DISCRIMINANT ANALYSIS : A REVIEW AND ITS APPLICATION TO THE CLASSIFICATION OF GRAPE CULTIVARS

BY

RENETTE JULIA BLIGNAUT

THESIS

Submitted in fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN THE DEPARTMENT
OF MATHEMATICAL STATISTICS OF THE
UNIVERSITY OF CAPE TOWN

SUPERVISORS : Prof. W. Zucchini
Prof. T.J. Stewart

JUNE, 1989

To Jacques

PREFACE

The aim of this study was to calculate a classification function for discriminating between five grape cultivars with a view to determine the cultivar of an unknown grape juice.

In order to discriminate between the five grape cultivars various multivariate statistical techniques, such as principal component analysis, cluster analysis, correspondence analysis and discriminant analysis were applied. Discriminant analysis resulted in the most appropriate technique for the problem at hand and therefore an in depth study of this technique was undertaken. Discriminant analysis was the most appropriate technique for classifying these grape samples into distinct cultivars because this technique utilized prior information of population membership.

This thesis is divided into two main sections. The first section (chapters 1 to 5) is a review on discriminant analysis, describing various aspects of this technique and matters related thereto.

In the second section (chapter 6) the theories discussed in the first section are applied to the problem at hand. The results obtained when discriminating between the different grape cultivars are given.

Chapter 1 gives a general introduction to the subject of discriminant analysis, including certain basic derivations used in this study.

Two approaches to discriminant analysis are discussed in Chapter 2, namely the parametrical and non-parametrical approaches. In this review the emphasis is placed on the classical approach to discriminant analysis. Non-parametrical approaches such as the K-nearest neighbour technique, the kernel method and ranking are briefly discussed.

Chapter 3 deals with estimating the probability of misclassification.

In Chapter 4 variable selection techniques are discussed.

Chapter 5 briefly deals with sequential and logistical discrimination techniques. The estimation of missing values is also discussed in this chapter.

A final summary and conclusion is given in Chapter 7.

Appendices A to D illustrate some of the obtained results from the practical analyses.

Finally, a list of references are given.

ACKNOWLEDGEMENTS

It is impossible to thank everyone who eased the way for me while I wrote this thesis. At the risk of slighting those worthy of specific distinction, I would however like to mention the following people by name :

i) my supervisors, Professors Walter Zucchini and Theo Stewart who guided me past the pitfalls of this thesis and whose constructive comments helped me master the subject;

ii) the Wine Research Department of Distillers Corporation (were it all started) for providing the data used in the practical application;

iii) my colleagues at the MRC for their support and encouragement;

iv) the Medical Research Council for the use of their computer facilities;

v) my parents and family for providing an intellectually stimulating environment in which to grow and learn.

Finally I wish to thank my husband, Jacques, who lovingly supported me through this thesis and whose valuable suggestions made the entire work more digestible (hopefully).

CONTENTS

CHAPTER 1

1. INTRODUCTION

Consider a situation in which it is sought to determine to which population of G mutually exclusive and exhaustive populations, $\pi_1, \pi_2, \ldots, \pi_G$, a particular item of unknown origin belongs. In the application under consideration the items are grape juice samples and the populations are cultivars.

The information available to obtain an answer consists of a number of measurements, say $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, on each of the items in G random samples taken from the G populations. Since the population membership of each of the items is known, these measurements can be examined to establish patterns that will assist in identifying the population to which an item of unknown origin belongs.

The purpose of discriminant analysis is to partition the sample space, $\Omega = R^m$, of measurements X into G regions, $\Omega_1, \Omega_2, \ldots, \Omega_G$, which will prescribe the classification rule : Namely, if $X \in \Omega_i$, then the corresponding item is classified as belonging to population π_i, $i=1,2,\ldots,G$.

In most applications there is no perfect classification rule, i.e. there is no partition of Ω which classifies all the items correctly. Each partition will be such that a certain proportion of items will be misclassified.

The above reference is to all the items in the population, not only those contained in the (training) sample. The training sample is the items of known population origin used to construct the classification rules. Since the number of classification rules in any application is infinite and since each of them will have different properties it is necessary to find some way to compare the different rules. A reasonable possibility will be to specify the (relative) expected cost associated with each type of misclassification.

Denote by C_{ij} the cost incurred if an item from population π_i is classified as belonging to population π_j, $i,j=1,2,\ldots,G$. Without loss of generality, assume $C_{ii}=0$, $i=1,2,\ldots,G$. In some applications it is inappropriate, or difficult, to determine the C_{ij} and the probabilities of misclassification are used to measure the quality of the classification rule. Essentially, this is a special case with $C_{ij}=1$ for $i,j=1,2,\ldots,G$, $i\neq j$.

To determine the classification rule associated with the minimum expected cost one needs to know:

i) the probability density function (pdf) of **X** in each population, denoted by $f_1(\mathbf{X}),f_2(\mathbf{X}),\ldots,f_G(\mathbf{X})$, and

ii) the probabilities are $P(\pi_1),P(\pi_2),\ldots,P(\pi_G)$, where $P(\pi_i)$ represents the prior probability that the

item to be classified belongs to population π_i, $i=1,2,\ldots,G$. The term 'prior' in this context means 'before any measurements are made', i.e. before X is determined.

It can be shown (see Anderson, 1958) that if $X \in \pi_i$, the expected cost associated with the rule defined by $\Omega_1,\ldots,\Omega_G$ is:

$$r_i = \sum_{j=1}^{G} C_{ij} \int_{\Omega_j} f_i(X)\, dX \qquad i=1,2\ldots,G.$$

The overall expected cost is thus

$$r = \sum_{i=1}^{G} r_i\, P(\pi_i)$$

$$= \sum_{i=1}^{G} \sum_{j=1}^{G} \int_{\Omega_i} C_{ij}\, P(\pi_i)\, f_i(X)\, dX$$

$$= \sum_{i=1}^{G} \int_{\Omega_i} \left[\sum_{j=1}^{G} C_{ij}\, P(\pi_i)\, f_i(X) \right] dX \,. \qquad 1.1$$

Thus the classification rule with the minimum expected cost assigns item X to population π_i, if

$$\sum_{k=1}^{G} P(\pi_k)\ C_{ik}\ f_k(X) < \sum_{k=1}^{G} P(\pi_k)\ C_{jk}\ f_k(X)\ ,\ j=1,\ldots,G;\ j\neq i$$

where $C_{jj} = 0$, $j=1,\ldots,G$. 1.2

This rule generates the partitions $\Omega_1, \Omega_2, \ldots, \Omega_G$ associated with the minimum expected cost of misclassification. When the costs of misclassification are all equal, item X is assigned to population π_i, if

$$P(\pi_i)\, f_i(X) = \max_{j=1,\ldots,G} P(\pi_j)\, f_j(X) \quad . \qquad 1.3$$

In the two population situation (G=2) rules (1.2) and (1.3) reduce to (1.4) and (1.5) respectively :

Assign item X to population π_1, if

$$P(\pi_2)\, C_{12}\, f_2(X) < P(\pi_1)\, C_{21}\, f_1(X)$$

and to population π_2, if

$$P(\pi_2)\, C_{12}\, f_2(X) \geq P(\pi_1)\, C_{21}\, f_1(X). \qquad 1.4$$

Assign item X to population π_1, if

$$P(\pi_2)\, f_2(X) < P(\pi_1)\, f_1(X)$$

and to population π_2, if

$$P(\pi_2)\, f_2(X) \geq P(\pi_1)\, f_1(X) \quad . \qquad 1.5$$

In practice the probability density functions are usually unknown and the consequences of this need to be considered.

These densities therefore need to be estimated by using either parametrical methods, such as the classical and Bayesian approaches, or non-parametrical methods, such as kernel densities, ranking of measurements and the K-nearest neighbour method.

Parametrical methods to estimate the unknown density functions $f_i(X)$, $i=1,2,..,G$, begin by postulating parametric models, $f_i(X,\theta_i)$, for these densities, where θ_i is a vector of unknown parameters for population π_i, $i=1,2,..,G$. Although it is not essential, it is usually the case that one uses the same family of models for all the densities but allows the parameters to vary over the different populations. It can be assumed that this will generally be the case but note that no special difficulties arise if different families of models are used for the different populations.

There are two basic approaches to estimate the unknown parameters, θ_i, $i=1,2,..,G$. In the classical or estimative approach these parameters are estimated using techniques such as maximum likelihood or the method of moments. The classification rules (1.2) to (1.5) are then applied using $\hat{f}_i(X) = f_i(X,\hat{\theta}_i)$ in place of $f(X)$.

In the predictive or Bayesian approach the classification rules 1.2 - 1.5 are obtained by assigning an item X to the population with the largest posterior probability. Since the prior probability of population π_i is $P(\pi_i)$

the posterior probability of π_1 by Bayes' theorem is:

$$P(\pi_1|X) = P(\pi_1, X) / P(X)$$

$$= \frac{P(\pi_1)f_1(X)}{P(\pi_1)f_1(X) + P(\pi_2)f_2(X)}$$

The posterior probability of π_j given data X is proportional to $P(\pi_1)f_1(X)$, so that the optimal classification rule is also rule 1.3. The Bayesian discriminant rule allocates an item X to the population for which $P(\pi_1)f_1(X)$ is maximized.

When all prior probabilities are equal the classification rules obtained by the maximum likelihood method and the Bayesian approach are equivalent.

In the case of discriminating between two population, G=2, the maximum likelihood discriminant rule is defined in terms of the discriminant function:

$$h(X) = \ln f_1(X) - \ln f_2(X).$$

The maximum likelihood rule is as follows:

Assign an item X to population π_1, if

$$h(X) > 0$$

and to population π_2, if

$$h(X) \leq 0.$$

1.6

In the Bayesian approach the effect of introducing prior probabilities is simply to shift the critical value of the discriminant function by an amount $\ln\{P(\pi_2/P(\pi_1)\}$. The classification rule 1.6 then becomes:

Assign an item X to population π_1, if

$$h(X) > \ln \{P(\pi_2) / P(\pi_1)\}$$

and to population π_2, if

$$h(X) \leq \ln \{P(\pi_2) / P(\pi_1)\}.$$

In the predictive or Bayesian approach the data are regarded as given whereas the unknown parameters are taken to be the random quantities. Begin by postulating prior distributions for the vectors of unknown parameters, say $g_i(\theta)$, $i=1,2,\ldots,G$. Then construct a joint distribution $f_i(X,T|\theta)$ $i=1,2,..,G$ for the unknown parameters to be classified, X, and the training sample T. The joint distribution of X,T and θ is $g_i(\theta)f_i(X,T|\theta)$ and by integrating θ out of the expression, the conditional distribution of $X \in \pi_i$, given T, is obtained. Criticism on this model usually centers on the difficulty of obtaining the prior distribution $g(\theta)$, and justifying a particular choice.

Particular emphasis will be placed on linear or quadratic discriminant analysis where items originate from a

multivariate normal distribution. Working under the assumption that the data come from a multivariate normal distribution, a more stable classification function can be computed. If this assumption is not met, one could attempt to normalize the data by transformations.

The unknown density functions $f_i(\mathbf{x})$, $i=1,2,..,G$ can also be estimated without having to assume specific models for them, i.e. they can be estimated non-parametrically. Parametrical and non-parametrical methods will be discussed in chapter 2.

To apply classification rules (1.2)-(1.5) one needs to specify the prior probabilities $P(\pi_1),P(\pi_2),..,P(\pi_G)$. In many situations there is little or no information on which to base this assignment. Conventionally, and for want of a better scheme, these probabilities are taken to be all equal in such cases.

If the expected cost (1.1) of the classification rule (1.2) is estimated by substituting $\hat{f}_i(X,\hat{\theta}_i)$ for $f_i(X)$ in (1.1) then the resulting estimate is biased. This is because the resulting regions $\hat{\Omega}_1,\hat{\Omega}_2,...,\hat{\Omega}_G$ having been estimated on the basis of $\hat{f}_i(X,\hat{\theta}_i)$ rather than $f_i(X)$, where the former are only estimates of the latter. In general this estimate of (1.1) tends to be lower than the true value, i.e. leads to an optimistic assessment of the expected total cost of the classification rule. The problem of obtaining an unbiased estimate of the

expected cost, or as a special case the misclassification rate, is discussed in chapter 3.

Another important issue which arises, especially where the number of measurements $X_1, X_2, \ldots, X_m$ is large, is that of variable selection. This problem, which is entirely analogous to variable selection in regression analysis, is as follows.

Some of the measurements in **X** will be less important than others for the purpose of discriminating items. By excluding measurements (variables) contributing only marginally to discriminating items, the number of parameters which need to be estimated in the models $f_i(X,\theta_i)$ and actually improve the performance of the classification rule can be reduced and thus reduce its expected cost. The inclusion of a variable gives rise to two opposing effects which are analogous to the properties of bias and standard error when one is considering an estimator of some parameter. Assuming that each of the m variables has <u>some</u> discriminating power, no matter how small, by excluding any of them the (potential) expected cost of misclassification is increased. In other words, if the $f_i(X)$ were known, reducing the number of variables will lead to a poorer classification rule. On the other hand, as already mentioned, the exclusion of variables reduces the number of parameters to be estimated and thus there is, on average, less sampling variation in the estimates $f_i(X,\hat{\theta})$. This consideration favours a reduction in the number of variables. Two competing

effects therefore have to be taken into account on deciding how many and which variables to use in the classification rule. This problem is discussed in chapter 4.

From the nature of a classification problem, where large numbers of variables are measured, some items may have incomplete data (values were not recorded for some measurements). Normally if only one measurement is missing, the only option in most computer packages is to delete all measurements from that item. To prevent the loss of valuable information, incomplete data values can be estimated. The EM (Expectation-Maximization) algorithm for estimating incomplete data values will be discussed in chapter 5.

In chapter 5 two other methods related to discriminant analysis are briefly mentioned. These are sequential and logistic discriminant analysis.

Finally, the above theory is applied to a practical problem where grape cultivars are classified into distinct populations, thereby illustrating the preceding theoretical analysis.

CHAPTER 2

2. Introduction

This chapter will deal briefly with three main developments in the theory of discriminant analysis. Firstly, methods such as the classical (estimative) and the predictive (Bayesian) approach to discriminant analysis will be considered. Particular emphasis will be placed on the situation where the measurements are assumed to follow a multivariate normal distribution. Non-normal data analyzed by means of Kernel estimates, K-nearest neighbours and ranking will also be considered. The chapter ends with a discussion on the relative merits of the above mentioned methods and some comments regarding the different practical situations in which each might be the most appropriate.

2.1 PARAMETRICAL METHODS

2.1.1 Classical discriminant analysis

In the first major work on discriminant analysis, the author, R.A. Fisher (1936), based his theory on the assumption of normality. This assumption, with minor relaxations, has formed the basis of much of the subsequent theoretical work and applications. It will therefore be considered in some detail. The derivation of the optimal discriminant rules is a good point of departure.

Suppose the existence of two populations, π_1 and π_2 with m measurements on an item. The item with measurements designated by **X** must be assigned to population π_1 or population π_2. A rule in terms whereof this item can be assigned to either of the populations is therefore required. If the parameters of the distribution of **X** in π_1 and π_2 are known, such knowledge may be used to construct an assignment rule.

If the parameters are unknown, random samples of size n_1 from population π_1 and size n_2 from population π_2 may be used to estimate the parameters.

Various criteria to test the validity of the classification rule have been proposed. Fisher (1936) suggested using the linear combinations of the measurements. These coefficients are chosen so that the ratio of the difference of the means of the linear combination to its variance is maximized. Welch (1939) suggested minimizing the total probability of misclassification. Von Mises (1945) suggested minimizing the maximum probability of misclassification in the two populations. Various other authors suggested that the total cost of misclassification be minimized.

2.1.1.1 Optimal classification rules:

For future reference the probability of misclassification within each group is defined as:

$$P_1 = \int_{\Omega_2} f_1(X)\, dX \qquad \text{and} \qquad P_2 = \int_{\Omega_1} f_2(X)\, dX \ ,$$

2.1

where $f_i(X)$ is the probability density function of X in population π_i. The sample space $\Omega=R^m$ is partitioned into G regions. If $X \in \Omega_i$ then the corresponding item is classified as belonging to population π_i, i=1,2,..,G.

Consider the special case were measurements from populations π_1 and π_2 are multivariate normal distributed with means μ_1 and μ_2 and common covariance matrix, Σ.

In population π_i the joint distribution of the measurements is

$$f_i(X) = (2\pi)^{-m/2} \, |\Sigma|^{-1/2} \, \exp\{-0.5(X-\mu_i)'\Sigma^{-1}(X-\mu_i)\} \ .$$

2.2

Thus

$$f_1(X) / f_2(X) = \exp\{-0.5(X-\mu_1)'\Sigma^{-1}(X-\mu_1) + 0.5(X-\mu_2)'\Sigma^{-1}(X-\mu_1)\}$$

$$= \exp[-0.5 \{ X'\Sigma^{-1}X + \mu'_1\Sigma^{-1}\mu_1 - 2\mu'_1\Sigma^{-1}X - X'\Sigma^{-1}X - \mu'_2\Sigma^{-1}\mu_2 + 2\mu'_2\Sigma^{-1}X \}]$$

$$= \exp[0.5 \{2\, X'\Sigma^{-1}\mu_1 - 2\, X'\Sigma^{-1}\mu_2 - \mu'_1\Sigma^{-1}\mu_1 + \mu'_2\Sigma^{-1}\mu_1 - \mu'_2\Sigma^{-1}\mu_1 + \mu'_2\Sigma^{-1}\mu_2 \}]$$

$$= \exp[\ \{X'\Sigma^{-1}(\mu_1-\mu_2) - 0.5\ \mu'_1\Sigma^{-1}(\mu_1-\mu_2) + 0.5\ \mu'_2\Sigma^{-1}(\mu_1-\mu_2)\ \}]$$

$$= \exp[\ X'\Sigma^{-1}(\mu_1-\mu_2) - 0.5(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\ ]$$

$$= \exp[\ (X - 0.5\ (\mu_1 + \mu_2))'\ \Sigma^{-1}\ (\mu_1 - \mu_2))\ ]. \qquad 2.3$$

By taking logarithms, the optimal classification rule is:

Assign an item to population π_1, if

$$\ln\{f_1(X)/f_2(X)\} = Z_{12}(X) = (X - 0.5\ (\mu_1 + \mu_2))'\ \Sigma^{-1}(\mu_1-\mu_2) \geq \ln\{P(\pi_1)/P(\pi_2)\} \qquad 2.4$$

or else, assign the item to population π_2.

The true discriminant function and its sample analogue are, respectively

$$Z_{12}(X) = (X - 0.5\ (\mu_1 + \mu_2))'\ \Sigma^{-1}(\mu_1-\mu_2) \qquad 2.5$$

and

$$Z_S(X) = (X - 0.5\ (\bar{X}_1 + \bar{X}_2))'\ S^{-1}(\bar{X}_1-\bar{X}_2)). \qquad 2.6$$

As is shown below, the coefficients of X are identical to Fisher's result for the linear discriminant function.

Since X is multivariate normally distributed, $Z_{12}(X)$, being a linear combination of X, is also normal. This fact assists us in calculating the error rates associated with the discriminant function $Z_{12}(X)$.

The mean of $Z_{12}(X)$, if X comes from population π_1, is

$$\begin{aligned} E\,(Z_{12}(X) \mid \pi_1) &= [\mu_1 - 0.5(\mu_1 + \mu_2)]'\Sigma^{-1}(\mu_1 - \mu_2) \\ &= 0.5(\mu_1 - \mu_2)'\,\Sigma^{-1}\,(\mu_1 - \mu_2) \\ &= 0.5\ MD_{12}^2\ , \end{aligned} \qquad 2.7$$

where $MD_{12}^2 = (\mu_1 - \mu_2)'\,\Sigma^{-1}\,(\mu_1 - \mu_2)$
is the Mahalanobis distance between the populations for known parameters.

The mean of $Z_{12}(X)$, if X comes from population π_2, is

$$\begin{aligned} E\,(Z_{12}(X) \mid \pi_2) &= [\mu_2 - 0.5(\mu_1 + \mu_2)]'\Sigma^{-1}(\mu_1 - \mu_2) \\ &= -\,0.5(\mu_1 - \mu_2)'\,\Sigma^{-1}\,(\mu_1 - \mu_2) \\ &= -\,0.5\ MD_{12}^2\ . \end{aligned} \qquad 2.8$$

The variance in each of the populations is :

$$\begin{aligned} & E\,[\,Z_{12}(X) - Z_{12}(\mu_1)]^2 \\ &= E\,[(X - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)]^2 \\ &= E\,[(\mu_1 - \mu_2)'\Sigma^{-1}(X - \mu_1)(X - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)] \\ &= (\mu_1 - \mu_2)'\Sigma^{-1}\ E\,[(X - \mu_1)(X - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)] \\ &= (\mu_1 - \mu_2)'\Sigma^{-1}\,(\mu_1 - \mu_2) \\ &= MD_{12}^2 \end{aligned} \qquad 2.9$$

The probabilities of misclassification can now be formulated:

$$P_1 = P\,[Z_{12}(X) < \ln\{P(\pi_2)/P(\pi_1)\}\,]$$

$$= P\left[\frac{Z_{12}(X) - MD_{12}^2/2}{MD_{12}} < \frac{\ln\{P(\pi_2)/P(\pi_1)\} - MD_{12}^2/2}{MD_{12}}\right]$$

$$= \Phi\left[\frac{\ln\{P(\pi_2)/P(\pi_1)\} - MD_{12}^2/2}{MD_{12}}\right]$$

Similarly,

$$P_2 = \Phi\left[\frac{-\ln\{P(\pi_2)/P(\pi_1)\} - MD_{12}^2/2}{MD_{12}}\right] \qquad 2.10$$

These probabilities of misclassification will be used in chapter 3.

The classification rules can now be generalized to the situation with more that two populations (G > 2) :

Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ be the vector of random variables representing the measurements taken on an item. Assume that an item belongs to population π_i and that $\mathbf{X}$ is multivariate normally distributed with mean μ_i and covariance matrix Σ_i, i=1,2,...,G. The probability density function of the measurements in population πi is therefore

$f_i(X) = 1/\{(2\pi)^{m/2}\,|\Sigma_i|^{1/2}\}\,\exp\{-0.5(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)\}$,

i=1,2,...G.

For convenience, assume that the costs of misclassification are all equal. In this case the expected cost may be represented by the probability of misclassification. Let $C_{ij} = 1$ if $i \neq j$ and $C_{ij} = 0$ if $i = j$.

