{\underline{x}}_1,\bar{\underline{x}}_2,S,\Pi_1] = (\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-1}\Sigma(S+kD)^{-1}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)$$

5.1.2.9

It can be shown now that

$$\text{var}[G(\underline{x})/\underline{\mathbb{X}}_1,\underline{\mathbb{X}}_2,S,\Pi_1] > \text{var}[G^*(\underline{x})/\underline{\mathbb{X}}_1,\underline{\mathbb{X}}_2,S,\Pi_1], k>0 \qquad 5.1.2.10$$

From 5.1.2.10 it follows now that the probability of misclassification may decrease when k>0.

The shift in location from $G(\underline{x})$ to $G^*(\underline{x})$ may however cause inconsistent results. This biasedness may be overcome consistently by the reduction in variance of the classification rule.

The probability of misclassification in the case where all population parameters are known is $\Phi(\frac{-\delta}{2})$ with

$$\delta^2 = (\underline{\mu}_2 - \underline{\mu}_1)'\Sigma^{-1}(\underline{\mu}_2 - \underline{\mu}_1) \qquad 5.1.2.11$$

In the case where only sample estimates are available the total probability of misclassification (PMC) is

$$\tfrac{1}{2}(1 - \Phi(z_1) + \Phi(z_2)) \qquad 5.1.2.12$$

where

$$z_i = \frac{\tfrac{1}{2}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'S^{-1}(\bar{\underline{x}}_2+\bar{\underline{x}}_1) - (\bar{\underline{x}}_2-\bar{\underline{x}}_1)'\underline{\mu}_i}{[(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'S^{-1}\Sigma S^{-1}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)]^{\frac{1}{2}}} \; ; \; i=1,2 \qquad 5.1.2.13$$

Under the biased correlation coefficient, let the biased probability of misclassification (PMC*) be

5.4/...

$$\tfrac{1}{2}(1 - \Phi(z_1^*) + \Phi(z_2^*)) \qquad 5.1.2.14$$

where for $k>0$

$$z_i^* = \frac{\tfrac{1}{2}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'\left[D^{\frac{1}{2}}(R+kI)D^{\frac{1}{2}}\right]^{-1}(\bar{\underline{x}}_2+\bar{\underline{x}}_1)-(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'\left[D^{\frac{1}{2}}(R+kI))D^{\frac{1}{2}}\right]^{-1}\underline{\mu}_i}{\left[(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'[D^{\frac{1}{2}}(R+kI)D^{\frac{1}{2}}]^{-1}\ \Sigma[D^{\frac{1}{2}}(R+kI)D^{\frac{1}{2}}]^{-1}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)\right]^{\frac{1}{2}}} \quad ; \ i=1,2$$

5.1.2.15

DiPillo used the sign of the rate of change of PMC* to determine whether PMC* increases or decreases for a given $k>0$. So

$$\frac{d}{dk}(PMC^*) = -\frac{e^{-\frac{1}{2}(z_1^*)^2}}{\sqrt{2\pi}}\frac{dz_1^*}{dk} + \frac{e^{-\frac{1}{2}(z_2^*)^2}}{\sqrt{2\pi}}\frac{dz_2^*}{dk} \qquad 5.1.2.16$$

with

$$dz_i^* = \frac{\tfrac{1}{2}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-1}(\bar{\underline{x}}_2+\bar{\underline{x}}_1-2\underline{\mu}_i).(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-2}\Sigma(S+kD)^{-1}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)}{\left[(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-1}\Sigma(S+kD)^{-1}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)\right]^{\frac{3}{2}}}$$

$$- \frac{\tfrac{1}{2}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)(S+kD)^{-2}(\bar{\underline{x}}_2+\bar{\underline{x}}_1-2\underline{\mu}_i)}{\left[(\bar{\underline{x}}_2-\bar{\underline{x}}_1)'(S+kD)^{-1}\Sigma(S+kD)^{-1}(\bar{\underline{x}}_2-\bar{\underline{x}}_1)\right]^{\frac{1}{2}}}, \quad i=1,2 \qquad 5.1.2.17$$

DiPillo found a marked increase in efficiency using computer runs on his biased procedure S+kI, not so much when the sample sizes are large, but especially when n is small and or the number of variables increases.

It should be illuminating to apply the biased procedure S+kD to the same data sets DiPillo used.

5.5/...

The optimum value of k is difficult to determine. It should obviously be where $\frac{d(PMC^*)}{dk}$ equals zero - not an easy equation to solve. Other methods depend on the maximisation of

$$\Delta_*^2 = (\underline{\bar{x}}_1-\underline{\bar{x}}_2)'(S+kD)^{-1}(\underline{\bar{x}}_1-\underline{\bar{x}}_2) = \underline{z}'\Lambda_*^{-1}\underline{z} = \sum_{i=1}^{P} z_i^2/(\lambda_i+ks_{ii}) = \sum_{i=1}^{P} c_i^2$$

5.1.2.18

as well as examination of reversal in direction of individual coefficients in the discriminant function.

A last remark is in order here namely that DiPillo (1979) was investigating the procedure as indicated above, using S+kD in stead of S+kI, but no results has appeared as far as the author is aware of.

5.2 Outlier detection in discriminant analysis

5.2.1 The outlier problem in general

The problem of outlier detection in univariate statistics has been studied quite thoroughly.

In the case of multivariate data it is not so easy to see an outlier physically as an outstanding "point" as in univariate statistics. Outliers can be of several types and some can be classified as follows. Hawkins (1980):

Assume the p-component sample vectors $\underline{x}_1, \ldots, \underline{x}_n$ from the multivariate normal distribution : $n(p, \underline{\mu}, \Sigma)$, i.e.

$$H_o : \underline{x}_i \sim n(p, \underline{\mu}, \Sigma) \quad , \quad i=1 \text{ to } n. \qquad 5.2.1.1$$

Without losing out on generality we assume that after having

5.6/...

permuted the observations the first k observations are outliers in one or other sense with respect to H_o, i.e.

$$\underline{x}_i \sim n(p, \underline{\mu}, \Sigma), \quad i = k+1 \text{ to } n \qquad 5.2.1.2$$

while $\underline{x}_i$, $i = 1$ to k differ from H_o. Some distributions that the outliers may follow are probably multivariate normal as follows

$$H_{1a} : \underline{x}_i \sim n(p, \underline{\mu}_i, \Sigma), \quad i = 1 \text{ to } k$$

where

$$H_{1b} : \underline{\mu}_i = \mu_2 = \ldots = \underline{\mu}_k \neq \underline{\mu} \qquad 5.2.1.3$$

$$H_{2a} : \underline{x}_i \sim n(p, \underline{\mu}, a_i\Sigma), \quad i = 1 \text{ to } k$$

where

$$H_{2b} : a_1 = a_2 = \ldots = a_k \neq 1 \qquad 5.2.1.4$$

$$H_{3a} : \underline{x}_i \sim n(p, \underline{\mu}, \Sigma_i), \quad i = 1 \text{ to } k$$

where

$$H_{3b} : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k \neq \Sigma \qquad 5.2.1.5$$

Assuming that the (b) part of the hypothesis is valid in each case the problem breaks down from a (k+1) group to a 2 group problem, although the third hypothesis implies extra difficulties which has had the result that it hasn't been studied as thoroughly as the other alternatives.

Note that outliers are values with high probabilities of occurring where the probability density of the true distribution is low and remote from the main body.

5.7/...

In a test for a single outlier, say $\underline{x}_j$, the following procedure may be followed if an H_1 case is expected:

$$\bar{\underline{x}} = \frac{1}{n}\sum_{i=1}^{n}\underline{x}_i, A = \sum_{i=1}^{n}(\underline{x}_i-\bar{\underline{x}})(\underline{x}_i-\bar{\underline{x}})', \quad S = \frac{1}{n-1}A$$

$$\bar{\underline{x}}_j = \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n}\underline{x}_i, \quad A_j = \sum_{\substack{i=1\\ i\neq j}}^{n}(\underline{x}_i-\bar{\underline{x}}_j)(\underline{x}_i-\bar{\underline{x}}_j)', \quad S_j = \frac{1}{n-2}A_j \qquad 5.2.1.6$$

Use the T^2 test or variants thereof to test for H_o against H_1, i.e.

$$T_j^2 = (\underline{x}_j-\bar{\underline{x}}_j)A_j^{-1}(\underline{x}_j-\bar{\underline{x}}_j)\cdot\frac{(n-1)(n+\nu-2)}{n} \qquad 5.2.1.7$$

where ν is the degrees of freedom associated with Σ. Now determine $T^2_{max} = \max_j\{T_j^2\}$ as test statistic.

Hawkins showed further that H_2 for a single outlier can also be tested using $T^2_{max} = \max_j\{T_j^2\}$.

H_1 and H_2 for the multiple outlier case are tested by using a (k+1) group one-way multivariate analysis of variance (MANOVA) approach. The test procedures include Wilks's lambda, Roy's largest-root criterion and Hotellings $T_o^2 = \sum_{i=1}^{P}\lambda_i$. Of these Wilks's Λ is recommended,

$$\Lambda = \prod_{i=1}^{P}(1+\lambda_i)^{-1} \qquad 5.2.1.8$$

H_o is rejected for small values of Λ. If it is not known a priori which of the n observations are outliers, one computes the test statistic for all $M = \frac{n}{k\ (n-k)}$ partitions of the data into k outliers and n-k "inliers". The most extreme of these M values is used as the outlier statistic and the corresponding partition is used to identify outliers. If Λ_k denotes the value

5.8/...

of Λ for a k-outlier model then the test for significance must be evaluated at the α/M level of the H_o distribution. (Wilks, 1963). Note that Λ_1 and Λ_2 cause no real problems, but more care is necessary for $k>2$.

Hawkins gave some alternative approaches, e.g. the use of the principal component residuals. He also pointed out that scaled principal residuals corresponding to the large γ_i, i.e. the large eigenvalue of Σ are the worst possible linear functions of $\underline{x}$ for detecting outliers on a single component and suggest their dismissal.

5.2.2 The influence function as an aid in outlier detection in discriminant analysis

As we can see from the preceding section one would intuitively try the same approach in discriminant analysis for suspect observations, but with respect to its "own" group distribution only, i.e. determine a statistic based on the group as a whole and also with the suspect observation removed. Evaluate the influence of this procedure on the statistic. From this procedure follows the name of the evaluation statistic, viz. the influence function. In discriminant analysis there is a variety of statistics which one can use in the influence function, e.g. Mahalanobis's Δ^2, a function of the coefficient vector of the discriminant function, the group means and probably more.

Campbell (1978) gave a number of references in this respect and then carried on to apply the influence function as an aid to outlier detection in discriminant analysis.

He distinguished between the theoretical and sample influence functions. Empirically it focuses on $\theta-\theta_{-m}$ where θ is an estimator based on n observations and θ_{-m} is an estimator, similar to θ, but determined without the mth observation. The influence function is then given by

5.9/...

$$I_m(\underline{x}, \theta) = \lim_{\varepsilon \to o} \frac{\hat{\theta}_{-m} - \theta}{\varepsilon} \qquad 5.2.2.1$$

where ε is taken as $-\frac{1}{n-1}$, i.e.

$$I_m(\underline{x}, \theta) = (n-1)(\hat{\theta} - \hat{\theta}_{-m}) \qquad 5.2.2.2$$

Theoretically the perturbed distribution $\bar{F}_k$, where the parameter $\theta = T(F_1, \ldots, F_k, \ldots, F_g)$, may be expressed as

$$\bar{F}_k = (1 - \varepsilon)F_k + \varepsilon\delta_{\underline{x}} \qquad 5.2.2.3$$

$\delta_{\underline{x}}$ being the distribution function which assigns unit probability to the point $\underline{x}$.

Note that as k refers to the group, we can ignore k and work with one group, say the first only, i.e. we will concentrate on $\bar{F}_1 = (1-\varepsilon)F_1 + \varepsilon\delta_{\underline{x}}$.

Assume a two population discriminant function $(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x} = \underline{\ell}'\underline{x}$ where $\underline{x} \sim n(p, \underline{\mu}_i, \Sigma)$. Given $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$, we have Mahalanobis's theoretical $\Delta^2 = \underline{\delta}'\Sigma^{-1}\underline{\delta}$ and the discriminant function coefficient vector $\underline{\ell} = \underline{\delta}'\Sigma^{-1}$.

In order to determine the influence functions of Δ^2 and $\underline{\ell}$, we must determine the influence of perturbing for $\underline{x}_m = \underline{x}$ on $\underline{\mu}_i$, $\underline{\delta}$, Σ and Σ^{-1}. Let

$$\Sigma = w_1\Sigma_{F_1} + w_2\Sigma_{F_2}, \quad w_1 + w_2 = 1, \; w_k > 0 \qquad 5.2.2.4$$

with

$$\Sigma_{F_1} = \int (\underline{x} - \underline{\mu}_i)(\underline{x} - \underline{\mu}_i)' dF_i, \quad \underline{\mu}_i = \int \underline{x} dF_i \qquad 5.2.2.5$$

For further derivations assume $\Sigma_{F_1} = \Sigma_{F_2}$, so that the weighting

5.10/...

factors make provision for unequal sample sizes only. So, if + indicates the perturbed parameter

$$\underline{\mu}_1 + (1-\varepsilon)\underline{\mu}_1 + \varepsilon\underline{x} = \underline{\mu}_1 + \varepsilon(\underline{x}-\underline{\mu}_1)$$

$$= \underline{\mu}_1 + \varepsilon\underline{z} \qquad 5.2.2.6$$

where $\underline{z} = \underline{x}-\underline{\mu}_1$.

$$\underline{\delta} + (1-\varepsilon)\underline{\delta} + \varepsilon(\underline{x}-\mu_2),$$

i.e. in the difference between the centroids, $\underline{x}$ will appear with probability 1, in stead of $\underline{\mu}_1$. So

$$\underline{\delta} + (1-\varepsilon)(\underline{\mu}_1-\underline{\mu}_2) + \varepsilon(\underline{x}-\underline{\mu}_2) = \underline{\mu}_1-\underline{\mu}_2 + \varepsilon(\underline{x}-\underline{\mu}_1)$$

$$= \underline{\delta} + \varepsilon\underline{z} \qquad 5.2.2.7$$

$$\Sigma_{F_1} + (1-\varepsilon)\Sigma_{F_1} + \varepsilon\underline{z}\underline{z}' \qquad 5.2.2.8$$

With the same reasoning as before 5.2.2.7 we have $\underline{z}\underline{z}'$ in stead of Σ_{F_1}, because $\underline{x}-\underline{\mu}_1 = \underline{z}$ appears with probability 1. Further, keeping in mind that $\Sigma_{F_1} = \Sigma_{F_2}$ and that a weighting factor w_1 is involved where $\Sigma = w_1\Sigma_{F_1} + w_2\Sigma_{F_2}$

$$\Sigma + (1-\varepsilon w_1)\Sigma + \varepsilon w_1\underline{z}\underline{z}' \qquad 5.2.2.9$$

From Press(1972) follows now that

$$\Sigma^{-1} + (1-\varepsilon w_1)^{-1}\left(\Sigma^{-1} - \frac{\varepsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}}{1-\varepsilon w_1 + w_1\underline{z}'\Sigma^{-1}\underline{z}}\right)$$

$$= (1+\varepsilon w_1)\Sigma^{-1} - \varepsilon w_1\Sigma^{-1}zz'\Sigma^{-1} \qquad 5.2.2.10$$

5.11/...

Now we can determine $I(\underline{x}, \Delta^2)$ and $I(\underline{x}, \underline{\ell})$.

(i) $I(\underline{x}, \Delta^2)$:

From 5.2.2.7 and 5.2.2.10 we have

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

$$= \underline{\delta}'\Sigma^{-1}\underline{\delta}$$

$$+ (\underline{\delta}+\varepsilon\underline{z})'\{(1+\varepsilon w_1)\Sigma^{-1} - \varepsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\}(\underline{\delta}+\varepsilon\underline{z}) \qquad 5.2.2.11$$

Let

$$\phi = (\underline{\mu}_1-\underline{\mu}_2)\Sigma^{-1}(\underline{x}-\underline{\mu}_1) = \underline{\delta}'\Sigma^{-1}\underline{z} \qquad 5.2.2.12$$

Then, ignoring terms in ε where the order is more than one it follows from 5.2.2.11 and 5.2.2.12 that

$$\Delta^2 \to \underline{\delta}'\Sigma\underline{\delta} + \varepsilon w_1\underline{\delta}'\Sigma\underline{\delta} + \varepsilon\underline{\delta}'\Sigma^{-1}\underline{z} + \varepsilon\underline{z}'\Sigma^{-1}\underline{\delta} - \varepsilon w_1\underline{\delta}\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\underline{\delta}$$

$$= \Delta^2 + \varepsilon w_1\Delta^2 + \varepsilon\phi + \varepsilon\phi - \varepsilon w_1\phi\phi'$$

$$= \Delta^2(1+\varepsilon w_1) + 2\varepsilon\phi - \varepsilon w_1\phi^2 \qquad 5.2.2.13$$

From this the influence function for Δ^2 can be determined, viz.

$$I(\underline{x}, \Delta^2) = \lim_{\varepsilon\to 0} \frac{\overline{\Delta^2}-\Delta^2}{\varepsilon}$$

$$= \lim_{\varepsilon\to 0} \frac{\Delta^2\varepsilon w_1+2\varepsilon\phi - \varepsilon w_1\phi^2}{\varepsilon}$$

$$= \Delta^2 w_1 + 2\phi - w_1\phi^2 \qquad 5.2.2.14$$

5.12/...

The coefficient vector can be standardised using $\sqrt{\Delta^2}$, so that the standarised vector $\underline{\ell}_s$ can be written as $\underline{\ell}_s = \Delta^{-1}\underline{\ell}$. Similarly ϕ can be standardised so that the standardised form of ϕ is $\phi_s = \Delta^{-1}\phi$. Now ϕ is distributed $N(0, \Delta^2)$, i.e. $\phi_s \sim N(0,1)$. Then 5.2.2.14 in terms of ϕ_s becomes:

$$I(\underline{x}, \Delta^2) = \Delta^2 w_1 + 2\Delta\phi_s - w_1\Delta^2\phi_s^2 \qquad 5.2.2.15$$

(ii) $I(\underline{x}; \underline{\ell})$

In the same way as before and using 5.2.2.7 and 5.2.2.10

$$\underline{\ell} = \Sigma^{-1}\underline{\delta} \rightarrow \left((1+\varepsilon w_1)\Sigma^{-1} - \varepsilon w_1 \Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\right)(\underline{\delta}+\varepsilon\underline{z})$$

$$= (1+\varepsilon w_1)\Sigma^{-1}\underline{\delta} - \varepsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\underline{\delta}$$

$$+ (1+\varepsilon w_1)\Sigma^{-1}\varepsilon\underline{z} - \varepsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\varepsilon\underline{z}$$

$$= \underline{\ell} + \varepsilon w_1\underline{\ell} - \varepsilon w_1\Sigma^{-1}\underline{z}\phi + \Sigma^{-1}\varepsilon\underline{z} \qquad 5.2.2.16$$

ignoring terms with ε^2 and smaller factors and taking 5.2.2.12 into account. Then the influence function is

$$I(\underline{x}; \underline{\ell}) = \lim_{\varepsilon\to 0} \frac{\bar{\underline{\ell}}-\underline{\ell}}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{(\underline{\ell} + \varepsilon w_1\underline{\ell} - \varepsilon w_1\Sigma^{-1}\underline{z}\phi + \Sigma^{-1}\varepsilon z) - \underline{\ell}}{\varepsilon}$$

$$= w_1\underline{\ell} - w_1\Sigma^{-1}\underline{z}\phi + \Sigma^{-1}\underline{z}$$

$$= w_1\underline{\ell} + (1 - w_1\phi)\Sigma^{-1}\underline{z} \qquad 5.2.2.17$$

And as before we can determine $I(\underline{x}; \underline{\ell}_s)$ and also $I(\underline{x}; \underline{\ell}'\underline{\ell})$; where $\underline{\ell}_s$ refers to the scaled vector:

$$I(\underline{x}; \underline{\ell}_s) = \{\tfrac{1}{2}w_1 - \tfrac{1}{2}w_1\phi(2-\phi)\Delta^{-2}\}\underline{\ell}_s - \Delta^{-1}\Sigma^{-1}\underline{z}(w_1\phi-1) \qquad 5.2.2.18$$

5.13/...

$$I(\underline{x};\ \underline{\ell}'\underline{\ell}) = 2w_1\underline{\ell}'\underline{\ell} + 2(1-w_1\Phi)\underline{\ell}'\Sigma^{-1}\underline{z} \qquad 5.2.2.19$$

In the sample analogues of the theoretical influence functions ε is replaced by $\frac{-1}{n-1}$ and terms of order (n^{-2}) are ignored. Further ϕ_s is replaced by $\hat{\phi}_s = \widehat{\Delta^{-1}\phi} = \widehat{\Delta^{-1}\underline{\ell}(\underline{x}_m-\underline{\mu}_1)} = \underline{c}_s'(\underline{x}_m-\bar{\underline{x}}_1)$ where $\underline{c}_s$ is the standardised vector of the sample discriminant function coefficients and $w_k = \frac{n_k}{n_1+n_2}$. Mahalanobis's Δ^2 is replaced by D^2 and $\underline{\ell}'\Sigma^{-1}(\underline{x}_m-\underline{\mu}_1)$ is replaced by $(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_m-\bar{\underline{x}}_1)$; S being the unbiased pooled cov. matrix based on n_1+n_2-2 degrees of freedom.

In an application on real data Campbell plotted the change in D^2 against the standardised discriminant score as a deviation from that for the species mean, i.e. he plotted $D^2-D^2_{-m}$ against $\underline{c}_s'(\bar{\underline{x}}_m-\bar{\underline{x}}_1)$. If one uses now the asymptotically normal distribution of the discriminant scores one can decide on possible outliers.

A second technique is to plot $D^2-D^2_{-m}$ against a gamma distribution with parameters estimated by means of the maximum likelihood method from the smallest 95 order statistics. In fact Campbell didn't use $D^2-D^2_{-m}$ as such, but

$$I_M(\underline{x},\ D^2) = I_{Max}(\underline{x},\ D^2) - I(\underline{x},\ D^2)$$

$$= w_1^{-1}\Big(1-2w_1D\underline{c}_s'\ (\bar{\underline{x}}_m-\bar{\underline{x}}_1) + w_1^2D^2\Big[\underline{c}_s'(\bar{\underline{x}}_m-\bar{\underline{x}}_1)\Big]^2\Big) \qquad 5.2.2.20$$

$$\doteq w_1D^2(\underline{c}_s'(\bar{\underline{x}}_m-\bar{\underline{x}}_1) - w_1^{-1}D^{-1})^2 \qquad 5.2.2.21$$

$\doteq \chi^{2'}$ with 1d.f. and non-centrality parameter $(w_1^2D^2)^{-1}$ which suggests a gamma distribution. He compared the moments for

5.14/...

$I_m(\underline{x}, D^2)$ with those for the $C\chi^2_\upsilon$ distribution, where $C = \upsilon^{-1}.(1+w_1^{-2}D^{-2})$ and $\upsilon = 1 + (2w_1^2D^2 + w_1^4D^4)^{-1}$.

In a similar way he used a third plot, viz. that of

$$I_M(\underline{x};\ \underline{c}'\underline{c}) = 2(1 - w_1 D\underline{c}_s'(\underline{x}_m-\bar{\underline{x}}_1))\underline{c}'S^{-1}(\underline{x}_m-\bar{\underline{x}}_1) \qquad 5.2.2.22$$

against the gamma quantiles as in the previous case.

All these techniques lead to the same indication of outliers in his example.

The study has shown that observations do influence D^2 rather assymetrically and also that the inclusion of an observation lying further from the mean of the other group decreases rather than increases D^2. All in all Campbell came to the conclusion that probability plots of the $D^2-D^2_{-m}$ values against the appropriate gamma distribution would seem to be preferable to probability plots of the discriminant scores against expected normal order statistics or similar appropriate normal plotting positions.

5.3 Discriminant Analysis and the Basic Structure Display of a Data Matrix

The application of the basic structure display of a data matrix (BSDM) on the linear discriminant analysis can be described as explained by Greenacre (1980). For the theoretical aspects see appendix A.

Assume n observations in g groups such that $n = n_1+n_2+ \ldots + n_g$. Then the display which tries to separate the groups maximally can be obtained using BSDM:

$$BSDM(Z\ ;\ \Omega,\ \phi;\ a, b) = BSDM(\bar{X}-\underline{1}\bar{\underline{x}}';\ D_n,\ A^{-1};\ 1,-) \qquad 5.3.1$$

5.15/...

where $\bar{X}$: gxm, the matrix of group means on the m variables; $D_n = \text{diag}(n_1, \ldots, n_g)$ and A is the pooled within groups sum of squares and cross products matrix. This provides the co-ordinates of the group centres of gravity in the discriminant subspace. The display of the cases themselves can be obtained using with $Z = \bar{X} - \underline{1}\underline{\bar{x}}'$:

$$F = ZA^{-\frac{1}{2}}V_1(D_{\mu_1}^{\frac{1}{2}})^{a-1} \qquad 5.3.2$$

where a = 1, so that

$$F = ZA^{-\frac{1}{2}}V_1 \qquad 5.3.3$$

where V_1 is the appropriate set of right basic vectors from 5.3.1, i.e. the eigenvectors of

$$Q = A^{-\frac{1}{2}}(\bar{X}-1\underline{\bar{x}}')D_n(\bar{X}-\underline{1}\underline{\bar{x}}')A^{-\frac{1}{2}} \qquad 5.3.4$$

The BSDM for correspondence analysis can be expressed similarly. If $X = [x_{ij}]$ such that $x_{ij}>0$ and P is defined as X divided by its total, then using the definitions as in Appendix E and the basic structure display theory in Appendix A the correspondence analysis of the rows is:

$$\text{BSDM } (D_{\underline{r}}^{-1}P - \underline{1}\underline{c}'; \quad D_{\underline{r}}, D_{\underline{c}}^{-1}; \quad 1, -) \qquad 5.3.5$$

and for the columns:

$$\text{BSDM}(PD_{\underline{c}}^{-1} - \underline{r}\underline{1}'; \quad D_{\underline{c}}, D_{\underline{r}}^{-1}; -, 1) \qquad 5.3.6$$

or for the simultaneous display:

$$\text{BSDM}(D_{\underline{r}}^{-1}PD_{\underline{c}}^{-1} - \underline{1}\underline{1}'; \quad D_{\underline{r}}, D_{\underline{i}}, 1, 1) \qquad 5.3.7$$

where 5.3.7 does not show a biplot as a+b≠1 and

$$F = D_{\underline{r}}^{-1}PGD_{\alpha_1}^{-1} \text{ and } G = D_{\underline{c}}^{-1}P'FD_{\alpha_1}^{-1} \qquad 5.3.8$$

5.16/...

5.4 Some other aspects and problems

Looking at the mass of literature on the topic of discriminant analysis it is immediately obvious that it is impossible to do more than only touch the surface of this subject in one volume. Not only is discriminant analysis related to so many fields of study, e.g. regression analysis, principal components, canonical variables, correspondence analysis and more, but the number of techniques involved are almost directly proportional to the number of types of distributions and/or variables one may encounter.

Fortunately we have seen in many cases that if one is willing to give up some accuracy one can get a long way with the basic methods, as many of these approaches are fairly robust for deviations from the assumed forms.

Despite the remark in the previous paragraph there are however, a number of problem areas which have not been emphasised or even pointed out at all. I shall just briefly summarise some of these aspects.

(i) The unequality of the var.-cov. matrices Σ_i can be quite problematical, especially when hypothesis tests must be applied. Kshirsagar (1972) discussed inter alia the Behrens-Fisher problem and gave some useful results. We have seen that when uncertainty exists with respect to $\Sigma_i = \Sigma_j$, $i \neq j$ it may often be better to assume them to be equal, e.g. the LDF gives better results than the QDF of Fisher when normality doesn't hold fairly well.

On the other hand some spesific forms of Σ may simplify calculations. If the common variance-covariance matrix Σ of two p-variate normal populations has the form

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \\ \rho & \rho & 1 & & \\ \dots & \dots & \dots & \dots & \dots \\ \rho & \dots & \dots & \rho & 1 \end{bmatrix}$$

5.17/...

then Penrose's size and shape factors can be used for discrimination and these are dependent on only two factors, viz. their variances and mean differences. Even if Σ does not have this special shape, it can be approximated by $\hat{\Sigma}'' = (1-\rho)I_p + \rho E_{pp}$ by first standarising the variates and then replacing each ρ_{ij} (the correlation between x_i and y_j) by ρ, the average correlation among all variables. A more detailled discussion can be found in Kshirsagar.

(ii) In all the discussions not much mention has been made of border line observations. This can be a field of study on its own. The programs in BMDP7M also make use of forced classification. It is however, possible to keep border line cases out of the classification cycle. One can use the misclassification rates to obtain confidence intervals for classification purposes. If an observation falls in an area of uncertainty according to the calculated/specified confidence limits it must be ignored. Another approach is to use the F distribution with $\frac{n-k-(p-1)}{np}.\frac{n_i}{n_i+1}D_i^2 = F_{p,n-(p-1)}$ to evaluate the probability of observing a distance as great as or greater than D_i^2, i=1, ..., k. These probabilities might be called typicality probabilities. Allocate the new individual to the population corresponding to the largest such probability. If these posterior probabilities for an observation are lower than a threshhold determined from F, it does not have to be allocated (see McKay and Campbell, 1982).

(iii) Almost all our methods depend on the fact that different groups have different means - even when we estimate a function by means of the modes as in kernel estimation. It is however, possible that distributions may be of the form, say $\underline{x}_i \sim n(p, \underline{\mu}, \Sigma_i)$, i.e. centroids are similar but the dispersion matrices are different. Possible cases may be $\Sigma_1 = \sigma_1^2 I_p$ and $\Sigma_2 = \sigma_2^2 I_p$ or $\Sigma_1 = (1-\rho_1)I_p + \rho_1 E_{pp}$ and $\Sigma_2 = \sigma^2[(1-\rho_2)I_p + \rho_2 E_{pp}]$. Kshirsagar discussed these problems and how to deal with them in detail.

5.18/...

(iv) One aspect which has not been mentioned yet is the effect(s) of initial misclassification, i.e. in the training sample, on the discriminant function(s).

Lachenbruch (1968, 1974, 1979) and others paid attention to this problem. Some discussion can also be found in Hand (1981). In his 1974 article Lachenbruch found that the true error rates of the LDF are only slightly affected by initial misclassification of the observations in a non-random manner. The apparent error rates as well as Mahalanobis's D^2 however, are severely affected. In the case of random misclassification the effects were less severe.

Lachenbruch (1979) found that in the case of the quadratic discrimination function the effects of misclassification are quite serious. Equal misclassification rates in both groups do not alleviate the problem as in the equal var.-cov. case. If it is known that a large number of observations might be misclassified, it may be better to resort to cluster analysis techniques where no initial classification is available and compare the results with the original classification.

There is a wide field open for a study of the effects of misclassification in the design set in the application of other discrimination methods, e.g. logistic functions, kernel functions etc.

(v) Another common problem is what to do with incomplete data. One simple option is to estimate substitute values for the absent items. This is done not to get more accurate results, but in an attempt to retain the simplicity of the existing complete data analysis procedures. In such a case care must be taken not to accept spurious information contributed by artificial observations. A common approach in estimating substitute data is to determine values which will minimise the residual sum of squares. (See Hand, 1981).

Another more realistic approach may be to run a marginal discriminant analysis on the data, using only those variables which are available in the vector(s) with missing values, i.e. we compare the class conditional densities (weighted by priors and costs) in the subspace spanned by these components. This is equivalent to ignoring the information in the unavailable components of the design set vectors.

When non-parametric methods are used like the kernel pdf estimators or nearest neighbourhood methods one can simply ignore irrelevant components of the design set elements.

For an interesting discussion on this subject and references as well as cautionary notes refer Hand.

These are the basic aspects and there are probably many more, but this section does give an indication of what types of problems are common and how to deal with some of them. Some other techniques have not even be mentioned at all, e.g. Fourier's approach, Bahadur's method, graphical methods, linear programming and the perception criterion (Hand), and so forth. Some of these approaches are fairly well known, while others are not well-known and further research should be done before they will be in general use.

5.5 Two typical problems in marketing research - a brief discussion

In the back I have included two computer runs on two typical problems which may face any company at any stage. The one problem is based on continuous ratio data where a firm can establish for itself whether it is on the downward slope to bankruptcy or not or for the assessment of the solvency of his clients - i.e. if it can get hold of the necessary data. The other problem has a bearing on marketing research where a firm can decide on its target population on the one hand for advertising purposes or given a set type modus operandi with respect to advertising can decide what the impact points of his advertising campaign are

After some runs in which it was obvious that standardisation of data has no real impact on the results - in any case much less than expected - I left the data in original form:

(i) Bankruptcy data: (See Johnson and Wichern, 1982)

Five variables were included in this run and an F to include/exclude was used in a stepwise discriminant analysis program - BMDP7M. The five variables are as follows:

1. CFTD = (cash flow)/(total debt)
2. NITA = (net income)/(total assets)
3. CACL = (current assets)/(current liabilities)
4. CANS = (current assets)/(net sales)
5. SOLVENCY = 1. BANKRUPT = Π_1
   2. SOUND = Π_2

Note that variables 1 and 2 can take on negative values when e.g. "total debt"=total creditors or net income = net outflow. I assume equal priors and equal costs. The training sample consisted of 44 observations of which 19 were classified as bankrupt and 25 as sound.

BANKRUPTCY DATA

Row	CFTD	NITA	CACL	CANS	SOLVENCY (Π_i)
1	-.4485	-.4106	1.0865	.4526	1.
2	-.5633	-.3114	1.5134	.1642	1.
3	.0643	.0156	1.0077	.3978	1.
4	-.0721	-.0930	1.4544	.2589	1.
5	-.1002	-.0917	1.5644	.6683	1.
6	-.1421	-.0651	.7066	.2794	1.
7	.0351	.0147	1.5046	.7080	1.
8	-.0653	-.0566	1.3737	.4032	1.
9	.0724	-.0076	1.3723	.3361	1.
10	-.1353	-.1433	1.4196	.4347	1.
11	-.2298	-.2961	.3310	.1824	1.
12	.0713	.0205	1.3124	.2497	1.

5.21/...

Row	CFTD	NITA	CACL	CANS	SOLVENCY (Π_i)
13	.0109	.0011	2.1495	.6969	1.
14	-.2777	-.2316	1.1918	.6601	1.
15	.1454	.0500	1.8762	.2723	1.
16	.3703	.1098	1.9941	.3828	1.
17	-.0757	-.0821	1.5077	.4215	1.
18	.0451	.0263	1.6756	.9494	1.
19	.0115	-.0032	1.2602	.6038	1.
20	.5135	.1001	2.4871	.5368	2.
21	.0769	.0195	2.0069	.5304	2.
22	.3776	.1075	3.2651	.3548	2.
23	.1933	.0473	2.2506	.3309	2.
24	.3248	.0718	4.2401	.6279	2.
25	.3132	.0511	4.4500	.6852	2.
26	.1184	.0499	2.5210	.6925	2.
27	-.0173	.0233	2.0538	.3484	2.
28	.2169	.0779	2.3489	.3970	2.
29	.1703	.0695	1.7973	.5174	2.
30	.1460	.0518	2.1692	.5500	2.
31	-.0985	-.0123	2.5029	.5778	2.
32	.1398	-.0312	.4611	.2643	2.
33	.1379	.0728	2.6123	.5151	2.
34	.1486	.0564	2.2347	.5563	2.
35	.1633	.0486	2.3080	.1978	2.
36	.2907	.0597	1.8381	.3786	2.
37	.5383	.1064	2.3293	.4835	2.
38	-.3330	-.0854	3.0124	.4730	2.
39	.4785	.0910	1.2444	.1847	2.
40	.5603	.1112	4.2918	.4443	2.
41	.2029	.0792	1.9936	.3018	2.
42	.4746	.1380	2.9166	.4487	2.
43	.1661	.0351	2.4527	.1370	2.
44.	.5808	.0371	5.0594	.1268	2.

The jackknife method was used to determine the percentage of correct classification rate. A 90,9% apparent correct classification rate was down to an 86,4% jackknifed classification rate using 2 of the variables only, viz. the 2nd and 3rd variables, NITA (net income)/

(total assets) and CACL (current assets)/(current liabilities) . Two classification functions were obtained and classification can be read from the printout by applying the posterior probability of an observation with respect to each group. (See section 5.4(ii)). The discrimination functions obtained were

SOUND = -2,88188 - 12,97369 NITA + 2,39554 CACL

BANKRUPTCY = -5,69076 - 1,48693 NITA + 3,88493 CACL

(ii) Advertising Campaign (See Jacobs, 1983)

The data in this training sample were obtained by means of a type of stratified sampling method. The Republic of South Africa was divided into 13 regions and the common usage by people of the public media was investigated. Although the original data were obtained with the aim of investigating some aspects of cinema attendance, the data are useful for further investigation, if one can assume that all these media make use of advertising, to cut on costs.

The question here was: If one has a certain population (market) in mind for advertising purposes, can one discriminate correctly by paying advertising fees only to certain types of papers, radio stations etc.? The question could be formulated differently: You are requested to investigate a company's advertising campaign, given its optimum target market, can one make any remarks with respect to its efficiency in the sense that the right market is being reached?

As in the last example a jackknifed classification was used. The data were not standardised as they were in percentage form in any case. In spite of the fact that the original data were of a discrete type based on unequal population sizes, which were eliminated in a sense by using percentages, an apparent rate of 96,7% of correct classification was obtained. Using the jackknifed classification rate the figure comes down to a still very useful 92,4% rate of correct classification.

The variables in this case were as follows:

1. CIN = percentage who attended a cinema in the last two weeks
2. ENGDAY = percentage reading English daily newspaper
3. AFRDAY = percentage reading Afrikaans daily newspaper
4. ANYW = percentage reading any weekly newspaper
5. MAGA = percentage reading any magazine
6. TELV = percentage watching prime television (20h00-21h30)
7. RADV = percentage listening to Radio 5
8. SPRI = percentage listening to Springbok radio
9. CLAGRO = class group:
    1. EURO = Europeans = Π_1
    2. COLOU = Coloureds = Π_2
    3. AFRI = Africans = Π_3
    4. ASIAN = Asian = Π_4

ADVERTISING MEDIA

Row	CIN	ENGDAY	AFRDAY	ANYW	MAGA	TELV	RADV	SPRI	CLAGRO (Π_i)
1.	20.3	45.4	34.7	86.0	98.8	60.1	15.8	24.9	1
2.	26.4	69.8	6.3	85.8	94.9	57.8	20.1	31.8	1
3.	23.5	40.6	33.2	79.2	91.9	56.2	17.1	27.2	1
4.	22.2	15.9	51.2	76.3	89.8	69.2	12.4	33.6	1
5.	23.9	56.4	27.3	83.2	90.9	56.9	21.0	27.7	1
6.	24.5	25.6	37.5	78.6	91.0	62.4	11.8	29.0	1
7.	17.8	18.0	41.5	81.8	93.7	58.4	7.5	29.2	1
8.	14.5	19.4	39.0	77.3	93.7	62.2	7.6	24.1	1
9.	15.3	42.1	33.5	81.0	91.7	64.0	11.6	28.4	1
10.	48.2	48.3	27.0	83.8	92.7	42.8	35.5	25.3	1
11.	46.1	39.3	32.1	82.3	95.4	48.3	37.0	30.5	1
12.	28.3	45.5	30.7	84.0	94.2	55.1	21.2	25.7	1
13.	12.9	48.5	34.0	83.2	92.1	63.0	8.9	23.8	1
14.	18.6	38.7	26.6	81.0	91.3	63.8	11.4	32.5	1
15.	20.4	64.8	2.7	83.5	94.4	61.5	15.5	42.3	1

5/24...

Row	CIN	ENGDAY	AFRDAY	ANYW	MAGA	TELV	RADV	SPRI	CLAGRO
16.	23.4	34.4	25.6	74.2	93.2	63.9	11.5	33.9	1
17.	16.3	16.8	53.0	67.0	95.3	61.8	6.5	42.0	1
18,	25.2	50.6	22.0	78.0	92.5	63.5	15.6	35.8	1
19.	20.3	17.1	28.2	75.6	91.9	64.0	7.4	32.9	1
20.	9.4	14.6	33.0	80.9	93.3	66.1	3.5	34.2	1
21.	10.1	17.6	29.2	67.9	96.5	59.3	2.6	35.3	1
22.	16.1	36.5	26.1	77.7	93.8	67.4	8.5	35.7	1
23.	50.0	46.9	27.4	79.1	96.9	45.5	33.8	31.3	1
24.	42.5	38.9	23.9	79.2	97.1	54.4	29.4	36.8	1
25.	22.6	38.2	27.0	77.1	95.9	63.6	13.1	32.9	1
26.	15.7	35.8	26.9	80.1	95.3	67.8	6.3	33.1	1
27.	12.8	29.1	13.0	58.8	48.7	27.1	9.8	21.6	2
28.	19.9	55.0	5.1	67.1	48.6	51.4	7.5	41.7	2
29.	16.2	55.1	8.0	77.6	60.4	45.5	13.3	28.4	2
30.	22.0	25.8	31.3	67.1	58.8	30.5	17.5	23.7	2
31.	13.0	4.3	16.9	51.9	43.2	13.9	4.7	21.2	2
32.	3.7	.2	10.4	23.3	23.8	3.5	1.7	13.1	2
33.	6.2	32.8	11.5	59.3	44.2	33.0	7.2	23.9	2
34.	26.1	29.6	15.3	63.1	59.3	23.4	14.4	23.2	2
35.	29.3	31.8	14.7	65.9	60.8	26.5	15.8	24.0	2
36.	8.4	35.5	11.5	60.5	54.2	30.2	11.7	17.9	2
37.	5.1	32.3	15.2	62.5	44.0	32.8	5.1	26.2	2
38.	7.5	26.5	10.0	57.3	60.9	28.9	9.4	31.1	2
39.	14.4	36.7	3.1	57.5	62.9	49.7	8.6	52.9	2
40.	10.1	47.6	6.2	72.1	72.7	47.0	13.9	37.5	2
41.	6.5	13.0	19.5	63.2	69.0	31.8	7.9	40.3	2
42.	8.5	2.0	19.8	46.4	55.0	10.5	2.1	26.0	2
43.	2.7	1.5	4.0	22.8	29.9	2.0	3.8	21.4	2
44.	4.8	28.4	8.1	55.0	58.2	36.1	5.8	34.9	2
45.	15.4	29.8	12.1	65.6	72.9	24.2	15.7	29.2	2
46.	15.1	30.2	10.8	63.7	77.5	26.5	16.8	28.3	2
47.	8.9	28.8	12.2	58.3	64.4	33.9	6.7	29.9	2
48.	1.4	28.2	8.8	56.6	54.5	33.7	5.7	38.2	2
49.	31.1	71.9	1.2	86.2	63.7	38.2	31.3	37.3	3
50.	41.7	69.4	1.0	78.8	60.2	50.3	14.8	24.9	3
51.	35.7	77.0	1.3	88.7	64.2	42.7	31.5	33.3	3

5.25/...