Classification rule (1.3) becomes:

Assign an item to population π_i, if

$$P(\pi_i)\ \{(2\pi)^{-m/2}\ |\Sigma_i|^{-1/2}\}\ \exp\{-0.5(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)\}$$

$$= \max_j P(\pi_j)\ \{(2\pi)^{-m/2}\ |\Sigma_j|^{-1/2}\}$$

$$\exp\{-0.5(X-\mu_j)'\Sigma_j^{-1}(X-\mu_j)\}. \qquad 2.11$$

By taking the logarithm, this rule simplifies to:

Assign an item to population π_i, if

$$2 \ln P(\pi_i) - \ln |\Sigma_i| - MD_i^2(X)$$

$$= \max_j 2 \ln P(\pi_j) - \ln |\Sigma_j| - MD_j^2(X) \qquad 2.12$$

where

$MD_i^2(X) = (X - \mu_i)'\ \Sigma_i^{-1}\ (X - \mu_i)$ is the Mahalanobis distance from X to μ_i.

If $\Sigma_i = \Sigma$ for i=1,2,..,G assign item X to population π_i, if

$$\ln P(\pi_i) + (X - \mu_i/2)'\Sigma^{-1}\mu_i = \max_j \ln P(\pi_j) + (X - \mu_j/2)'\Sigma^{-1}\mu_j \quad . \qquad 2.13$$

Most standard statistical packages use this equation when assigning an item to i=1,2,..,G populations. One should be cautious of non-robustness when unequal covariance matrices or non-normality occurs. (See 2.1.1.6.)

Since parameters are not always known, estimates are needed to calculate classification functions. The asymptotic estimates for unknown parameters are :

Estimate Σ by the pooled covariance matrix, where

$$\hat{\Sigma} = S = \frac{1}{\Sigma(n_\cdot - 1)} \sum_{i=1}^{G} \sum_{j=1}^{n_i} (\bar{X}_{ij} - \bar{X}_{i\cdot})(\bar{X}_{ij} - \bar{X}_{i\cdot})' \qquad 2.14$$

and estimate μ_i by

$$\hat{\mu}_i = \bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \qquad 2.15$$

2.1.1.2 Quadratic discriminant function

In practice the assumption of equal covariance matrices is seldom satisfied. A marginal difference between the covariance matrices will not affect the result measurably and one could assume equality (see 2.3). However, when

the covariance matrices differ substantially and normality holds, the optimal classification rule for the two-group case is:

Assign an item to population π_1, if

$$Q_{12}(X) = \ln \{f_2(X)/f_1(X)\} > \ln \{P(\pi_2)/P(\pi_1)\} \qquad 2.16$$

or else, assign the item to population π_2,

with

$$\begin{aligned} Q_{12}(X) &= 0.5 \ln(|\Sigma_2|/|\Sigma_1|) - 0.5 (X-\mu_1)'\Sigma_1^{-1} (X-\mu_1) \\ &\quad + 0.5 (X-\mu_2)'\Sigma_2^{-1} (X-\mu_2) \\ &= 0.5 \ln(|\Sigma_2|/|\Sigma_1|) - 0.5\{ X' (\Sigma_1^{-1} - \Sigma_2^{-1})X \\ &\quad - 2X'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)\} - 0.5(\mu_1' \Sigma_1^{-1}\mu_1) \\ &\quad + 0.5(\mu_2' \Sigma_2^{-1}\mu_2) \\ &= -0.5 X'(\Sigma_1^{-1} - \Sigma_2^{-1}) X + X'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) \\ &\quad -0.5 (\mu_1' \Sigma_1^{-1}\mu_1 - \mu_2' \Sigma_2^{-1}\mu_2) + 0.5 \ln(|\Sigma_2|/|\Sigma_1|). \end{aligned} \qquad 2.17$$

The function $Q_{12}(X)$ is a quadratic function since $\Sigma_1^{-1} - \Sigma_2^{-1}$ does not vanish. Care is needed when applying this function in practice, as this procedure is not robust to non-normality (see 2.3).

2.1.1.3 Linear Discriminant Analysis by means of Fisher's method :

This approach to discriminate between populations is based on Fisher's (1936) original method.

In his first work on discriminant analysis R.A. Fisher specifically considered the problem of discriminating between two populations. In the second part of this section the results are generalized to the G population situation.

The two population situation :

Let

$$X_1 = \begin{bmatrix} X_{111} & \dots & X_{11m} \\ X_{121} & \dots & X_{12m} \\ \vdots & & \\ X_{1n_11} & \dots & X_{1n_1m} \end{bmatrix} \quad \text{and} \quad X_2 = \begin{bmatrix} X_{211} & \dots & X_{21m} \\ X_{221} & \dots & X_{22m} \\ \vdots & & \\ X_{2n_21} & \dots & X_{2n_2m} \end{bmatrix},$$

where X_{ijk} ($i=1,2$; $j=1,2..n_i$ and $k=1,2...,m$) represents a measurement, k, of the j-th item, coming from the i-th population.

As previously mentioned the result of the discriminant rule should be such that items with similar measurements are classified in the same population and items with dissimilar measurements are classified in different populations. The measure that Fisher suggested to quantify the similarity or dissimilarity between two values is the square of their difference. The object of this measure

Let $z_{ij} = Z_{ij} - \bar{Z}_{i.} \quad i = 1,2 \; ; \quad j = 1,2,\ldots n_i,$

where $\bar{Z}_{i.} = \sum_{j=1}^{n_i} Z_{ij} / n_i$

$$= \bar{X}_{i.}u$$

and

$$\bar{Z}_{..} = \sum_{i=1}^{2} n_i \bar{Z}_{i.} \Big/ \sum_{i=1}^{2} n_i = \bar{X}_{..}u$$

$$= (n_1\bar{Z}_{1.} + n_2\bar{Z}_{2.})/(n_1+n_2)$$

The within-groups sum of squares is given by

$$SS_w = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left(\sum_{k=1}^{m} X_{ijk} u_k - \sum_{k=1}^{m} \bar{X}_{i.k} u_k \right)$$

$$\left(\sum_{k=1}^{m} X_{ijk} u_k - \sum_{k=1}^{m} \bar{X}_{i.k} u_k \right)'$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{n_i} [(X_{ij1} - \bar{X}_{i.1})u_1 + (X_{ij2} - \bar{X}_{i.2})u_2 + \ldots$$

$$+ (X_{ijm} - \bar{X}_{i.m})u_m] [(X_{ij1} - \bar{X}_{i.1})u_1$$

$$+ (X_{ij2} - \bar{X}_{i.2})u_2 + \ldots + (X_{ijm} - \bar{X}_{i.m})u_m]'$$

$$= \sum_{i=1}^{2} [u_1' S_{(i)11} u_1 + \ldots + u_m' S_{(i)mm} u_m$$

$$+ 2u_1' S_{(i)12} u_2 + \ldots + 2u_{m-1}' S_{(i)m-1m} u_m]$$

$$= \sum_{i=1}^{2} u' S_{(i)**} u \quad , \qquad 2.19$$

where:

$$S_{(i)st} = \sum_{j=1}^{n_i} (X_{ijs} - \bar{X}_{i.s})(X_{ijt} - \bar{X}_{i.t})' \quad s,t = 1,2,\ldots,m$$

and

$$S_{(i)**} = \begin{bmatrix} S_{(i)11} & S_{(i)12} & \ldots & S_{(i)1m} \\ \vdots & & & \\ S_{(i)m1} & S_{(i)m2} & \ldots & S_{(i)mm} \end{bmatrix}$$

By defining $W = S_{(1)**} + S_{(2)**}$, SS_w can be written as

$$SS_w = u'(S_{(1)**} + S_{(2)**})u = u'Wu. \qquad 2.20$$

The difference between group means on variable X_k is

$d_k = \bar{X}_{1.k} - \bar{X}_{2.k}$. Further define $D = \bar{Z}_{1.} - \bar{Z}_{2.}$

$= u'\bar{X}_{1.} - u'\bar{X}_{2.}$

$= u'd.$

The between-groups sum of squares is derived as follows:

$$SS_B = n_1(\bar{Z}_{1.} - \bar{Z}_{..})^2 + n_2(\bar{Z}_{2.} - \bar{Z}_{..})^2$$

$$= n_1\bar{Z}_{1.}^2 - 2n_1\bar{Z}_{1.}\bar{Z}_{..} + n_1\bar{Z}_{..}^2$$

$$+ n_2\bar{Z}_{2.}^2 - 2n_2\bar{Z}_{2.}\bar{Z}_{..} + n_2\bar{Z}_{..}^2$$

$$= n_1\bar{Z}_{1.}^2 - 2n_1\bar{Z}_{1.}(n_1\bar{Z}_{1.} + n_2\bar{Z}_{2.})/(n_1+n_2)$$

$$+ n_1[(n_1\bar{Z}_{1.} + n_2\bar{Z}_{2.})/(n_1+n_2)]^2$$

$$+ n_2\bar{Z}_{2.}^2 - 2n_2\bar{Z}_{2.}(n_1\bar{Z}_{1.} + n_2\bar{Z}_{2.})/(n_1+n_2)$$

$$+ n_2[(n_1\bar{Z}_{1.} + n_2\bar{Z}_{2.})/(n_1+n_2)]^2$$

$$= n_2^2n_1\bar{Z}_{1.}^2 + n_1n_2^2\bar{Z}_{2.}^2 + n_1^2n_2\bar{Z}_{2.}^2 - 2n_1^2n_2\bar{Z}_{1.}\bar{Z}_{2.}$$

$$- 2n_1n_2^2\bar{Z}_{1.}\bar{Z}_{2.} + n_1^2n_2\bar{Z}_{1.}^2$$

$$= n_2^2n_1(\bar{Z}_{1.} - \bar{Z}_{2.})^2 + n_1^2n_2(\bar{Z}_{1.} - \bar{Z}_{2.})^2$$

$$= (\bar{Z}_{1.} - \bar{Z}_{2.})^2(n_2^2n_1 + n_1^2n_2)/(n1 + n_2)^2$$

$$= D^2n_1n_2/(n_1+n_2) \qquad 2.21$$

Finally, the ratio of between-groups sum of squares to within-groups sum of squares is given by

$$\tau=[(u'd)\ n_1n_2/(n_1+n_2)]/u'Wu.$$

This τ is the criterion which Fisher proposed should be maximized.

The coefficients which yield the maxima are obtained by differentiating

$$\tau = h(u'd)^2/u'Wu$$

with respect to u, where $h=n_1n_2(n_1+n_2)$;

$$\delta\tau/\delta u = \frac{\delta\{h(u'd)^2/(u'Wu)\}}{\delta u}$$

$$= \frac{(u'Wu)\,\delta h(u'd)^2/\delta u - h(u'd)^2\,\delta(u'Wu)/\delta u}{(u'Wu)^2}$$

$$= 2h(u'd)d/(u'Wu) - h(u'd)^2 2Wu/(u'Wu)^2$$

$$= 2h(u'd)/(u'Wu)[d - \{(u'd)/(u'Wu)\}(Wu)]$$

Thus $2h(u'd)/(u'Wu)\,[d - \{(u'd)/(u'Wu)\}\,(Wu)] = 0$ is the necessary condition for maximization.

A solution of these equations is $u=W^{-1}d$. This can be seen by substituting this expression into the above equation.

This solution is unique only up to a scalar constant, as any non-zero scalar multiple of $W^{-1}d$ (say $hW^{-1}d$) is also a solution: this can be seen by substitution. However, different choices of the constant h simply correspond to the scaling of u. Thus u is proportional to $W^{-1}d$. Conventionally the scaling is chosen such that $u'd/u'Wu = 1$.

The discriminant rule based on $u'X$ thus classifies the item to population π_1, if

$$|u'X - u'\bar{X}_1| < |u'X - u'\bar{X}_2|$$

or equivalently, if

$$u'X > 0.5\ [\ u'(\ \bar{X}_1 + \bar{X}_2\ )].$$

Since $d = (\bar{X}_1 - \bar{X}_2)$ and W is proportional to the pooled estimate of the covariance, i.e. S, the weight vector u is proportional to $S^{-1}(\ \bar{X}_1 - \bar{X}_2\ )$.

The rule thus classifies an item X to population π_1, if

$$(\ \bar{X}_1 - \bar{X}_2\ )'\ S^{-1}X > 0.5\ (\ \bar{X}_1 - \bar{X}_2\ )'\ S^{-1}(\ \bar{X}_1 + \bar{X}_2).$$

This equation corresponds to equation 2.4 when the prior probabilities are equal.

Tests of significance

In some applications one might wish to establish that the measurements from items in different populations do indeed follow different distributions. If they do not, it is pointless to discriminate between the populations on the basis of such measurements.

The above requirement can be satisfied by a test of significance when using D as the test statistic. The method used is analogous to univariate analysis of variance testing the hypothesis H_o: $u_1 - u_2 = 0$.

Since $(u'Wu)/u'd = 1$,

it means that $SS_w = u'Wu = u'd = D$ (by the optimal condition).

The analysis of variance table of testing the significance of D.

Source	Sum of Squares	DF	Mean Square	F
Between Groups	$SS_B = \frac{n_1 n_2 D^2}{(n_1+n_2)}$	m	$MS_B = SS_B/m$	$\frac{MS_B}{MS_w}$
Within Groups	$SS_w = D$	n_1+n_2-m-1	$MS_w = \frac{D}{(n_1+n_2-m-1)}$	

The above illustrated Fisher's suggestion of selecting coefficients that maximize the ratio of the between-groups sum of squares to the within-groups sum of squares. The extension to the general G population situation follows.

2.1.1.4 Linear discriminant analysis in the multiple population situation as based on Fisher's original method.

The problem of discriminating between G populations, using measurements from G random samples, will now be considered.

Let

$$X_i = \begin{bmatrix} X_{i11} & X_{i12} & \dots & X_{i1m} \\ X_{i21} & X_{i22} & \dots & X_{i2m} \\ \vdots & & & \\ \vdots & & & \\ X_{in_i 1} & X_{in_i 2} & \dots & X_{in_i m} \end{bmatrix} \qquad i=1,2,\dots,G$$

The linear function for discriminating between G populations by means of a linear combination or canonical vector of m measurements is defined as:

$$Z_{ij} = \sum_{k=1}^{m} X_{ijk}\, u_k, \qquad 2.22$$

where Z_{ij} represents the linear combination of the m measurements from the j-th item in population i.

Similar to the two population situation discussed above, select the coefficients of the linear discriminant function $u=(u_1, u_2, \ldots, u_m)'$ such that the ratio of the among-groups sum of squares to the within-groups sum of squares is maximized. The derivation of the coefficients which maximize this ratio will now be outlined.

The discriminant scores for these sets of observations (for given coefficients u) can be expressed as

$$Z_i = X_i u = \begin{bmatrix} Z_{i1} \\ \vdots \\ Z_{in_i} \end{bmatrix} \qquad i=1,2,\ldots,G$$

The means of the random variable Z for the i-th group may be denoted by

$$\bar{Z}_{i.} = \sum_{j=1}^{n_i} Z_{ij} \,/\, n_i \qquad i=1,2,\ldots,G$$

$$= \bar{X}_{i.} u$$

and

$$\bar{Z}_{..} = \sum_{i=1}^{G} n_i \bar{Z}_{i.} \Big/ \sum_{i=1}^{G} n_i$$

$$= \bar{X}_{..}u \ .$$

Let $z_{ij} = Z_{ij} - \bar{Z}_{i.} \qquad i=1,2,\ldots,G \quad j=1,2,\ldots,n_i.$

The within-groups sum of squares is given by

$$SS_w = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \Big(\sum_{k=1}^{m} X_{ijk} u_k - \sum_{k=1}^{m} X_{i.k} u_k \Big)'$$

$$\Big(\sum_{k=1}^{m} X_{ijk} u_k - \sum_{k=1}^{m} X_{i.k} u_k \Big)$$

$$= \sum_{i=1}^{G} \sum_{j=1}^{n_i} [(X_{ij1} - X_{i.1})u_1 + \ldots + (X_{ijm} - X_{i.m})u_m]'$$

$$[(X_{ij1} - X_{i.1})u_1 + \ldots + (X_{ijm} - X_{i.m})u_m]$$

$$= \sum_{i=1}^{G} [u_1'S_{(i)11}u_1 + \ldots + u_m'S_{(i)mm}u_m$$

$$+ 2u_1'S_{(i)12}u_2 + \ldots + 2u_{m-1}'S_{(i)m-1m}u_m]$$

$$= \sum_{i=1}^{G} u'S_{(i)**}u \ , \qquad 2.23$$

where

$$S_{(i)st} = \sum_{j=1}^{n_i} (X_{ijs} - \bar{X}_{i.s})(X_{ijt} - \bar{X}_{i.t})' \quad s,t = 1,2,\ldots,m.$$

By defining $W = S_{(1)**} + S_{(2)**} + \ldots + S_{(G)**}$,

write

$$SS_w = u'(S_{(1)**} + S_{(2)**} + \ldots S_{(G)**})u = u'Wu. \qquad 2.24$$

The variability is expressed by the sum of squares among group means :

$$\begin{aligned} SS_A &= \sum_{i=1}^{G} n_i \, (\bar{Z}_{i.} - \bar{Z}_{..})(\bar{Z}_{i.} - \bar{Z}_{..})' \\ &\quad - \sum_{i=1}^{G} n_i \, (\bar{X}_{i.}u - \bar{X}_{..}u)(\bar{X}_{i.}u - \bar{X}_{..}u)' \\ &\quad - u' \sum_{i=1}^{G} n_i \, (\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})' \, u \\ &= u'Au \end{aligned} \qquad 2.25$$

where $A = \sum_{i=1}^{G} n_i \, (\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})'$

Finally, the ratio of among-groups sum of squares to within-groups sum of squares is given by

$\tau = u'Au/u'Wu.$

This τ is the criterion which should be maximized. The coefficients which yield the maxima are obtained by differentiating $\tau = u'Au/u'Wu.$

The partial derivative of τ with respect to u is obtained as follows:

$$\delta\tau/\delta u = \frac{(u'Wu)\,\delta(u'Au)/\delta u - (u'Au)\,\delta(u'Wu)/\delta u}{(u'Wu)^2}$$

$$= \frac{2[(u'Wu)(Au) - (u'Au)(Wu)]}{(u'Wu)^2}$$

By setting the above equation equal to zero, obtain:

$$2[(u'Wu)(Au) - (u'Au)(Wu)] = 0$$

$$2[(Au) - \tau(Wu)] = 0$$

$$(W^{-1}A - \tau I)\,u = 0.$$

(Note that W^{-1} exists since W is a square, non-singular matrix.)

The characteristic equation of the matrix $W^{-1}A$ is :

$$| W^{-1}A - \tau I | = 0 . \qquad 2.26$$

The solutions to equation 2.26 are the eigenvalues of $W^{-1}A$.

The number of roots of this equation is equal to the rank of $W^{-1}A$.

Since W is a non-singular, square matrix, the rank of W^{-1} must be equal to m, the number of X's. It is further know that the rank of A is G-1, since

$$A = \sum_{i=1}^{G} n_i (\bar{X}_{i.} - \bar{X}_{..}) (\bar{X}_{i.} - \bar{X}_{..})'$$

which is ordinarily less than m. The rank of $W^{-1}A$ and hence the number of values of τ, is equal to the smaller of m or G-1, i.e. usually G-1.

When solving equation (2.26) each obtained value of τ is associated with an eigenvector u. This vector u is multiplied by the corresponding measurements in X when constructing the corresponding discriminant function Z.

The first function, Z_1, defines a dimension on which the groups differ maximally. The second function, Z_2, defines a dimension uncorrelated with the first, on which group differences are second in magnitude. All succeeding functions are uncorrelated. Thus, discriminant analysis produces a number (the smaller of m or G-1) of discriminant functions, all mutually uncorrelated and ordered from the greatest to the least in terms of the extent in which they discriminate between the groups.

By re-scaling, it can be shown that if r vectors are used, the rule becomes :

Assign an item to population π_i, if

$$\sum_{s=1}^{r} [u_s' (X - \bar{X}_{i.})]^2 = \min_{s} \sum_{j=1}^{r} [u_s' (X - \bar{X}_{j.})]^2 \qquad 2.27$$

2.1.1.5 Summary

This canonical vector approach to discriminant analysis (see 2.1.1.4) may be of interest for several reasons:

1) the dimension can be reduced from a large number of measurements, m, to relatively few linear combinations or canonical vectors. This has the advantage of summarizing the between-population variation;

2) plotting the first few canonical vectors can be useful as an explorative technique;

3) the eigenvalues can be used to test the hypothesis that the means between the populations are equal.

A disadvantage of both the optimal classification rule and the canonical variates is that the assumption of normality and equal covariances should be met. The robustness of the discriminant function now requires attention.

2.1.1.6 Robustness of the discriminant function

The linear discriminant function is the optimal assignment rule when the following assumptions are true:

1. $f_i(X)$, i=1,2,...,G, are multivariate normal,

2. the covariance matrices in the different populations are equal,
3. the a priori probability, $P(\pi_i)$, that an item comes from a given population is known,
4. the means, μ_i, and the covariance matrix, Σ, are known.

If these assumptions do not hold, the linear discriminant function calculated will not be the optimal assignment rule.

Provided that a randomly generated training sample is not too small, the a prior probability, means and covariance matrix can be estimated, if unknown. The first two assumptions are now discussed:

When the distributions of the populations are non-normal, linear discriminant analysis does in general not produce the optimal assignment rule. In some situations transformations on the data can be used to obtain normality. (See below)

The second assumption requires equal covariance matrices for the populations. In the presence of a small amount of heteroscedasticity the linear discriminant function can still be applied. When a marked difference between the covariance matrices occurs the quadratic discriminant function is superior to the linear discriminant function. Note that the quadratic discriminant function is less

attractive if the number of measurements is quite large or when the number of populations, G, is large or when the number of items, n_i, is small.

2.1.1.1 Transformations

When marked departures from the normal distribution occur, the analyst, in order to obtain optimal classification rules, should consider transformations on the data measurements prior to computing the linear discriminant functions. The primary question is whether the transformation can appreciably reduce the misclassification probabilities. When these transformations are unsatisfactory, one should consider using non-parametrical methods such as the kernel method, ranking or the K-nearest neighbour technique. (See paragraph 2.2)

Transformation techniques have been introduced and tested by various authors. In chapter 6 the results of the logarithmic, square root and reciprocal transformation techniques, used to obtain normality of the data measurements, are demonstrated.

The extent of the damage on the misclassification probability caused by the application of a linear discriminant function or a quadratic discriminant function to data that should have been transformed was studied by Beauchamp, Folkert and Robson (1980).