Row	CIN	ENGDAY	AFRDAY	ANYW	MAGA	TELV	RADV	SPRI	CLAGRO
52.	21.6	49.4	1.1	68.5	46.7	22.4	23.8	37.4	3
53.	18.7	71.2	.7	85.2	56.9	42.6	21.7	35.4	3
54.	59.5	73.4	2.0	86.3	77.2	36.8	43.5	36.4	3
55.	58.1	72.2	3.1	85.7	76.0	34.5	41.9	35.0	3
56.	28.7	73.6	.3	90.3	70.6	39.2	35.2	42.6	3
57.	16.1	73.9	.0	88.2	59.1	47.4	17.8	28.7	3
58.	17.9	46.6	.0	69.5	54.2	42.9	25.9	37.9	3
59.	32.2	43.2	.5	65.8	59.5	51.7	13.4	38.5	3
60.	22.4	50.7	.1	72.3	59.1	48.5	27.1	38.4	3
61.	10.7	25.6	.0	49.0	33.4	20.7	15.5	31.7	3
62.	16.3	45.7	.0	70.0	53.5	46.3	18.8	42.5	3
63.	37.9	59.2	.3	82.6	76.1	42.8	46.7	27.3	3
64.	30.7	51.8	.2	80.9	75.7	37.8	40.6	32.2	3
65.	16.6	52.8	.0	75.4	58.4	49.2	22.1	49.1	3
66.	13.3	43.0	.0	63.2	46.4	51.8	12.4	39.6	3
67.	8.3	11.9	.8	21.3	24.4	4.8	2.4	2.5	4
68.	10.5	9.9	.2	46.6	42.0	3.5	3.1	10.6	4
69.	21.4	17.6	1.1	22.7	35.2	3.9	2.5	1.9	4
70.	10.0	9.0	4.8	14.9	34.2	3.2	.4	.9	4
71.	25.3	28.0	1.7	41.8	44.7	6.5	4.1	3.3	4
72.	35.0	18.7	2.9	29.2	47.7	6.4	2.8	3.2	4
73.	25.8	17.9	2.9	29.2	50.0	7.6	2.3	2.3	4
74.	3.8	3.8	.3	15.9	22.6	1.2	1.3	4.1	4
75.	8.2	13.1	1.2	21.5	24.8	3.2	1.3	2.4	4
76.	24.5	15.6	1.1	32.7	46.3	5.1	3.7	5.3	4
77.	26.0	10.1	1.4	29.6	46.9	4.0	4.4	4.6	4
78.	22.7	25.0	1.7	37.3	47.9	6.7	2.7	4.9	4
79.	5.9	14.8	1.3	23.4	24.7	3.3	1.6	2.2	4
80.	2.6	3.6	.5	10.7	16.1	1.7	.6	.8	4
81.	3.0	4.3	.0	33.4	33.8	.4	2.2	.9	4
82.	3.0	5.3	.6	11.7	24.7	2.4	.6	.8	4
83.	.3	4.0	1.5	5.4	22.2	1.8	1.2	.2	4
84.	7.1	17.4	1.5	34.4	44.7	5.2	2.3	2.6	4
85.	4.6	6.0	2.6	26.1	42.3	10.0	.9	1.6	4

5.26/...

Row	CIN	ENGDAY	AFRDAY	ANYW	MAGA	TELV	RADV	SPRI	CLAGRO
86.	.8	4.5	.0	15.2	36.1	1.7	2.5	1.2	4
87.	1.4	1.0	.2	10.5	16.8	.2	.6	.2	4
88.	1.3	3.9	.3	11.5	16.0	1.8	.7	.7	4
89.	6.3	6.3	1.1	24.7	44.3	2.2	2.1	1.1	4
90.	6.4	4.9	.9	20.2	42.4	1.9	2.3	1.2	4
91.	2.0	5.7	.6	18.0	23.1	2.6	1.4	.9	4
92.	1.2	5.7	.4	17.7	21.7	1.7	.1	.9	4

What is very interesting in this printout is the fact that six variables were viewed as significant by the F in/out as presented by the BMDP7M-package, but that variable 2 - English daily newspaper cannot be used as a discriminator variable. This does not necessarily mean that one should not advertise in this newspaper. It may obviously mean that "approximately" the same proportion of each of the population groups reads this paper. What the results can be used for, is to decide which media can be ignored if one wants to advertise with say Europeans in mind; then variable 7 with a discriminant function coefficient of 0,0520 can be ignored, or for Coloureds magazine advertisements can be ignored (discriminant function coefficient of 0,08709), i.e. if these t tests are non-significant - see Appendix D and also section 1.5.

A further interesting point is that if one wants to reach the whole population one cannot ignore any of these media.

These two elementary examples shown and briefly discussed are only two of a large assortment of typical applications of discriminant analysis in marketing research on a variety of data types. In order to see more examples of this sort one just has to page through copies of the Journal of Marketing Research, The Journal of Finance or similar journals.

APPENDIX

A.1 The basic structure display of a data matrix

The basic structure display of a data matrix (BSDM) is also known as the "canonical form" (Eckart and Young, 1936) or the singular "decomposition" (Good, 1969). This display is also known as the "Eckart-Young decomposition" (see Kristof, 1970).

Greenacre (1980), like Green and Carroll (1976) preferred the term "basic structure" in his research paper where he summarised his description of basic structure in his thesis - see Greenacre (1978).

The basic structure of a matrix is the decomposition of the matrix into elements of sample structure with an immediate geometric appeal. Given a matrix A and using its basic structure one can find a least-squares approximation $\hat{A}$ of A with the feature that rank $(\hat{A})$ < rank (A). This $\hat{A}$ now provides a graphical display of the orgininal A.

A.2 Basic Structure - Greenacre, 1980

Any real matrix A_{nxm} may be expressed in the basic structure as

$$A_{nxm} = U_{nxr} D_{\alpha_{rxr}} V'_{rxm} \qquad A.2.1$$

$$= \sum_{k=1}^{r} \alpha_k \underset{nxl}{u_k} \underset{lxm}{v'_k} \qquad A.2.2$$

where $D_\alpha = \text{diag}(\alpha_1, \ldots, \alpha_r)$, $\alpha_i > 0$, $i = 1, \ldots, r$; $r \leq \min(n,m)$ = rank (A) and $U'U = I = V'V$. Call α_k the kth basic value, $\underset{-}{u}_k$ the kth left basic vector and $\underset{-}{v}_k$ the kth right basic vector.

A.2/...

The column vectors $\underline{u}_k$, k = 1, ..., r of U form an orthonormal basis for the columns of A and the column vectors $\underline{v}_k$, k = 1, ..., r of V form an orthonormal basis for the transposed rows of A. The matrices U and V thus determine the multidimensional subspace in which A is contained. The basic values in D_α determine the "magnitude" of A in each of its r basic dimensions.

In the special case when A is a symmetric matrix, say $A = B_{nxn}$ with rank (B) = r ⩽ n the basic structure of B is

$$B_{nxn} = U_{nxr}\, D_{\lambda,rxr}\, U'_{rxn}$$

$$= \sum_{k=1}^{r} \lambda_k \underline{u}_k \underline{u}'_k \qquad \text{A.2.3}$$

If it is assumed that the basic values are arranged in descending order so that $\alpha_1 \geqslant \alpha_2 \geqslant \ldots \geqslant \alpha_r > 0$ with the basic vectors of U and V correspondingly then the basic structure is uniquely determined so that one can approximate A by $\hat{A}_{[p]}$ where

$$\hat{A}_{[p]} = \sum_{k=1}^{p} \alpha_k \underline{u}_k \underline{v}'_k \qquad \text{A.2.4}$$

$\hat{A}_{[p]}$ is the nxm matrix formed from the first p (i.e. the largest) basic values and corresponding basic vectors of the matrix A of rank r where p < r. $\hat{A}_{[p]}$ is called the "best rank p approximation" of A in the sense that it minimises

$$\| A - A_{[p]} \|^2 = \text{trace} \left[(A - A_{[p]})(A - A_{[p]})' \right] \qquad \text{A.2.5}$$

for all rank p matrices $A_{[p]}$.

In matrix form $\hat{A}_{[p]}$ can be expressed as

$$\hat{A}_{[p]} = U_1\, D_{\alpha_1}\, V'_1 \qquad \text{A.2.6}$$

A.3/...

from

$$A = U_1 D_{\alpha_1} V_1' + U_2 D_{\alpha_2} V_2'$$

$$= \hat{A}_{[p]} + (A - \hat{A}_{[p]}) \qquad \text{A.2.7}$$

where

$$U = \underset{p}{\left[U_1 \vdots U_2\right]}\ n \qquad \underset{r-p}{}$$

$$V = \underset{p}{\left[V_1 \vdots V_2\right]}\ m \qquad \underset{r-p}{}$$

$$D_\alpha = \begin{matrix}\left[\begin{matrix} D_{\alpha_1} & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & D_{\alpha_2} \end{matrix}\right] & \begin{matrix} p \\ \\ r-p \end{matrix} \\ \underset{p}{} \qquad \underset{r-p}{} & \end{matrix}$$

$\hat{A}_{[p]}$ is called the rank p basic structure of A where A is the rank r basic structure. A measure of the"fit" of $\hat{A}_{[p]}$, the "least squares estimate" (Referring to the matrix norm) to A is given by

$$\tau_{[p]} = \frac{\| \hat{A}_{[p]} \|^2}{\| A \|^2} = \frac{\text{trace } (\hat{A}_{[p]} \hat{A}_{[p]}')}{\text{trace } (AA')}$$

$$= \frac{\sum_{k=1}^{p} \alpha_k^2}{\sum_{k=1}^{r} \alpha_k^2} \qquad \text{A.2.8}$$

so that $0 \leqslant \tau_{[p]} \leqslant 1$ and the error of approximation is given by

A.4/...

$$1 - \tau_{[p]} = \frac{\| A - \hat{A}_{[p]} \|^2}{\| A \|^2}$$

$$= \frac{\text{trace}\left[(A - \hat{A}_{[p]})(A - \hat{A}_{[p]})' \right]}{\sum_{k=1}^{r} \alpha_k^2}$$

$$= \frac{\sum_{k=p+1}^{r} \alpha_k^2}{\sum_{k=1}^{r} \alpha_k^2} \qquad \text{A.2.9}$$

Computation of the basic structure can be accomplished by the algorithm of Golub and Reinsch (1971) or by using the fact that if

$$A = U D_\alpha V' \qquad \text{A.2.10}$$

then

$$A'A = V D_\alpha U' U D_\alpha V'$$

$$= V D_\alpha^2 V' \qquad \text{A.2.11}$$

which is the eigenstructure of the mxm symmetric matrix A'A with eigenvalues the squared values α_k^2, k = 1, ..., r and eigen vectors the right basic vectors $\underline{v}_k$, k = 1, ..., r of A (refer 2.3). If $m \leq n$, find the structure in 2.11, i.e. V and D_α^2 and therefore $D_\alpha = + \sqrt{D_\alpha^2}$. Then from 2.10 and 2.1

$$U = A V D_\alpha^{-1} \qquad \text{A.2.12}$$

If m > n one could use AA' for computational purposes.

A.5/...

A.3 The Generalised Basic Structure

In more general terms the basic structure could be determined using the "generalised Fröbenius norm" in stead of the "Fröbenius norm" where the latter is

$$\| A \|^2 = tr[AA'] = \sum_i^n \sum_j^m a_{ij}^2$$

$$= \sum_{i=1}^{n} \underline{a}_i' \underline{a}_i \qquad \text{A.3.1}$$

where $\underline{a}_i$ is the ith row vector of A written as a column vector.

Use the generalised Euclidian norm to define ϕ:

$$\| \underline{a}_i \|_\phi^2 = \underline{a}_i' \phi \underline{a}_i \qquad \text{A.3.2}$$

where ϕ is positive definite.

The generalised Fröbenius norm will then be

$$\| A \|_{D_\omega, \phi}^2 \equiv \sum_{i=1}^{n} \omega_i \| \underline{a}_i \|_\phi^2$$

$$= tr\ (D_\omega A \phi A') \qquad \text{A.3.3}$$

where $\underline{\omega}$ is just a weighting vector.

The solution to our problem of finding the least-squares lower rank approximation of the matrix A, i.e. $\hat{A}_{[p]}$ where $\hat{A}_{[p]}$ is the solution to the expression:

$$\text{Minimise } \| A - A_{[p]} \|_{\Omega, \phi}^2 = tr\left[\Omega (A - A_{[p]}) \phi (A - A_{[p]})'\right] \qquad \text{A.3.4}$$

A.6/...

can be determined using the generalised basic structure of A.

The latter can be obtained from the ordinary basic structure of the appropriately transformed matrix, i.e.

$$A_{nxm} = N_{nxr}\, D_{\alpha}{}_{rxr}\, M'_{rxm} \qquad A.3.5$$

where $N'\Omega N = I = M'\phi M$ gives rise to $\hat{A}_{[p]} = N_1 D_{\alpha_1} M_1'$, because if $N = \Omega^{-\frac{1}{2}} U$ and $M = \phi^{-\frac{1}{2}} V$, then the transformed matrix

$$\Omega^{\frac{1}{2}} A \phi^{\frac{1}{2}} = (\Omega^{\frac{1}{2}} N) D_{\alpha} (M' \phi^{\frac{1}{2}}) \qquad A.3.6$$

where $(\Omega^{\frac{1}{2}}N)'(\Omega^{\frac{1}{2}}N) = I = (\phi^{\frac{1}{2}}M)'(\phi^{\frac{1}{2}}M)$, i.e. $A = ND_{\alpha}M'$ where $N'\Omega N = I = M'\phi M$ as stated. This means that $U_1 D_{\alpha_1} V_1'$ is replaced by $\Omega^{\frac{1}{2}} N_1 D_{\alpha_1} M_1' \phi^{\frac{1}{2}}$ so that the error

$$\| \Omega^{\frac{1}{2}} A \phi^{\frac{1}{2}} - \Omega^{\frac{1}{2}} N_1 D_{\alpha_1} M_1' \phi^{\frac{1}{2}} \|^2 = tr\left[\Omega (A - N_1 D_{\alpha_1} M_1') \phi (A - N_1 D_{\alpha_1} M_1')' \right]$$

$$= \| A - N_1 D_{\alpha_1} M_1' \|^2_{\Omega,\phi} \qquad A.3.7$$

is minimised.

As a result $N_1 D_{\alpha_1} M_1'$ is implied as the rank p approximation of A in the general norm.

A.4 Basic Structure Display

A graphical display which approximates the higher dimensional rectangular data matrix can now be obtained using features of the basic structure.

Assume a data matrix which is preprocessed, for example to "centre" the data and call this processed matrix Z. Then find the lower rank p matrix through the generalised basic structure and call this matrix $\hat{Z}_{[p]}$ where

A.7/...

$$\hat{Z}_{[p]} = N_1{}_{nxp} D_{\alpha_1}{}_{pxp} M_1'{}_{pxm} \qquad A.4.1$$

The object is to represent the rows of $\hat{Z}_{[p]}$ as points in a p-dimensional Euclidian space, i.e. with p axes, so that the between points distances in the display are exactly the between rows distances in the metric ϕ. These displayed distances are approximations of the true distances in the metric ϕ between rows of the original Z.

Let the coordinates of the row points in the display be contained in the rows of matrix F_{nxp}. The linear structure of F is then the set of scalar products, i.e.

$$\begin{aligned} FF' &= \hat{Z}_{[p]}\phi\hat{Z}_{[p]}' \\ &= N_1 D_{\alpha_1} M_1' \phi M_1 D_{\alpha_1} N_1' \\ &= (N_1 D_{\alpha_1})(N_1 D_{\alpha_1})' \end{aligned} \qquad A.4.2$$

so that one can take F to be $N_1 D_{\alpha_1}$.

Let p be equal to 2, then the basic concept of a biplot permits the columns of $\hat{Z}_{[p]}$ to be represented by

$$G = M_1 \qquad A.4.3$$

i.e.

$$\begin{aligned} \hat{Z}_{[2]} &= N_1 D_{\alpha_1} M_1' \\ &= DM_1' \\ &= FG' \end{aligned} \qquad A.4.4$$

i.e.

$$\hat{Z}_{ij} = \underline{f}_i' \underline{g}_j \qquad A.4.5$$

A.8/...

where $\underline{f}_i$ and $\underline{g}_j$ are the ith and jth rows of F and G respectively written as column vectors and $\hat{z}_{ij}$ is an approximation of the (i,j) th element of the original Z.

Other factorisations may be possible:

$$\hat{Z}_{[2]} = FG' \text{ where } F = N_1 \text{ and } G = M_1 D_{\alpha_1} \qquad \text{A.4.6}$$

and

$$\hat{Z}_{[2]} = FG' \text{ where } F = N_1 D_{\alpha_1}^{\frac{1}{2}} \text{ and } G = M_1 D_{\alpha_1}^{\frac{1}{2}} \qquad \text{A.4.7}$$

both having the biplot property that $\hat{z}_{ij} = \underline{f}_i' \underline{g}_j$, but with different meanings. Note further that from $Z = ND_\alpha M'$ with $N'\Omega N = I = M'\phi M$ follow that

$$\begin{aligned} Z &= ND_\alpha M' \\ Z\phi M &= ND_\alpha M'\phi M \\ Z\phi M &= ND_\alpha \\ Z\phi M D_\alpha^{-1} &= N \end{aligned} \qquad \text{A.4.8}$$

and

$$\begin{aligned} Z &= ND_\alpha M' \\ N'\Omega Z &= N'\Omega N D_\alpha M' \\ N'\Omega Z &= D_\alpha M' \\ D_\alpha^{-1} N'\Omega Z &= M' \\ Z'\Omega N D_\alpha^{-1} &= M \end{aligned} \qquad \text{A.4.9}$$

A.9/...

so that for example when $\hat{Z}_{[2]} = FG'$ with $F = N_1 D_{\alpha_1}$ and $G = M_1$ we have

$$N_1 = \hat{Z}_{[2]}\phi M_1 D_{\alpha_1}^{-1} \quad \text{and} \quad M_1 = \hat{Z}'_{[2]}\Omega N_1 D_{\alpha_1}^{-1}$$

$$N_1 D_{\alpha_1} = \hat{Z}_{[2]}\phi M_1 \quad \text{and} \quad M_1 = \hat{Z}'_{[2]}\Omega N_1 D_{\alpha_1} D_{\alpha_1}^{-2}$$

$$F = \hat{Z}_{[2]}\phi G \qquad G = \hat{Z}'_{[2]}\Omega F D_{\alpha_1}^{-2} \qquad \text{A.4.10}$$

from A.4.7.

A.5 Computation of the coordinates

First symmetrise the matrix to be diagonalised by pre-multiplying by $\phi^{\frac{1}{2}}$. Then, if one assumes that $m \leqslant n$, solve for M first by setting up the eigen equation

$$\phi^{\frac{1}{2}}(Z'\Omega Z\phi)M = \phi^{\frac{1}{2}}MD_\alpha^2$$

that is

$$(\phi^{\frac{1}{2}}Z'\Omega Z\phi^{\frac{1}{2}})\phi^{\frac{1}{2}}M = \phi^{\frac{1}{2}}MD_\alpha^2 \qquad \text{A.5.1}$$

where $M'\phi M = I$, i.e. $V'V = (\phi^{\frac{1}{2}}M)'\phi^{\frac{1}{2}}M = I$ and then $M = \phi^{-\frac{1}{2}}V$. Use A.4.8 to determine the left basic vectors N.

A symmetric argument for the basic vectors N leads to

$$(Z\phi Z'\Omega)N = ND_\alpha^2 \qquad \text{A.5.2}$$

In 5.2 if $\Omega = I$ then $N = S$, the scalar products in the metric ϕ of the rows of Z : $s_{ij} = \underline{z}_i'\phi\underline{z}_j$.

A.10/...

If $\Omega = D_\omega$ then $N = SD_\omega = [s_{ij}\omega_j]$, ω_j being the weights assigned to the row points.

A.6 Notation

The notation BSDM (Z; Ω, ϕ; a, b) for the generalised basic structure display of the data matrix summarises the procedure with the following meanings attached:

(i) Ω_{nxn} defines a norm on the columns of Z, or alternatively a set of weights on the rows.

(ii) ϕ_{mxm} defines a norm on the rows of Z, or alternatively a set of weights on the columns.

(iii) If $Z = ND_\alpha M'$ is the generalised basic structure of Z with Ω and ϕ defined as above, then the coordinate matrices F and G of the row and column points are: $F = N_1 D^a_{\alpha_1}$ and $G = M_1 D^b_{\alpha_1}$ with N_1, M_1 and D_{α_1} as defined in earlier sections.

(iv) a+b = 1 indicates a biplot interpretation as in section A.4.

Approximate Euclidian distances between rows of Z which are unweighted would be obtained by having $\phi = I$ and $\Omega = I$ respectively, i.e. BSDM (Z; I, I; 1, 0).

Note the difference between BSDM (Z; I, I; 1, 0) and BSDM (Z; I, I; 1, -) where the first indicates that the column points are going to be plotted with the biplot interpretation between row and column points (since a+b=1). The latter indicates that only row points are going to be displayed.

A.7 Computation and the BSDM analysis

Assume $m \leq n$; then the following algorithm is:

A.11/...

(i) Read data matrix X

(ii) Transform X to Z

(iii) Perform BSDM (Z; ϕ, Ω; a,b) with ϕ, Ω, a and b specified, i.e.

(a) Compute the symmetric matrix

$$\phi^{\frac{1}{2}}Z'\Omega Z\phi^{\frac{1}{2}} = Q \qquad \text{A.7.1}$$

where ϕ is diagonal; if not compute $\phi^{\frac{1}{2}}$ using the eigenstructure of ϕ:

$$\begin{aligned}\phi &= UD_\lambda U' \\ &= UD_\lambda^{\frac{1}{2}}U'UD_\lambda^{\frac{1}{2}}U'\end{aligned} \qquad \text{A.7.2}$$

so that

$$\phi^{\frac{1}{2}} = UD_\lambda^{\frac{1}{2}}U' \qquad \text{A.7.3}$$

(b) Determine the eigenstructure of Q

$$Q = VD_\mu V' \qquad \text{A.7.4}$$

(c) Find F and G in p dimensions:

$$G = \phi^{-\frac{1}{2}}V_1(D_\mu^{\frac{1}{2}})^b \qquad \text{A.7.5}$$

and

$$F = Z\phi^{\frac{1}{2}}V_1(D_{\mu_1}^{\frac{1}{2}})^{a-1} \qquad \text{A.7.6}$$

where D_{μ_1} and V_1 contain the largest p eigenvalues of Q and corresponding eigenvectors respectively.

(d) Complete the plotting routine for row and/or column points in selected pairs of dimensions.

so that from 1.9

$$H = BZZ'B'$$

$$= BB' \qquad \text{B.1.12}$$

We will first find the distribution of B (the Bartlett decomposition of the Wishart matrix H) and from B.1.12 the distribution of H. After this procedure it is straightforward to find the distribution of D (in B.1.4).

$$\text{As } \underline{y}_i = b_{i1}\underline{y}_1 + b_{i2}\underline{y}_2 + \ldots + b_{ii}\underline{y}_i, \quad i = 1, \ldots, p \qquad \text{B.1.13}$$

it follows that

$$b_{ij} = \underline{z}_j'\underline{y}_i, \quad j = 1, \ldots, i; \quad i = 1, \ldots, p \qquad \text{B.1.14}$$

$$h_{ii} = \underline{y}_i'\underline{y}_i = b_{i1}^2 + b_{i2}^2 + \ldots + b_{ii}^2, \quad i=1, \ldots, p \qquad \text{B.1.15}$$

so that

$$b_{ii}^2 = \underline{y}_i'\underline{y}_i - \sum_{j=1}^{i-1} b_{ij}^2, \quad i = 1, \ldots, p \qquad \text{B.1.16}$$

An incomplete random orthogonal transformation specifies i-1 new variables where the transformation is from $\underline{y}_i$ to b_{i1}, b_{i2}, ..., $b_{i,i-1}$, i.e.

$$\begin{bmatrix} b_{i1} \\ b_{i2} \\ \cdot \\ \cdot \\ \cdot \\ b_{i,i-1} \end{bmatrix} = \begin{bmatrix} \underline{z}_1' \\ \underline{z}_2' \\ \cdot \\ \cdot \\ \cdot \\ \underline{z}_{i-1}' \end{bmatrix} \underline{y}_i \qquad \text{B.1.17}$$

so that $b_{ij} \sim N(0,1)$ for $j = 1, \ldots, i-1$ and $b_{ii}^2 = \underline{y}_i'\underline{y} - \sum_{j=1}^{i-1} b_{ij}^2$

B.4/...

has the χ^2 distribution with n-(i-1) degrees of freedom and they are unconditional distributions as $y_1, \ldots, y_{i-1}$ do not appear in these variates. (See Kshirsagar, 1972).

The following theorem follows:

If the pxn matrix $Y = [y_{ir}]$ represents a random sample of size n, (n>p) from the $n(p, \varrho, I_p)$ population, and if YY' = BB', where $B = [b_{ij}]$, i = 1, ..., p; j=1, ..., i is a lower triangular matrix (with $b_{ii}>0$, i=1, ..., p) then the variates b_{ij}, i, j=1, ..., p; i>j, are N(0,1), the variates b_{ii}^2, i=1, ..., p are independently distributed as χ^2 variates with n-(i-1) degrees of freedom and are also independent of the b_{ij}'s (i>j).

Then the distribution of matrix B follows as

$$\prod_{i=1}^{p}\prod_{j=1}^{i-1}\left\{\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}.b_{ij}^2}\, db_{ij}\right\}\prod_{k=1}^{p}\left\{\chi^2_{n-(k-1)}(b_{kk}^2)\, d(b_{kk}^2)\right\}$$

$$-\infty<b_{ij}<\infty \text{ for } i>j$$

$$0<b_{ij}<\infty \text{ for } i=j \qquad \text{B.1.18}$$

From B.1.12 transform from B to H using the transformation Jacobian (see Deemer and Olkin, 1951)

$$J(B\to H) = \frac{1}{J(H\to B)} = 2^p \prod_{i=1}^{p} (b_{ii})^{i-p-1} \qquad \text{B.1.19}$$

and

$$|H| = \prod_{i=1}^{p} b_{ii}^2 \qquad \text{B.1.20}$$

$$\text{tr } H = \text{tr } BB' = \sum_{i=1}^{p}\sum_{j=i}^{p} b_{ij}^2 \qquad \text{B.1.21}$$

B.5/...

The probability density function of H then becomes

$$W(p, n, I) = \begin{cases} K(p,n)|H|^{\frac{n-p-1}{2}} e^{-\frac{1}{2}trH}, & H>O \\ O & \text{otherwise} \end{cases} \qquad \text{B.1.22}$$

where

$$K(p,n) = \left[2^{\frac{1}{2}np}\pi^{p(p-1)/4}\prod_{i=1}^{p}\left\{\Gamma[\tfrac{1}{2}(n+1-i)]\right\}\right]^{-1} \qquad \text{B.1.23}$$

with A>O (meaning that A is positive - definite).

So from B.1.11 YY' has the W(p, n, I) distribution.

From B.1.4 and B.1.6 it now also follows that

$$H = C^{-1}D(C^{-1})', \qquad \text{B.1.24}$$

Transform now to D using the above-mentioned reference for Jacobians again, as well as the second part of B.1.5, then

$$J(A\to D) = |C^{-1}|^{p+1} = |\Sigma|^{-\frac{1}{2}(p+1)} \qquad \text{B.1.25}$$

$$|H| = |C^{-1}||D||C^{-1}| = |D|/|\Sigma| \qquad \text{B.1.26}$$

so that the distribution of D comes out as

$$W(p, n,\Sigma) = \begin{cases} \dfrac{K(p,n)}{|\Sigma|^{\frac{1}{2}n}}|D|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}tr\Sigma^{-1}D}, & D>O \\ O & \text{otherwise} \end{cases} \qquad \text{B.1.27}$$

And this is the distribution of D which we were after right from the start.

When p = 1, then D and Σ are both scalar quantities so that the distribution of D reduces to Σ (a scalar) times a χ^2-variate with

B.6/...

n degrees of freedom. The Wishart distribution is thus the multivariate generalisation of a χ^2 distribution and plays the same role in multivariate statistics as the univariate χ^2 distribution plays in univariate statistics.

Although only one solution to the problem is given, one can find numerous different solutions in literature (See Kshirsagar).

Kshirsagar went on from this point and generalised this result from the theoretical population parameters, μ and Σ and proved that $\bar{\underline{x}}$ and A as given in B.1.2 and B.1.3 are independently distributed as follows

$$\sqrt{n}\,\bar{\underline{x}} \sim n(p, \sqrt{n}\,\underline{\mu}, \Sigma) \qquad \text{B.1.28}$$

$$A \sim W(p, n-1, \Sigma), \qquad \text{B.1.29}$$

i.e. the distribution of the maximum likelihood estimator of Σ, is that of $\frac{1}{n}A$.

Note that A is the matrix of s.s. and s.p. of sample observations, measured from the sample means and have n-1 degrees of freedom and not n, like the matrix D where the sample observations are measured from the true means $\underline{\mu}$.

It is also useful to note that if

$$\underline{u} = \sqrt{n}\, C^{-1} (\bar{\underline{x}}-\underline{\mu}), \quad \Sigma = CC' \qquad \text{B.1.30}$$

so that

$$\underline{u}'\underline{u} = n(\bar{\underline{x}}-\underline{\mu})'\Sigma^{-1}(\bar{\underline{x}}-\bar{\underline{\mu}}) \qquad \text{B.1.31}$$

then $\underline{u} \sim n(p,\underline{o},I)$ and $\underline{u}'\underline{u} \sim \chi^2$ with p degrees of freedom which provides us with a test for the hypothesis

$$H_o : \underline{\mu} = \underline{\mu}_o \qquad \text{B.1.32}$$

B.7/...

when a sample of size n is available from a p-variate normal population with unknown mean $\underline{\mu}$ but known variance covariance matrix .

To conclude this section a summary of lemmas is given as presented by Kshirsagar:

Lemma 1: If a symmetric pos.-def. matrix D is distributed $W(p,n,\Sigma)$, then the matrix HDH', where H is mxp with constant elements and rank m has the $W(m,n,H\Sigma H')$ distribution.

Lemma 2: Let $D \sim W(p,n,\Sigma)$ and $\underline{h}$, a pxl vector be independently distributed, then $u = \underline{h}'D\underline{h}/\underline{h}'\Sigma\underline{h} \sim \chi^2_n(u)$ independent of $\underline{h}$.

Lemma 3: Let $D \sim W(p,n,\Sigma)$, then $|D|/|\Sigma|$ is distributed as the product of p independent χ^2 variates with n, n-1, ..., n-p-1 d.f. respectively.

Lemma 4: Let $D \sim W(p,n,\Sigma)$ then σ^{pp}/d^{pp} has the $\chi^2_{n-(p-1)}$ distribution where d^{pp}, σ^{pp} are the last elements of D^{-1} and Σ^{-1} respectively.

Lemma 5: Let $D \sim W(p,n,\Sigma)$ and $\underline{h}$, a pxl vector be independently distributed, then $\underline{h}'\Sigma^{-1}\underline{h}/\underline{h}'D^{-1}\underline{h} \sim \chi^2_{n-(p-1)}$ and is independent of $\underline{h}$.

Lemma 6: $\int_{XX'=D} f(XX')dx = (2\pi)^{\frac{1}{2}np} K(p,n)|D|^{\frac{1}{2}(n-p-1)} f(D)dD$ where

X is a pxn matrix (not necessarily normal), n>p, D>O.

Lemma 7: $\int_{D>O} |D|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}tr\Sigma^{-1}D} dD = \frac{|\Sigma|^{\frac{1}{2}n}}{K(p,n)}$ where D and Σ are pxp symmetric positive - definite and n>p. K(p,n) is as defined before.

B.8/...

Lemma 8: Given the matrix A which is pxp symmetric where $A \sim W(p,n,I)$ in the canonical form, then A is invariant for a transformation of the type $A^* = HAH'$ where H is any pxp orthogonal matrix with constants or variables as elements, distributed independently of A. If it is independent then HAH' is distributed independently of H.

Lemma 9: If $A \sim W(p,n,I)$ then

(i) $E(A^k) = c(k,n,p)I_p$

(ii) $E(A^{-k}) = d(k,n,p)I_p$

where $c(1,n,p) = n$; $c(2,n,p) = n(n+p+1)$, $d(1,n,p) = \frac{1}{n-p-1}$

Lemma 10: If D has the $W(p,n,\Sigma)$ distribution, then

(i) $E(D) = n\Sigma$

(ii) $E(D^{-1}\Sigma D^{-1}) = d(2,n,p)\Sigma^{-1}$

(iii) $E(D^{-1}) = \frac{1}{n-p-1}\Sigma^{-1}$ if $n-p-1 > 0$.

Lemma 11: If $D_i \sim W(p,n_i,\Sigma_i)$, $i=1, \ldots, k$, then $D = \sum_{i=1}^{k} D_i \sim W(p, \sum_{i=1}^{k} n_i, \Sigma)$ even if $n<p$, or even $n_i<p$ for some i, in which case we are dealing with the pseudo-Wishart distribution.

Lemma 12: Let X ; pxn be such that the columns of X, i.e. $\underline{x} \sim n(p,\underline{o},\Sigma)$, then for A : nxn idempotent of rank m we find that $XAX' \sim W(p,m,\Sigma)$ if $m>p$ and pseudo-Wishart if $m \leq p$.

Lemma 13: Let X : pxn be such that the columns of X, i.e. $\underline{x} \sim n(p,\underline{m},\Sigma)$ where $\underline{m}$ is the corresponding column of M, then $D=XX'$ is called the non-central Wishart distribution. Further if A

B.9/...

is idempotent of rank m and order n we have that $(X-M)A(X-M)' = U \sim W(p,m,\Sigma$ and $E(U) = m\Sigma$ and $E(XAX') = m\Sigma + MAM'$ which holds even when $m < p$.

B.2 The distribution of D_p^2 and other results

Let $\underline{u} \sim n(p,\underline{\mu},\Sigma)$ and $D \sim W(p,f,\Sigma)$ independent of $\underline{u}$, then

$$T^2 = f\underline{u}'D^{-1}\underline{u} \qquad \text{B.2.1}$$

which is known as Hotelling's T^2 based on f degrees of freedom with parameter of non-centrality $\lambda^2 = \underline{\mu}'\Sigma^{-1}\underline{\mu}$. (See Kshirsagar, 1972). If $\underline{u} \sim n(p,\underline{o},\Sigma)$ and D is independent $W(p,f,\Sigma)$ then T^2 has the central distribution. From this we have

$$H_p'(T^2/f,\lambda^2)dT^2$$

$$= e^{-\frac{1}{2}\lambda^2} \Sigma \frac{(\frac{1}{2}\lambda^2)^r}{r!} B^{-1}(\tfrac{1}{2}p+r,\ \tfrac{1}{2}(f-p+1))$$

$$\cdot \frac{(T^2/f)^{\frac{1}{2}p+r-1}}{(1+\frac{T^2}{f})^{\frac{1}{2}(f+1)+r}} d(\frac{T^2}{f}),\quad T^2>0 \qquad \text{B.2.2}$$

or

$$H_p(T^2/f,\lambda^2=0)dT^2 = B^{-1}(\tfrac{1}{2}p,\ \tfrac{1}{2}(f-p+1)) \cdot \frac{(T^2/f)^{\frac{1}{2}p-1}}{(1+\frac{T^2}{f})^{\frac{1}{2}(f+1)}} d(\frac{T^2}{f}),\ T>0$$

B.2.3

In the case of B.2.3 we have that $\frac{f-p+1}{p}.\frac{T^2}{f}$ has an F distribution with p and f-p+1 degrees of freedom.

Let us define $\bar{\underline{x}}$, $\bar{\underline{y}}$, $n = n_1+n_2$ and $A = A_x+A_y$, A being the "pooled" matrix of corrected s.s. and s.p. from both samples in a 2 group

B.10/...

discriminant analysis and S_x, S_y, S the corresponding var.-cov. matrices in the usual way: $\underline{x} \sim n(p, \underline{\mu}_x, \Sigma)$; $\underline{y} \sim n(p, \underline{\mu}_y, \Sigma)$.

In discriminant analysis we apply Hotelling's T^2 with $\lambda^2 = 0$, i.e. the central distribution to test the hypothesis

$$H_o : \underline{\mu}_x - \underline{\mu}_y = 0 \qquad \text{B.2.4}$$

i.e. we use

$$T^2 = \frac{n_1 n_2}{n_1+n_2} (n_1+n_2-2) (\bar{\underline{x}}-\bar{\underline{y}})' A^{-1} (\bar{\underline{x}}-\bar{\underline{y}})$$

$$= \frac{n_1 n_2}{n_1+n_2} (\bar{\underline{x}}-\bar{\underline{y}})' S^{-1} (\bar{\underline{x}}-\bar{\underline{y}}) \qquad \text{B.2.5}$$

The test is therefore whether

$$\lambda^2 = \frac{n_1 n_2}{n_1+n_2} (\underline{\mu}_x - \underline{\mu}_y)' \Sigma^{-1} (\underline{\mu}_x - \underline{\mu}_y) \qquad \text{B.2.6}$$

equals zero or not.

Now Mahalanobis's sample distance measure squared is

$$D^2 = (\bar{\underline{x}}-\bar{\underline{y}})' S^{-1} (\bar{\underline{x}}-\bar{\underline{y}}) \qquad \text{B.2.7}$$

so that from B.2.5 and B.2.6

$$T^2 = \frac{n_1 n_2}{n_1+n_2} D^2 \qquad \text{B.2.8}$$

and similarly

$$\lambda^2 = \frac{n_1 n_2}{n_1+n_2} \Delta^2 \qquad \text{B.2.9}$$

so that H_o implies that $\Delta^2 = (\underline{\mu}_x - \underline{\mu}_y)' \Sigma^{-1} (\underline{\mu}_x - \underline{\mu}_y) = 0$.

B.11/...

The distribution of D^2, when $\mu_x \neq \mu_y$ is therefore from B.2.1 and B.2.2.

$$H'_p\left(\frac{n_1n_2}{n_1+n_2} D^2/n_1+n_2-2, \frac{n_1n_2}{n_1+n_2} \Delta^2\right) d\left(\frac{n_1n_2}{n_1+n_2} D^2\right) \qquad \text{B.2.10}$$

The null distribution of D^2 is obtained by putting $\Delta^2 = 0$ in B.2.10, so it becomes

$$H_p\left(\frac{n_1n_2}{n_1+n_2} D_2/n_1+n_2-2\right) d\left(\frac{n_1n_2}{n_1+n_2} D^2\right) \qquad \text{B.2.11}$$

So, when $\Delta^2 = 0$, i.e. $\mu_x = \mu_y$, the F-test can be used in terms of D^2 by using $T^2 = \frac{n_1n_2}{n_1+n_2} D^2$, or

$$\frac{n_1+n_2-p-1}{p} \cdot \frac{n_1n_2}{n_1+n_2} \cdot \frac{D^2}{n_1+n_2-2} = F_{p,n_1+n_2-p-1} \qquad \text{B.2.12}$$

For an expansion on this for comparisons with respect to Δ_p^2 and Δ_k^2 and other combinations refer to Kshirsagar, where the discussion was taken further to include the Fisher-Behrens problem in the multivariate case.

C. Wilks's Λ Criterion

Kshirsagar (1972) gave a detailed discussion of Wilks's Λ criterion and its applications. Only some of the characteristics of this criterion will be mentioned in this section.

Wilks's Λ criterion in multivariate analysis can be compared to the F distribution in univariate statistics. In multivariate statistics we have seen that it is often possible to construct matrices A and B such that $A \sim W(p,n-q,\Sigma)$ and independently distributed of B where $B \sim W(p,q,\Sigma)$. Under certain zero hypotheses B will be central Wishart; if H_o is not satisfied then B will be non-central Wishart. The criterion

$$\Lambda = \frac{|A|}{|A+B|} \qquad C.1$$

was proposed to test H_o. We can therefore define this criterium as follows:

If $A \sim W(p,n-q,\Sigma)$ independently of $\underline{z}_r$, r=1, ..., q which themselves are independently distributed as $\underline{z}_r \sim n(p,\underline{o},\Sigma)$ then

$$\Lambda(n,p,q) = \frac{|A|}{|A+B|} \qquad C.2$$

where

$$B = \sum_{r=1}^{q} \underline{z}_r \underline{z}_r' \qquad C.3$$

where $\Lambda(n,p,q)$ can be used to test $H_o: E(\underline{z}_r) = \underline{o}$, r = 1, ..., q.

Note that n, p and q, the parameters of Λ are respectively the d.f. of A+B, the order of A and B and the d.f. of B.

Again comparing to univariate statistics, A is called the error s.s. and s.p. matrix and B is called the hypothesis s.s. and s.p. matrix.

B.2/...

where

$$B(L|\tfrac{n-q}{2}, \tfrac{q}{2})dL = \frac{1}{B(\frac{n-q}{2}, \frac{q}{2})}|L|^{\frac{1}{2}(n-q-p-1)}|I-L|^{\frac{1}{2}(q-p-1)}dL,$$

$$L>0; \quad I-L>0$$

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

where

$$B(\tfrac{f_1}{2}, \tfrac{f_2}{2}) = \pi^{\frac{1}{4}(p(p-1))} \prod_{i=1}^{p} \left\{ \frac{\Gamma(\frac{f_1+f_2+1-i}{2})}{\Gamma(\frac{f_1+1-i}{2})\Gamma(\frac{f_2+1-i}{2})} \right\}^{-1}$$

L being positive definite and I is the identity matrix.

It can be shown that

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|CLC'|}{|CC'|}$$

$$= \frac{|C||L||C'|}{|C||C|}$$

$$= |L|$$

$$= |TT'|$$

$$= |T|^2$$

$$= \prod_{i=1}^{p} t_{ii}^2 \quad , \qquad \text{C.9}$$

i.e. Wilks's $\Lambda(n,p,q)$ distribution, when $p \leq q$, is the distribution $\prod_{i=1}^{p} t_{ii}^2$, where t_{ii}^2, i=1, ..., p are independently distributed as $B(t_{ii}^2|\frac{1}{2}(n-q-i+1), \frac{1}{2}q)dt_{ii}^2$.

<u>When p>q</u>:

B is not Wishart distributed. Accordingly let

$$G = A+ZZ' \qquad \text{C.10}$$

C.4/...

where Z is according to C.3. Let

$$U = G^{-\frac{1}{2}}Z, \quad H^* = U'U, \quad L^* = I-H^*, \qquad C.11$$

then it can be shown that as $|L_i^*| = \prod_{j=1}^{i} t_{jj}^2$,

$$t_{ii}^{*2} = \frac{|L_i^*|}{|L_{i-1}^*|} \sim B(t_{ii}^{*2} | \tfrac{1}{2}(n-p-i+1), \tfrac{1}{2}p)\, dt_{ii}^{*2} \qquad C.12$$

and all t_{ii}^{*2} are distributed independently for i=1, ..., q.

Again

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|A|}{|A+ZZ'|} = \frac{|G-ZZ'|}{|G|} = |I-G^{-\frac{1}{2}}ZZ'G^{-\frac{1}{2}}|$$

$$= |I-UU'| = |L^*|$$

$$= \prod_{i=1}^{q} \frac{|L_i^*|}{|L_{i-1}^*|} = \prod_{i=1}^{q} t_{ii}^{*2}, \quad (|L_o^*| = 1), \qquad C.13$$

i.e. Wilks's $\Lambda(n,p,q)$, when p>q, is distributed like $\prod_{i=1}^{q} t_{ii}^{*2}$ where t_{ii}^{*2}, i=1, ..., q are independently distributed as $B(t_{ii}^{*2} | \tfrac{1}{2}(n-p-i+1), \tfrac{1}{2}p)\, dt_{ii}^{*2}$.