Beauchamp et al.(1980) showed that even when the log transformation is required to achieve normality in the univariate situation, the probability of misclassification for the untransformed data is in many cases not appreciably different from the optimal misclassification probabilities if the form of the discriminant function is changed from a linear discriminant function to a quadratic discriminant function. The linear discriminant function is robust against mild departures from normality (Lachenbruch 1975).

Beauchamp and Robson (1986) found when using non-negative variables, it is likely in practice that a transformation is needed when 1/(coefficient of variation) exceeds 2.

Other than transformations, extreme values can be deleted from the data set by means of **trimming and Huberizing**, in order to obtain at least approximate normality.
The effect of Huberizing and trimming the quadratic discriminant function was studied by Broffitt, Clarke and Lachenbruch (1980). In their study they used different methods of trimming and Huberizing. Broffitt suggested that trimming and Huberizing should be performed on Mahalanobis distances

$$MD^2(X_{ij}) = (X_{ij} - \mu_i)' \Sigma^{-1} (X_{ij} - \mu_i).$$

Trimming is performed by eliminating the αN_i values of X with the largest distances $MD^2(X_{ij})$.

Huberizing is performed by eliminating 100(α)% cutoffs from $MD^2(X_{ij})$.

Q-trimming was introduced by Clarke (1975). This procedure eliminates the Ni(α/2) observations with the highest quadratic discriminant scores and the $N_i(\alpha/2)$ values with the lowest quadratic discriminant scores.

Broffitt et al. (1980) concluded that trimming, Q-trimming and Huberizing the quadratic discriminant function did not produce satisfactory improvements to the misclassification probabilities for the ordinary quadratic discriminant function for non-normal distributions. Data transformations seem to be the obvious alternative to this problem as the desired normality can be achieved.

In the following section the predictive or Bayesian approach to discriminant analysis is discussed.

2.1.2 Predictive or Bayesian discriminant analysis

It was stated above that the Bayesian (predictive) approach to discriminant analysis does not base the estimate of the density $p(X|X_1, X_2, \ldots, X_n)$ on a simple estimate of θ but it forms a weighted combination of densities $p(X|\theta)$ with weights given by a function $g(\theta|X_1, X_2, \ldots, X_n)$.

In this approach the data are regarded as given and the unknown parameters are integrated out of the model.

Let

$$p(X|X_1,\ldots,X_n) = \int p(X|\theta)\ g(\theta|X_1,\ldots,X_n)\ d\theta$$

where $g(\theta|X_1,\ldots,X_n)$ is regarded as a density function for possible values of θ.

The estimation of function $g(\theta|X_1,\ldots,X_n)$ may lead to criticism of this model, since the calculations of the unknown parameters are difficult.

The estimation of the function $g(\theta|X_1,\ldots,X_n)$ is done by starting from an initial prior density $g(\theta)$, updated to take the sample $\{X_1,\ldots,X_n\}$ into account.

The predictive or Bayesian approach also leads to classification rules 1.1 to 1.5, but this is reached by assigning an item X to the population with the largest posterior probability.

By definition the conditional density of X given population π_i, is $f_i(X)$. Since the a priori probability of population π_i, is $P(\pi_i)$, the posterior probability of population π_i by Bayes' theorem is :

$$P(\pi_i|X) = P(\pi_i,X) \ / \ P(X)$$

$$= \frac{P(\pi_i)\ f_i(X)}{P(\pi_1)\ f_1(X) + P(\pi_2)\ f_2(X)} \qquad i=1,2. \qquad 2.28$$

Assign an item to population π_1, if

$$P(\pi_1|X) > P(\pi_2|X) . \qquad 2.29$$

This is equivalent to the rule that minimizes the total probability of misclassification. These posterior probabilities are useful when estimating the risk of an item as belonging to population π_1.

Two different Bayesian approaches have been applied to the discriminant analysis problem.

The discriminant function when parameters are known, is

$$Z_{12}(X) = [X - 0.5(\mu_1 - \mu_2)]'\Sigma^{-1} (\mu_1 - \mu_2).$$

The 'noninformative' prior for μ_1, μ_2 and Σ^{-1} is

$$g(\mu_1, \mu_2, \Sigma^{-1}) \alpha |\Sigma|^{(m+1)/2} . \qquad 2.30$$

By using 2.30 Geisser (1966) showed that the posterior mean of $Z_{12}(X)$ is

$$E(Z_{12}(X)|\bar{X}_1, \bar{X}_2, S) = Z_{12}(X) + 0.5\ m\ (1/n_2 + 1/n_1). \qquad 2.31$$

Now consider the case of unequal covariance matrices. The discriminant function when parameters are known, is:

$$Z_{12}(X) = 0.5 \left[\ln \left[\frac{|\Sigma_1^{-1}|}{|\Sigma_2^{-1}|} \right] + (X - \mu_2)' \Sigma_2^{-1} (X - \mu_2) - (X - \mu_1)' \Sigma_1^{-1} (X - \mu_1) \right]. \qquad 2.32$$

By using the posterior density

$$g(\mu_1, \Sigma^{-1}) \quad \alpha \quad | \Sigma_1^{-1} |^{(m+1)/2} ,$$

Enis and Geisser (1970) showed that the posterior mean of $Z_{12}(X)$ is:

$$Z_B(X) = V + h(m, n_1, n_2)$$

with

$$V = 0.5 \left[\ln \left[\frac{|S_1^{-1}|}{|S_2^{-1}|} \right] + (X - \bar{X}_2)' S_2^{-1} (X - \bar{X}_2) - (X - \bar{X}_1)' S_1^{-1} (X - \bar{X}_1) \right]$$

and

$$h(m, n_1, n_2) = 0.5 \sum_{i=1}^{2} \sum_{k=1}^{m} (-1)^i \{ \log(n_i - 1) + n_i^{-1} - \delta[0.5(n_i - k)] \} \qquad 2.33$$

It is thus clear that $h(m,n,n)=0$. Hence, when the samples sizes are equal, $Z_{12}(X)$ is unbiased for V (a posteriori).

Another Bayesian approach is to use the posterior distribution of X for assignment. The predictive density of X, given the data, is:

$$f(X|(\bar{X}_1, \bar{X}_2, S, \pi_1)$$

$$= \int f(X|\mu_1, \mu_2, \Sigma^{-1}, \pi_1)\ g(\mu_1, \mu_2, \Sigma^{-1}\ |\bar{X}_1, \bar{X}_2, S)\ d\mu_1\ d\mu_2\ d\Sigma^{-1} \qquad 2.34$$

Assignment is based on the statistic

$$W = \ln \frac{f(X|\bar{X}_1, \bar{X}_2, S, \pi_1)}{f(X|\bar{X}_1, \bar{X}_2, S, \pi_2)}, \qquad 2.35$$

where

$$W = \frac{m}{2} \ln \frac{n_1}{n_2} + \frac{v-m+1}{2} \ln \frac{n_1+1}{n_2+1}$$

$$+ [(v+1)/2] \ln \left[\frac{(n_1+1)v + n_2(X-\bar{X}_2)'S^{-1}(X-\bar{X}_2)}{(n_2+1)v + n_1(X-\bar{X}_1)'S^{-1}(X-\bar{X}_1)} \right] \qquad 2.36$$

with $v=n_1+n_2-2$.

2.1.2.1 Predictive probability of misclassification

If $P(\pi_i|\pi_j)$ is the predictive probability that an item X was classified as belonging to population π_j when in fact it belonged to population π_i, the predictive probability of misclassification is :

$$\sum_{i=1}^{G} P(\hat{\pi}_i|\pi_i) = 1 - \sum_{i=1}^{G} P(\pi_i|\pi_i) \quad , \qquad 2.37$$

where

$$P(\hat{\pi}_i|\pi_i) = P(\pi_i)\ (1 - \int_{\Omega_i} f(Z|X,\phi,\pi_i)dZ)$$

and

$$P(\pi_i|\pi_i) = P(\pi_i) \int_{\Omega_i} f(Z|X,\phi,\pi_i)dZ). \qquad 2.38$$

2.1.2.2 Remarks

Problems exist when using the Bayesian approach:

- The selection of the prior density functions could be difficult.
- As in the classical approach, the assumption of normality must be met.

An advantage of the Bayesian approach (equation 2.31) is that it gives a theoretical justification for using $D_{ij}(X)$. The classical approach offers no such justification.

2.1.3 Discussion on parametrical approaches

In paragraph 2.1 the parametrical methods were divided into two main types:

i) Classical or estimative approach

ii) Bayesian or predictive approach.

Little or no difference exists between the classification rules obtained when using either of these two approaches.

The main difference between these approaches is that the posterior probabilities can differ substantially. Hawkins et.al. (1982) advised practitioners of discriminant analysis to study the posterior probabilities of all the populations π_i, since the estimative calculations can be misleading.

The Bayesian of predictive approach admits that the true value of $\hat{\theta}_i$ is unknown. Instead of using a single estimate, a distribution $p(\theta|X)$ is used. In this approach the data are regarded as given and the unknown parameters are integrated out of the model.

The Bayesian and classical methods assume normality of the data, whereas the non-parametrical approaches relax this assumption.

In the following paragraph the classification function is obtained by using non-parametrical methods.

2.2 Non-parametrical approach to discriminant analysis

As shown above, both the classical and predictive approaches to discriminant analysis are optimal only when the data measurements are normally distributed. Where marked departures from normality occur, an attempt should be made to transform the data to obtain at least approximate normality. Should such transformation be unsuccessful, non-parametrical methods such as the Kernel method, the K-nearest neighbour technique or ranking can be applied to minimize the total probability of misclassification.

The **K-nearest neighbour** technique fixes a number, k, of the design set points and finds the volume which contains the k nearest items. From this number and volume the density function can be estimated.

The Kernel method calculates a probability density estimate based on the proportion of sample points falling in a specific volume. The choice of a kernel shape and corresponding smoothing parameter, τ, determines the applicable density estimate. The method of choice will be discussed below.

In **ranking** the ordinary linear or quadratic discriminant function is applied after the original data have been replaced by corresponding ranks.

In the following paragraphs these techniques will be discussed in detail. It should be borne in mind, however, that the population conditional distribution estimates and decision surfaces are more flexible in non-parametrical discriminant analysis than in the parametrical methods.

2.2.1 The Kernel Method

In the kernel method the probability density functions, $f_i(X)$, $i=1,2,\ldots,G$, (as defined in classification rules 1.1 - 1.5) are estimated directly from training samples by means of kernel functions. The choice of the kernel shape and corresponding smoothing parameter, τ, determines the applicable density estimate. For a fixed kernel shape, τ determines how much each sample point contributes to the estimate at any point X. τ is thus the spread or smoothing parameter.

The first kernel probability density estimates were formulated by Parzen (1962). Cacoullos (1966) extended the development to multivariate distributions.

Parzen proposed a class of estimates of the form

$$\hat{f}(X) = \frac{1}{n\tau} \sum K\left[\frac{(X - X_{ij})}{\tau}\right] \qquad 2.39$$

where X_{ij} are the observed data points and $K(X)$ and $\tau(n)$ are functions satisfying:

1. $\sup_{\infty < z < \infty} |K(z)| < \infty$

2. $\lim_{z \to \infty} |zK(z)| = 0$

3. $\int_{-\infty}^{\infty} |K(z)| \, \delta z < \infty$

4. $\int_{-\infty}^{\infty} K(z) \, \delta z = 1$

5. $\lim_{n \to \infty} \tau(n) = 0$

6. $\lim_{n \to \infty} n\tau(n) = \infty$

The estimates of $f_i(X)$ are consistent and asymptotically normal. The above conditions (1-4) satisfy a wide variety of functions, $K(z)$.

For a function to qualify as a kernel function the above mentioned assumptions must be met.

Given training samples from the G populations, let $K(.)$ be a kernel function with G values $\tau_1, \tau_2, \ldots, \tau_G$ of the smoothing parameter τ. Assign an item X of unknown origin to population π_i, i=1,2,..,G, by using one of the classification rules 1.1 - 1.5, with $f_i(X)$ replaced by

$$\hat{f}_i(X) = 1/n_i \sum_{j=1}^{n_i} K(X|X_{ij}, \tau_i) \qquad 2.40$$

In estimating $f_i(X)$, first choose the function $K(.)$ and then find the smoothing parameters τ_i.

When using continuous data, the most commonly used kernel function is the spherical normal density function

$$K(X|X_{ij},\tau_1,..,\tau_\Theta) = (2\pi)^{-1/2\,\Theta} / \sqrt{(\tau_1,..,\tau_\Theta)} \quad x \quad \exp\{-1/2 \sum_{t=1}^{k} (X_t - X_{ijt})^2 / \tau_t\} \qquad 2.41$$

where X_t and X_{ijt} are the t-th variables from the unknown item and the ij-th item from the data set, respectively, and τ_t is the smoothing parameter associated with the t-th variable.

Note that the kernel method performs poorly as a density function estimate when departures from the assumptions of equal variance and no correlation between the components are violated. One should therefore multi-standardize for the components to have the same spread with no correlation between them.

Aitchison and Aitken (1976) used the following kernel function for linear data :

$$k(X|X_{ij},\tau_1,..,\tau_\Theta) = \prod_{t=1}^{G} \tau_t^{z_t} (1-\tau_t)^{1-z_t} \qquad 2.43$$

where $z_t = 0$ for $X_{ijk} \neq X_t$ and $z_t = 1$ for $X_{ijk} = X_t$.

For determination of the smoothing parameters τ_i different methods are used. Habbema et.al. (1974) used the Jackknife modification of the maximum likelihood method and found a value of τ_i that maximizes the function

$$L(\tau_i|T) = \prod_{j=1}^{n_i} \hat{f}(X_{ij}|\ T-X_{ij}, \tau_i)$$

$$= \prod_{j=1}^{n_i} (n_i-1)^{-1} \sum_{r=1, j\neq r}^{n_i} (\tau_i)^{-G}\ k\{(X_{ij} - X_{ir})/\tau_i\}, \qquad 2.44$$

where T is the training sample and $T-X_{ij}$ denotes the training sample with item X_{ij} removed.

Murthy (1972) showed that if $k_1(X_1), k_2(X_2), \ldots, k_G(X_G)$ are G one dimensional kernel functions, then

$$k(X_1, X_2, \ldots, X_G) = k_1(X_1).k_2(X_2).k_3(X_3)\ldots k_G(X_G) \qquad 2.45$$

is a G dimensional kernel function. Thus, by multiplying the various kernel functions corresponding to each variable, one could find a kernel function for estimating $f_i(X)$.

The kernel function has the ability to simultaneously handle various types of data when classifying an item of unknown origin into population π_i which it most likely resembles.

Van Ness and Simpson (1976) concluded that non-parametrical

methods should not be used when available parametrical methods are appropriate. If the assumptions for the parametrical methods do not hold, the kernel method with minimal assumptions may be considered. Note that the kernel method only utilizes data in the neighbourhood of an item X and does not base the estimate of the density function on the entire data sample, as is done in parametrical methods.

2.2.2 The K-Nearest Neighbour Procedure

The kernel method calculates a probability estimate based on the proportion of sample points falling in a specific volume around the item to be assigned. The K-nearest neighbour technique fixes the proportion, k, of the design set points and then finds the volume which contains the k nearest items. From this number and volume the density can be estimated.

The non-parametrical estimates of $f_1(X)/f_2(X)$ are derived as follows:

Let $X_{11},\ldots,X_{1n_1}$ be a sample from population π_1 and $X_{21},\ldots,X_{2n_2}$ be a sample from population π_2. Let X be an item to be assigned to either population π_1 or population π_2.

In classifying **X** as belonging to population π_1, two methods of distance can be used, namely :

i) the Mahalanobis distance, based on the total covariance matrix Σ between X_{ij} and X,

$$D^2 (X_{ij},X) = (X_{ij} - X)' \Sigma^{-1} (X_{ij} - X) \text{ and} \qquad 2.46$$

ii) the Euclidean distance

$$D^2 (X_{ij},X) = (X_{ij} - X)' (X_{ij} - X). \qquad 2.47$$

Using one of the above mentioned distance functions, $D^2 (X_{ij},X)$, order the values $D^2(X_{ij},X)$.

Choose some integer, K, and let K_i be the number of observations from π_i to the K closest items to X. Then assign the item X to population π_i, if

$$K_1/n_1 > K_2/n_2. \qquad 2.48$$

In the general G population situation,
classify an item **X** as belonging to population π_i, if

$$K_i/n_i = \max_j (K_j/n_j) \quad i=1,2,..,G \quad . \qquad 2.49$$

A simple generalization when the rule is to account for unequal a priori probabilities is:

Assign an item **X** to population π_i, if
$(K_1/n_1) / (K_2/n_2) > P(\pi_2)/P(\pi_1)$. 2.50

These estimates are consistent and the error rate tends to the error rate of the maximum likelihood rule when $n_i \dashrightarrow \infty$.

In paragraph 2.3 various discrimination methods are compared. Goldstein (1975) concluded that the results derived from both the kernel and K-nearest neighbour methods compared favourably with the results obtained by using parametrical methods.

Note that the K-nearest neighbour method is applied in chapter 6. Although reasonable results were obtained, the application indicated that the linear discriminant function would have been more appropriate.

2.2.3 Ranking

The ranking of data for classification purposes was introduced by Lachenbruch (1975) and Moore and Smith (1975). This procedure may be used when classifying items from two or more populations:

Replace each of the m measurements from the n_i observed items by a rank, where the smallest rank is given to the smallest value. By interpolation an item **X** will be compared to all measurements and a rank will be denoted. Ordinary parametrical methods are then applied but the original data is replaced by corresponding ranks.

Canover and Iman (1980) found that this procedure compared favourably with various parametrical procedures.

2.2.4 Remarks

Non-parametrical procedures have the advantage of not requiring an exact distribution of measurements. Obvious disadvantages are that computation is lengthy, computer programs are not always readily available and procedures are less powerful than the parametric procedures.

2.3 Comparisons between different discriminant analysis approaches .

Various authors compared discriminant analysis procedures, eg. Koffler and Penfield (1979), Remme et.al. (1980), Titterington et.al (1981), Hawkins (1982), Knoke (1982), Schmitz et.al. (1983), Nakaniski and Sato (1985) and Schmitz et.al. (1986).

After studying the posterior probabilities the above mentioned authors reached similar conclusions. It is respectfully suggested that their conclusions are correct. They reached the following conclusions :

The linear discriminant function was superior to any of the other procedures when the data were multivariate normally distributed with equal covariance matrices. When

the normality condition holds and when unequal covariance matrices are present, the quadratic discriminant function gave the best result in terms of the misclassification rate. The quadratic discriminant function is, however, inappropriate for small sample sizes. When small sample sizes occur, the linear and quadratic discriminant functions' classification can be improved by using the Bayesian or predictive approach.

The kernel method produces extremely good results, but it does not outperform the linear discriminant function when the needed assumptions of normality and equal covariance matrices are met.

Schmitz et al. (1983) found that for unequal covariance matrices the kernel method was better than or equal to the quadratic discriminant function.

Nakanishi and Sato (1985) investigated the classification of observations from non-normal distributions on the basis of the misclassification rate. They showed that the sign of the Skewness of each population and the Kurtosis has an essential effect on the linear and quadratic discriminant functions. For the use of the linear or the quadratic discriminant functions the following factors should be considered:

- In skew data, the misclassification rates vary less when applying the linear discriminant function in

comparison to the quadratic discriminant function.

- If the mean and variance for each variable in population 2 is larger the that of population 1 and if the distribution has a positive Skewness, the misclassification rates for the linear discriminant function are lower than the results obtained when applying the quadratic discriminant function. However, if the Skewness is negative, the quadratic discriminant function should be applied.

- Be cautious to use either the linear or the quadratic discriminant function when the kurtosis is small, even in large sample sizes.

Avoid the exclusive use of the linear discriminant function, since this procedure is optimal only when the assumption of normality and equal covariance matrices are met. In practice there is no best method when applying discriminant analysis. The circumstances will indicate the most applicable method.

CHAPTER 3

3.1 Estimating the Probability of Misclassification

The probability of misclassification (also known as the error rate of misclassification) is a measure of the accuracy of a discriminant function when used for classifying items of unknown origin. This chapter deals with estimating the probability of misclassification. For simplicity's sake the analysis will be restricted to measuring the probability of misclassification where the costs of misclassification are set equal to one. The methods outlined below can be extended to deal with the more general case, namely that in which the costs of misclassification are not all equal.

A discriminant function is evaluated by determining its performance when classifying future items of unknown origin. Let $f_1(X)$ be the density function of item X if this item comes from population π_1 and let $f_2(X)$ be the density function of X if it comes from population π_2. Let $P(\pi_1)$ be the a priori probability that an item X comes from population π_1 and let $P(\pi_2)$ be the a priori probability that an item X comes from population π_2. Further suppose that item X is assigned to population π_1 if item X is in a region Ω_1 and to population π_2 if item X is in a region Ω_2. These regions are mutually exclusive and their union includes the entire space Ω.

The total probability of misclassification is :

$$T(\Omega,f) = P(\pi_1) \int_{\Omega_2} f_1(X)dX + P(\pi_2) \int_{\Omega_1} f_2(X)dX \quad . \qquad 3.1$$

The first argument (Ω) refers to the classification region while the second argument (f) is the presumed distribution of the items that will be classified. The quantity is minimized if Ω_1 is chosen so that $P(\pi_2)f_2(X) - P(\pi_1)f_1(X) < 0$ for all points in Ω_1.

The probability of misclassification is used to assess the quality of the classification function, eg. whether it is economically worthwhile implementing the method. Naive estimates of the probability of misclassification are biased and therefore poor decisions are made. Methods for obtaining unbiased estimates are now considered.

In discriminant analysis the training sample is used for two purposes:

i) to estimate an optimal classification rule, and
ii) to estimate the probability of misclassification.

As the same data used to estimate the optimal classification rule are also used to estimate the probability of misclassification, care must be taken to avoid obtaining optimistically biased assessments of these quantities. The bias associated with some estimators of the probability of misclassification is such that it favours complete

models (eg. models with large numbers of parameters). In particular when selecting variables for inclusion into the classification rule (a problem that is discussed in chapter 4) such estimates are biased to the point where all the variables are always selected, even variables which might lead to an increase in the probability of misclassification. Thus obtaining unbiased estimators of the probability of misclassification is of particular importance for the purpose of variable selection.

If $f_1(X)$ is a multivariate normal distribution with known means μ_1 and covariance matrix Σ, the probabilities of misclassification within each population in the two population situation are calculated as follows :

$$P_1 = \int_{\Omega_2} f_1(X)dX \qquad \text{and} \qquad P_2 = \int_{\Omega_1} f_2(X)dX \ , \qquad 3.2$$

where Ω_1 and Ω_2 are given in equation 3.1.

In chapter 2 the special case where populations π_1 and π_2 are multivariate normally distributed, Z_{12}, is a linear discriminant function corresponding to population π_1 and population π_2. If item X is from population π_1, Z_{12} is normally distributed with mean $(0.5\ \hat{MD}^2_{12})$ and variance $\hat{MD}^2_{12}$, whereas if item X comes from population π_2, Z_{12} has mean $(-0.5\ \hat{MD}^2_{12})$ and variance $\hat{MD}^2_{12}$ where

$$\hat{MD}^2_{12} = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \qquad 3.3$$

is the Mahalanobis distance between the two populations.

As derived in chapter 2, the probabilities of misclassification in the two group situation with equal covariance matrices are given by

$$P_1 = P[\text{Misclassification} \mid X \text{ from } \pi_1]$$

$$= \Phi\left[\frac{\ln\{P(\pi_2)/P(\pi_1)\} - 0.5\ MD_{12}^2}{MD_{12}}\right]$$

and

$$P_2 = P[\text{Misclassification} \mid X \text{ from } \pi_2]$$

$$= \Phi\left[\frac{-\ln\{P(\pi_2)/P(\pi_1)\} - 0.5\ MD_{12}^2}{MD_{12}}\right] \qquad 3.4$$

where Φ is the standard normal distribution function.

The expected probability of misclassification for a randomly chosen item from population π_1 or population π_2 is:
$P = P(\pi_1)P_1 + P(\pi_2)P_2$ and is also called the error rate of the classification rule.

For $P(\pi_1) = P(\pi_2)$, both misclassification probabilities in equation 3.4 are equal, so that the error rate becomes:

$$P = \Phi(-0.5\ MD_{12}). \qquad 3.5$$

The computation of misclassification probabilities in the two group situation containing unequal covariance matrices is more complicated, but Bayne and Tan(1981) developed a theory for this situation. They examined the effect of unequal covariances and population distances on the classification probabilities in the presence of known population parameters. Their approximation of misclassification probabilities is limited to the discrimination between two bivariate normal populations. Bayne and Beauchamp (1984) developed a program to compute these misclassification probabilities.

Bayne and Tan's quadratic misclassification probability for two bivariate normal populations containing unequal covariance matrices is thus :

$$P(Q) = P(\pi_1)P(2/1) + P(\pi_2)P(1/2) \quad , \qquad 3.6$$

where P(i/j) is the probability of assigning an item from population j to population i.

Classify an item **X** into one of two populations. If $N(\mu_i, \Sigma_i)$; i=1,2 is the density function of the random mx1 vector **X** in population i, then the logarithm of the likelihood ratio function $\tilde{Q}_{12}(X)$, minimizes P(Q) by the rules :

Assign an item **X** into population π_1, if

$$\tilde{Q}_{12}(X) < K$$

and to population π_2, if

$$\tilde{Q}_{12}(X) \geq K \qquad 3.7$$

where $\tilde{Q}_{12}(X)$ is called Fisher's quadratic discriminant function defined by :

$$\tilde{Q}_{12}(X) = (X - \mu_1)'\Sigma_1^{-1}(X - \mu_1) - (X - \mu_2)'\Sigma_2^{-1}(X - \mu_2)$$

and K is :

$$K = \ln (|\Sigma_2|/|\Sigma_1|) + 2 \ln \{P(\pi_1)/P(\pi_2)\}. \qquad 3.8$$

When the covariance of population π_1 and population π_2 are equal (i.e. $\Sigma = \Sigma_1 = \Sigma_2$), the quadratic discriminant function reduces to Fisher's linear discriminant function.

For the situation containing equal covariance matrices and $G > 2$ populations, Bonferroni's first inequality is used to obtain an upper bound for the probabilities of misclassification :

$$P_i = P[\text{Misclassification} \mid X \text{ from } \pi_i]$$

$$\leq \sum_{\substack{j=1 \\ j\neq i}}^{G} \Phi\left[\frac{\ln\{P(\pi_j)/P(\pi_i)\} - 0.5\, \overline{MD}_{ij}^{2}}{MD_{ij}}\right] \qquad 3.9$$

$$= \sum_{\substack{j=1 \\ j\neq i}}^{G} \Phi\,(-0.5\, MD_{ij}) \text{ if } P(\pi_j)=P(\pi_i)\ ,\ j=1,\ldots G\ .$$

3.2 Probabilities of misclassification (Equal Covariance matrices)

The function $T(\Omega,f)$ defines the error rates. The classification region is represented by the first argument (Ω), while the second argument (f) is the presumed distribution of the items to be classified.

There are three probabilities of misclassification that can be considered:

1. The total probability of misclassification :

$$T(\Omega,f) = P(\pi_1) \int_{\Omega_2} f_1(X)dX + P(\pi_2) \int_{\Omega_1} f_2(X)dX \quad . \qquad 3.10$$

This function gives the true probability of misclassification under the population-based classification rules.

2. The conditional probability (actual error rate) :

$$T(\hat{\Omega},f) = P(\pi_1) \int_{\hat{\Omega}_2} f_1(X)dX + P(\pi_2) \int_{\hat{\Omega}_1} f_2(X)dX \quad . \quad 3.11$$

The actual error rate gives the true probability of misclassification for the estimated discriminant function. (Note that $T(\hat{\Omega},f)$ depends on the training sample and is therefore a random sample.)

3. The expected probability for discriminant functions based on samples of n_1 from population π_1 and n_2 from population π_2 is :

$$E(T(\hat{\Omega},f)) = E\left[P(\pi_1) \int_{\hat{\Omega}_2} f_1(X)dX + P(\pi_2) \int_{\hat{\Omega}_1} f_2(X)dX\right] . \quad 3.12$$

This is the expectation of the random variable in 2, i.e. the expectation over the training sample of the same composition.

The conditional probability of misclassification (actual error rate) based on the training sample measures the performance of a particular discriminant function and should be estimated.

The simplest estimate of the actual error rate is obtained by replacing the parameters f_1 and f_2 by estimates obtained

from the training sample. This "plug-in" estimate when using estimated parameters for f_1 and f_2 is :

$$\hat{T}(\hat{\Omega},\hat{f}) = P(\pi_1) \int_{\hat{\Omega}_2} \hat{f}_1(X)dX + P(\pi_2) \int_{\hat{\Omega}_1} \hat{f}_2(X)dX \quad . \quad 3.13$$

For the two population situation with equal prior probabilities, this yields :

$$\hat{P}_1 = \hat{P}_2 = \Phi(-0.5\ D_{12}) \qquad 3.14$$

where

$$\hat{D}^2_{12} = (X_{1.} - X_{2.})' S^{-1} (X_{1.} - X_{2.})$$

is the sample-based Mahalanobis distance between populations π_1 and π_2. This estimator can also be obtained by estimating MD^2_{12} by $\hat{D}^2_{12}$ in the probability defined in equation 3.5.

However, this estimator is biased for moderate sample sizes and gives a too favourable impression of the true probability of misclassification.

Hills (1966) proved that :

$$E[\Phi(-0.5\ D_{12})] < \Phi(-0.5\ MD_{12}). \qquad 3.15$$

Thus the expected value in equation 3.14 is less than

the optimum probability defined in equation 3.5.

An empirical estimator, the apparent error rate, is defined as the fraction of items from the training sample which are misclassified by the sample discriminant function. (See below.)

The estimation of error rates received considerable attention in the past. The apparent error rate proposed by Smith (1949) has the advantage of being easy to calculate and does not require any distributional assumptions. It is seriously biased however and underestimates the expected error rate. For this reason the Jackknife estimate was proposed by Lachenbruch (1967). Efron (1979) proposed the bootstrap as an alternative estimate of the error rate. Efron found that although the bootstrap and jackknife estimates have similar bias the bootstrap estimates are less variable.

3.3 The following estimation methods will now be considered:

1. The apparent or count estimate,
2. cross-validation and the jackknife estimate and
3. the bootstrap estimate.

3.3.1 The apparent or count estimate

The count estimate or apparent error is the proportion of the training sample that is misclassified.

For either the linear or quadratic functions the discriminant function can be written as :

$$0.5\ [\ Q_2\ (X) - Q_1(X)\ ] + 0.5 \ln |\Sigma_2| / |\Sigma_1|, \qquad 3.16$$

where $Q_i(X) = (X-\mu_i)' \Sigma_i^{-1} (X-\mu_i)$.

In the equal covariance matrix situation the term involving the logarithm is zero and the quadratic terms in X vanish.

If $\Sigma_1 \neq \Sigma_2$, we estimate μ_i by $\bar{X}_i$ and Σ_i by S_i, the within-group means and covariance matrices, for the sample quadratic discriminant function.

To obtain the apparent error rate, we calculate

$$Q_{11}(X_i) = (X_i - \bar{X}_1)' S_1^{-1} (X_i - \bar{X}_1) ,$$
$$Q_{22}(X_i) = (X_i - \bar{X}_2)' S_2^{-1} (X_i - \bar{X}_2) \text{ and}$$

$$Q_e(X_i) = 0.5\ [\ Q_{22}(X_i) - Q_{11}(X_i)] + 0.5 \ln |S_2|/|S_1| \qquad 3.17$$

$Q_e(X_i)$ is the empirically derived discriminant function; i.e. X_i is classified to population π_i if $Q_e(X_i) > 0$ and vice versa.

This is repeated for each item. One obtains the apparent error by counting the number of items misclassified.

For the linear discriminant function

$$Q_{12}(X_i) = (X_i - \bar{X}_1)' S^{-1} (X_i - \bar{X}_2) ,$$

$$Q_{11}(X_i) = (X_i - \bar{X}_1)' S^{-1} (X_i - \bar{X}_1) \quad \text{and}$$

$$Q_{22}(X_i) = (X_i - \bar{X}_2)' S^{-1} (X_i - \bar{X}_2)$$

are needed where S is the pooled covariance matrix.

The apparent error rate can then be calculated by first evaluating,

$$Q_A(X_i) = 0.5 [Q_{22}(X_i) - Q_{11}(X_i)], \qquad 3.18$$

for each Xi. The apparent error rate is then obtained by counting the number of misclassified items.

The following paragraphs will illustrate that other estimates which are preferred to the apparent error rate exist. (Note: Equations 3.17 and 3.18 are used in 3.19 and 3.20 respectively.)

3.3.2 Cross-validation and the Jackknife estimate

As has been mentioned, the apparent error rate tends to be a biased assessment of the true error rate. A more

accurate assessment is obtained by dividing the data set in half and using the one half to calculate a classification function while the other half is then classified with the calculated function. This method, **cross validation**, can be used to obtain a more realistic estimate of the true error rate. A serious setback of this method is that only a portion of the available information is utilized when obtaining the classification function. Important information therefore goes to waste. Where only a limited number of observations are available one can obviously not afford to discard information in this manner. A method that uses available information to a greater extent is therefore needed.

More information is utilized when an item is classified by the classification rule obtained when using the full data set but omitting only the item that needs to be classified from the sample. This procedure is repeated for all the items and the proportion of misclassified items is determined. This refinement of cross validation is often called the **jackknife** estimate.

Lachenbruch (1967) proposed the jackknife procedure. By repeating the procedure outlined above for all the items from population π_i he subsequently obtained the observed proportion of misclassified items. (See Lachenbruch(1975) p36,37.)

For the jackknife estimate, using 3.17, we first calculate

$$Q_J(X_i) = Q_A(X_i) - 0.5 \left[\frac{Q_{11}(X_i) + \bar{Q}_{11}(X_i)}{n_1 - 1 - Q_{11}(X_i)} + \ln\{1 - (n_1-1)^{-1} Q_{11}(X_i)\} + m \ln\{n_1/(n_1-1)\} \right]$$

if an item X_i comes from population π_1.

We then calculate the proportion of items from population π_1 in the training set which would be misclassified using a discriminant rule based on $Q_J(X_i)$. Note that since $Q_J(X_i) < Q_A(X_i)$ this proportion will be larger than the apparent error rate. The jackknife estimate of the proportion misclassified from population π_2 is similarly calculated.

$$Q_J(X_i) = Q_A(X_i) + 0.5 \left[\frac{Q_{22}(X_i) + \bar{Q}_{22}(X_i)}{n_2 - 1 - Q_{22}(X_i)} + \ln\{1 - (n_2-1)^{-1} Q_{22}(X_i)\} + m \ln\{n_2/(n_2-1)\} \right]. \qquad 3.19$$

For the linear discriminant function where S is the pooled covariance matrix, we calculate the jackknife estimate in similar fashion, using 3.18, by

$$Q_J(X_1) = 0.5\left[\frac{v-1}{v}Q_{22}(X_1) + \frac{C_1(v-1)}{v^2}\frac{[Q_{12}(X_1)]^2}{1-(C_1/v)Q_{11}(X_1)}\right.$$

$$\left. - C_1^2\left[\frac{v-1}{v}Q_{11}(X_1) + \frac{C_1(v-1)}{v^2}\frac{Q_{11}^2(X_1)}{1-(C_1/v)Q_{11}(X_1)}\right]\right]$$

if an item X_1 comes from population π_1 and if an item X_1 comes from population π_2,

$$Q_J(X_1) = 0.5\left[C_2^2\left[\frac{v-1}{v}Q_{22}(X_1) + \frac{C_2(v-1)}{v}\frac{Q_{22}(X_1)}{1-(C_2/v)Q_{22}(X_1)}\right.\right.$$

$$\left.\left. - \frac{v-1}{v}Q_{11}(X_1) - \frac{C_2(v-1)}{v^2}\frac{[Q_{12}^2(X_1)]^2}{1-(C_2/v)Q_{22}(X_1)}\right]\right], \qquad 3.20$$

where $C_1=n_1/(n_1-1)$ and $v=n_1+n_2-2$.

3.3.2.1 Remarks

Lachenbruch and Mickey (1968) showed that the jackknife estimate is superior to the apparent error rate in the two population situation. When the assumptions of normality and homoscedasticity are violated, estimates of the error rate based on these assumptions become unreliable and

Lachenbruch (1975) recommends using the jackknife estimate under such circumstances.

The jackknife estimate does not justify the use of the linear or quadratic discriminant function in any application. It estimates the error rate of the classification rule, whether or not the rule is appropriate.

3.3.3 The Bootstrap estimate

An alternative method for estimating error rates is the **bootstrap** method. Efron (1983) showed that the bootstrap method is essentially a non-parametrical maximum likelihood estimate of the true error rate. As noted above, Efron (1979) found that the bootstrap procedure estimates with similar bias but with less variability than the jackknife procedure.

Efron (1983) constructed a prediction rule for estimating the error rate in classifying future observations when the training sample size is small. By using the double bootstrap (bootstrapping the bootstrap) the bias of the ordinary bootstrap can be corrected.

In terms of the mean squared error, Efron´s bootstrap error rate estimator is computed as follows :

$$\text{(apparent error rate)} - E(B) \qquad 3.21$$

where B is the bootstrap estimator of the bias in using the apparent error to estimate the true error rate. B is computed as follows (the two population situation is considered for convenience's sake) :

For population π_i (i=1,2), from the training set (the data set of size n_i from which the classification rule will be calculated) draw with replacement a random sample of size n_i. Call this sample BS_i. Call the remaining undrawn sample from population π_i, BS_{i*}.

By using BS_1 and BS_2 as training sets, classify all the n_i items in the training set and let the proportion of misclassified items be $\hat{P}_i$. Using the classification functions obtained from BS_1 and BS_2, classify all the BS_{i*} items. Let

the proportion of misclassified items herein be $\hat{P}_i^*$.

Then $B = \hat{P}_i - \hat{P}_i^*$. 3.22

Continue this procedure and draw repeated samples with replacement for each of the populations. Calculate the discriminant function and estimate the error rate in each case and then average the error rates over all replications to obtain the bootstrap estimate of the error rate.

3.3.3.1 Remarks

Although this method involves extensive computation, the estimated error is less variable than the corresponding estimate obtained by the jackknife procedure. According to Efron (1983) replications in the order of 25-200 seem adequate to calculate the bootstrap error rate.

3.4 Summary

The probability of misclassification is an essential and important measure to evaluate discriminant functions. Since the apparent error rate results in very optimistic probabilities of misclassification, various other methods for obtaining unbiased estimates of the true error rate have been proposed. The bootstrap estimate of the error rate is less variable than the corresponding estimate obtained when using the jackknife method but involves extensive computations which could be expensive.

CHAPTER 4

4.1 VARIABLE SELECTION IN DISCRIMINANT ANALYSIS

For the purpose of classifying an item as belonging to a certain population one needs to make measurements on this item. There is an unlimited number of measurements which could be made on an item but for the purpose of classification the measures are restricted to those providing information relevant for classification purposes. The question arises whether all of the measurements (m) are needed for classification purposes or whether a subset q of m (where $q<m$) would be sufficient.

Assume that each of the m variables has some discriminatory power, no matter how small. By excluding any of the measures the potential expected cost of misclassification is increased. In other words, if the $f_i(X)$ were known, reducing the number of variables would lead to a poorer classification rule. On the other hand, some variables are less important for the purpose of discriminant analysis. By excluding measurements (variables) with a marginal contribution, the number of parameters to be estimated in the models $f_i(X,\theta_i)$ can be reduced. The performance of the classification rule will therefore increase and the expected cost will be reduced, since there is on average less sampling variation in the estimation of $f_i(X,\hat{\theta})$. This consideration favours a reduction in the number of variables. The selection of

variables further reduces the number of measurements needed in future. Consequently the cost of obtaining information is reduced substantially.

In deciding how many and which measurements to use in the classification rule, two competing effects have to be taken into account. No unique statistical procedure exists to determine which measurements should be selected. A great deal of judgement is therefore required. The fact that different variable selection methods do not necessarily lead to the same solution when applied to the same problem further confuses the issue.

Be cautious when implementing univariate statistical methods to eliminate individual variables that do not significantly contribute to the discrimination between the groups. Implementation of univariate statistical methods could lead to incorrect selections, since not only individual variables but intercorrelations between variables should be taken into account when selecting variables. Two or more individual variables, taken on their own, might not be good discriminators. Used together, they could prove to be highly effective. Such multivariate relationships are difficult to determine and therefore variable selection techniques are used in the search for a suitable subset of variables.

4.2 Selection criteria

Selection is based on the following philosophy: use the simplest model which is not inconsistent with the data.

Suppose variables $X_q=(X_1,X_2,..,X_q)$ have already been selected. It is now questioned whether to include a further set of variables $X_{q'}=(X_{q+1},...,X_m)$. Using the above selection philosophy, additional variables will only be included if it can be proved, in the usual statistical sense, that they add to the discriminating power of the test.

The first decision is what measure of discriminating power, or criterion, to use. In this section a few selection criteria are discussed.

When using hypothesis testing as the criterion for selection, the null hypothesis tests whether the Mahalanobis distance between two populations is the same, irrespective of whether only X_q or both $(X_q,X_{q'})$ are used. The additional variables are included only if the null hypothesis can be rejected, otherwise the simpler model based on X_q is used.

The use of criteria based on hypothesis testing is criticized for placing emphasis on populations that are far apart. In practical applications this is not always the case.

In an effort to overcome this problem, Hawkins (1982) suggested :

- consider the variable, that after inclusion, maximizes the minimum between-groups Mahalanobis distance and

- include it if it provides significant additional discrimination between the populations.

Estimation of the error rate, or more generally, the cost of misclassification was suggested by Habbema and Hermans (1977) to be a more appropriate criterion for variable selection. To test whether a variable should or should not be included in the discriminant function, the error rate is estimated in both cases with and without the subset included in the discriminant function. The difference in the error rates will determine whether a subset should be included. Note that the computation of these error rates can be extremely expensive, since methods such as the jackknife and bootstrap estimates are used to determine the true error rate. These methods are extremely computer intensive. (Refer to chapter 3 for estimates of the error rates.)

Mardia (1979) proposed a rule of thumb for inclusion of variables in the discriminant function. Retain enough measurements for the squared multiple correlation, using q measurements, to be at least 90-95% of the squared multiple correlation, using all m measurements.

A Bayesian approach to variable selection was proposed by Menzefricke (1981): An additional subset of variables is included for future classification purposes if the additional measurement costs for this subset are lower than the resulting reduction in expected misclassification costs.

Rao's criteria for two populations, Wilk's lambda and the generalization of Rao's criteria by Kshirsagar will now be formulated.

4.2.1 Selection criteria based on test of hypothesis

4.2.1.1 Rao's selection criteria

Rao (1972) suggested that the Mahalanobis distance between the populations be used for the selection of variables. The null hypothesis tested for two populations is that the Mahalanobis distance between the two populations is the same whether one uses X_q only or both $(X_q, X_{q'})$. The additional variables are included only if the above hypothesis can be rejected; otherwise the simpler model based on X_q is used.

The null hypothesis can also be formulated as follows : The coefficients of the variables $(X_{q+1}, \ldots, X_m)$ in the linear discriminant function are zero.

If $MD^2_{(q)}$ is the sample Mahalanobis distance between the two populations, based on q measurements, and $MD^2_{(m)}$ is the corresponding distance based on all m variables in X, then Rao (1972) derived the following F test statistic :

$$F = \frac{(v - m + 1)/(m - q)\,[n_1 n_2/(n_1 + n_2)]\,(MD^2_{(m)} - MD^2_{(q)})}{v + (n_1 n_2/(n_1 + n_2))\,MD^2_{(q)}} \qquad 4.1$$

with $v=n_1+n_2-2$.

Under the null hypothesis this F statistic has a F distribution with (m-q) and (v-m+1) degrees of freedom.

If q=m-1, Rao's statistic tests whether a single measurement can be dropped without affecting the overall discrimination power.

Rao's statistic then becomes:

$$F = \frac{(v - m + 1)/\,[n_1 n_2/(n_1 + n_2)]\,(MD^2_{(m)} - MD^2_{(m-1)})}{v + (n_1 n_2/(n_1 + n_2))\,MD^2_{(m-1)}} \; . \qquad 4.2$$

Under the null hypothesis this statistic has a squared t distribution with (v-m+1) degrees of freedom .

4.2.1.2 Wilk's lambda

Before generalizing Rao's criteria, Wilk's Lambda needs to be defined. This is the likelihood ratio statistic for testing the hypothesis that the means of the population on the selected measurements are equal. Furthermore, a statistic will be given to test whether the chosen measurements provide significant discrimination between the groups.