We can summarise the results as follows:

$\Lambda(n,p,q)$ and $\Lambda(n,q,p)$ have exactly the same distribution, i.e. the distribution of $\prod_{i=1}^{p} t_{ii}^2$ if p≤q and of $\prod_{i=1}^{q} t_{ii}^{*2}$ if p>q where t_{ii}^2 are independent $B(t_{ii}^2 | \tfrac{1}{2}(n-q-i+1), \tfrac{1}{2}q)$ variables and t_{ii}^{*2} are independent $B(t_{ii}^{*2} | \tfrac{1}{2}(n-p-i+1), \tfrac{1}{2}p)$ variables.

C.5/...

When p = 1,2 we find that

$$\frac{(1-\Lambda(n,1,q)/q}{\Lambda(n,1,q)/(n-q)} \sim F_{q,n-q} \qquad \text{C.14}$$

$$\frac{(1-\sqrt{\Lambda(n,2,q)})/q}{\sqrt{\Lambda(n,2,q)}/(n-q-1)} \sim F_{2q,2(n-q-1)} \qquad \text{C.15}$$

When q = 1,2 interchange p and q in C.14 and C.15, then

$$\frac{(1-\Lambda(n,p,1))/p}{\Lambda(n,p,1)/(n-p)} \sim F_{p,n-p} \qquad \text{C.16}$$

$$\frac{(1-\sqrt{\Lambda(n,p,2)})/p}{\sqrt{\Lambda(n,p,2)}/(n-p-1)} \sim F_{2p,\,2(n-p-1)} \qquad \text{C.17}$$

If however p and q take on other values these expressions are much more difficult to handle. Bartlett derived an approximation for Wilks's Λ in a transformed form:

$$\chi^2 = -(n-\tfrac{1}{2}(p+q+1))\ \log_e \Lambda \sim \chi^2_{pq} \qquad \text{C.18}$$

If $\frac{1}{3}[n-\frac{1}{2}(p+q+1)] \geq p^2+q^2$ the approximation is accurate to at least three decimal places, i.e. if n is not too small, all one needs is the χ^2 tables and Bartlett's correction factor: $n-\frac{1}{2}(p+q+1)$ after $-\log_e \Lambda$ has been calculated.

D. Testing the significance of individual coefficients in the linear discriminant function - the standard deviations

Instead of a forward, backward or mixture selection procedure for inclusion/exclusion of a variable or variables, one might estimate the standard error for all p variates and then discard those whose coefficients are not significantly different from zero.

The motivation for the above-mentioned procedure is the fact that the F-test[1)] is computationally cumbersome, unless direct estimations of the misclassification probabilities as a function of D^2 are being used and the distributions are normal (Cochran, 1964). Constanza and Afifi(1979) concluded that a forward selection procedure in some non-normal situations may be feasible if the significance level is in the range 0,10 - 0,25. Murray (1977) however, warned that stepwise procedures may lead to highly over-optimistic assessments of error rates if the direct estimates are used.

Kendall, Stuart and Ord (1982) gave a brief deriviation of the large-sample standard errors of the coefficients.

The discriminant function is $\underline{\ell}'\underline{x}$ where $\underline{\ell}$ is the coefficient vector, which is $\underline{\ell}' = (\ell_1, \ldots, \ell_p)$ for p variables. Let us find the standard deviation of ℓ_i using the covariance between ℓ_k and ℓ_m as calculation step.

Denote the elements of S^{-1}, the inverse of the pooled sample var-cov. matrix, by s^{jk} and $S = [s_{jk}]$

$$\underline{\ell} = S^{-1}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) \qquad \text{D.1}$$

1) $$F = \frac{B(n_1+n_2-p-1)(D_p^2 - D_k^2)}{(p-k)(1 + BD_k^2)} \quad \text{with } B = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}, \quad k<p$$

is distributed $F(p-k, n_1+n_2-p-1)$ under H_o.

D.2/...

where S^{-1} is symmetrical. From E.1 we have now

$$\ell_k = \sum_j s^{jk}(\bar{x}_{1j}-\bar{x}_{2j}) \qquad D.2$$

This leads to

$$d\ell_k = \sum_j \left\{ (\bar{x}_{1j}-\bar{x}_{2j})ds^{jk}+s^{jk}d(\bar{x}_{1j}-\bar{x}_{2j}) \right\} \qquad D.3$$

and

$$d\ell_m = \sum_r \left\{ (\bar{x}_{1r}-\bar{x}_{2r})ds^{rm}+s^{rm}d(\bar{x}_{1r}-\bar{x}_{2r}) \right\} \qquad D.4$$

so that

$$d\ell_k d\ell_m = \sum_{j,r} \left\{ (\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r})ds^{jk}ds^{rm}+(\bar{x}_{1j}-\bar{x}_{2j})s^{rm}ds^{jk}d(\bar{x}_{1r}-\bar{x}_{2r}) \right.$$
$$\left. + (\bar{x}_{1r}-\bar{x}_{2r})s^{jk}ds^{rm}d(\bar{x}_{1j}-\bar{x}_{2j})+s^{jk}s^{rm}d(\bar{x}_{1j}-\bar{x}_{2j})d(\bar{x}_{1r}-\bar{x}_{2r}) \right\} \qquad D.5$$

but means and covariances are independent for normal variations so

$$\operatorname{cov}(\ell_k,\ell_m) = \sum_{j,r} \left[(\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r})\operatorname{cov}(s^{jk}, s^{rm}) \right.$$
$$\left. + s^{jk}s^{rm}\operatorname{cov}\left\{(\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r})\right\} \right] \qquad D.6$$

Let the sample sizes be n_1 and n_2 for Π_1 and Π_2 respectively, then

$$\operatorname{cov}\left\{(\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r})\right\} = \operatorname{cov}(\bar{x}_{1j},\bar{x}_{1r}) + \operatorname{cov}(\bar{x}_{2j},\bar{x}_{2r})$$
$$= \left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{jr} \qquad D.7$$

Let Γ_{jk} be the co-factor of s_{jk} in $|S|$ so that $s^{jk} = \Gamma_{jk}/|S|$.

D.3/...

Now we can find the covariance of s^{jk} and s^{rm}:

$$ds^{jk} = \frac{1}{|S|} d\Gamma_{jk} - \frac{\Gamma_{jk}}{|S|^2} d|S|$$

$$= \frac{1}{|S|} \sum_{\alpha,\beta} \Gamma_{jk,\alpha\beta} ds_{\alpha\beta} - \frac{\Gamma_{jk}}{|S|^2} \sum_{\alpha,\beta} \Gamma_{\alpha\beta}\, ds_{\alpha\beta} \qquad \text{D.8}$$

with $\Gamma_{jk.\alpha\beta}$ the co-factor of $s_{\alpha\beta}$ in Γ_{jk}, i.e.

$$ds^{jk} = \frac{1}{|S|^2} \sum_{\alpha,\beta} \left\{ |S|\, \Gamma_{jk,\alpha\beta} ds_{\alpha\beta} - \Gamma_{jk}\, \Gamma_{\alpha\beta} ds_{\alpha\beta} \right\}, \qquad \text{D.9}$$

but according to Jacobi's theorem $|S|\, \Gamma_{jk,\alpha\beta} = \Gamma_{jk}\, \Gamma_{\alpha\beta} - \Gamma_{j\beta}\, \Gamma_{\alpha k}$, hence D.9 becomes

$$ds^{jk} = - \frac{1}{|S|^2} \sum_{\alpha,\beta} \Gamma_{j\beta}\, \Gamma_{k\alpha} ds_{\alpha\beta}$$

$$= - \sum_{\alpha,\beta} s^{j\beta} s^{k\alpha} ds_{\alpha\beta} \qquad \text{D.10}$$

So

$$\text{cov}(s^{jk}, s^{rm}) = \frac{1}{n_1+n_2} (s^{jm} s^{kr} + s^{jr} s^{km}) \qquad \text{D.11}$$

Substitute D.7 and D.11 in D.6 and then

$$\text{cov}(\ell_k, \ell_m) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sum_{j,k} s^{jk} s^{rm} s_{jr}$$

$$+ \frac{1}{n_1+n_2} \sum_{j,r} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r})(s^{jm} s^{kr} + s^{jr} s^{km})$$

$$= \left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^{mk} + \frac{1}{n_1+n_2} \sum_{j,r} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r}) s^{jm} s^{kr}$$

D.4/...

$$+ \frac{1}{n_1+n_2} s^{km} \sum_{j,r} (\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r}) s^{jr} \qquad \text{D.12}$$

$$= (\frac{1}{n_1} + \frac{1}{n_2}) s^{mk} + \frac{1}{n_1+n_2} \ell_m \ell_k + \frac{1}{n_1+n_2} s^{km} \sum_r \ell_r (\bar{x}_{1r}-\bar{x}_{2r})$$

$$= (\frac{1}{n_1} + \frac{1}{n_2}) s^{mk} + \frac{1}{n_1+n_2} \ell_m \ell_k + \frac{1}{n_1+n_2} s^{km} \underline{\ell}' (\bar{\underline{x}}_1-\bar{\underline{x}}_2)$$

D.13

Let $k = m = j$, then

$$\text{var } \ell_i = (\frac{1}{n_1} + \frac{1}{n_2}) s^{jj} + \frac{\ell_j^2}{n_1+n_2} + \frac{1}{n_1+n_2} \underline{\ell}' (\bar{\underline{x}}_1-\bar{\underline{x}}_2) s^{jj} \qquad \text{D.14}$$

E. The algebra of correspondence analysis with special reference to discriminant analysis

In this section the formal algebra of correspondence analysis is given as a background to the discussion of the application of correspondence analysis in discriminant analysis (Chapter 4). This section comes almost directly from Greenacre (1984).

Let N be a matrix of non-negative numbers excluding trivial vectors where sums of rows and columns are zero. The correspondence matrix P is defined as the matrix of elements of N devided by the grand total of N. The row and column sums of P are denoted by vectors $\underline{r}$ and $\underline{c}$. From the latter vectors the diagonal matrices $D_{\underline{r}}$ and $D_{\underline{c}}$ are formed.

Therefore, if the data matrix is $N = [n_{ij}]_{IxJ}$, $n_{ij} \geq 0$, then the correspondence matrix is

$$P = \frac{1}{n..} N \ , \quad n.. = \underline{1}'N\underline{1} \qquad \text{E.1}$$

and

$$\underline{r} = P\underline{1} \ ; \quad \underline{c} = P'\underline{1} \ ; \quad r_i, c_j > 0; \quad i=1, \ldots, I; \ j=1, \ldots, J \qquad \text{E.2}$$

$$D_{\underline{r}} = \text{diag } (\underline{r}) \text{ and } D_{\underline{c}} = \text{diag } (\underline{c}) \qquad \text{E.3}$$

The row and column profiles of the correspondence matrix are defined as the vectors of rows and columns of P divided by their respective sums, i.e. if $\underline{r}' = (r_1, r_2, \ldots, r_I)$ and $\underline{c}' = (c_1, c_2, \ldots, c_J)$ then

$$\underline{r}' = \frac{1}{r_i} \frac{1}{n..} (n_{i1}, n_{i2}, \ldots, n_{iJ}), \quad i=1, \ldots, I$$

$$= \frac{1}{r_i} (p_{i1}, p_{i2}, \ldots, p_{iJ}) \qquad \text{E.4}$$

and

E.2/...

$$\underline{c}_j' = \frac{1}{c_j}\frac{1}{n..}(n_{1j}, n_{2j}, \ldots, N_{Ij}), \quad j=1, \ldots, J$$

$$= \frac{1}{c_j}(p_{1j}, p_{2j}, \ldots, p_{Ij}) \qquad \text{E.5}$$

i.e. the matrices of row and column profiles are

$$R = D_{\underline{r}}^{-1} P = \begin{bmatrix} \underline{r}_1' \\ \underline{r}_2' \\ \cdot \\ \cdot \\ \cdot \\ \underline{r}_I' \end{bmatrix} \quad ; \quad C = D_{\underline{c}}^{-1} P = \begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{c}_J \end{bmatrix} \qquad \text{E.6}$$

Note this could have been written in terms of N as well, since the analysis is concerned of relative values only and is therefore invariant with respect to n.. .

The row and column profiles define two clouds of points in respective J- and I- dimensional weighted Euclidian spaces.

	Row cloud	Column cloud
Points:	I points $\underline{r}_1, \ldots, \underline{r}_I$ in J-dimensional space where $\underline{r}_i$ is as in E.4	J points in $\underline{c}_1, \ldots, \underline{c}_J$ in I-dimensional space where $\underline{c}_i$ is as in E.5
Masses:	The I elements of $\underline{r}$	The J elements of $\underline{c}$
Metric:	Weighted Euclidian with dimension weights defined by the inverses of the elements of $\underline{c}$, i.e. $D_{\underline{c}}^{-1}$.	Weighted Euclidian with dimension weights defined by the inverses of the elements of $\underline{r}$ i.e. $D_{\underline{r}}^{-1}$.

The last remarks say that the centroids of the row and column

E.3/...

clouds in their respective spaces are $\underline{c}$ and $\underline{r}$ respectively, because the row centroid is

$$\underline{r}'R/\underline{r}'1 = \underline{r}'R = \underline{r}'D_{\underline{r}}^{-1}P = \underline{1}'P = \underline{c} \qquad \text{E.7}$$

and the column centroid is

$$C'\underline{c}/\underline{1}'\underline{c} = PD_{\underline{c}}^{-1}\underline{c} = P\underline{1} = \underline{r} \qquad \text{E.8}$$

The overall spatial variation of each cloud of points is quantified by its total inertia, i.e. the weighted sum of squared distances from the points to their respective centroids, or analytically

Total inertia of row points

$$in(I) = \sum_i r_i(\underline{r}_i-\underline{c})'D_{\underline{c}}^{-1}(\underline{r}_i-\underline{c}) = \text{trace}\left[D_{\underline{r}}(R-\underline{1}\underline{c}')D_{\underline{c}}^{-1}(R-\underline{1}\underline{c}')'\right]$$

Total inertia of column points

$$in(J) = \sum_j c_j(\underline{c}_j-\underline{r})'D_{\underline{r}}^{-1}(\underline{c}_j-\underline{r}) = \text{trace}\left[D_{\underline{c}}(C-\underline{1}\underline{r}')D_{\underline{r}}^{-1}(C-\underline{1}\underline{r}')'\right] \qquad \text{E.9}$$

and further, the total inertia is similar in both clouds.

$$\begin{aligned} in(I) &= \sum_i r_i \sum_i (p_{ij}/r_i - c_j)^2/c_j \\ &= \sum_{ij}\sum (p_{ij} - r_i c_j)^2/r_i c_j \\ &= \sum_j c_j \sum_i (p_{ij}/c_j - r_i)^2/r_i \\ &= in(J) \end{aligned} \qquad \text{E.10}$$

i.e.

$$in(I) = in(J) = \frac{\chi^2}{n..} = \frac{1}{n..}\sum_{ij}\sum\left(\frac{n_{ij}-e_{ij}}{e_{ij}}\right)^2 \qquad \text{E.11}$$

E.4/...

because $p_{ij} = \frac{n_{ij}}{n..}$, so that

$$e_{ij} = \frac{(\Sigma_i n_{ij})(\Sigma_j n_{ij})}{n..}$$

$$= \frac{(n.. c_j)(n.. r_i)}{n..}$$

$$= n.. r_i c_j \qquad \text{E.12}$$

i.e.

$$\Sigma\Sigma_{ij}(p_{ij} - r_i c_j)^2/r_i c_j$$

$$= \Sigma\Sigma_{ij}(\frac{n_{ij}}{n..} - r_i c_j)^2/\frac{e_{ij}}{n..}$$

$$= \frac{1}{n..}\Sigma\Sigma_{ij}(n_{ij} - r_i c_j n..)^2/e_{ij}$$

$$= \frac{1}{n..}\Sigma\Sigma_{ij}(n_{ij} - e_{ij})^2/e_{ij} \qquad \text{E.13}$$

At this stage we want to find the K*-dimensional subspaces of the row and column clouds respectively which are closest to the points in terms of weighted sums of squared distances. These subspaces are defined by the K* right and left generalised singular vectors of $P-\underline{r}\underline{c}'$, in the metrics $D_{\underline{r}}^{-1}$ and $D_{\underline{c}}^{-1}$ corresponding to the K* largest singular values, i.e. the right and left singular values define the principal axes of the row and column clouds respectively.

When we look at the cloud of column points defined by the column

profiles of $C = PD_c^{-1}$ then we see that from section A if the centered column profiles

$$C - \underline{r}\underline{1}' = PD_c^{-1} - \underline{r}\underline{1}' = LD_\phi M', \text{ where } L'D_r^{-1}L = M'D_c M = I \qquad \text{E.14}$$

then the columns of L define the principal axes, and the rows of $G = MD_\phi$ define the co-ordinates. Also from E.14

$$P - \underline{r}\underline{c}' = LD_\phi (D_c M)', \text{ where } L'D_r^{-1}L = (D_c M)'D_c^{-1}(D_c M) = I \qquad \text{E.15}$$

In a similar way it can be shown that the principal axes of the row cloud, defined in J-dimensional space are given by the columns of Z in

$$D_r^{-1}P - \underline{1}\underline{c}' = YD_\psi Z', \text{ where } Y'D_r Y = Z'D_c^{-1}Z = I \qquad \text{E.16}$$

i.e. in

$$P - \underline{r}\underline{c}' = (D_r Y)D_\psi Z' \text{ where } (D_r Y)'D_r^{-1}(D_r Y) = Z'D_c^{-1}Z = I \qquad \text{E.17}$$

We can therefore define: Let the generalised singular value decomposition of $P - \underline{r}\underline{c}'$ be

$$P - \underline{r}\underline{c}' = AD_\mu B' \text{ where } A'D_r^{-1}A = B'D_c^{-1}B = I \qquad \text{E.18}$$

with $\mu_1 \geq \mu_2 \ldots \geq \mu_k > 0$, then the columns of A and B define the principal axes of the column and row clouds respectively.

Note that the sets of singular values μ_i, ϕ_i and ψ_i, i=1, ..., k

E.6/...

are identical and uniquely defined up to reflections only, assuming that all singular values are different. So the principal axes L of the column cloud are identical to the columns of A up to reflections. The subspace defined by L is the same as that of A.

It can further be shown that the respective co-ordinates of the row and column profiles with respect to their own principal axes (i.e. the principal co-ordinates) are related to the principal axes of the other cloud of profiles by simple rescalings.

Principal co-ordinates of row profiles

Let $F = (D_r^{-1}P - \underset{IxJ}{1}\underset{JxJ}{c'})D_c^{-1}\underset{JxK}{B}$ E.19

$\quad IxK$

be the co-ordinates of the row profiles with respect to principal axes B in the chi-square metric D_c^{-1}, then

$$F = D_r^{-1}AD_\mu$$

Principal co-ordinates of column profiles

Let $G = (D_c^{-1}P' - \underset{JxI}{1}\underset{IxI}{r'})D_r^{-1}\underset{IxK}{A}$ E.19

$\quad JxK$

be the co-ordinates of the column profiles with respect to principal axes A in the chi-squared metric D_r^{-1}, then

$$G = D_c^{-1}BD_\mu \quad ^{1)} \qquad \text{E.20}$$

The co-ordinates of individual points are contained in the rows of F and G. The co-ordinates of the points with respect to optimal K*-dimensional subspaces are contained in the rows of the first K* columns of F and G. If we write $F_{(2)}$ and $G_{(2)}$ as the first two columns of F and G respectively, then the rows of $F_{(2)}$ and $G_{(2)}$ define the projections of the row and column profiles onto respectively optimal planes.

The co-ordinates of F and G are related as follows - using E.18 and E.20

$^{1)}$ e.g. $G = D_c^{-1}(P-cr')D_r^{-1}A = D_c^{-1}(AD_\mu B')'D_r^{-1}A$

$= D_c^{-1}BD_\mu(A'D_r^{-1}A) = D_c^{-1}BD_\mu I$

$= D_c^{-1}BD_\mu$

$$G = D_{\underline{c}}^{-1}P'FD_{\underline{\mu}}^{-1} = CFD_{\underline{\mu}}^{-1} \; ; \quad F = D_{\underline{r}}^{-1}PGD_{\underline{\mu}}^{-1} = RGD_{\underline{\mu}}^{-1}$$

$$\text{or} \quad D_{\underline{\mu}}G = D_{\underline{c}}^{-1}P'F \; ; \qquad ; \; D_{\underline{\mu}}F = D_{\underline{r}}^{-1}PG \qquad \text{E.21}$$

With reference to the principal axes, the respective clouds of row and column profiles have centroids at the origin. The weighted variance (moment of inertia, sum of squares of the points' co-ordinates) along the k th principal axes in each cloud is equal to μ_k^2 , which is denoted by λ_k and called the k th principal inertia. The weighted covariance is zero.

<u>Centroid of rows of F</u> | <u>Centroid of rows of G</u>

$$\underline{r}'F = \underline{0}' \qquad \underline{c}'G = \underline{0}' \qquad \text{E.22}$$

<u>Principal inertias of row cloud</u> | <u>Principal inertias of column cloud</u>

$$F'D_{\underline{r}}F = D_{\underline{\mu}}^2 = D_{\underline{\lambda}} \qquad G'D_{\underline{c}}G = D_{\underline{\mu}}^2 = D_{\underline{\lambda}} \qquad \text{E.23}$$

The centerings in E.22 follow because the rows of F and G are merely the respective sets of centered profiles with respect to new reference systems of axes, e.g. $\underline{r}'(D_{\underline{r}}^{-1}P - \underline{1}\underline{c}') = \underline{1}'P - \underline{c}' = \underline{c}' - \underline{c}' = \underline{0}'$. These results follow immediately from the standardisation of the principal axes in E.18 as well as from E.20.

To be able to express the results graphically we draw up a table of columns which expresses the contributions of the rows and columns respectively to the inertia of an axis. This results from E.9, E.10 and E.23, i.e. the total inertia of each cloud of points is decomposed along the principal axes and among the points themselves.

		Decomposition of inertia axes				
		1	2	.....	K	Total
Rows	1	$r_1 f_{11}^2$	$r_1 f_{12}^2$	.....	$r_1 f_{1K}^2$	$r_1 \sum_{k=1}^{K} f_{1k}^2$
	2	$r_2 f_{21}^2$	$r_2 f_{22}^2$	.....	$r_2 f_{2K}^2$	$r_2 \sum_{k=1}^{K} f_{1k}^2$
	.	.	.	.....	.	.
	.	.	.	.....	.	.
	.	.	.	.....	.	.
	I	$r_I f_{I1}^2$	$r_I f_{I2}^2$	.....	$r_I f_{IK}^2$	$r_I \sum_{k=1}^{K} f_{Ik}^2$
	Total	$\lambda_1=\mu_1^2$	$\lambda_2=\mu_2^2$	.....	$\lambda_K=\mu_K^2$	inertia (I) = inertia (J)
Columns	1	$c_1 g_{11}^2$	$c_1 g_{12}^2$	.....	$c_1 g_{1K}^2$	$c_1 \sum_{k=1}^{K} g_{1k}^2$
	2	$c_2 g_{21}^2$	$c_2 g_{22}^2$	.....	$c_2 g_{2K}^2$	$c_2 \sum_{k=1}^{K} g_{2k}^2$
	.	.	.	.....	.	.
	.	.	.	.....	.	.
	.	.	.	.....	.	.
	J	$c_J g_{J1}^2$	$c_J g_{J2}^2$	.....	$c_J g_{JK}^2$	$c_J \sum_{k=1}^{K} g_{JK}^2$

Figure E.1

Note each of these contributions can be expressed as a proportion of the respective inertia λ_k in order to interpret the axis itself. These proportions are often called "absolute contributions",

E.9/...

because they are affected by the mass of each point. Each row of these tables contains the contributions of the axis to the inertia of the respective profile point. Again we can express each of these as proportions of the points' inertia in order to interpret how well the point is represented on the axes. These are often called "relative contributions", because the masses are divided out.

Note that the columns of F and G are the (non-trivial) eigenvectors of the respective matrices RC and CR, standardised according to E.23. The non-trivial eigenvalues of both these matrices are the principal inertias.

Row co-ordinates as eigenvectors	Column co-ordinates as eigenvectors
$(RC)F = FD_{\underline{\lambda}}$	$(CR)G = GD_{\underline{\lambda}}$
i.e. $(D_{\underline{r}}^{-1}PD_{\underline{c}}^{-1}P')F = FD_{\underline{\lambda}}$	i.e. $(D_{\underline{c}}^{-1}P'D_{\underline{r}}^{-1}P)G = GD_{\underline{\lambda}}$
	E.24
with standarisation	with standardisation
$F'D_{\underline{r}}F = D_{\underline{\lambda}}$ (or $D_{\underline{\mu}}^2$)	$G'D_{\underline{c}}G = D_{\underline{\lambda}}$ (or $D_{\underline{\mu}}^2$)
	E.25

Note for example that from $(D_{\underline{c}}^{-1}P')F = GD_{\underline{\mu}}$ (see E.21) and $R = D_{\underline{r}}^{-1}P$ it follows that

$$(RC)F = (D_{\underline{r}}^{-1}P)(D_{\underline{c}}^{-1}P')F = (D_{\underline{r}}^{-1}P)GD_{\underline{\mu}} = ((D_{\underline{r}}^{-1}P)G)D_{\underline{\mu}} = FD_{\underline{\mu}}^2 = FD_{\underline{\lambda}}$$

E.26

Note further that the above eigenequations should not be used

E.10/...

separately to obtain F and G, because it will be wasteful computationally and will also lead to differences in signs of the corresponding eigenvectors, seeing that the signs of eigenvector solutions are not identified.

For further discussion of the reconstitution formula of the correspondence matrix P which is useful for imputing missing values in the data matrix, the "standard co-ordinate" standardisation technique - unit inertias along principal axes, and in the last place the principle of "distributional equivalence" - the merging of points - see Greenacre (1984).

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. 1975. Discrete multivariate analysis: Theory and practice. Cambridge: The MIT Press.

Broffitt, J.D., Randles, R.H. and Hogg, R.V. 1976. Distribution-free partial discriminant analysis. Journal of the American Statistical Association, Volume 71, 1976; pp. 931-939.

Campbell, N.A. 1978. The influence function as an aid in outlier detection in discriminant analysis. Applied Statistics, Volume 27, No. 3, 1978; pp. 251-258.

*Chatfield, C. and Collins, A.J. 1980. Introduction to multivariate analysis. London: Chapman and Hall.

*Chanda, K.C. 1980. Asymptotic properties of classification rules based on Wilcoxon-type statistics. Journal of the American Statistical Association, Volume 75, No. 371, 1980; pp. 326-328.

Chernoff, H. 1973. Using Faces to represent points in K-dimensional space graphically. Journal of the American Statistical Association, Volume 68, no. 342, 1973; pp. 361-368.

Cochran, W.G. 1964. Approximate significance levels of the Behrens-Fisher test. Biometrics, Volume 20, 1964; pp. 191-195.

Cochran, W.G. 1964. On the performance of the linear discriminant function. Technometrics, Volume 6, 1964; pp. 179-190.

Constanza, M.C. and Afifi, A.A. 1979. Comparison of stopping rules in forward stepwise discriminant analysis. Journal of American Statistitical Association, Volume 74, 1979; pp. 777+.

Cochran, W.G. and Hopkins, C.E. 1961. Some classification problems with multivariate qualitative data. Biometrics, Volume 17, 1961; pp. 10-32.

4/...

*Cooley, W.W. and Lohnes, P.R. 1962. Multivariate properties for the behavioural science. New York. Wiley.

*Cooper, R.A. and Weekes, A.J. 1983. Data, models and statistical analysis. Southampton : Philip Allan.

Cox, D.R. 1970. The analysis of binary data. London : Methuen

Crash, A.R. and Perrault, W.D.(jr.) 1977. Validation of discriminant analysis in marketing research. Journal of Marketing Research, Volume XIV, 1977; pp. 60-68.

Day, N.E. and Kerridge, D.F. 1967. A general maximum likelihood discriminant. Biometrics, Volume 23, 1967; pp. 313.

Deemer, W. and Olkin, I. 1951. The Jacobians of certain matrix transformations useful in multivariate analysis. Biometrika, Volume 38, 1951; pp. 345.

Dillon, W.R. 1979. The performance of the linear discriminant function in non-optimal situations and the estimation of classification error rates: A review of recent findings. Journal of Marketing Research, Volume XVI, 1979; pp. 370-381.

Dillon, W.R., Goldstein, M. and Schiffman, L.G. 1978. Appropriateness of linear discriminant and multinomial classification analysis in marketing research. Journal of Marketing Research, Volume XV, 1978; pp. 103-112.

Dillon, W.R. and Western, S. 1982. Scoring frequency data for discriminant analysis: Perhaps discrete procedures can be avoided. Journal of Marketing Research, Volume XIX, 1982; pp. 44-56.

DiPillo, P.J. 1976. The application of bias to discriminant analysis. Communications in Stat. - Theor. Meth. Volume A 5(9), 1976; pp. 843-854.

5/...

Geisser, S. 1966. Predicture discriminantion in Krishnaiah, P.R., editor. Multivariate Analysis - Proceedings of an International Symposium held in Daytona, Ohio, June 14-19, 1965. New York and London : Academic Press.

Gilbert, E.S. 1968. On discrimination using qualitative variables. Journal of the American Statistical Association, Volume 63, 1968; pp. 1399-1418.

Gitlow, H.S. 1979. Discrimination procedures for the analysis of nominally scaled data sets. Journal of Marketing research, Volume XVI, 1979; pp. 387-393.

Glick, N. 1972. Sample based classification procedures derived from density estimates. Journal of the American Statistical Association, Volume 67, 1972; pp. 116-122.

Glick N. 1973. Sample based multinomial classification. Biometrics, Volume 29, 1973; pp. 241-256.

Glick, N. 1978. Additive estimators for probabilities of correct classification. Pattern Recognition, Volume 10, 1978; pp. 211-220.

*Gokhale, and Kullback, 1978. The information in contingency tables. New York : Marcel Dekker.

Goldstein, M. and Dillon, W.R. 1978. Discrete discriminant analysis. New York : Wiley.

Goldstein, M. and Rabinowitz, M. 1975. Selection of variates for the two-group multinomial classification problem. Journal of the American Statistical Association, Volume 70, 1975; pp. 776-781.

Golub, G.H. and Reinsch, C. 1971. Singular value decomposition and least squares solution. Linear Algebra II. Springer-verlag, Berlin.

7/...

Good, I.J. 1969. Some applications of the singular decomposition of a matrix. Technometrics, Volume 11, 1969; pp. 823-831.

*Gray, H.L. and Schucany, W.R. 1972. The generalised jackknife statistic. New York : Marcel Dekker.

*Graybill, F.A. 1961. An introduction to linear statistical models. Volume I, New York : McGraw-Hill.

Graybill, F.A. 1969. Introduction to matrices with applications in statistics. Belmont : Wadsworth Publishing Co.

Green, P.E. 1964. Bayesian classification procedures in analysing customers' characteristics. Journal of Marketing Research, 1964; pp. 44-50.

Green, P.E. 1978. Analysing multivariate data. Hinsdale, Illinois : The Dryden Press.

*Green, P.E. 1966. Bayesian classification procedures in analysing multivariate data. Journal of Marketing Research, Volume 1966; pp. 44-50.

Green, P.E. and Carroll, J.D. 1976. Mathematical tools for applied multivariate analysis. New York : Academic Press.

Greenacre, M.J. 1978. Some objective methods of graphical display of a data matrix. Doctoral thesis (Universeté Pierre et Marie Curie, Paris) published as a special report by the University of South Africa.

Greenacre, M.J. 1980. Basic structure display of a data matrix. Research report 80/2, October 1980 : University of South Africa.

Greenacre, M.J. 1981. Practical correspondence analysis in Barnett, V. editor. Interpreting multivariate data. Chichester: Wiley.

8/...

Greenacre, M.J. 1984. Theory and Applications of correspondence analysis. London : Academic Press.

*Greenacre, M.J. and Underhill, L.G. 1982. Scaling a data matrix in a low-dimensional Euclidian space in Hawkins, D.M. editor. Topics in Applied multivariate analysis. Cambridge : Cambridge University Press.

Habbema, J.D.F., Hermans, J. and Van den Broek, K. 1974. A stepwise discriminant analysis program using density estimation. Proceedings in computational statistics in Wien in Compstat, 1974; pp. 101-110.

Haberman, S.J. 1974. Log-linear models for frequency tables with ordered classifications. Biometrics, Volume 30, 1974; pp. 589-600.

*Hamburg, M. 1970. Statistical Analysis for decision making. New York : Harcourt, Brace and World.

Hand, D.J. 1981. Discrimination and classification. New York: Wiley.

Hand, D.J. 1982. Kernel discriminant analysis. Chichester: Research Studies Press - John Wiley and Sons.

Hawkins, D.M. 1976. The subset problem in multivariate analysis of variance. Journal of the Royal Statistical Society B, Volume 38, 1976; pp. 132-139.

Hawkins, D.M. 1980. Identification of outliers. London : Chapman and Hall.

Hawkins, D.M. 1981. A new test for multivariate normality and homocedasticity. Technometrics, Volume 23, 1981; pp. 105-110.

9/...

Hill, M.O. 1974. Correspondence analysis: A neglected multivariate method. Applied Statistics, Volume 23, No. 3, 1974; pp. 340-354.

Hill, M.O.1973. Reciprocal averaging : An eigenvector method of ordination. Journal of Ecology, Volume 61, 1973; pp. 237-251.

Hills, M. 1967. Discrimination and allocation with discrete data. Applied Statistics, Volume 16, No. 3, 1967; pp. 237-250.

Hirschfeld, H.O. 1935. A connection between correlation and contingency. Cambridge Philosophical Society Proceedings. (Math. Proc.), Volume 31, 1935; pp. 520-524.

*Hora, C. 1974. Sample size determination in Bayesian discrimination analysis. Journal of the American Statistical Association, Volume 73, No. 363, September 1974; pp. 569-572.

Hora, S.C. and Wilcox, J.B. 1982. Estimation of error rates in several population discriminant analysis. Journal of Marketing Research, Volume XIX, 1982; pp. 57-61.

*Jackson, E.C. 1968. Missing values in linear multiple discriminant analysis. Biometrics, Volume 24, 1968; pp. 835+

Jacobs, M. 1983. Linear regression techniques for identifying influential data and applications in commercial data analysis. Unpublished Ph.D. Thesis at University of Cape Town, 1983.

James, A.T. 1955. The non-central Wishart distribution in Proceedings of the Royal Society A, Volume 229; pp. 364.

Jennrich, R.A. 1977. Stepwise discriminant analysis in Enslein, K., Ralston, A. and Wilf, H.S., editor. Statistical methods for digital Computers. New York: Wiley.

Johnson, N.L. and Kotz, S. 1972. Distributions in Statistics: Continuous multivariate distributions,Volume 4. New York: Wiley.

10/...

Johnson, R.A. and Wichern, D.W. 1982. Applied multivariate statistical analysis. Englewood Cliffs: Prentice-Hall.

Kendall, M.G. 1963. A course in multivariate analysis. London: Charles Griffin.

Kendall, M.G. 1966. Discrimination and classification in Krishnaiah, P.R., editor. Multivariate analysis - Proceedings of an International Symposium held in Daytona, Ohio, June 14-19, 1965. New York and London : Academic Press.

Kendall M. and Stuart, A. 1976. The advanced theory of statistics, Volume 1, London and Wycombe: Charles Griffin.

Kendall, M. and Ord, J.K. 1981. The advanced theory of statistics, Volume II. London and Wycombe: Charles Griffin.

Kendall, M. 1982. The advanced theory of statistics, Volume III. London and Wycombe: Charles Griffin.

*Kotze, T.J. v. W. 1982. The log-linear model and its applications to multiway contingency tables in Hawkins, D.M., editor. Topics in applied multivariate analysis. Cambridge : Cambridge University Press.

Kristof, W. 1970. A theorem on the trace of certain matrix products. Journal of Mathematical Psychology, Volume 7, 1970; pp. 515-530.

Krzanowski, W.J. 1975. Discrimination and classification using both binary and continuous variables. Journal of the American Statistical Association, Volume 70, No. 352, 1975; pp. 782-790.

Krzanowski, W.J. 1977. The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics, Volume 19, No. 2, 1977; pp. 191-200.

11/...

Kshirsagar, A.M. 1972. Multivariate analysis. New York: Marcel Dekker.

*Kshirsagar, A.M. and Arseven, E. 1975. A note on the equivalency of two discrimination procedures. The American Statistician, Volume 29, No. 1, 1975; pp. 38-39.

Lachenbruch, P.A. 1967. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, Volume 23, 1967; pp. 639-645.

Lachenbruch, P.A. 1968. On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficent. Biometrics, Volume 24, 1968; pp. 823-834.

Lachenbruch, P.A. 1974. Discriminant analysis when the initial samples are misclassified II. Non-random misclassification models. Technometrics, Volume 16, No. 3, August 1974; pp. 419-424.

Lachenbruch, P.A. 1979. Note on initial misclassification effects on the quadratic discrimination function. Technometrics, Volume 21, No. 1, February 1979; pp. 129-132.

Lachenbruch, P.A. and Goldstein, M. 1979. Discriminant analysis. Biometrics, Volume 35, 1979; pp. 69-85.

Lachenbruch, P.A. and Mickey, M.R. 1968. Estimation of error rates in discriminant analysis. Technometrics, Volume 10, No. 1, 1968; pp. 1-11.

Marks, S. and Dunn, O.J. 1974. Discrimination functions when covariance matrices are unequal. Journal of the American Statistical Association, Volume 69, No. 346, 1974; pp. 555-559.

12/...

Marshall, A.W. and Olkin, I. 1968. A general approach to some screening and classification procedures. Journal of the Royal Statistical Society, B, Volume 30, 1968; pp. 407+

McKay, R.J. and Campbell, N.A. 1982. Variable selection techniques in discriminant analysis I. Description. British Journal of Mathematical and Statistical Psychology, Volume 35, 1982; pp. 1-29.

McKay, R.J. and Campbell, N.A. 1982. Variable selection techniques in discriminant analysis II. Allocation. British Journal of Mathematical aand Statistical Psychology, Volume 35, 1982; pp. 30-41.

McLachlan, G.J. 1979. A comparison of the estimative and predictive methods of estimating posterior probabilities. Communications in Statistics, Volume A 8(9), 1979; pp. 919-929.

Meisel, W.S. 1972. Computer oriented approach to pattern recognition. New York : Academic Press.

Montgomery, D.B. 1975. New product distribution: An analysis of supermarket buyer decisions. Journal of Marketing Research, Volume XII, 1975; pp. 255-264.

Moore, D.H. 1973. Evaluation of five discriminant procedures for binary variables. Journal of the American Statistical Association, Volume 68, 1973; pp. 399-404.

Moore, D.S., Whitsell, S.J. and Lundgrebe, D.A. 1976. Variance comparisons for unbiased estimators of probability of correct classification. IEEE Transactions on Information Theory, Volume 22, 1976; pp. 102-105.

*Morrison, D.G. 1969. On the interpretation of discriminant analysis. Journal of Marketing Research, Volume 6, 1969; pp. 156-163.

13/...

Morrison, D.F. 1976. Multivariate Statistical methods, 2nd Edition. New York: McGraw-Hill.

*Morrison, D.F. 1983. Applied linear statistical methods. Englewood Cliffs: Prentice-Hall.

Mosteller, F. and Tukey, W.T. 1968. Data analysis, including statistics in Lindsey, G. and Aronson, E., editors. The Handbook of Social Psychology, Volume 2, 1968. Reading, Massachusetts: Addison-Wesley; pp. 80-203.

*Muirhead, R.J. 1982. Aspects of multivariate statistical theory. New York: Wiley.

Murray, G.D. 1977. A cautionary note on selection of variables in discriminant analysis. Applied Statistics, Volume 26, 1977; pp. 246+.

*Murthy, M.N. 1957. Ordered and unordered estimators in sampling without replacement. Sankhya, Volume 18, 1957; pp. 379-390.

Murthy, V.K. 1966. Non-parametric estimation of multivariate densities with applications in Krishnaiah, P.R., editor. Multivariate analysis - Proceedings of an International Symposium held in Daytona, Ohio, June 14-19, 1965. New York and London : Academic Press.

*Murthy, M.N. 1967. Sampling theory and methods. Statistical Publishing Society, Calcutta, India.

Nishisato, S. 1980. Analysis of categorical data: Dual scaling and its applications. University of Toronto Press, Toronto.

Okamoto, M. 1963. An asymptotic expansion for the distribution of the linear discriminant function. Annals of Mathematical Statistics, Volume 34, 1963; pp. 1386-1401.

14/....

Olkin, I. and Tate, R.F. 1961. Multivariate correlation models with mixed discrete and continuous variables. Annals of Mathematical Statistics, Volume 32, 1961; pp. 448-465.

O'Neill, T.J. 1980. The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. Journal of the American Statistical Association, Volume 75, No. 369, 1980; pp. 154-160.

Parzen, E. 1962. On estimation of a probability density function and mode. Annals of Mathematical Statistics, Volume 33, No. 3, 1962; pp. 1065-1077.

Press, S.J. 1972. Applied multivariate analysis. New York: Holt, Rinehart and Winston.

Press, S.J. and Wilson, S. 1978. Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, Volume 73, No. 364, 1978; pp. 699-705.

Raath, E.L. and Hawkins, D.M. 1980. Predictive discrimination with an informative prior for the covariance matrix. Technical Report of the National Research Institute for Mathematical Sciences of the C.S.I.R. Twisk 147. Pretoria.

Randles, R.H., Broffitt, J.D., Ramberg, J.S. and Hogg, R.V. 1978. Discriminant analysis based on ranks. Journal of the American Statistical Association, Volume 73, No. 362, 1978; pp. 379-384.

Randles, R.H., Broffitt, J.D., Ramberg, J.S. and Hogg, R.V. 1978. Generalised linear and quadratic discriminant functions using robust estimates. Journal of the American Statistical Association, Volume 73, No. 363, 1978; pp. 564-568.

15/...

Rao, C.R. 1966. Covariance adjustment and related problems in Krishnaiah, P.R., editor. Multivariate analysis - Proceedings of an International Symposium held in Daytona, Ohio, June 14-19, 1965. New York and London: Academic Press.

Rao, C.R. 1973. Linear Statistical inference and its applications, 2nd edition. New York: Wiley.

Rencher, A.C. and Larson, S.F. 1980. Bias in Wilks's Λ in stepwise discriminant analysis. Technometrics, Volume 22, No. 3, 1980; pp. 349-356.

*Rigby, R.A. 1982. A credibility interval for the probability that a new observation belongs to one of two multivariate normal populations. Journal of the Statistical Society B, Volume 44, No. 2, 1982; pp. 212-220.

Sayre, J.W. 1980. The distribution of the actual error rates in linear discriminant analysis. Journal of the American Statistical Association, Volume 75, No. 369, 1980; pp. 201-205.

Schatzoff, M. 1966. Exact distribution of Wilks's likelihood ratio criterion. Biometrika, Volume 53, 1966; pp. 347+.

*Tabachnick, B.G. and Fidell, L.S. 1983. Using multivariate statistics. New York: Harper and Row.

Troskie, C.G. 1980. The distributions of the ratios of latent roots (condition numbers) and their applications in principal components or ridge regression. Technical Report, No. 6. University of Cape Town, October 1980.

Tukey, J.W. 1958. Bias and confidence in not quite large samples, (Abstract). Annals of Mathematical Statistics, Volume 20, 1958; pp. 614+.

16/...

Van der Merwe, C.A. and Kotze, T.J.v.W. 1980. Bekendstelling van verdelingsvrye klassifikasieprosedure gebaseer op die skatting van waarskynlikheidsverdelings deur gebruik te maak van die sleutelfunksiemetode (kernel function method). Technical Report, No. 3, 1980. Institute for Biostatistics, P.O. Box 70, Tygerberg, South Africa.