To obtain this likelihood ratio, refer to known formulae (See paragraph 2.1.1.4). The within-groups sum of squares

matrix is defined as

$$W = \sum_{i=1}^{G} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(X_{ij} - \bar{X}_{i.})' \quad , \qquad 4.3$$

and the between-groups sum of squares matrix is

$$A = \sum_{i=1}^{G} n_i (\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})'. \qquad 4.4$$

The matrix containing the total groups sum of squares is

$$T = \sum_{i=1}^{G} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})(X_{ij} - \bar{X}_{..})' \qquad 4.5$$

$$= \sum_{i=1}^{G} \sum_{j=1}^{n_i} \{(X_{ij} - \bar{X}_{..}) - (\bar{X}_{i.} - \bar{X}_{..})$$
$$(X_{ij} - \bar{X}_{..}) - (\bar{X}_{i.} - \bar{X}_{..})'\} +$$
$$\sum_{i=1}^{G} n_i (\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})'$$

$$= \sum_{i=1}^{G} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(X_{ij} - \bar{X}_{i.})' +$$
$$\sum_{i=1}^{G} n_i (\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})'.$$

$$= W + A.$$

The aim is to use **T**, **A**, and **W** as the basis for a separability measure. A univariate function that can be optimized therefore needs to be found. A univariate function based on these matrices is $|W|/|T|$. The statistic $|W|/|T|$ is known as Wilks's Lambda which is an overall measure of

the between-group differences. The aim is now to minimize $|W| / |T|$.

By minimizing $|W|/|T|$ (which is equivalent to maximizing $|I+W^{-1}A|$) we can obtain τ an eigenvalue of $W^{-1}A$, with v the corresponding eigenvector.

By setting

$$U = |W|/|A+W| = |W|/|W||W^{-1}A + I| = |W^{-1}A + I|^{-1}$$

this ratio has the Wilk's lambda distribution [U(p,m,n)] with parameters p,m and n (Mardia (1979)).

It follows that

$$W^{-1}A\, v - \tau I\, v = 0 ,$$

which can be written as

$$(W^{-1}A + I)v - (\tau+1)Iv = 0. \qquad 4.6$$

It thus follows that, if τ is an eigenvalue of $W^{-1}A$, then $(\tau+1)$ is an eigenvalue of $(W^{-1}A + I)$.

Further, let r be the number of non-zero eigenvalues of $(W^{-1}A + I)$. Write U in terms of its r roots as :

$$1/U_1 = (1+\tau_1)(1+\tau_2) \ldots (1+\tau_r)$$

$$= \prod_{i=1}^{r} (1+\tau_i) \quad . \qquad 4.7$$

Bartlett's (1947) V statistic can now be defined as :

$$V_1 = -[n-1-(m+G)/2] \ln U_1$$

$$= [n-1-(m+G)/2] \ln 1/U_1. \qquad 4.8$$

V_1 provides a test that τ_1 through τ_r are equal to zero. V_1 has approximately a chi-square (m(G-1)) distribution. If this hypothesis is rejected, it can be concluded that at least one of the parameters corresponding to the τ's is greater than zero. Because τ_1 is the maximum root, it can now be considered statistically significant.

If V_1 results in rejection of the null hypothesis, remove the first root τ_1 and obtain

$$1/U_2 = (1+\tau_2) \ldots (1+\tau_r) = \prod_{i=2}^{r} (1+\tau_i) \quad . \qquad 4.9$$

The corresponding Bartlett's V is :

$$V_2 = [n-1-(m+G)/2] \ln (1/U_2). \qquad 4.10$$

This formula (4.10) tests whether τ_2 through τ_r are all equal to zero. Rejecting this hypothesis, it can be concluded that τ_2 is significant.

In general, a test of significance for the k-th root τ_k is:

$$V_k = [n-1-(m+G)/2] \ln (1/U_k) \qquad 4.11$$

in which

$$1/U_k = \prod_{i=k}^{r} (1-\tau_i) \qquad 4.12$$

with V_k distributed chi-square $(m-k+1)(G-k)$ assuming a multivariate normal distribution of the m X's.

When this hypothesis is not significant it can be concluded that no values smaller than τ_k will be significant and no further tests need to be done. Thus only the eigenvalues greater than τk provide significant discrimination between the populations and only those measurements corresponding to the eigenvalues should be included when calculating the discriminant function.

4.2.1.3 Kshirsagar's criterion

Rao's criteria were generalized by Kshirsagar (1972) to the k-population situation by using the factorization of Wilk's lambda, the likelihood ratio statistic :

$$\frac{|W|}{|T|} = \frac{W_{11}\ W_{22.1}\ W_{33.1,2}\ldots W_{mm.1,2,\ldots,m-1}}{t_{11}\ t_{22.1}\ t_{33.1,2}\ldots t_{mm.1,2,\ldots,m-1}} = \prod_{i=1}^{m} LR_i$$

4.13

where T=W+B is the total sum of squares. $w_{i i.1,2\ldots,i-1}$ is the first diagonal element in the within-groups sum of squares matrix $W_{22.1} = W_{22} - W_{21}W_{11}^{-1}W_{12}$.

W has been partitioned according to the first i-1 and the last m-i+1 variables:

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}.$$

$t_{i i.1,2\ldots,i-1}$ is similarly defined, and

$LR_i = w_{i i.1,2\ldots,i-1} / t_{i i.1,2\ldots,i-1}$.

The F statistic is then :

$$F_i = (LR_i^{-1} - 1)(n-i+1)/(G+1) \qquad 4.14$$

and

$$n = \sum_{i=1}^{G} n_i - G.$$

Under the null hypothesis the F statistic with (G-1) and (n-q+1) degrees of freedom tests whether the measurements X_q provide any additional discrimination power to the q-1 variables already included in the discriminant function.

In the above mentioned criterion an attempt is made to prove (in the usual statistical sense) whether extra variables improve the discrimination.

4.2.2 The Bayesian decision-theoretic approach to variable selection

Menzefricke (1981) described a decision-theoretic approach to variable selection. A small subset of variables measured on individuals of known origin can be used to classify individuals of unknown origin. In this approach to variable selection an additional subset of variables is included for future classification purposes if the additional measurement costs for this subset are lower than the resulting reduction in expected misclassification costs.

Let X_q, $q=(1,2,..,q)$, be a set of variables used for the classification of an item of unknown origin into one of two populations. The cost of misclassifying an item from population π_1 into population π_2 is c_{21} and the cost of misclassifying an item from population π_2 into population π_1 is c_{12} and the measurement cost incurred when using X_q is c_q. Denote the decision to use the variables corresponding to the variable set $\mathbf{X}q$ for discrimination purposes as action a_q. (The expected loss function $EL(a_q)$ will be derived below.) The posterior probability of X_q coming from population π_j is :

$$p(\pi_j|X_q) = p(X_q|\pi_j)\ P(\pi_j)\ /\ p(X_q), \qquad 4.15$$

with $p(X_q) = \Sigma\ P(\pi_j)\ p(X_q|\pi_j)$ and the predictive

distribution of X_q when X_q comes from population π_j is

$$p(X_q|\pi_j) = \int f(X_q|\theta_j)\ p(\theta_j)\ d\theta_j \quad , \tag{4.16}$$

with $f(X_q|\theta_j)$ the density when X_q comes from population π_j, where θ_j is a set of parameters with available information from the training sample with known origin. This information is expressed in a distribution for θ_j, $p(\theta_j)$.

The minimum expected loss for a_q, given X_q, is

$$EL(a_q|X_q) = c_q + \min\{c_{12}\ p(\pi_2|X_q)\ ,\ c_{21}\ p(\pi_1|X_q)\}\ . \tag{4.17}$$

Assign item X_q to population π_1, if

$$c_{21}\ p(\pi_1|X_q) > c_{12}\ p(\pi_2|X_q) \tag{4.18}$$

or if

$$\frac{p(X_q|\pi_1)}{p(X_q|\pi_2)} > \frac{P(\pi_2)c_{12}}{P(\pi_1)c_{21}} = K. \tag{4.19}$$

The following relationship is obtained from the difference between the expected loss of using only vector X_q, $EL(a_q)$, and the expected loss of using additional variables $EL(a_{qr})$ and action a_r is based

on using $X_w = (X_a, X_s)$ which contains additional variables X_s.

The following relationship is derived :

$$
\begin{aligned}
EL(a_a) &= \int p(X_a)\, EL(a_a|X_a)\, dX_a \\
&= \int p(X_a)\, \min\{c_{12}\, p(\pi_2|X_a)\, , c_{21}\, p(\pi_1|X_a)\} dX_a + c_a \\
&= \int p(X_a)\, \min\{c_{12} \int p(\pi_2|X_a, X_s)\, p(X_s|X_a)\, dX_s\, , \\
&\qquad c_{21} \int p(\pi_1|X_a, X_s)\, p(X_s|X_a)\, dX_s\} dX_a + c_a \\
&\geq \iint p(X_a)\, p(X_s|X_a)\, \min\{c_{12}\, p(\pi_2|X_a, X_s)\, , \\
&\qquad c_{21}\, p(\pi_1|X_a, X_s)\}\, dX_s\, dX_a + c_a \\
&= \int p(X_w)\, EL(a_w|X_w)\, dX_w - (c_w - c_a) \\
&= EL(a_w) - (c_w - c_a). \qquad 4.20
\end{aligned}
$$

The inequality between the third and forth lines follows from the well known fact that if B_1 and B_2 are random variables, then

$$E\,[\min(B_1, B_2)] \leq \min\,[E(B_1), E(B_2)].$$

Menzefricke (1981) showed that, if $p(X_w) = p(X_a, X_s) = p(X_a)p(X_s|X_a)$ and if the additional variables, X_s, are observed without cost, one can never be worse off when including these variables for discri-

mination purposes.

Define $A = \{X_q: p(X_q|\pi_1)/p(X_q|\pi_2) > K\}$, where A is the set of all X_q variables that are assigned to population π_1. A more convenient expression for $EL(a_q)$ can then be derived :

$$EL(a_q) = \int_A p(X_q)\, c_{12}\, p(\pi_2|X_q)\, dX_q +$$

$$\int_{AC} p(X_q)\, c_{21}\, p(\pi_1|X_q)\, dX_q + c_q$$

$$= P(\pi_2)\, c_{12} \{ \int_A p(X_q|\pi_2)\, dX_q \} +$$

$$P(\pi_1)\, c_{21} \{ \int_{AC} p(X_q|\pi_1)\, dX_q \} + c_q \qquad 4.21$$

The expressions in braces are the probabilities of misclassifying an item from population π_2 into population π_1 and of misclassifying an item from population π_1 into population π_2.

Let $P_{ij}(q)$ be the probability of misclassifying an item from population j into population i when using X_q for discrimination purposes. The expected loss, using X_q, is

$$EL(a_q) = c_1 + P(\pi_1)\, c_{21}\, p_{21}(q) + P(\pi_2)\, c_{12}\, p_{12}(q). \qquad 4.22$$

Include the subset X_s for discrimination purposes if $EL(a_q) > EL(a_w)$, where $w=q \cup s$, or if

$$c_w - c_q < P(\pi_1)\, c_{21}\, [p_{21}(q) - p_{21}(q \cup s)] + P(\pi_2)\, c_{12}\, [p_{12}(q) - p_{12}(q \cup s)] . \quad 4.23$$

The right side of the equation represents the expected reduction in misclassification costs when the additional subset of measurements is included. Such inclusion is justified if the expected reduction in the misclassification cost is larger than the increase of added measurement cost.

The decision whether to include the additional subset depends on the corresponding reduction in the two probabilities of misclassification given in brackets in equation 4.23.

In the decision-theoretic approach a larger sample size reduces the uncertainty about parameter estimate values. If the additional contribution of measurements to the discriminating power is known to be small, however, its measurement cost must be low to warrant inclusion, even if the sample size is large. If the cost of misclassification is high relative to the measurement cost, then all variables should be included since measuring all the variables is then essentially free and free information should be used.

A number of formal variable selection procedures (i.e. algorithms) suggested in the literature will now be discussed. These selection procedures can be evaluated by using any of the above mentioned criteria. In the following discussion only one of the possible criteria will be used as an example to illustrate the methodology of the procedures.

4.3 Variable selection procedures

To find the "best" set of q of m ($q<m$) available measures for discriminating between populations, each of the possible sets could be evaluated. For m large this is practically not feasible and even for m moderate the computations can be expensive. The search can be accelerated by the use of various selection procedures. In this section algorithms for finding subsets of variables to discriminate between the various populations are discussed. The selected variables are then tested by means of any of the mentioned criteria for their significance to discriminate between populations.

The following procedures will be discussed :

1) Exhaustive search
2) Accelerated search
3) Forward selection
4) Backward elimination
5) Stepwise selection.

4.3.1 Exhaustive search

The exhaustive search method can be implemented when the number of measurements ,m, is small. Mckay (1976) presented a method to find all subsets of measurements whose discrimination power is not significantly worse than the complete set of measurements. His method determines the probability of the overall type I error and the significance level of each individual test. It is important to determine the probability of the overall Type I error because this procedure involves multiple tests.

Testing the complete set of m measurements, the hypothesis that the two populations are the same is rejected, if

$$T^2_m > \frac{(n_1 + n_2 - 2)\, m}{n_1+n_2-1-m} F^{\alpha}_{m,\, n_1+n_2-1-m} \qquad 4.24$$

where T^2_m is the sample Hotelling's T^2 in m dimensions,

$$T^2_m = [n_1 n_2/(n_1+n_2)]\, (\bar{X}_1 - \bar{X}_2)' W^{-1} (\bar{X}_1 - \bar{X}_2) \qquad 4.25$$

where

$$W = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(X_{ij} - \bar{X}_{i.})'$$

and $F^{\tau}m,n_1+n_2-1-m$ is the τ level of the F distribution with m and (n_1+n_2-1-m) degrees of freedom.

If this initial hypothesis is not rejected, the variables in the complete set do not discriminate between the populations and the analysis is completed.

If the hypothesis is rejected, test subsets of measurements to see if they discriminate adequately between the various populations.

If

$$T^2_q < \frac{(n_1+n_2-1-m)\ T^2_m}{n_1+n_2-1-m + m(F^{\tau}m,n_1+n_2-1-m)} - \frac{(n_1+n_2-2)\ m\ F^{\tau}m.n_1+n_2-1-m}{} \qquad 4.26$$

the hypothesis that a subset of measurements $(q < m)$, is as effective as the original set of measurements (m) is rejected.

Where computationally feasible, an exhaustive search can be applied in conjunction with any of the above mentioned criteria.

4.3.2 Accelerated search

A branch and bound method, which considers all variable sets but does not explicitly evaluate them, reduces the computations required for variable selection.

Hand(1981) implements this method to find the best m′ subset of measurements from the complete set of m measurements.

Start with the set containing all m available measurements and then construct a tree by successively deleting variables.

From the node corresponding to the complete measurement set, m new nodes from which a single measure has been deleted can be generated. From each of these nodes (m-1) further nodes can be generated.

eg.

$[x_1\ x_2\ x_3\ x_4]$

$[x_2\ x_3\ x_4]$ $[x_1\ x_3\ x_4]$ $[x_1\ x_2\ x_4]$ $[x_2\ x_3\ x_4]$

$[x_2\ x_3]$ $[x_2\ x_4]$ $[x_3\ x_4]$ ..... etc.

Continue until each node contains m′ measures. The only constraint is that all final nodes should contain m′ measures. Proceeding down the tree, measures are always discarded and never added.

Suppose that a branch of the tree was followed to its final node and that a criterion value $T^2_{j'}$ has been computed for that node. If the criterion value from another branch is $T^2_{j''}$ and less than $T^2_{j'}$, there is no point in proceeding any further down this branch. The aim is to find the final node with the

largest T^2_j value. Select a branch and continue down until either reaching a terminal node or T^2_j becomes less than T^2_{j-1}. If a terminal node with a value T^2_j larger that T^2_{j-1} is reached, this replaces T^2_{j-1} for future stages of the search.

Hand (1981) described this procedure. It can, however, be inefficient and time consuming when numerous measurements need to be tested.

4.3.3 Forward selection procedure

In the forward, backward and stepwise procedures measurements are selected for inclusion in or deletion from a model by evaluating the significance level of an F test from an analysis of covariance where the measurement under consideration is the dependent variable and the measurements already included in the model act as covariates. The F value is equal to the ratio of the mean square of the model corrected for the covariates to the mean square of the error. The F test partitions the variation and tests the amount of variation resulting from the difference between the groups. The probability value for the test should be compared to the reference probability value decided on before running the test.

The forward selection procedure starts with no measurements in the model. At each step the

measurement that contributes most to the discrimination power of the model as measured by Kshirsagar's criterion is entered. For the remaining (m-1) measures the F statistics are recalculated and the measure contributing the most to the discrimination power is then entered. Prior to the implementation of the procedure a stopping rule (a minimum F-statistic value) for measure inclusion into the model is defined. This procedure will stop when no measure has a F-statistic value greater than the inclusion criterion. When a measure has been included in the model, it can not be removed. When the stopping rule is reached only measurements included in the model are used when calculating the discriminant function.

Constanza and Afifi (1979) remarked that a moderate significance level in the range of 10% - 25% should be used as the stopping rule or the procedure will tend to stop before a sufficient number of measurements has been included.

According to Constanza (1979) doubling the sample size improves selection slightly. The main effect of doubling the sample size is to make the stopping rule more sensitive to changes in terms of sizes of the 'best' subsets.

Constanza (1980) examined the classification performance of the forward selection procedure by using small reference samples (both equal and unequal). He concluded that the classification performance was improved by variable selection even when small sample sizes were used.

4.3.4 Backward elimination procedure

The calculations of backward elimination are similar to that of the forward procedure except that all the variables are included in the model at the onset of the procedure. The variables are then excluded from the model, one by one. At each step the variable with the smallest F-value, showing the least contribution to the model, is discarded. The procedure stops when all the remaining variables produce significant F-statistics. The stopping rule's F-value is chosen prior to the implementation of the procedure. The remaining variables in the model significantly contribute in discriminating between the various populations and these variables are then used to compute the discriminant function.

4.3.5 Stepwise procedure

The stepwise procedure is the most commonly used method for selecting variables in discriminant analysis.

This method is a modification of the forward and the backward selection techniques.

This procedure starts similar to the forward procedure with no variables in the initial model. By using Kshirsagar's criterion, the measurements are tested individually for possible inclusion in the model. As in the forward procedure, the measurement contributing most to the discriminating power is entered into the model, with the main difference that the measurements included in the model do not necessarily remain in the model. All the measurements included in the model are tested for the significance of their respective contributions to the model. If a measurement's contribution is not significant it is deleted from the model. Only after this test for deletion could have been resolved can another measurement be added to the model. Measurements not included in the discriminant function are then tested for possible inclusion. A variable with the highest F-statistic, greater than the prior defined minimum F-statistic, is entered into the model. If the F-statistic is less than the prior defined minimum F-statistic, the procedure stops. The variables included in the model are then used to calculate the discriminant function which is in turn used to classify new observations of unknown origin.

If cycling occurs between inclusion and exclusion

of a variable, it can be avoided by making the inclusion significance level stricter than the exclusion level.

The main advantage of the stepwise method is that variables are tested for possible deletion only after their inclusion in the model could have been rendered redundant by further inclusions.

4.4 Conclusion

As mentioned, the exclusion of certain variables reduces the number of parameters which have to be estimated. On average there is therefore less sampling variation in the estimates $\hat{f}_i(\mathbf{X},\hat{\theta})$. Since there exists no unique statistical procedure for determining the selection of measurements for discrimination purposes, the final selection of measurements greatly depends on the statistician's judgement. The statistician must determine whether the reduction in the expected costs when selecting a subset of variables is greater than the increase in expected cost of misclassification when excluding information from some measurements.

This chapter explored a few criteria for testing whether discrimination between populations can be achieved equally well by a subset of measurements as by all of the observed measurements. It further dealt with different selection techniques for finding the 'best' subset of variables for discriminant analysis.

As seen, the disadvantage of the forward selection method is that a chosen variable can not be removed should additions render it redundant.

The disadvantage of the backward elimination procedure is that it requires more demanding computations. Compared to the forward procedure, it has the advantage that the complete variable set can be judged since all the variables are included in the model when the procedure starts. A further disadvantage is that the inclusion of all the variables could lead to an ill-conditioned, or even a singular, sample covariance matrix.

The stepwise method is a modification of the forward and backward selection techniques and has the advantage that variables can be deleted after their inclusion if additions render their inclusion redundant.

Selection results should be interpreted with caution as the selection of a subset of measurements does not necessarily produce the 'best' subset. Various combinations might perform equally well. Large correlations among variables or large correlations between linear combinations of variables can further contribute to selection problems.

The variables selected for the discriminant function should be evaluated with a sample to test the validity of these selected variables. The data set can be divided

into two sets X_A and X_B. The stepwise method can be implemented on X_A. X_B can then be classified by the obtained discriminant function based on X_A. This process can be repeated with X_B as the data set from which a discriminant function will be derived. The first data set, X_A, must then be classified by this new function obtained from X_B. Repeat this procedure a number of times. The occurrence of a variable appearing in the selected subset provides a measure of the importance of that variable for future classification purposes.

In the decision-theoretic approach to variable selection an additional subset of variables is included for future classification purposes if the additional measurement costs for this subset are lower than the resulting reduction in expected misclassification costs. Thus, if the cost of misclassification is high relative to the measurement cost, all variables should be included since measuring all the variables is then essentially free and free information should be used.

It is important to note that if an adequate subset of variables can be found to calculate a discriminant function, only those chosen variables need be measured on future items. This could save cost and time when obtaining future information.

CHAPTER 5

Matters related to discriminant analysis

In this chapter three matters related to discriminant analysis will be touched on briefly. At first the problem of **missing** data values will be discussed and the EM (Expectation-Maximization) algorithm for estimating incomplete data values will be mentioned.

The second section deals with **sequential discriminant analysis**. In this section three different sequential methods will be referred to. By using sequential discriminant analysis the classification rule is gradually built up.

In the final section **logistic discriminant analysis** is briefly mentioned. The advantage of this discriminant analysis procedure is that it simultaneously deals with both continuous and discrete variables.

5.1 Incomplete data values

From the nature of classification problems (where large numbers of variables are measured) some items may have incomplete data (i.e. values not recorded for all variables). Normally even if only one measurement is missing the entire item will be excluded from the analysis since most computer programs on discriminant analysis are not able to deal with incomplete data. To prevent the problem

of losing valuable information, incomplete data values can be estimated. The EM algorithm for estimating incomplete data values will now be outlined:

5.1.1 The EM Algorithm

In this section the EM algorithm for estimating incomplete data values for the case where the observation can be assumed to follow a multivariate normal distribution is outlined. This section is based on Little and Rubin (1987) who discussed the EM algorithm in general and its application to discriminant analysis.

Measure a m-variate item, $\mathbf{X}=(X_1, X_2, .. X_m)$, which has a normal distribution with mean $\mu=(\mu_1, \mu_2, .., \mu_m)$ and covariance $\Sigma=(\sigma_{jk})$. Further, write $X = (X_{obs}, X_{mis}) = (X_1, X_2, .., X_m)$, where X_{obs} is the set of observed values and X_{mis} is the set of missing data values.

The distribution of the complete data **X** can be factorized as the density of the joint distributions of X_{obs} and X_{mis}.

$$f(X|\theta) = f(X_{obs}, X_{mis} | \theta) = f(X_{obs} | \theta) f(X_{mis} | X_{obs}, \theta) \qquad 5.1$$

where $f(X_{obs} | \theta)$ is the density of the observed X_{obs} with parameter $\theta=(\mu, \Sigma)$ and $f(X_{mis} | X_{obs}, \theta)$ is the conditional density of the missing values given the observed values X_{obs}.

The log-likelihood functions can be written as:

$$l(\theta|X) = l(\theta|\ X_{obs}, X_{mis}) = l(\theta|\ X_{obs}) + \ln f(X_{mis}\ |X_{obs}, \theta). \qquad 5.2$$

For the fixed X_{obs} θ is estimated by maximizing the incomplete data likelihood $l(\theta|X_{obs})$ with respect to θ.

The iterative process of the EM algorithm consists of the following steps:

1) replace the missing values by estimated values (eg. means and other estimates discussed below) ,

2) estimate θ by maximizing the log-likelihood $l(\theta|X)$,

3) re-estimate the missing values from the probability density function, assuming that the new parameter estimates are correct and return to step 2.

Continue this iterative process until convergence to a stationary value of $l(\theta|X_{obs})$. Note that convergence is to a local maximum or saddle point of $l(\theta|X_{obs})$ (see Little and Rubin (1987)). In other words even if the algorithm converges this does not guarantee convergence to a global maximum. A disadvantage of the EM algorithm is that convergence can be quite slow if many missing values exist.

In the multivariate situation with some values X_{ij} missing, numerous methods can be used to obtain initial estimates:

$\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m)$ and

$$S = S_{jk} = (n-1)^{-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k).$$

Derive the initial estimates for the parameters of the iteration algorithm as follows :

- select only complete data items containing no missing values when calculating these initial estimates,

- include only those items where variables of interest are present when calculating the initial estimates,

- use methods of imputing. If X_j contains missing values and X_k is highly correlated with X_j, then use X_k to predict the missing values of X_j. Substitute (impute) the predicted values in the analysis involving X_j. The initial estimates can then be estimated.

Since the initial estimates of the parameters have been obtained, the iteration of the EM algorithm can now start. The EM iteration consists of an E step (expectation step) and a M step (maximization step).

Given the observed data and the current estimated parameters, the E step finds the conditional expectation of the "missing

data" and then substitutes these expectations for the "missing data".

When implementing this EM algorithm, note that the "missing values" are not X_{mis} but functions of X_{mis} appearing in the complete-data log-likelihood, $l(\theta|X)$.

The M step performs a maximum likelihood estimate of θ just as if there were no missing data. The M step of EM uses similar computational methods as the maximum likelihood estimation of $l(\theta|X)$.

To derive the EM algorithm assume that the hypothetical complete data set **X** belongs to the regular exponential family with sufficient statistics

$$S = \{ \sum_{i=1}^{n} X_{ij},\ j=1,2,..,m \text{ and } \sum_{i=1}^{n} X_{ij}X_{ik},\ j,k=1,2,..,m\}. \qquad 5.3$$

At the t-th iteration , let $\theta^t = (\mu^t, \Sigma^t)$ denote the current estimates of the parameters. The E step consists of calculating

$$E\left(\sum_{i=1}^{n} X_{ij} \mid X_{obs}, \theta^t \right) = \sum_{i=1}^{n} X_{ij}^t, \quad j=1,1,..,m$$

$$E\left(\sum_{i=1}^{n} X_{ij} X_{ik} \mid X_{obs}, \theta^t \right) = \sum_{i=1}^{n} \left(X_{ij}^t X_{ik}^t + c_{jki}^t \right), \quad j,k = 1,\ldots m, \qquad 5.4$$

where

$$X_{ij}^{t} = \begin{cases} X_{ij} & \text{- if } X_{ij} \text{ is observed} \\ E(X_{ij} \mid X_{obs,i}, \theta^{t}) & \text{- if } X_{ij} \text{ is missing,} \end{cases}$$

and

$$c_{jki}^{t} = \begin{cases} 0 & \text{- if } X_{ij} \text{ and } X_{ik} \text{ are observed} \\ Cov(X_{ij}, X_{ik} \mid X_{obs,i}, \theta^{t}) & \text{- if } X_{ij} \text{ or } X_{ik} \text{ are missing} \end{cases}$$

The M step is straightforward: The new estimates θ^{t+1} of the parameters are estimated from the estimated complete-data sufficient statistics. That is,

$$\mu_j^{t+1} = n^{-1} \sum_{i=1}^{n} X_{ij}^{t}, \qquad j=1,\ldots,m; \qquad 5.5$$

$$\sigma_{jk}^{t+1} = n^{-1} E\left[\sum_{i=1}^{n} X_{ij} X_{ik} \mid X_{obs} \right] - \mu_j^{t+1} \mu_k^{t+1}$$

$$= n^{-1} \sum_{i=1}^{n} [(X_{ij}^{t} - \mu_j^{t+1})(X_{ik}^{t} - \mu_k^{t+1}) + c_{jki}^{t}] \qquad j,k=1,\ldots,m. \qquad 5.6$$

The iteration of the EM algorithm will continue until convergence to a stationary value of $l(\theta \mid X_{obs})$.

5.1.2 Remarks

It is important to study the occurrence of missing values. By using the EM algorithm missing values can be estimated. If missing values occur in any of the variables an entire item could be lost. If a large number of items are lost due to incomplete data fields and if these missing values follow some kind of pattern, an incorrect classification function could be obtained.

5.2 Sequential discrimination

By using sequential discriminant analysis the classification function is gradually built up. This is the main difference between sequential discriminant analysis and any of the previously mentioned methods. It does not necessarily require large data sets, since items are classified as soon as enough information is obtained from the measurements. A further advantage is that the classifier can adjust itself, should the nature of the newly obtained measurements change.

Various methods of sequential discrimination have been proposed. Three different approaches will be briefly discussed:

5.2.1 Sequential probability ratio

Lachenbruch (1975) proposed a method whereby independent

observations can be made on an item to be classified. In the two population situation with equal covariance matrices, avoid having more than ϵ_1 proportion of errors in population π_1 and ϵ_2 proportion of errors in population π_2. By using a sequential probability ratio test, one could decide whether to assign an item to one of the two populations or whether to take another measurement from the item.

Suppose the discriminant function is normally distributed with mean $0.5\ MD^2_{12}$ in population π_1, mean $-0.5\ MD^2_{12}$ in population π_2 and variance MD^2_{12}, where

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

is the Mahalanobis distance between two populations. (Refer to chapter 2 for these results.)

Then the hypothesis H_0: $X \in \pi_1$ vs H_1: $X \in \pi_2$ is tested by means of a sequential likelihood ratio test.

Observe the first measurement, X_1, and calculate

$$A = \frac{1 - \epsilon_2}{\epsilon_1}, \qquad B = \frac{\epsilon_2}{1 - \epsilon_1} \quad \text{and}$$

$$\tau_1 = \frac{f_2(Z_{12}(X_1) ; MD^2_{12})}{f_1(Z_{12}(X_1) ; MD^2_{12})} = \exp(-Z_{12}(X_1)) \qquad 5.7$$

where

$$Z_{12}(X_1) = \{X_1 - 0.5\ (\mu_1 + \mu_2)\}'\ \Sigma^{-1}\ (\mu_1 - \mu_2).$$

If $\tau_1 \leq B$, assign the item to population π_1 and
if $\tau_1 \geq A$, assign the item to population π_2 . 5.8

Should neither of the above be satisfied, obtain another measurement and calculate

$$\tau_2 = \prod_{i=1}^{2} \frac{f_2\ (Z_{12}(X_1)\ ;\ MD^2_{12})}{f_1\ (Z_{12}(X_1)\ ;\ MD^2_{12})} \quad \text{and} \qquad 5.9$$

compare τ_2 to A and B. Continue taking measurements until τ_1 is less than B or greater than A.

Generally

$$\tau_i = \exp\ (-\sum_{j=1}^{i} Z_{12}(X_j)) \ = \ \exp\ (-\ i\ Z_{12}(\bar{X}))\ . \qquad 5.10$$

The sequential probability ratio is obtained by measuring m measurements and by calculating their mean, $\bar{X}m$, where

$$Z_{12}(\bar{X}_m) = \{\bar{X}_m - 0.5\ (\mu_1 + \mu_2)\}'\ \Sigma^{-1}\ (\mu_1 - \mu_2)$$

The rule can be simplified as follows :

Assign an item to population π_1, if after m measurements

$Z_{12}(\bar{X}_m) \geq -\ 1/m\ \ln B$

or to population π_2 if

$$Z_{12}(\bar{X}_m) \leq -1/m \ln A \quad . \qquad 5.11$$

If neither of the above is satisfied, obtain another measurement from the item.

In practical situations Σ^{-1}, μ_1 and μ_2 can be replaced by sample estimates.

5.2.2 Sequential variable inclusion

Mallows (1953) studied sequential discriminant analysis from a different viewpoint. Instead of assuming the possible replication of the entire vector **X** he considered the situation in which the measurements **X** were obtained sequentially in an increasing order of cost. After each observed variable a decision is made whether to assign an individual to a population or to observe the next measurement.

In the two population situation with equal covariance matrices and with ϵ_1 and ϵ_2 the maximum misclassification probabilities for populations π_1 and π_2 respectively, Mallows (1953) proposed the following sequential rule:

Let $Z_{12}(X_q)$ denote the linear discriminant function based on vector X_q of the first q measurements. Then, if

$$Z_{12}(X_q) \geq \ln (1 - \epsilon_1) / \epsilon_2 \text{ , assign to } \pi_1 \text{ or}$$

$$\leq \ln \epsilon_1 / (1 - \epsilon_2) \text{ , assign to } \pi_2 \qquad 5.12$$

or else observe the next measurement (q+1).

If **X** is normally distributed, then

$$\ln \frac{f_2(X_1,\ldots,X_q)}{f_1(X_1,\ldots,X_q)} = -[X_q - 0.5(\mu_1 + \mu_2)]'\Sigma^{-1}(\mu_1 - \mu_2)$$

$$= -Z_{12}(X_q) \qquad 5.13$$

where **X** has q measurements.

5.2.3 Simple sequential rule

Kendall and Stuart (1966) based their sequential method on order statistics of individual measurements. Their method orders all the values of a measurement and then divides the range of variation into three mutually exclusive regions. Suppose that on measurement X_1 all items less than a_1 belong to population π_1 and all items greater than b_1 belong to population π_2. The first region contains only items from population π_1 and the second region contains only items from population π_2 while the third region contains items from a mixture of both π_1 and π_2.

By using the following rule, assign an item to population π_1, if

$X_1 < a_1$ and to population π_2, if

$X_1 < b_1$ 5.14

or else obtain measurement X_2 if $a_1 < X_1 < b_1$.

Thus the first measurement is classified into the applicable region. If the observation falls into the third region the next measurement should be used. Continue until all items are allocated or all measurements used.

Although this method is easy to understand, it may leave a number of items unassigned. Note, furthermore, that this method does not take the joint distribution of variables into account.

5.2.4 Remarks

Although sequential techniques can be quick, the disadvantage exists that an item may remain unclassified after all possible variables have been measured. Sequential methods assume an infinite number of measurements which do not exist in practice. Obviously sequential methods are irrelevant if only a priori measurements are made.

5.3 Logistic discriminant analysis

Logistic discriminant analysis is based on the assumption that the posterior probabilities of belonging to each

population, given a particular item, have linear logistic forms.

According to Bayes' theorem the probability of being an item from population π_i, given X, is

$$P(\pi_i \mid X) = \frac{P(X \mid \pi_i)\, P(\pi_i)}{P(X \mid \pi_1)P(\pi_1) + P(X \mid \pi_2)P(\pi_2)} \qquad 5.15$$

if G=2.

The probability of an item belonging to population π_i, for G>2, is

$$P(\pi_i \mid X) = \frac{P(X \mid \pi_i)\, P(\pi_i)}{\sum_{i=1}^{G} P(X \mid \pi_i)P(\pi_i)} \quad . \qquad 5.16$$

If $P(X \mid \pi_i)$ is multivariate normally distributed with (μ_i, Σ) write 5.15 as

$$\left.\begin{aligned} P(\pi_1 \mid X) &= \exp(\alpha_0 + \beta' X)\, P(\pi_2 \mid X) \text{ and} \\ P(\pi_2 \mid X) &= 1/(1 + \exp(\alpha_0 + \beta' X)) \end{aligned}\right] \qquad 5.17$$

and write 5.16 as

$$\left.\begin{aligned} &P(\pi_i \mid X) = \exp(\alpha_{0i} + \beta_i' X)\, P(\pi_G \mid X) \text{ for } i=1,2,..G-1 \\ &\text{and} \\ &P(\pi_G \mid X) = 1/\left(1 + \sum_{i=1}^{G-1} \exp(\alpha_{0i} + \beta_i' X)\right) . \end{aligned}\right] \qquad 5.18$$

while the logistic discriminant function is calibrated directly from the initial training sample, by estimating α and β using logistic regression techniques. (See equations 5.17 and 5.18).

The logistic discriminant function can by used in a broader class of distributions than the classical linear discriminant function. Furthermore, both binary and continuous variables can be simultaneously dealt with in the same data set.

5.3.1 Remarks

The logistic discriminant function will marginally outperform the linear discriminant function when the assumptions of normality and homoscedasticity are violated. As shown by Efron (1975), if the assumptions of normality and homoscedasticity are met, the linear discriminant function's asymptotic error rate will be 1½ to two times lower than the asymptotic error rate of the logistic discriminant function.

Thus, one should first test whether the assumptions of normality and homoscedasticity are met. If so, the linear discriminant function should be used. On the other hand, if these assumptions are violated and the posterior probabilities have a linear logistic form, the logistic discriminant function should be considered.

CHAPTER 6

DISCRIMINATION BETWEEN FIVE DIFFERENT GRAPE CULTIVARS WITH A VIEW TO DETERMINING THE CULTIVAR OF AN UNKNOWN GRAPE JUICE.

6.1 INTRODUCTION

Pattern recognition has recently achieved a certain prominence and methods of multivariate analysis have contributed greatly to solving analytical chemistry problems.

The aim of this study was to find a classification function which could classify five different grape cultivars into distinct cultivar populations. The object was further to use this derived function to classify items of unknown cultivar origin into the population it most likely resembles. In the search to obtain an appropriate classification function, different multivariate techniques were tested, namely Cluster analysis, Correspondence analysis, Principal component analysis and Discriminant analysis.

In the previous chapters the theoretical basis of discriminant analysis was explored. This chapter illustrates the practical application of those theories.

Approximate normality of the data was obtained by means of the logarithmic transformation. This proved to be

the most successful transformation technique when applied to the data.

A test of homogeneity of the within groups covariance matrices was applied. The chi-square test value was not significant. The covariance matrices were therefore all assumed equal and a pooled covariance matrix could be used in the calculation of the discriminant functions. The assumptions for linear discriminant analysis were met and could thus be applied.

The calculated classification function was evaluated by examining the expected probabilities of misclassification. Variable selection techniques were applied to obtain an unbiased assessment of the true error rate.

Finally, the classification function was tested by classifying new items of unknown origin by using the derived classification function.

Usually, if normality can not be achieved by applying various transformation techniques on the data, non-parametrical methods are applied. In this application non-parametrical techniques were used purely for comparative purposes as the desired results were achieved by using linear discriminant analysis.

6.2. DATA USED

The data consisted of five different grape cultivars, used as the training sample and one other cultivar, used as the test data set. For each of the cultivars twenty different farms were visited and grapes were harvested at a specific sugar level. The blended juice of each cultivar sample was analyzed chemically on an Auto Amino Analyzer to determine the amino acid values present in the sample. The multivariate measurements (amino acid values) of the five cultivars were then used to calculate the classification function.

The Auto Amino Analyzer establishes a chromatogram of the amino acid values present in the sample (Refer to diagram 6.1 for an example of a chromatogram). This is a trace of chemical compounds present in a solution. The area under a peak is representative of the amount of a specific chemical compound present in the solution. (It can be seen as molecule counts.) For each individual sample used, 23 different amino acids were measured.

6.3 UNIVARIATE (DESCRIPTIVE) STATISTICS

Different statistical techniques were applied to test the data for the required assumptions before applying specific multivariate techniques.

Univariate statistics such as the mean, standard deviation,

minimum and maximum values and the coefficient of variation were calculated for each of the 23 amino acid values of the 6 different cultivars.

As mentioned in the previous chapters the results obtained when using linear discriminant analysis are optimal only when the assumptions of normality and equal covariance matrices are met. Violation of these assumptions could result in unreliable classification functions. In order to obtain approximate normality, several transformation techniques were applied to the data.

Normality is tested by hypotheses tests. The null hypothesis tests whether the data values come from a normal distribution, eg. a Shapiro-Wilk statistic can be computed. The Shapiro-Wilk's statistic is the ratio of the best estimate of the variance to the usual corrected sum of squares estimate of the variance. When probability is close to zero, the data are not normally distributed and the hypothesis is rejected.

Normality can also be tested by inspecting the Skewness and the Kurtosis. The overall size of the deviations from the mean is measured by the variance. Skewness, on the other hand, is a measure of the tendency of the deviations to be larger in one direction than in the other. Sample Skewness is derived by:

$$n/(n-1)(n-2) \sum_{i=1}^{n} (X_i - \bar{X})^3 / S^3 \qquad 6.1$$

Kurtosis is a measure of the heaviness of the tails of a population and is very unreliable in small samples.

The sample Kurtosis is derived by:

$$n(n+1) \,/\, (n-1)(n-2)(n-3) \sum_{i=1}^{n} (X_i - \bar{X})^4 \,/\, S^4$$

$$- 3(n-1)(n-1)/(n-2)(n-3) \qquad 6.2$$

In both the above mentioned formulas the Skewness and Kurtosis should be close to zero when data are normally distributed.

The inspection of Bar charts, Normal probability plots, Box-and-Whisker plots and Stem-and-leaf diagrams can also be of great assistance in the evaluation of normality.

For each of the amino acid values from the different cultivars, univariate statistics were calculated. (See appendix B.1, TABLE B.1) Populations 1 to 5 are the 5 grape cultivars sampled in 1987 and used to calculate the classification function. Population 0 represents the cultivar used to test the derived classification function. (PK1 to PK23 represent the amino acid values.) Note that these univariate statistics do not take into account any correlations between two or more measurements (variables). Chapter 4 showed that the multivariate structure is very important, since a variable on its own might not show any significant contribution to the classification

function but in conjunction with other variables it could have a highly significant contribution in classifying items into distinct populations.

6.4 MISSING VALUES

Chapter 5 showed that the occurrence of missing values should be treated with circumspection.

In this application missing values occurred only in certain amino acid samples. Certain software packages delete an entire item if any of the variables contain a missing value. When having only a few items per cultivar, using all the possible information is of extreme importance when calculating the classification function.

Incomplete data in certain amino acid samples were caused by one of two reasons:

- the values were either really absent or
- were too small to be detected by the Auto Amino Analyzer.

Examination of the missing values revealed that they occurred only where the obtained measurements were small. After discussions with the chemist who had prepared the samples, it was concluded that the values were present but the quantities were too small for the machine to detect. These missing values were consequently replaced by zero's.

An alternative method would be to replace the missing values with their corresponding population means. The discriminant function is then calculated on the entire data set. The estimation of missing values by means of regression equations was not a feasible solution for the problem at hand as values smaller than zero may have been obtained. In chapter 5.1 the use of the EM algorithm for incomplete data values was discussed.

The following missing values occurred in the five different cultivars.

TABLE 6.1

VARIABLE	CULTIVAR					TOTAL
	1	2	3	4	5	
P4	9	5	8	12	17	51
P6	1	0	0	0	0	1
P9	2	0	8	7	5	22
P14	0	0	1	0	0	1
P15	2	0	2	0	0	4
P18	4	0	7	0	0	11
P20	0	1	0	0	0	1
P23	0	0	1	0	0	1

This table only reveals the measurements containing missing values. A large number of missing values occurred only in three variables, namely: P4, P9 and P18.

6.5 KRUSKAL-WALLIS TESTS

Before applying multivariate techniques the univariate statistics were examined for significant differences between populations. This could be of some assistance when selecting variables for calculating the discriminant function. Univariate statistics must be applied cautiously for variable selection, since valuable information contained in the multivariate structure of the measurements is not taken into account.

A Kruskal_Wallis test was used to perform an analysis of variance on the ranks of the response variables among the different populations. Dunn´s multiple comparison procedure was subsequently applied to compare the mean ranks for specific population differences. Theoretically these tests are not strictly speaking necessary when applying multivariate analyses. In this case they were applied and are mentioned for completeness´ sake.

In the following table the cultivar numbers are ordered by increasing average ranks. Simultaneous underlinement represents no significant difference between the cultivars on a 5 % level of significance .

TABLE 6.2

VARIABLE	CULTIVAR	VARIABLE	CULTIVAR
P1	3 2 1 4 5	P13	4 5 3 2 1
P2	2 1 3 4 5	P14	4 1 2 3 5
P3	2 1 3 4 5	P15	3 4 1 2 5
P4	2 1 3 4 5	P16	5 4 1 3 2
P5	1 2 3 4 5	P17	2 1 5 4 3
P6	2 1 3 4 5	P18	2 1 4 3 5
P7	4 2 1 5 3	P19	2 1 5 3 4
P8	2 1 3 4 5	P20	2 4 1 3 5
P9	1 2 4 5 3	P21	2 1 4 5 3
P10	5 4 2 3 1	P22	2 4 1 3 5
P11	2 4 1 3 5	P23	5 4 1 2 3
P12	5 2 4 3 1		

It can be seen that between the different cultivars, all the amino acid samples differ and they therefore all have a greater or lesser discrimination ability. (Only measurements present in all the cultivars can contribute towards the determination of the classification function.)