Van Ness, J. 1979. On the effects of dimension in discriminant analysis for unequal covariance populations. Technometrics, Volume 21, No. 1, 1979; pp. 119-127.

Van Ness, J.W. and Simpson, C. 1976. On the effects of dimension in discriminant analysis. Technometrics, Volume 18, 1976; pp. 175-181.

Vinoid, H.D. 1976. Application of new ridge regression methods to a study of Bell System scale economics. Journal of the American Statistical Association, Volume 71, 1976; pp. 835+.

Wahba, G. 1975. Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. Annals of Statistics, Volume 3, 1975; pp. 15-29.

Wegman, E.J. 1972. Nonparametric probability density estimation: I. A summary of available methods. Technometrics, Volume 14, No. 3, 1972; pp. 533-546.

Welch, B.L. 1939. A note on discriminant functions. Biometrika, Volume 31, 1939; pp. 218-220.

Wildt, A.R., Lambert, Z.V. and Durand, R.M. 1982. Applying the Jackknife statistic in testing and interpreting canonical weights, loadings and cross-loadings. Appendix A. Journal of Marketing Research, Volume XIX, 1982; pp. 99-107.

17/...

Wilks, S.S. 1963. Multivariate Statistical Outliers. Sankhya, Volume 25, 1963; pp. 407-426.

*Zellner, A. 1971. An introduction to Bayesian inference in econometrics. New York: Wiley.

```
PAGE   1
BMDP7M  -  STEPWISE DISCRIMINANT ANALYSIS.
BMDP STATISTICAL SOFTWARE, INC.
1964 WESTWOOD BLVD. SUITE 202
(213) 475-5700
PROGRAM REVISED APRIL     1982
MANUAL REVISED -- 1981
COPYRIGHT (C) 1982 REGENTS OF UNIVERSITY OF CALIFORNIA

TO SEE REMARKS AND A SUMMARY OF NEW FEATURES FOR
THIS PROGRAM, STATE NEWS. IN THE PRINT PARAGRAPH.

PROGRAM CONTROL INFORMATION

/PROBLEM TITLE IS 'BANKRUPTCY'.
/INPUT VARIABLES ARE 5.
        FORMAT IS FREE.
/VARIABLE NAMES ARE CFTD,NITA,CACL,CANS,SOLVENCY.
       GROUPING IS SOLVENCY.
/GROUP CODES(5) ARE 1 TO 2.
        NAMES(5) ARE BANKRUPT, SOUND.
/END

PROBLEM TITLE IS
BANKRUPTCY

NUMBER OF VARIABLES TO READ IN. . . . . .          5
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .    0
TOTAL NUMBER OF VARIABLES . . . . . .              5
NUMBER OF CASES TO READ IN. . . . . . TO END
CASE LABELING VARIABLES . . . . . .
MISSING VALUES CHECKED BEFORE OR AFTER TRANS. . NEITHER
BLANKS ARE. . . . . MISSING
INPUT UNIT NUMBER . . . . . .                      5
REWIND INPUT UNIT PRIOR TO READING. . DATA. . .   NO
NUMBER OF WORDS OF DYNAMIC STORAGE. . . . . .  14998

VARIABLES TO BE USED
     1 CFTD        2 NITA        3 CACL        4 CANS        5 SOLVENCY

INPUT FORMAT IS
FREE

MAXIMUM LENGTH DATA RECORD IS   80 CHARACTERS.

TOLERANCE. . . . . .     .010
F-TO-ENTER . . . . . .  4.000    4.000
F-TO-REMOVE. . . . . .  3.996    3.996
METHOD . . . . . .            1
MAXIMUM FORCED LEVEL . . .    0
MAXIMUM NUMBER OF STEPS. .   10
GROUPING VARIABLE. . . . .    5
NUMBER OF GROUPS . . . . .    2
PRIOR PROBABILITIES. . . .  .50000    .50000

                                                                    INTERVAL RANGE
  VARIABLE       MINIMUM   MAXIMUM   MISSING CATEGORY CATEGORY   GREATER LESS THAN
  NO. NAME        LIMIT     LIMIT     CODE     CODE     NAME       THAN    OR = TO

    5 SOLVENCY
                                                  1.00000 BANKRUPT
                                                  2.00000 SOUND
```

NUMBER OF CASES READ. 44

MEANS

GROUP =	BANKRUPT	SOUND	ALL GPS.
VARIABLE			
1 CFTD	-.06756	.23536	.10455
2 NITA	-.08181	.05505	-.00405
3 CACL	1.38430	2.59389	2.07157
4 CANS	.44853	.42658	.43606
COUNTS	19.	25.	44.

STANDARD DEVIATIONS

GROUP =	BANKRUPT	SOUND	ALL GPS.
VARIABLE			
1 CFTD	.21103	.21760	.21481
2 NITA	.13898	.04810	.09798
3 CACL	.42337	1.02272	.82129
4 CANS	.21166	.16174	.18479

COEFFICIENTS OF VARIATION

GROUP =	BANKRUPT	SOUND	ALL GPS.
VARIABLE			
1 CFTD	-3.12349	.92456	2.05462
2 NITA	-1.69896	.87370	-24.22033
3 CACL	.30583	.39428	.39646
4 CANS	.47190	.37915	.42378

STEP NUMBER 0

VARIABLE	F TO REMOVE	FORCE LEVEL	TOLERANCE	*	VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 1	43		*		DF= 1	42	
				*	1 CFTD	21.468	1	1.000000
				*	2 NITA	21.061	1	1.000000
				*	3 CACL	23.417	1	1.000000
				*	4 CANS	.152	1	1.000000

STEP NUMBER 1
VARIABLE ENTERED 3 CACL

VARIABLE	F TO REMOVE DF= 1 42	FORCE LEVEL	TOLERANCE	*	VARIABLE	F TO ENTER DF= 1 41	FORCE LEVEL	TOLERANCE
3 CACL	23.417	1	1.000000	*	1 CFTD	6.330	1	.884536
				*	2 NITA	8.149	1	.950823
				*	4 CANS	1.329	1	.954479

U-STATISTIC OR WILKS' LAMBDA .6420356 DEGREES OF FREEDOM 1 1 42
APPROXIMATE F-STATISTIC 23.417 DEGREES OF FREEDOM 1.00 42.00

F - MATRIX DEGREES OF FREEDOM = 1 42

	BANKRUPT
SOUND	23.42

CLASSIFICATION FUNCTIONS

GROUP =	BANKRUPT	SOUND
VARIABLE		
3 CACL	2.05230	3.84559
CONSTANT	-2.11365	-5.68067

STEP NUMBER 2
VARIABLE ENTERED 2 NITA

VARIABLE	F TO REMOVE DF= 1 41	FORCE LEVEL	TOLERANCE	*	VARIABLE	F TO ENTER DF= 1 40	FORCE LEVEL	TOLERANCE
2 NITA	8.149	1	.950823	*	1 CFTD	.192	1	.355158
3 CACL	9.985	1	.950823	*	4 CANS	1.989	1	.936181

U-STATISTIC OR WILKS' LAMBDA .5355808 DEGREES OF FREEDOM 2 1 42
APPROXIMATE F-STATISTIC 17.776 DEGREES OF FREEDOM 2.00 41.00

F - MATRIX DEGREES OF FREEDOM = 2 41

	SOUND
BANKRUPT	17.78

CLASSIFICATION FUNCTIONS

GROUP =	BANKRUPT	SOUND
VARIABLE		
2 NITA	-12.97369	-1.48693
3 CACL	2.39554	3.88493
CONSTANT	-2.88188	-5.69076

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - BANKRUPT	SOUND
BANKRUPT	84.2	16	3
SOUND	96.0	1	24
TOTAL	90.9	17	27

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - BANKRUPT	SOUND
BANKRUPT	78.9	15	4
SOUND	92.0	2	23
TOTAL	86.4	17	27

SUMMARY TABLE

STEP NUMBER	VARIABLE ENTERED	VARIABLE REMOVED	F VALUE TO ENTER OR REMOVE	NUMBER OF VARIABLES INCLUDED	U-STATISTIC	APPROXIMATE F-STATISTIC	DEGREES OF FREEDOM	
1	3 CACL		23.4169	1	.6420	23.417	1.00	42.00
2	2 NITA		6.1494	2	.5356	17.776	2.00	41.00

	INCORRECT CLASSIFICATIONS	MAHALANOBIS D-SQUARE FROM AND POSTERIOR PROBABILITY FOR GROUP -			
GROUP	BANKRUPT	BANKRUPT		- SOUND	
CASE					
1		11.4	.997	23.2	.003
2		6.0	.984	14.2	.016
3		1.5	.756	3.7	.244
4		.0	.847	3.4	.153
5		.1	.822	3.1	.178
6		.8	.924	5.8	.076
7		1.0	.598	1.8	.402
8		.1	.804	2.9	.196
9		.6	.701	2.3	.299
10		.4	.912	5.1	.088
11		5.5	.997	16.9	.003
12		1.2	.650	2.4	.350
13	SOUND	1.3	.400	.5	.600
14		2.3	.976	9.7	.024
15	SOUND	1.9	.364	.8	.636
16	SOUND	3.9	.194	1.1	.806
17		.0	.819	3.0	.181
18		1.2	.503	1.3	.497
19		.8	.725	2.9	.275
GROUP	SOUND	BANKRUPT		SOUND	
CASE					
20		4.4	.115	.3	.885
21		1.4	.400	.6	.600
22		7.4	.036	.8	.964
23		2.3	.252	.2	.748
24		12.8	.013	4.1	.987
25		14.2	.012	5.4	.988
26		3.0	.180	.0	.820
27		1.5	.373	.5	.627
28		3.4	.170	.2	.830
29		2.4	.339	1.1	.661
30		2.3	.266	.3	.734
31		2.0	.315	.5	.685
32	BANKRUPT	1.9	.923	6.8	.077
33		3.9	.128	.0	.872
34		2.5	.237	.2	.763
35		2.5	.234	.1	.766
36		2.1	.351	.9	.649
37		4.2	.132	.5	.868
38		4.2	.333	2.8	.667
39		3.4	.478	3.3	.522
40		14.0	.008	4.3	.992
41		2.9	.255	.7	.745
42		7.0	.042	.8	.958
43		2.6	.223	.1	.777
44		20.1	.006	9.8	.994

EIGENVALUES

.86713

CUMULATIVE PROPORTION OF TOTAL DISPERSION

1.00000

CANONICAL CORRELATIONS

.68148

VARIABLE	COEFFICIENTS FOR CANONICAL VARIABLES
2 NITA	-6.25390
3 CACL	-.81089
CONSTANT	1.65451

GROUP	CANONICAL VARIABLES EVALUATED AT GROUP MEANS
SOUND	1.04360
BANKRUPT	-.79314

POINTS TO BE PLOTTED

GROUP	MEAN COORDINATES		SYMBOL FOR CASES	SYMBOL FOR MEAN
BANKRUPT	1.04	.00	S	1
SOUND	-.79	.00	B	2

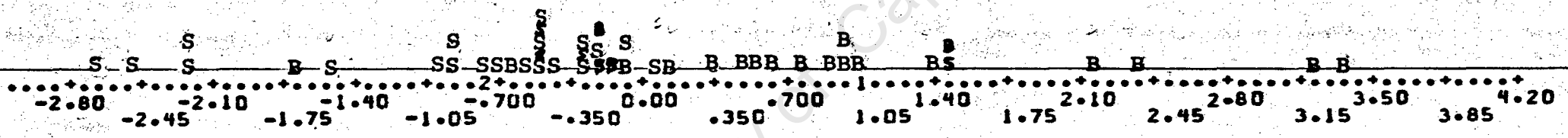
GROUP BANKRUPT

CASE	CAN.V	CASE	CAN.V
1	3.34	11	3.24
2	2.37	12	.46
3	.74	13	-.10
4	1.06	14	2.14
5	.96	15	-.18
6	1.49	16	-.65
7	.34	17	.95
8	.89	18	.13
9	.59	19	.65
10	1.40		

GROUP SOUND

CASE	CAN.V	CASE	CAN.V	CASE	CAN.V
20	-.99	30	-.43	40	-2.52
21	-.09	31	-.30	41	-.46
22	-1.67	32	1.48	42	-1.57
23	-.47	33	-.92	43	-.55
24	-2.23	34	-.51	44	-2.68
25	-2.27	35	-.52		
26	-.70	36	-.21		
27	-.16	37	-.90		
28	-.74	38	-.25		
29	-.24	39	.08		

HISTOGRAM OF CANONICAL VARIABLE



NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM 3807
CPU TIME USED 4.289 SECONDS

BMDP7M - STEPWISE DISCRIMINANT ANALYSIS.
BMDP STATISTICAL SOFTWARE, INC.
1964 WESTWOOD BLVD. SUITE 202
(213) 475-5700
PROGRAM REVISED APRIL 1982
MANUAL REVISED -- 1981

TO SEE REMARKS AND A SUMMARY OF NEW FEATURES FOR
THIS PROGRAM, STATE NEWS. IN THE PRINT PARAGRAPH.

PROGRAM CONTROL INFORMATION

```
/PROBLEM TITLE IS 'ADVERTISING PREFERENCE AREA'.
/INPUT VARIABLES ARE 9.
      FORMAT IS FREE.
/VARIABLE NAMES ARE CIN,ENGDAY,AFRDAY,ANYW,MAGA,TELV,RADV,SPRI,CLAGRO.
       GROUPING IS CLAGRO.
/GROUP CODES(9) ARE 1 TO 4.
      NAMES(9) ARE EURO,COLOU,AFRI,ASIAN.
/DISCRIMINANT
      JACKKNIFE.
/PRINT
      WITHIN.
/END
```

PROBLEM TITLE IS
ADVERTISING PREFERENCE AREA

NUMBER OF VARIABLES TO READ IN.	9
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS.	0
TOTAL NUMBER OF VARIABLES	9
NUMBER OF CASES TO READ IN.	TO END
CASE LABELING VARIABLES	
MISSING VALUES CHECKED BEFORE OR AFTER TRANS.	NEITHER
BLANKS ARE.	MISSING
INPUT UNIT NUMBER	5
REWIND INPUT UNIT PRIOR TO READING. DATA.	NO
NUMBER OF WORDS OF DYNAMIC STORAGE.	14998

VARIABLES TO BE USED

1 CIN	2 ENGDAY	3 AFRDAY	4 ANYW	5 MAGA
6 TELV	7 RADV	8 SPRI	9 CLAGRO	

INPUT FORMAT IS
FREE

MAXIMUM LENGTH DATA RECORD IS 80 CHARACTERS.

TOLERANCE.	.010			
F-TO-ENTER	4.000	4.000		
F-TO-REMOVE.	3.996	3.996		
METHOD	1			
MAXIMUM FORCED LEVEL	0			
MAXIMUM NUMBER OF STEPS.	18			
GROUPING VARIABLE.	9			
NUMBER OF GROUPS	4			
PRIOR PROBABILITIES.	.25000	.25000	.25000	.25000

VARIABLE	MINIMUM	MAXIMUM	MISSING CATEGORY	CATEGORY	INTERVAL RANGE GREATER LESS THAN

NO.	NAME	LIMIT	LIMIT	CODE	CODE	NAME	THAN	OR = TO
9	CLAGRO							
					1.00000	EURO		
					2.00000	COLOU		
					3.00000	AFRI		
					4.00000	ASIAN		

NUMBER OF CASES READ. 92

MEANS

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN	ALL GPS.
1 CIN	23.63462	11.72727	28.28889	10.28462	17.92500
2 ENGDAY	37.14231	27.46364	58.36667	10.30769	31.39674
3 AFRDAY	29.98462	12.15909	.65556	1.21538	11.85326
4 ANYW	79.37692	57.99091	77.03333	23.31923	57.96196
5 MAGA	93.43077	55.63182	60.60556	33.28462	60.97174
6 TELV	59.96154	29.18636	41.43333	3.57692	33.04239
7 RADV	15.10000	9.32273	26.88889	1.92692	12.30217
8 SPRI	31.53462	28.84545	36.31111	2.35769	23.52174
COUNTS	26.	22.	18.	26.	92.

STANDARD DEVIATIONS

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN	ALL GPS.
1 CIN	11.16179	7.54303	14.26934	10.22263	10.86250
2 ENGDAY	15.36097	15.40982	14.87236	7.14680	13.45142
3 AFRDAY	10.62502	6.23704	.84660	1.09643	6.46795
4 ANYW	4.66993	13.10256	11.12787	10.38148	10.08481
5 MAGA	2.09757	13.23881	11.76738	11.99292	10.52073
6 TELV	6.57164	13.24539	8.93881	2.38232	8.43702
7 RADV	9.61345	4.86287	11.92862	1.15466	7.46819
8 SPRI	4.96129	9.09425	5.85902	2.22066	5.89591

COEFFICIENTS OF VARIATION

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN	ALL GPS.
1 CIN	.47226	.64323	.50441	.99397	.60600
2 ENGDAY	.41357	.56113	.25481	.69335	.42843
3 AFRDAY	.35435	.51295	1.29142	.90212	.54567
4 ANYW	.05883	.22594	.14446	.44519	.17399
5 MAGA	.02245	.23797	.19416	.36031	.17255
6 TELV	.10960	.45382	.21574	.66602	.25534
7 RADV	.63665	.52162	.41016	.59923	.60706
8 SPRI	.15733	.31528	.16270	.94188	.25066

WITHIN COVARIANCE MATRIX

		CIN 1	ENGDAY 2	AFRDAY 3	ANYW 4	MAGA 5	TELV 6	RADV 7	SPRI 8
CIN	1	117.99432							
ENGDAY	2	68.77939	180.94083						
AFRDAY	3	.11711	-39.79859	41.83438					
ANYW	4	51.24705	96.88607	-1.44623	101.70346				
MAGA	5	60.21884	59.59827	2.65093	76.30783	110.68575			
TELV	6	-9.47966	43.73734	-2.30502	37.36813	29.65213	71.18334		
RADV	7	58.16221	48.14007	-5.29601	33.01884	35.26939	-14.99194	55.77395	
SPRI	8	-3.55629	10.43424	-8.47772	10.57266	17.41446	25.50587	-3.55058	34.76176

STEP NUMBER 0

VARIABLE	F TO REMOVE	FORCE LEVEL	TOLERANCE	*	VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 3 89			*		DF= 3 88		
				*	1 CIN	14.531	1	1.000000

*	2	ENGDAY	47.630	1	1.000000
*	3	AFRDAY	109.549	1	1.000000
*	4	ANYW	162.805	1	1.000000
*	5	MAGA	144.415	1	1.000000
*	6	TELV	201.399	1	1.000000
*	7	RADV	42.500	1	1.000000
*	8	SPRI	160.583	1	1.000000

STEP NUMBER 1
VARIABLE ENTERED 6 TELV

VARIABLE	F TO REMOVE DF= 3 88	FORCE LEVEL	TOLERANCE
6 TELV	201.399	1	1.000000

VARIABLE	F TO ENTER DF= 3 87	FORCE LEVEL	TOLERANCE
1 CIN	7.768	1	.989301
2 ENGDAY	24.673	1	.871154
3 AFRDAY	51.220	1	.998216
4 ANYW	21.308	1	.857119
5 MAGA	14.513	1	.888406
7 RADV	30.253	1	.943388
8 SPRI	55.979	1	.737094

U-STATISTIC OR WILKS' LAMBDA .1271317 DEGREES OF FREEDOM 1 3 88
APPROXIMATE F-STATISTIC 201.399 DEGREES OF FREEDOM 3.00 88.00

F - MATRIX DEGREES OF FREEDOM = 1 88

	EURO	COLOU	AFRI
COLOU	158.55		
AFRI	51.30	20.86	
ASIAN	580.61	109.79	214.14

CLASSIFICATION FUNCTIONS

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN
6 TELV	.84235	.41002	.58207	.05025
CONSTANT	-26.64070	-7.36974	-13.44474	-1.47616

STEP NUMBER 2
VARIABLE ENTERED 8 SPRI

VARIABLE	F TO REMOVE DF= 3 87	FORCE LEVEL	TOLERANCE
6 TELV	74.242	1	.737094
8 SPRI	55.979	1	.737094

VARIABLE	F TO ENTER DF= 3 86	FORCE LEVEL	TOLERANCE
1 CIN	7.308	1	.989292
2 ENGDAY	17.552	1	.867417
3 AFRDAY	40.881	1	.943594
4 ANYW	11.012	1	.804075
5 MAGA	14.045	1	.872151
7 RADV	20.465	1	.941067

U-STATISTIC OR WILKS' LAMBDA .0433852 DEGREES OF FREEDOM 2 3 88
APPROXIMATE F-STATISTIC 110.228 DEGREES OF FREEDOM 6.00 174.00

F - MATRIX DEGREES OF FREEDOM = 2 87

	EURO	COLOU	AFRI
COLOU	94.36		
AFRI	50.71	11.78	
ASIAN	307.20	123.17	188.66

CLASSIFICATION FUNCTIONS

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN
6 TELV	.70182	.15288	.28609	.03520
8 SPRI	.39222	.71763	.82603	.04200
CONSTANT	-28.61153	-13.96750	-22.18619	-1.49876

STEP NUMBER 3
VARIABLE ENTERED 3 AFRDAY

VARIABLE	F TO REMOVE DF= 3 86	FORCE LEVEL	TOLERANCE	VARIABLE	F TO ENTER DF= 3 85	FORCE LEVEL	TOLERANCE
3 AFRDAY	40.881	1	.943594	1 CIN	6.828	1	.989280
6 TELV	27.743	1	.731679	2 ENGDAY	7.623	1	.646506
8 SPRI	45.007	1	.696761	4 ANYW	10.769	1	.803785
				5 MAGA	6.318	1	.864874
				7 RADV	18.381	1	.928608

U-STATISTIC OR WILKS' LAMBDA .0178829 DEGREES OF FREEDOM 3 3 88
APPROXIMATE F-STATISTIC 98.318 DEGREES OF FREEDOM 9.00 209.45