6.6 TRANSFORMATIONS

The following transformations were applied to normalize the data values departing from the assumptions of normality:

1. Logarithmic transformation --- **LPKx** = log(0.5 + **PKx**)

2. Square root transformation --- SQ_PKx = sqrt(PKx)
3. Reciprocal transformation --- RES_PKx = 1/(0.5+PKx)
4. Trimming of extreme observations.

The normality of the transformed data was compared to the untransformed data by measures such as the Shapiro-Wilk Statistic, Skewness and Kurtosis. The most satisfactory transformation result was obtained when applying the logarithmic transformation. (See appendix B.2, Tables B.2.1 to B.2.4)

Note that although trimming some of the extreme measurement values produced relatively favourable results, valuable information obtained from the very small sample sizes was excluded and this technique trimming was thus not a feasible solution to obtain normality.

The **logarithmically transformed data** were used in all further analyses requiring the assumption of normality.

6.7 MULTIVARIATE TECHNIQUES APPLIED

In complex data sets with large numbers of measurements, multivariate analysis techniques provide powerful tools to investigate the relationships between the measurements and among the populations. Graphical presentations of these techniques visually display the multivariate structure in two or three dimensions. With major developments in the computer environment practical limitations of computa-

tional procedures have disappeared and the implementation of multivariate techniques have progressed extremely rapidly.

The following multivariate techniques were applied:

1. PRINCIPAL COMPONENT ANALYSIS
2. CLUSTER ANALYSIS
3. CORRESPONDENCE ANALYSIS
4. DISCRIMINANT ANALYSIS

The results of methods 1 to 3 are discussed and illustrated in Appendix D. Discriminant analysis produced the best result. This is so because discriminant analysis utilized prior information of population membership when classifying the items into distinct populations. None of the other methods mentioned above did this. For this reason the theory of discriminant analysis was examined in chapters 2 to 5.

6.7.1 DISCRIMINANT ANALYSIS

By using quantitative measurements (transformed amino acid values with a multivariate normal distribution), discriminant functions which best revealed the differences between the populations were calculated. The purpose of the discriminant function is to reduce the dimension from a large number of characteristics to relatively few linear combinations. Such linear combinations can then

be graphically displayed in three dimensions. (See below.)

6.7.1.1 **Results obtained utilizing all measurements**

Originally all the available information was used to calculate the classification function. Evaluation of the classification rule was done by examining the expected probabilities of misclassification. As described in chapter 3, the probability of misclassification is a measure of the classification function's expected performance when classifying items.

A test of homogeneity of the within covariance matrices was applied after obtaining approximate normality of the data. The chi-square test value (calculated, using a test proposed by Kendall and Stuart (1961)) was not significant. The covariance matrices were therefore all assumed equal and a pooled covariance matrix could be used in the calculations of the discriminant function. Linear discriminant analysis could then be applied as the required assumptions had been met.

Discriminant functions were developed by using a measure of generalized squared distances. This procedure assumes a multivariate normal distribution within each population. Each item was classified into the cultivar from which it had the smallest generalized squared distance.

Table 6.3 shows that cultivar 5 has the largest generalized

square distance from any of the other cultivars.

Pairwise squared generalized distances between populations were calculated by :

$$D^2(I|J) = (\bar{X}_I - \bar{X}_J)' \, COV^{-1} \, (\bar{X}_I - \bar{X}_J) \qquad 6.3$$

TABLE 6.3

GENERALIZED SQUARED DISTANCE TO POPULATION

FROM POPULATION	1	2	3	4	5
1	0	47.5	64.4	63.1	130.1
2		0	81.4	75.5	165.1
3			0	73.4	166.6
4				0	100.8
5					0

The obtained linear discriminant functions and the corresponding graphical display are given in TABLE 6.4 and GRAPH 6.1 respectively. (See end of chapter.) These linear discriminant functions were then used to determine the probability of misclassification in each population.

Appendix C, TABLE C.1.1 displays the standardized canonical coefficients which are normalized to give the canonical variates with unit within-population variance. These coefficients are used to obtain the graphical representations of the data.

Posterior probability of membership in each population is

$$PR(J|X) = \exp(-0.5\ \tilde{D}_J(X)) \Big/ \sum_{I=1}^{G} \exp(-0.5\ \tilde{D}_I(X)) \qquad 6.4$$

where the generalized squared distance function to population j is:

$$\tilde{D}_J(X) = (X - \bar{X}_J)'\ COV^{-1}\ (X - \bar{X}_J) \qquad 6.5$$

PERCENTAGE OF CORRECT CLASSIFICATIONS:

CULTIVAR	PERCENTAGE
1	100.00
2	100.00
3	100.00
4	100.00
5	100.00

An item was classified into the cultivar that produced the smallest generalized squared distance value or the largest posterior probability. As mentioned in chapter 3 an optimistic estimation of the true error rate could result when using the same data to determine and evaluate the classification function. The Jackknife and Bootstrap approaches were therefore applied to compare the obtained estimates. (See below.)

6.7.1.2 VARIABLE SELECTION

Measurements were selected by using the selection methods described in chapter 4. The number of parameters in the model was reduced by eliminating measurements with a marginal contribution when discriminating between populations.

The following selection techniques were implemented:

1. Forward selection
2. Backward elimination
3. Stepwise selection

The subset selections were interpreted with caution. The obtained subset might not necessarily have been the 'best' subset. To test whether the 'best' subset was obtained the data were divided into separate sections and the stepwise selection method was implemented on these sections. This process was repeated for a few separate sections. The appearance of a measurement in the selected subset denotes the importance of that measurement for future classification purposes.

TABLE 6.5

SUMMARY OF VARIABLES SELECTED

STEPWISE ALL OBS	STEPWISE ID >10	STEPWISE ID < 10	STEPWISE 5>=ID>15	FORWARD ALL OBS	BACKWARD ALL OBS
1				1	1
2	2	2	2	2	2
4		4	4	4	4
		5	5		
6				6	6
7	7			7	7
8	8	8	8	8	8
10	10	10		10	10
11	11		11	11	11
12	12	12	12	12	12
13	13			13	13
14		14	14	14	14
			15		
16	16		16	16	16
17	17	17	17	17	17
18		18		18	18
19	19			19	19
		20			
21	21	21	21	21	21
22	22	22	22	22	22
23	23	23	23	23	23

The stepwise, forward selection and backward elimination procedures all resulted in the same subset selection when using all items. In practice these three methods will not always produce identical results as in this example. The result in this specific instance could be a mere coincidence or because the chosen subset is a clear winner. This subset was used in calculating the linear discriminant function. Appendix B.3, TABLES B.3.1 to B.3.6 summarize the various variable selection methods. Note that the order of variable inclusion in the various

methods differs.

TABLE B.3.1 - Stepwise selection, all items included.

TABLE B.3.2 - Forward selection, all items included.

TABLE B.3.3 - Backward elimination, all items included.

TABLE B.3.4 - Stepwise selection, items with ID > 10 included.

TABLE B.3.5 - Stepwise selection, items with ID < 10 included

TABLE B.3.6 - Stepwise selection, items with ID $\leq$ 5 and ID > 15 included.

6.7.1.3 LINEAR DISCRIMINANT FUNCTION WITH SELECTED VARIABLES

When using only selected variables, less parameters need to be estimated. The performance of the classification rule is generally improved.

The linear discriminant functions were recalculated with only the 18 selected variables present. The results follow:

TABLE 6.6

The pairwise squared generalized distance between populations:

	GENERALIZED SQUARED DISTANCE TO POPULATION				
FROM POPULATION	1	2	3	4	5
1	0	46.2	56.3	60.4	117.9
2		0	77.9	74.7	153.4
3			0	68.8	151.2
4				0	92.0
5					0

These distances were slightly smaller than the distances computed when all measurements had been included in the calculation.

The obtained linear discriminant functions and the corresponding graphical display are given in TABLE 6.7 and GRAPH 6.2 respectively. (See end of chapter.) Standardized input variables (zero mean and unit variance) were used to compute canonical variables from standardized coefficients.

The smallest value of either the number of variables, m, or the number of populations minus one, (G-1), determines the number of canonical components. In this application only the first three canonical components were calculated for graphical display. (See Appendix C, TABLE C.1.2.)

Results calculated: (The prior proportions are taken to be the empirical proportions in the training samples; i.e. n_i/n , i=1,2,...,5 .)

CULTIVAR	FREQUENCY	PRIOR PROPORTION
1	19	0.195876
2	19	0.195876
3	19	0.195876
4	20	0.206186
5	20	0.206186

ITEMS USED :	97
NUMBER OF MEASUREMENTS:	18
NUMBER OF POPULATIONS :	5

TABLE 6.8 (see end of chapter) shows the result of the squared multiple correlation coefficient, R^2, the measure of the amount of between-group variation accounted for by the selected variables. The R^2 values range from 0 to 1 and the larger the R^2 value the better the model fit. R^2 is the ratio of the sum of squares for the corrected total and is given by:

$$R^2 = (1 - SSE) / TSS_i \qquad 6.6$$

TABLE 6.8 further shows the F-value which tested how well the model as a whole accounted for the dependent

variable's behavior. The significance associated with F is given by PR>F. By examining the probabilities associated with the univariate R^2 values in Table 6.8 it can be seen that only one amino acid, LPK12, did not show a significant difference between the cultivars on a five percent level of significance.

The technique for analyzing the relationship between canonical components is known as canonical correlations. In TABLE 6.9 the eigenvalues are calculated by CANRSQ/(1-CANRSQ), where CANRSQ represents the corresponding squared canonical correlation which is the ratio of the between-population variation to the within-population variation. It can be seen that the first three eigenvalues explain 90% of the variation present in the data set. It can further be seen that the squared canonical correlations are much higher than either univariate R^2 value.

TABLE 6.9

	CANONICAL CORRELATIONS	SQUARED CANONICAL CORRELATIONS	EIGENVALUES	PROPORTION EXPLAINED
1	0.976	0.952	19.77	0.52
2	0.945	0.892	8.29	0.22
3	0.927	0.858	6.06	0.16
4	0.894	0.799	3.97	0.10

In TABLE 6.10 the likelihood ratio for all canonical correlations was calculated by Wilks' lambda. It can be seen that the populations differ significantly on all the linear combinations calculated. The likelihood ratio, approximate F statistic value with degrees of freedom and the probability is given.

TABLE 6.10

	LIKELIHOOD RATIO	APPROXIMATE F	NUM DF	DEN DF	PR > F
1	0.000147	34.7880	72	297.276	0.0
2	0.003063	26.6595	51	227.07	0.0001
3	0.028459	23.7144	32	154	0.0001
4	0.201047	20.6646	15	78	0.0001

6.8 EXPECTED PROBABILITIES OF MISCLASSIFICATION

The number of correct classifications when calculating the posterior probability of membership in each population was:

CULTIVAR	PERCENTAGE
1	100.00
2	100.00
3	100.00
4	100.00
5	100.00.

As seen in chapter 3 these posterior probabilities tend to give an optimistic assessment of the expected probabilities of misclassification. By calculating probabilities of misclassification by procedures such as the Jackknife or bootstrap procedures, the expected probabilities of misclassification are less biased. (See chapter 3.1.2 & 3.1.3)

NUMBER OF CORRECT CLASSIFICATIONS BY THE JACKKNIFE METHOD:

CULTIVAR	PERCENTAGE
1	100.00
2	94.70
3	100.00
4	100.00
5	100.00

The Jackknife procedure resulted in a slight reduction in correct classifications in Cultivar 2. Nevertheless, the expected probabilities of misclassification for the calculated classification function remained extremely small.

When using the Bootstrap procedure to estimate the expected probability of misclassification the bootstrap sample was drawn as follows :

A random bootstrap sample, with replacement, was drawn from each population by sampling the id_numbers (numbers

given to each of the items). This sample was the same size as the original population. The discriminant function calculated from the bootstrap sample was used to classify the remaining unsampled id_numbers. The undrawn samples were then used to test the validity of the classification function. This procedure was repeated a number of times. An average percentage of items correctly classified in all the samples was obtained and the bootstrap estimate was calculated.

Draw a bootstrap sample as follows :

eg. SAMPLE ID NO	BOOTSTRAP SAMPLE ID NO	REMAINING SAMPLE ID NO
1. 11	1. 11	1. 22
2. 22	2. 33	2. 55
3. 33	3. 44	3. 66
4. 44	4. 33	
5. 55	5. 88	
6. 66	6. 00	
7. 77	7. 99	
8. 88	8. 88	
9. 99	9. 77	
10. 00	10. 33	

The averages of the percentage correctly classified items by means of the bootstrap procedure were :

CULTIVAR 1	98.33
CULTIVAR 2	90.07
CULTIVAR 3	100.00
CULTIVAR 4	100.00
CULTIVAR 5	100.00

The bootstrap procedure resulted in slight reductions in the percentage of correctly classified items in the first two cultivars. Cultivars 3 to 5 yielded the same results when compared to the other methods.

Examining the various probabilities of misclassification, it can be seen that the obtained classification function accurately classified the various grape cultivars.

6.9 TESTING THE CLASSIFICATION FUNCTION BY USING NEW ITEMS :

The validity of the classification function derived above was then tested by classifying new items of unknown origin. This was done by classifying newly measured amino acid samples (harvested in 1988) by using the derived discriminant function.

The univariate statistics of the new data are given in TABLE B.1, appendix B. (Group = 0.)

THE DISCRIMINANT FUNCTION OBTAINED WHEN USING THE 5 ORIGINAL CULTIVARS (1987 data) WILL BE USED TO CLASSIFY NEW ITEMS OF UNKNOWN ORIGIN (1988 data):

VARIABLES : 18 selected variables

POPULATIONS TO CALCULATE CLASSIFICATION FUNCTION :
5 Cultivars - 1987 data

POPULATION CLASSIFIED BY DISCRIMINANT FUNCTION :
1 Cultivar - 1988 data

The **1988 cultivar is similar to Cultivar 2** of the original (1987) data, but harvested and analyzed a year later.

THE PERCENTAGE OF CORRECTLY CLASSIFIED ITEMS OF THE 1988 SAMPLE WHEN USING THE CALCULATED DISCRIMINANT FUNCTION WAS:

	PERCENTAGE	
POPULATION	ALL VARIABLES	SELECTED VARIABLES
1	0.0	0.0
2	94.74	94.74
3	0.0	0.0
4	5.26	5.26
5	0.0	0.0

As shown above, a very satisfactory result was obtained in both the category containing all the variables and the category containing only the selected variables. It showed that the discriminant function was able to classify 94.74% of the unknown items into the correct cultivar, namely cultivar 2. The graphical representations are given in Graphs 6.3 and 6.4. In these graphical displays

cultivar 2 and cultivar 6 are similar cultivars, harvested in 1987 and 1988 respectively.

6.10 NON-PARAMETRICAL METHODS

Non-parametrical methods are used if the application of transformation techniques can not achieve normality of the data.

The results obtained when using the k-nearest neighbour technique or ranking are shown.

6.10.1 k-NEAREST NEIGHBOUR APPROACH

This is a non-parametric technique to classify the data into populations containing the k-nearest neighbours. The Mahalanobis distances based on the total covariance matrix were used to determine proximity. The items were classified into the cultivar containing the highest proportion of the k-nearest neighbours. The percentage of correctly classified items from the 5 original cultivars were (using k=19 and with no data transformations) :

CULTIVAR	ALL VARIABLES PRESENT	18 SELECTED VARIABLES
1	78.95	78.95
2	57.89	63.16
3	78.95	100.00
4	80.00	95.00
5	90.00	85.00

An overall improvement in most of the proportion correctly classified items was obtained by using the 18 selected variables. The results clearly show that linear discriminant analysis performs better than the k-nearest neighbour technique when the data are approximately normally distributed with equal covariance matrices.

6.10.2 RANKING VARIABLES FOR CLASSIFICATION

Each of the original amino acid values was ranked by giving the smallest rank to the smallest value. Thereafter linear discriminant analysis was applied on the ranks. The ranking procedure was performed on two data sets; when all the measurements were present and when only the selected measurements were present.

The obtained linear discriminant functions when all variables are present and when only selected variables are used are given in TABLES 6.12 and 6.13 respectively. The corresponding graphical displays are given in GRAPHS 6.5 and 6.6. (See end of chapter.) The canonical coefficients are given in Appendix C, TABLES C.2.1 and C.2.2 respectively.

The percentages of correctly classified items, when calculating the apparent error rate for the ranked data were similar to the percentages obtained when using the original transformed data.

CULTIVAR	PERCENTAGE
1	100.00
2	100.00
3	100.00
4	100.00
5	100.00

6.11 SUMMARY

The linear discriminant function, when applied to logarithmically transformed data, performed extremely well as a classification technique. When a priori information of population origin is given and the assumptions of equal covariance matrices and normality of the data are met, linear discriminant analysis is an appropriate classification technique.

The probabilities of misclassification were extremely low and therefore the obtained discriminant function could be used to accurately classify items of unknown origin into one of the populations. The 1988 data were satisfactorily classified by using the classification function derived from the 1987 data. The distinct population classifications can be seen in the graphical displays.

The k-nearest neighbour technique classified with higher probabilities of misclassification when compared to the results obtained by using linear discriminant analysis. This proved that linear discriminant analysis is a more appropriate technique

when the data are normally distributed with equal covariance matrices.

The ranking procedure classified the items extremely well. Linear discriminant analysis is to be preferred to ranking as the former procedure uses the obtained data values in the calculations and the original data are not replaced by corresponding ranks.

DIAGRAM 6.1

EXAMPLE OF A CHROMATOGRAM

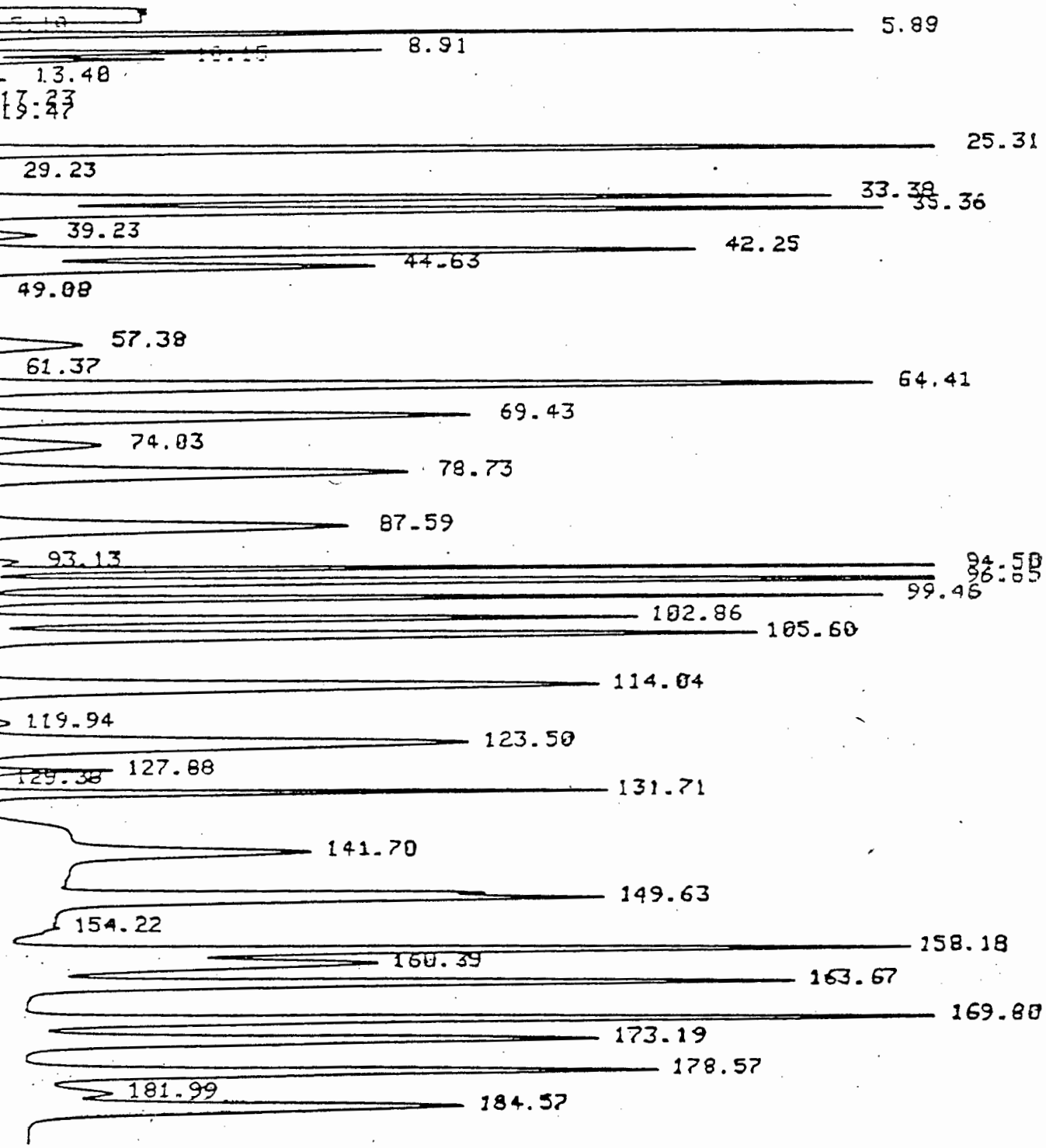


TABLE 6.4 - ALL VARIABLES LOGARITHMIC TRANSFORMED

DISCRIMINANT ANALYSIS — LINEAR DISCRIMINANT FUNCTION

$$\text{CONSTANT} = -.5\ \bar{X}_J'\ \text{COV}^{-1}\ \bar{X}_J \qquad \text{COEFFICIENT VECTOR} = \text{COV}^{-1}\ \bar{X}_J$$