F - MATRIX DEGREES OF FREEDOM = 3 86

	EURO	COLOU	AFRI
COLOU	88.41		
AFRI	95.57	16.43	
ASIAN	321.23	109.40	128.23

CLASSIFICATION FUNCTIONS

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN
3 AFRDAY	.88481	.47107	.21071	.04186
6 TELV	.63579	.11773	.27037	.03208
8 SPRI	.65645	.85831	.88895	.05450
CONSTANT	-44.06353	-18.34737	-23.06253	-1.53335

STEP NUMBER 4
VARIABLE ENTERED 7 RADV

VARIABLE	F TO REMOVE	FORCE LEVEL	TOLERANCE	*	VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 3 85			*		DF= 3 84		
3 AFRDAY	37.793	1	.931130	*	1 CIN	1.516	1	.473450
6 TELV	29.362	1	.695880	*	2 ENGDAY	2.487	1	.360787
7 RADV	18.381	1	.928638	*	4 ANYW	4.624	1	.485492
8 SPRI	37.858	1	.696438	*	5 MAGA	11.928	1	.560378

U-STATISTIC OR WILKS' LAMBDA .0108463 DEGREES OF FREEDOM 4 3 88
APPROXIMATE F-STATISTIC 84.974 DEGREES OF FREEDOM 12.00 225.18

F - MATRIX DEGREES OF FREEDOM = 4 85

	EURO	COLOU	AFRI
COLOU	78.07		
AFRI	71.47	28.10	
ASIAN	288.48	91.69	147.09

CLASSIFICATION FUNCTIONS

GROUP =	EURO	COLOU	AFRI	ASIAN
VARIABLE				
3 AFRDAY	.96588	.51367	.30093	.04910
6 TELV	.77430	.19035	.42415	.04441
7 RADV	.61107	.32109	.67992	.05451
8 SPRI	.63723	.84821	.86757	.05278
CONSTANT	-53.73320	-21.01724	-35.03400	-1.61029

STEP NUMBER 5
VARIABLE ENTERED 5 MAGA

VARIABLE	F TO REMOVE DF= 3 84	FORCE LEVEL	TOLERANCE	*	VARIABLE	F TO ENTER DF= 3 83	FORCE LEVEL	TOLERANCE
3 AFRDAY	19.131	1	.895958	*	1 CIN	1.238	1	.424373
5 MAGA	11.928	1	.560378	*	2 ENGDAY	2.846	1	.355553
6 TELV	17.615	1	.582637	*	4 ANYW	12.532	1	.355724
7 RADV	25.829	1	.691673	*				
8 SPRI	38.272	1	.675433	*				

U-STATISTIC OR WILKS' LAMBDA .3076061 DEGREES OF FREEDOM 5 3 88
APPROXIMATE F-STATISTIC 75.186 DEGREES OF FREEDOM 15.00 232.29

F - MATRIX DEGREES OF FREEDOM = 5 84

	EURO	COLOU	AFRI
COLOU	67.68		
AFRI	80.60	28.52	
ASIAN	229.65	74.02	130.52

CLASSIFICATION FUNCTIONS

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN
3 AFRDAY	.82379	.42291	.26812	-.06772
5 MAGA	.56864	.36324	.13127	.46748
6 TELV	.49343	.01113	.35938	-.18625
7 RADV	.15235	.02807	.57462	-.32261
8 SPRI	.47672	.74569	.83051	-.07918
CONSTANT	-63.76137	-25.10908	-35.56845	-8.38790

STEP NUMBER 6
VARIABLE ENTERED 4 ANYW

VARIABLE	F TO REMOVE	FORCE LEVEL	TOLERANCE	*	VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 3 83			*		DF= 3 82		
3 AFRDAY	18.874	1	.894653	*	1 CIN	1.338	1	.422166
4 ANYW	12.532	1	.355724	*	2 ENGDAY	1.339	1	.222479
5 MAGA	21.532	1	.410593	*				
6 TELV	19.444	1	.445723	*				
7 RADV	12.335	1	.517595	*				
8 SPRI	34.161	1	.644236	*				

U-STATISTIC OR WILKS' LAMBDA .0062349 DEGREES OF FREEDOM 6 3 68
APPROXIMATE F-STATISTIC 75.633 DEGREES OF FREEDOM 18.00 235.24

F - MATRIX DEGREES OF FREEDOM = 6 83

	EURO	COLOU	AFRI
COLOU	64.46		
AFRI	80.64	24.39	
ASIAN	189.50	73.65	126.41

CLASSIFICATION FUNCTIONS

GROUP = VARIABLE	EURO	COLOU	AFRI	ASIAN
3 AFRDAY	.83282	.44545	.29544	-.06147
4 ANYW	.23957	.59865	.72531	.16578
5 MAGA	.45813	.08709	-.20330	.39101
6 TELV	.36943	-.29873	-.01604	-.27206
7 RADV	.05210	-.22245	.27050	-.39198
8 SPRI	.54216	.90920	1.02864	-.03389
CONSTANT	-64.79959	-31.59192	-45.08461	-8.68505

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - EURO	COLOU	AFRI	ASIAN
EURO	100.0	26	0	0	0
COLOU	86.4	0	19	2	1
AFRI	100.0	0	0	18	0
ASIAN	100.0	0	0	0	26
TOTAL	96.7	26	19	20	27

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -			
		EURO	COLOU	AFRI	ASIAN
EURO	96.2	25	1	0	0
COLOU	77.3	0	17	3	2
AFRI	94.4	0	1	17	0
ASIAN	100.0	0	0	0	26
TOTAL	92.4	25	19	20	28

SUMMARY TABLE

STEP NUMBER	VARIABLE ENTERED	VARIABLE REMOVED	F VALUE TO ENTER OR REMOVE	NUMBER OF VARIABLES INCLUDED	U-STATISTIC	APPROXIMATE F-STATISTIC	DEGREES OF FREEDOM	
1	6 TELV		201.3986	1	.1271	201.399	3.00	88.00
2	8 SPRI		55.9787	2	.0434	110.228	6.00	174.00
3	3 AFRDAY		43.6856	3	.0179	98.318	9.00	209.45
4	7 RADV		18.3814	4	.0108	84.974	12.00	225.18
5	5 MAGA		11.9278	5	.0076	75.186	15.00	232.29
6	4 ANYW		12.5323	6	.0052	70.633	18.00	235.24

GROUP		INCORRECT CLASSIFICATIONS	JACKKNIFED MAHALANOBIS D-SQUARE FROM AND POSTERIOR PROBABILITY FOR GROUP -							
			EURO		COLOU		AFRI		ASIAN	
CASE										
1			3.7	1.000	39.1	.000	51.1	.000	97.2	.000
2			19.2	.972	26.9	.021	28.9	.008	67.8	.000
3			.8	1.000	35.5	.000	50.5	.000	89.9	.000
4			20.3	1.000	85.4	.000	101.8	.000	177.7	.000
5			1.9	1.000	31.1	.000	38.5	.000	85.8	.000
6			1.9	1.000	43.9	.000	61.4	.000	104.1	.000
7			6.1	1.000	42.7	.000	69.6	.000	107.7	.000
8			5.6	1.000	53.4	.000	80.2	.000	100.6	.000
9			1.4	1.000	41.5	.000	56.8	.000	97.8	.000
10			12.9	1.000	31.6	.000	33.9	.000	86.5	.000
11			14.4	1.000	45.3	.000	44.0	.000	114.1	.000
12			1.9	1.000	34.7	.000	46.0	.000	88.0	.000
13			4.7	1.000	44.4	.000	64.3	.000	93.0	.000
14			.9	1.000	31.9	.000	44.2	.000	89.2	.000
15		COLOU	27.9	.156	25.5	.511	26.4	.333	78.6	.000
16			1.8	1.000	37.4	.000	51.7	.000	89.3	.000
17			33.4	1.000	90.8	.000	129.6	.000	193.9	.000
18			2.2	1.000	32.0	.000	39.7	.000	89.0	.000
19			1.6	1.000	35.9	.000	54.1	.000	89.1	.000
20			3.9	1.000	38.7	.000	60.6	.000	101.9	.000
21			8.8	1.000	41.5	.000	73.6	.000	89.0	.000
22			2.1	1.000	37.7	.000	52.1	.000	94.9	.000
23			10.4	1.000	34.9	.000	39.5	.000	94.4	.000
24			7.6	1.000	35.4	.000	37.2	.000	100.0	.000
25			1.1	1.000	39.0	.000	53.4	.000	91.8	.000
26			3.3	1.000	39.9	.000	57.6	.000	93.3	.000
GROUP	COLOU		EURO		COLOU		AFRI		ASIAN	
CASE										
27			38.7	.000	2.6	.999	16.3	.001	33.6	.000
28		AFRI	49.1	.000	19.2	.004	8.1	.996	91.9	.000
29		AFRI	32.5	.000	10.7	.199	8.0	.801	59.4	.000
30			26.1	.006	16.0	.990	27.4	.003	76.0	.000
31			61.8	.000	7.5	1.000	34.2	.000	37.3	.000
32		ASIAN	92.5	.000	24.2	.005	57.9	.000	13.6	.995
33			40.7	.000	4.6	.989	13.5	.011	40.1	.000
34			35.2	.000	2.0	1.000	16.7	.000	37.6	.000
35			32.2	.000	2.0	.999	15.4	.001	40.2	.000
36			32.4	.000	6.5	.998	19.4	.002	27.2	.000
37			44.7	.000	5.8	.995	16.3	.005	52.1	.000
38			34.2	.000	1.0	1.000	18.5	.000	38.1	.000
39		AFRI	49.2	.000	27.8	.061	22.4	.939	106.4	.000
40			23.1	.000	8.0	.754	10.2	.245	60.3	.000
41			36.6	.000	10.1	1.000	31.2	.000	77.4	.000
42			63.9	.000	15.4	1.000	58.1	.000	43.3	.000
43		ASIAN	103.7	.000	24.8	.060	56.7	.000	19.3	.940
44			32.4	.000	2.4	.999	16.8	.001	42.7	.000
45			34.7	.000	5.1	1.000	24.9	.000	42.6	.000
46			29.3	.000	8.5	1.000	28.8	.000	38.1	.000

47		25.9	.000	2.3	1.000	22.5	.000	36.8	.000
48		42.5	.000	3.3	.998	15.9	.002	54.9	.000
GROUP AFRI		EURO		COLOU		AFRI		ASIAN	
CASE									
49		60.5	.000	21.2	.000	2.8	1.000	92.0	.000
50		39.0	.000	19.6	.140	15.9	.860	55.1	.000
51		54.3	.000	23.1	.000	2.5	1.000	87.6	.000
52		90.7	.000	20.0	.067	14.7	.933	83.8	.000
53		61.5	.000	19.3	.001	4.7	.999	85.3	.000
54		50.7	.000	28.5	.000	7.1	1.000	94.4	.000
55		49.5	.000	24.5	.000	6.3	1.000	88.3	.000
56		66.2	.000	27.7	.000	6.3	1.000	110.7	.000
57		55.1	.000	21.6	.005	11.0	.995	73.1	.000
58		51.2	.000	19.2	.000	1.6	1.000	78.3	.000
59		34.2	.000	13.4	.072	8.3	.928	63.1	.000
60		45.4	.000	23.0	.000	3.1	1.000	82.8	.000
61	COLOU	81.4	.000	12.8	.807	15.7	.193	53.1	.000
62		58.7	.000	17.0	.001	3.0	.999	83.3	.000
63		49.4	.000	49.1	.000	22.4	1.000	89.8	.000
64		43.4	.000	25.3	.000	7.5	1.000	75.4	.000
65		61.3	.000	26.0	.000	7.1	1.000	110.7	.000
66		49.6	.000	19.7	.003	7.8	.997	76.1	.000
GROUP ASIAN		EURO		COLOU		AFRI		ASIAN	
CASE									
67		93.4	.000	36.9	.000	67.0	.000	1.8	1.000
68		95.0	.000	24.7	.096	56.5	.000	20.3	.904
69		91.5	.000	41.5	.000	78.0	.000	.2	1.000
70		92.2	.000	48.9	.000	91.1	.000	2.5	1.000
71		84.5	.000	31.7	.000	65.7	.000	5.3	1.000
72		83.3	.000	41.7	.000	82.5	.000	2.8	1.000
73		84.3	.000	46.8	.000	89.2	.000	4.5	1.000
74		99.4	.000	38.3	.000	71.2	.000	1.9	1.000
75		102.8	.000	36.0	.000	62.6	.000	7.6	1.000
76		84.2	.000	34.9	.000	72.6	.000	2.4	1.000
77		85.6	.000	38.7	.000	77.7	.000	2.4	1.000
78		82.8	.000	33.8	.000	71.6	.000	3.9	1.000
79		95.1	.000	35.5	.000	67.1	.000	1.7	1.000
80		107.4	.000	47.7	.000	79.1	.000	4.4	1.000
81		101.8	.000	37.8	.000	72.9	.000	4.0	1.000
82		102.6	.000	50.2	.000	86.2	.000	2.3	1.000
83		107.4	.000	58.7	.000	95.1	.000	6.4	1.000
84		87.7	.000	37.8	.000	76.3	.000	2.6	1.000
85		83.5	.000	45.0	.000	83.5	.000	2.4	1.000
86		100.8	.000	54.1	.000	94.5	.000	2.9	1.000
87		110.1	.000	49.1	.000	81.9	.000	3.7	1.000
88		107.6	.000	47.2	.000	78.2	.000	4.4	1.000
89		96.8	.000	49.8	.000	93.2	.000	3.0	1.000
90		97.4	.000	51.9	.000	95.0	.000	2.9	1.000
91		99.8	.000	42.1	.000	74.7	.000	1.4	1.000
92		103.3	.000	42.2	.000	75.8	.000	1.5	1.000

EIGENVALUES

14.20638 5.83113 .83897

CUMULATIVE PROPORTION OF TOTAL DISPERSION

.68050 .95981 1.00000

CANONICAL CORRELATIONS

	.96656	.92391	.67544

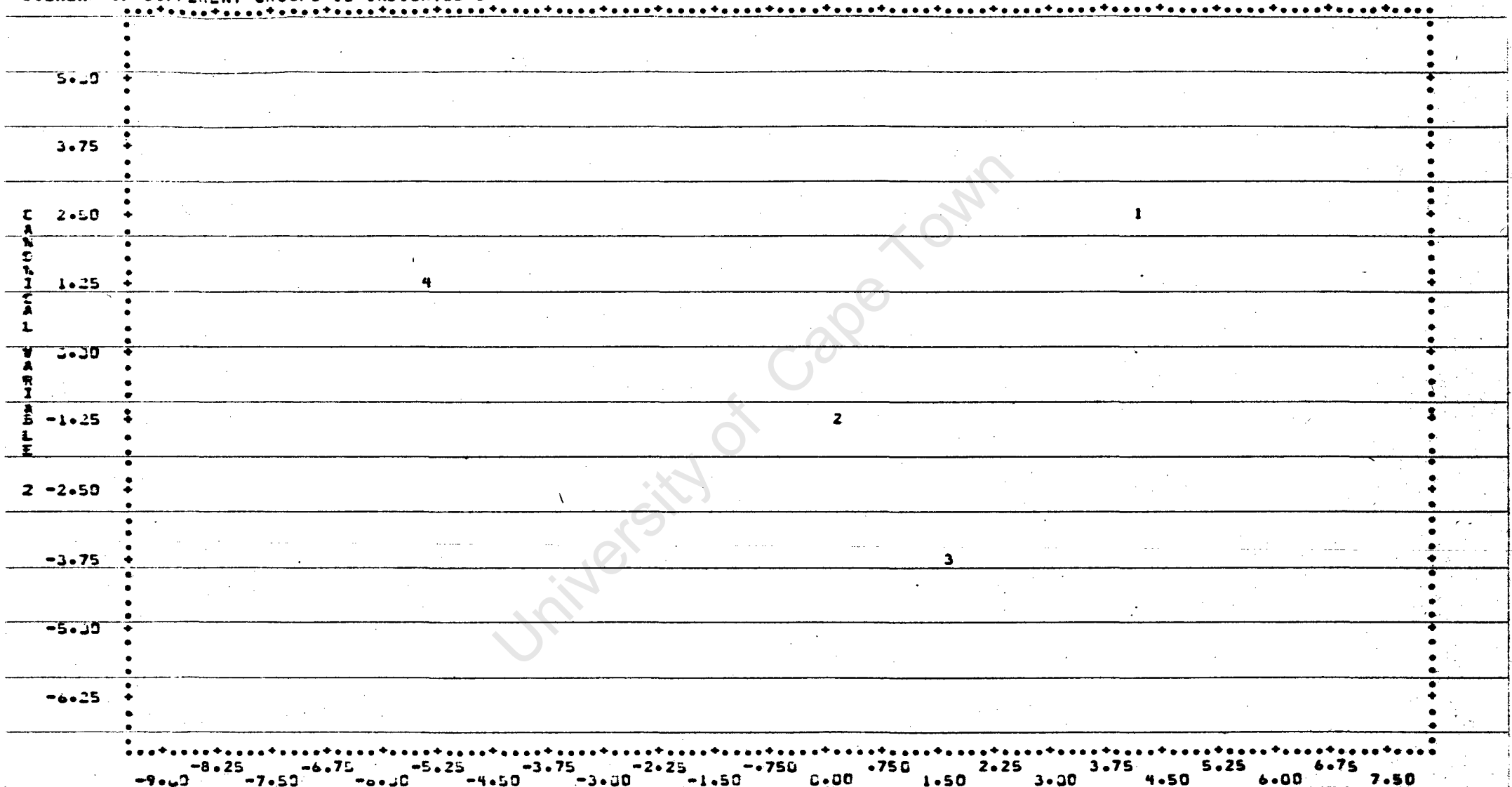
VARIABLE	COEFFICIENTS FOR CANONICAL VARIABLES		
3 AFRDAY	.08636	.04229	.07163
4 ANYW	.02075	-.08945	.05509
5 MAGA	-.00931	.10577	.00053
6 TELV	.06135	.05153	-.12585
7 RADV	.05468	-.04019	-.11929
8 SPRI	.07704	-.11887	.12444
CONSTANT	-6.17106	-.17725	-1.37523

GROUP	CANONICAL VARIABLES EVALUATED AT GROUP MEANS		
EURO	4.12999	2.60633	-.22930
COLOU	.08735	-1.26625	1.52389
AFRI	1.70671	-3.85677	-.99347
ASIAN	-5.38547	1.13518	-.37267

POINTS TO BE PLOTTED

GROUP	MEAN COORDINATES		SYMBOL FOR CASES	SYMBOL FOR MEAN
EURO	4.13	2.61	A	1
COLOU	.09	-1.27	B	2
AFRI	1.71	-3.86	C	3
ASIAN	-5.39	1.14	D	4

OVERLAP OF DIFFERENT GROUPS IS INDICATED BY *



POINTS TO BE PLOTTED

GROUP	MEAN COORDINATES		SYMBOL FOR CASES	SYMBOL FOR MEAN
EURO	4.13	2.61	A	1
COLOU	.09	-1.27	B	2
AFRI	1.71	-3.86	C	3
ASIAN	-5.39	1.14	D	4

GROUP EURO

CASE	X	Y	CASE	X	Y	CASE	X	Y
1	4.24	2.60	11	4.76	1.28	21	3.36	3.95
2	2.37	.84	12	3.87	2.50	22	4.17	2.78
3	3.96	2.84	13	3.82	3.62	23	3.99	1.42
4	6.51	3.73	14	4.00	2.32	24	4.41	1.27
5	3.84	1.94	15	2.85	-.03	25	4.02	3.05
6	4.56	3.30	16	3.87	2.93	26	3.98	3.18
7	4.48	3.41	17	6.29	4.08			
8	4.02	4.50	18	3.99	1.95			
9	4.30	3.15	19	3.84	3.36			
10	3.55	1.04	20	4.37	3.05			

GROUP COLOU

CASE	X	Y	CASE	X	Y	CASE	X	Y
27	-.42	-1.30	37	.34	-2.10	47	.24	-.14
28	1.99	-3.43	38	.00	-1.52	48	.58	-2.14
29	1.27	-1.96	39	2.30	-2.61			
30	2.03	-.59	40	1.72	-1.27			
31	-1.29	-1.53	41	1.67	-1.18			
32	-3.69	-.75	42	-1.24	-.33			
33	-.10	-1.75	43	-3.66	-1.47			
34	-.07	-1.05	44	.35	-1.12			
35	.24	-1.14	45	.15	-.68			
36	-.56	-.41	46	.09	.10			

GROUP AFRI

CASE	X	Y	CASE	X	Y
49	2.06	-4.82	59	1.55	-2.20
50	.83	-1.78	60	2.20	-3.54
51	2.09	-4.29	61	-.91	-4.35
52	.47	-5.57	62	1.93	-4.20
53	1.66	-4.64	63	2.14	-2.42
54	2.51	-3.83	64	1.84	-2.91
55	2.27	-3.74	65	2.86	-4.93
56	2.68	-5.23	66	1.62	-3.46
57	1.23	-3.50			
58	1.73	-4.00			

GROUP ASIAN

CASE	X	Y	CASE	X	Y	CASE	X	Y
67	-5.27	.39	77	-5.03	1.68	87	-6.03	.63

68	-4.38	-1.10	78	-4.76	1.28	88	-5.85	.48
69	-5.41	1.44	79	-5.34	.24	89	-5.64	2.24
70	-5.46	2.35	80	-5.86	.56	90	-5.72	2.32
71	-4.73	.66	81	-5.58	.24	91	-5.66	.65
72	-4.97	2.22	82	-5.86	1.42	92	-5.79	.53
73	-5.01	2.65	83	-5.94	1.77			
74	-5.56	.33	84	-5.13	1.40			
75	-5.31	-.66	85	-5.01	2.36			
76	-4.93	1.33	86	-5.86	2.13			

OVERLAP OF DIFFERENT GROUPS IS INDICATED BY *