	GROUP 1	2	3	4	5
CONSTANT	-866.38751379	-737.28181765	-862.09196478	-743.65146662	-874.03762079
LPK1	136.81685914	125.62099885	138.74451642	129.27430060	133.86111877
LPK2	105.19601020	109.03505101	135.71458759	82.32029172	48.24985437
LPK3	-106.28783263	-107.23349433	-119.88770150	-101.23254226	-104.07358575
LPK4	-0.00595486	0.16155487	1.46551974	0.07992687	-3.02211599
LPK5	44.23977197	40.06677635	33.07231567	37.64801364	33.57496093
LPK6	-5.60397888	-2.58905821	-0.02335999	-3.07889518	-2.98269784
LPK7	-98.95280038	-91.66817015	-112.54150984	-97.75673173	-119.63729696
LPK8	34.64446955	25.17184573	55.22480303	28.67155476	49.39637210
LPK9	-5.30468641	-6.17480279	-7.27657379	-6.20036160	-5.88094275
LPK10	7.24187246	24.41716237	6.69947195	25.27535104	54.05857512
LPK11	-133.98747288	-104.99483379	-127.12503731	-114.76680475	-142.06675574
LPK12	91.67397681	49.26366867	81.97606809	58.40781440	86.64747266
LPK13	-1.13204833	20.93894978	31.17644230	27.06606292	14.78884692
LPK14	24.16475288	16.50367027	19.26859563	23.27740902	15.36848763
LPK15	-11.80031251	-11.17384998	-10.86033305	-10.25721607	-8.60276160
LPK16	148.28276282	128.02324887	133.27141288	144.20718539	144.70524240
LPK17	97.07618266	100.59664193	76.16045934	82.43733876	122.47307715
LPK18	-4.30541668	-4.35490337	-4.25877994	-3.01811744	-6.43979625
LPK19	4.79962569	3.22237757	4.19645379	-0.78662781	8.62986355
LPK20	-16.17795039	-14.40121609	-13.11090517	-14.95654198	-23.52797759
LPK21	9.18143247	5.03605388	-15.92983351	-12.45282665	-10.57763870
LPK22	-53.84656445	-57.55663465	-57.78161857	-35.73610954	-54.55702759
LPK23	-15.36964102	-16.48416886	-10.24639676	-8.11189069	3.81636031

TABLE 6.7 - SELECTED VARIABLES LOGARITHMIC TRANSFORMED

DISCRIMINANT ANALYSIS

$$\text{CONSTANT} = -.5\ \bar{X}_J'\ \text{COV}^{-1}\ \bar{X}_J$$

LINEAR DISCRIMINANT FUNCTION

$$\text{COEFFICIENT VECTOR} = \text{COV}^{-1}\ \bar{X}_J$$

	GROUP				
	1	2	3	4	5
CONSTANT	-735.44987067	-612.34116160	-729.74261772	-629.72269522	-751.09406788
LPK1	137.72953729	124.23376236	131.50532296	127.73355742	129.42987359
LPK2	68.90461078	70.18035730	87.64724851	46.39795155	15.01048292
LPK4	-2.07344921	-1.91858275	-0.75441253	-1.84827645	-4.64941576
LPK6	-18.96265939	-15.81667705	-14.07998144	-15.50582468	-14.99430717
LPK7	-96.11638118	-89.88186107	-113.08103751	-96.07405202	-118.69097274
LPK8	16.17102930	5.05650125	28.62510449	9.31631029	25.77174201
LPK10	-8.87529020	7.58938115	-13.01802793	8.85966498	34.84195041
LPK11	-115.47132486	-86.46199832	-106.70095256	-96.99402149	-122.54091690
LPK12	77.68858289	38.64721114	75.69801809	47.57320919	69.09144635
LPK13	6.93429139	28.21626058	37.91766608	34.77130750	27.45319907
LPK14	12.08334741	5.53172085	9.28381046	13.35636921	7.26604067
LPK16	132.38096778	113.81413317	120.90030640	130.65821479	129.92089306
LPK17	77.15244748	81.88317986	57.90788983	64.41955234	101.99820414
LPK18	-4.55007861	-4.37133549	-3.89305119	-2.95391740	-6.11250676
LPK19	1.17346116	-0.60692892	-0.21828759	-4.79611299	2.93744407
LPK21	24.08696738	17.38724745	-6.25557952	-1.06359732	2.29973657
LPK22	-68.78468304	-71.04840296	-68.34664045	-49.29860656	-70.35597847
LPK23	-12.66357865	-14.46091552	-9.45345092	-6.39849210	4.50540792

TABLE 6.8 - SELECTED VARIABLES LOG TRANSFORMED

UNIVARIATE STATISTICS

ARIABLE	MEAN	TOTAL STD	WITHIN STD	BETWEEN STD	R-SQUARED	RSQ/(1-RSQ)	F	PROB > F
PK1	3.94529410	0.42090811	0.29199921	0.33725304	0.539655	1.172	25.790	0.0001
PK2	4.49459664	0.48435788	0.30048287	0.41994334	0.631870	1.716	37.761	0.0001
PK4	1.19378383	1.76337085	1.37740727	1.24107395	0.416377	0.713	15.696	0.0001
PK6	4.65697407	1.02347274	0.79386489	0.72733144	0.424513	0.738	16.228	0.0001
PK7	1.50388567	0.37384179	0.29322048	0.26158390	0.411552	0.699	15.387	0.0001
PK8	4.97019009	0.48630401	0.40563930	0.30676469	0.334483	0.503	11.057	0.0001
PK10	3.27768998	0.35888912	0.32888344	0.17362521	0.196736	0.245	5.388	0.0002
PK11	1.85989560	0.65540533	0.39064392	0.58083765	0.660189	1.943	42.742	0.0001
PK12	2.70994824	0.33315648	0.32606516	0.10516928	0.083765	0.091	2.011	0.0826
PK13	3.27891728	0.34103813	0.31676773	0.15550945	0.174778	0.212	4.659	0.0007
PK14	2.51098808	0.54169468	0.45492685	0.33701603	0.325365	0.482	10.610	0.0001
PK16	4.49008940	0.45847323	0.25245670	0.42135193	0.709972	2.448	53.855	0.0001
PK17	4.34349160	0.61147180	0.34704042	0.55476152	0.691893	2.246	49.404	0.0001
PK18	2.69515047	1.24964634	1.15775893	0.57662398	0.178974	0.218	4.796	0.0005
PK19	1.97268676	0.62593886	0.42042334	0.51475272	0.568476	1.317	28.982	0.0001
PK21	3.38987053	0.51369772	0.38559530	0.38044807	0.461057	0.855	18.821	0.0001
PK22	4.17283322	0.61332537	0.43845753	0.47827677	0.511159	1.046	23.004	0.0001
PK23	3.31542027	0.92704793	0.49689169	0.86107735	0.725202	2.639	58.059	0.0001

AVERAGE R-SQUARED: UNWEIGHTED = 0.4470008 WEIGHTED BY VARIANCE = 0.4374519

TABLE 6.12 - ALL VARIABLES RANKED

DISCRIMINANT ANALYSIS — LINEAR DISCRIMINANT FUNCTION

$$\text{CONSTANT} = -.5\ \bar{X}_J'\ \text{COV}^{-1}\ \bar{X}_J \qquad \text{COEFFICIENT VECTOR} = \text{COV}^{-1}\ \bar{X}_J$$

	GROUP				
	1	2	3	4	5
CONSTANT	-35.10990192	-31.42637736	-30.95490263	-29.19146383	-56.36733588
RPK1	0.14474786	0.04092030	0.19953103	0.13507310	0.17258980
RPK2	0.39359850	0.37742624	0.81947625	-0.31739316	-0.36940398
RPK3	-0.43613320	-0.44494428	-0.54328808	0.07571531	-0.08492583
RPK4	0.06752036	0.12448788	0.11664915	0.09809749	0.00420692
RPK5	0.28422212	0.16136857	0.05592826	0.10979988	0.18025818
RPK6	-0.13113245	0.05573889	0.12712229	-0.10303864	-0.21347392
RPK7	-0.04262396	0.07555121	-0.18191269	-0.05435933	-0.15314173
RPK8	0.03188444	-0.01613740	0.29766112	-0.22004622	0.06433025
RPK9	0.17768620	0.12290888	0.05161535	0.10485131	0.17999421
RPK10	0.02086873	0.05465103	-0.02476011	-0.14670082	0.00477359
RPK11	-0.33108216	0.21588817	-0.22756707	0.06789820	-0.34786581
RPK12	0.19219909	-0.07323434	0.02713598	-0.01194427	0.37145370
RPK13	-0.27305888	-0.02730570	0.17990186	0.11899271	-0.13918034
RPK14	0.11350042	-0.01339895	0.08257546	0.17268209	-0.17860511
RPK15	0.04986074	-0.07900341	0.10500040	0.03184727	-0.00472780
RPK16	0.47377541	0.10112813	0.19207773	0.39796072	0.53552962
RPK17	0.37157288	0.36770140	-0.06580140	-0.00644153	0.57629831
RPK18	0.04140856	0.03652192	0.06785099	0.08490917	-0.01721273
RPK19	0.05331818	0.03594874	0.07754481	-0.05847783	0.03604267
RPK20	-0.20719930	-0.05492057	-0.14093811	-0.07134574	-0.28971674
RPK21	-0.09732354	-0.02521978	-0.43773127	-0.25028067	-0.17026126
RPK22	0.09827645	-0.14698176	-0.13824279	0.25912062	0.13502082
RPK23	-0.07568593	-0.26074752	-0.01442123	0.27445456	0.67881121

TABLE 6.13 - SELECTED VARIABLES RANKED

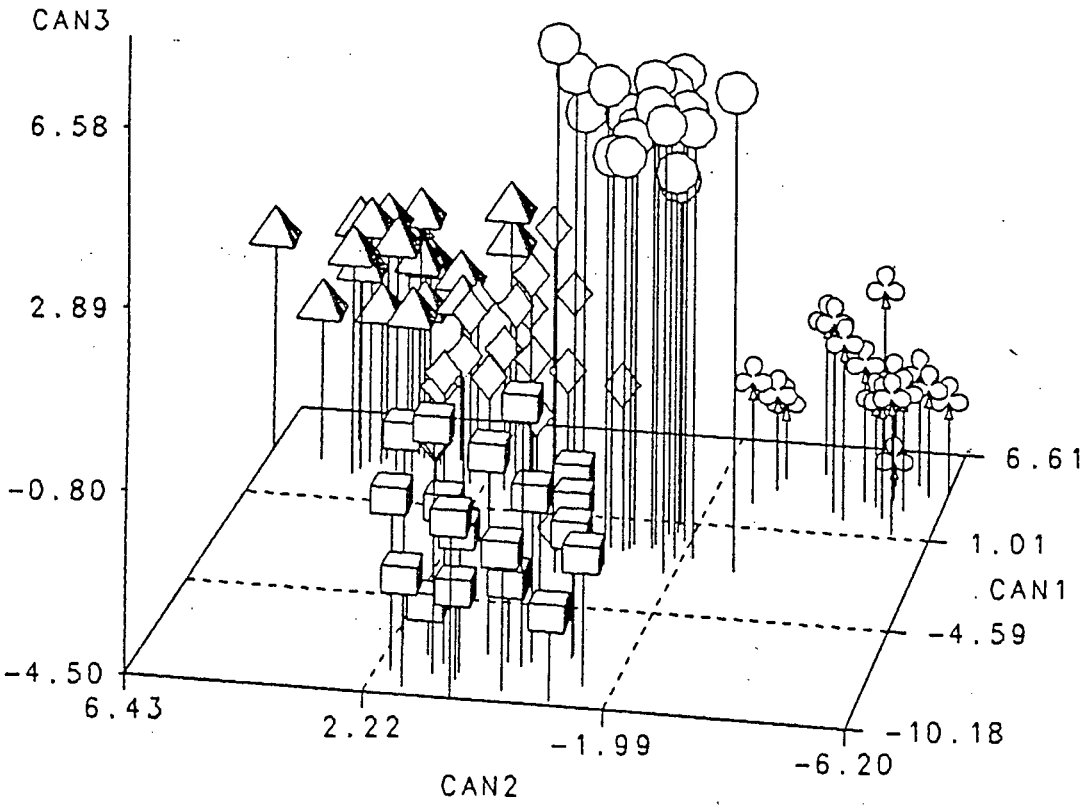
DISCRIMINANT ANALYSIS — LINEAR DISCRIMINANT FUNCTION

$$\text{CONSTANT} = -.5\ \bar{X}_J'\ \text{COV}^{-1}\ \bar{X}_J \qquad \text{COEFFICIENT VECTOR} = \text{COV}^{-1}\ \bar{X}_J$$

	GROUP				
	1	2	3	4	5
CONSTANT	-25.04589472	-26.15290237	-24.84503899	-27.13934700	-46.96538164
RPK1	0.23769724	0.11260007	0.17627213	0.18062161	0.23038247
RPK2	0.18111072	0.11885305	0.49254409	-0.23925252	-0.41041449
RPK4	0.06210017	0.11530080	0.12504557	0.09666987	-0.00249752
RPK6	-0.16759998	-0.02820941	0.02871835	-0.04787864	-0.14099386
RPK7	-0.03415064	0.09333107	-0.19986123	-0.04986172	-0.14308129
RPK8	0.00597302	-0.11942894	0.14094743	-0.14211536	0.10039399
RPK10	-0.09875207	-0.06660276	-0.14939449	-0.13221069	-0.01338628
RPK11	-0.26980105	0.26306270	-0.11086498	0.04019957	-0.33096754
RPK12	0.13432049	-0.06236206	0.05038437	-0.07158617	0.21002404
RPK13	-0.18685984	0.01410468	0.22215467	0.14405410	-0.05398040
RPK14	0.11926329	-0.07193991	0.07766342	0.20128338	-0.17789361
RPK16	0.35356948	0.04530781	0.16004459	0.34807403	0.40303912
RPK17	0.24617757	0.26828714	-0.14885254	-0.02926542	0.50687747
RPK18	0.03226060	0.02430979	0.07764054	0.07636588	-0.04289275
RPK19	0.03269362	0.06047054	0.02339084	-0.05698968	0.00904005
RPK21	0.11425407	0.10528160	-0.31034170	-0.17172470	-0.02036953
RPK22	-0.05744843	-0.17503151	-0.18055549	0.19523334	-0.04024765
RPK23	-0.04766801	-0.21604675	0.00785221	0.26299896	0.66156582

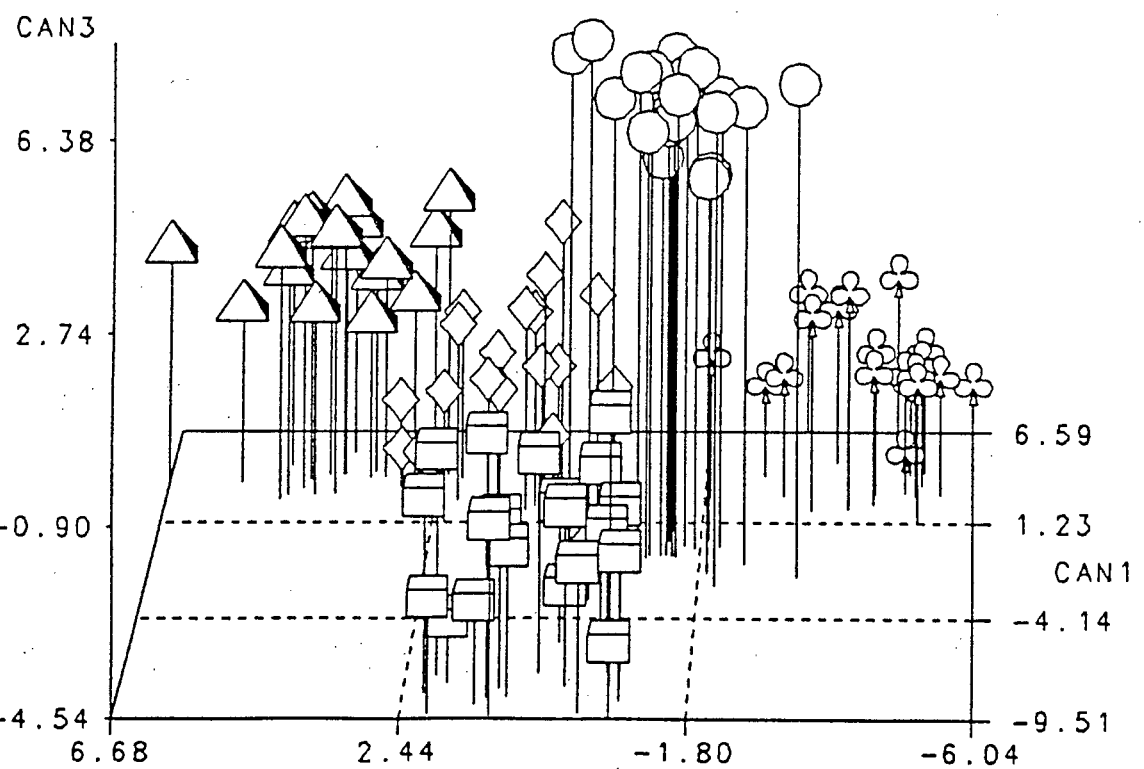
DISCRIMINANT ANALYSIS ON TRANSFORMED DATA
ALL VARIABLES USED FROM 5 GRAPE CULTIVARS

GRAPH 6.1



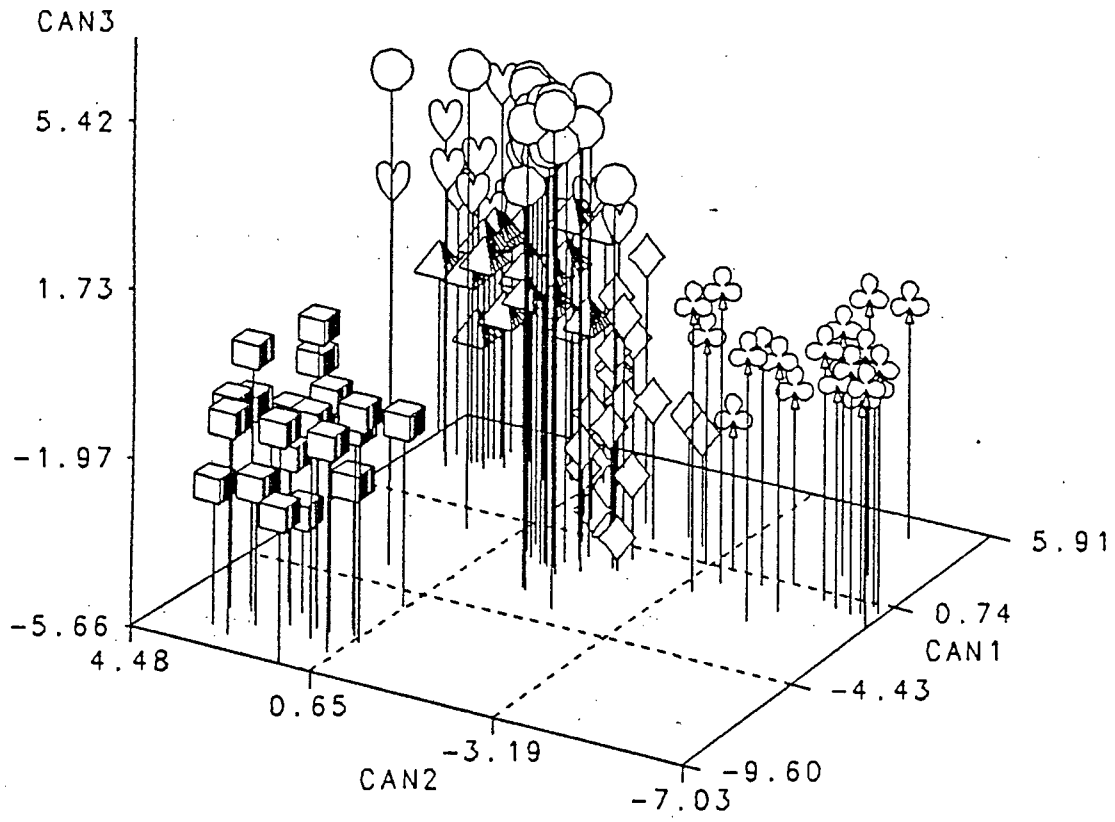
DISCRIMINANT ANALYSIS ON TRANSFORMED DATA
SELECTED VARIABLES FROM 5 GRAPE CULTIVARS

GRAPH 6.2



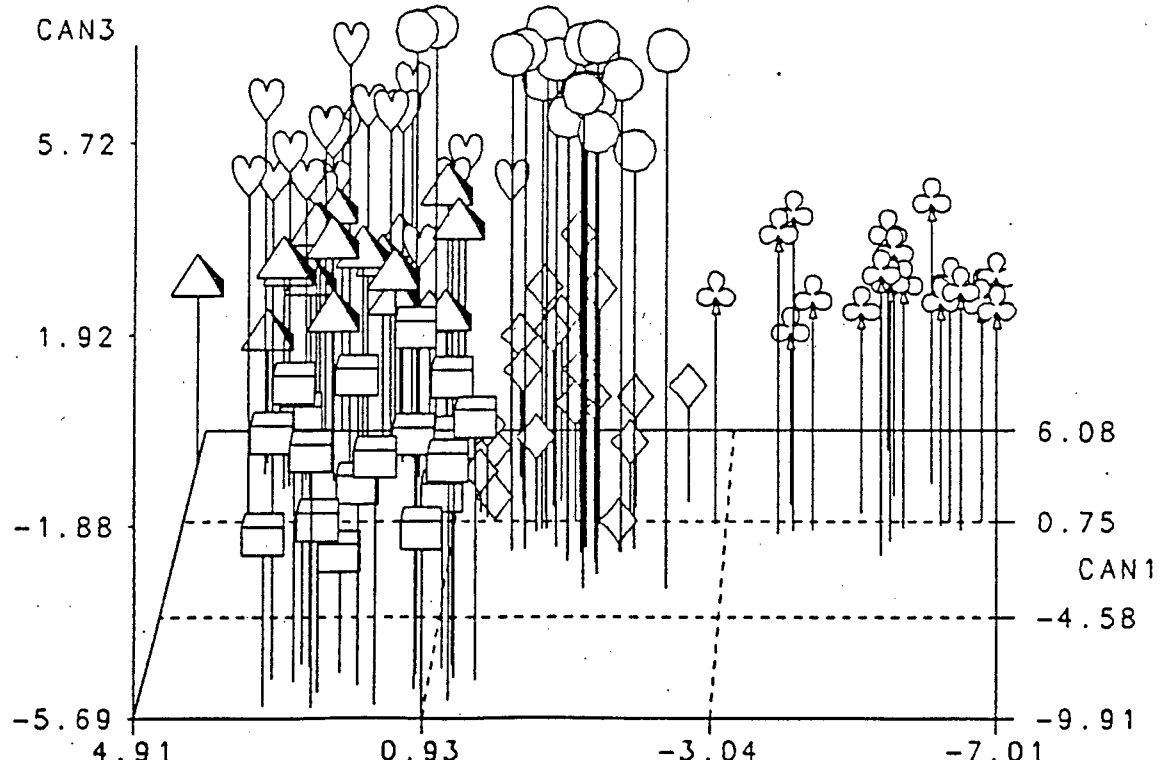
DISCRIMINANT ANALYSIS ON TRANSFORMED DATA
ALL VARIABLES USED FROM 6 GRAPE CULTIVARS

GRAPH 6.3



DISCRIMINANT ANALYSIS ON TRANSFORMED DATA
SELECTED VARIABLES USED FROM 6 GRAPE CULTIVARS

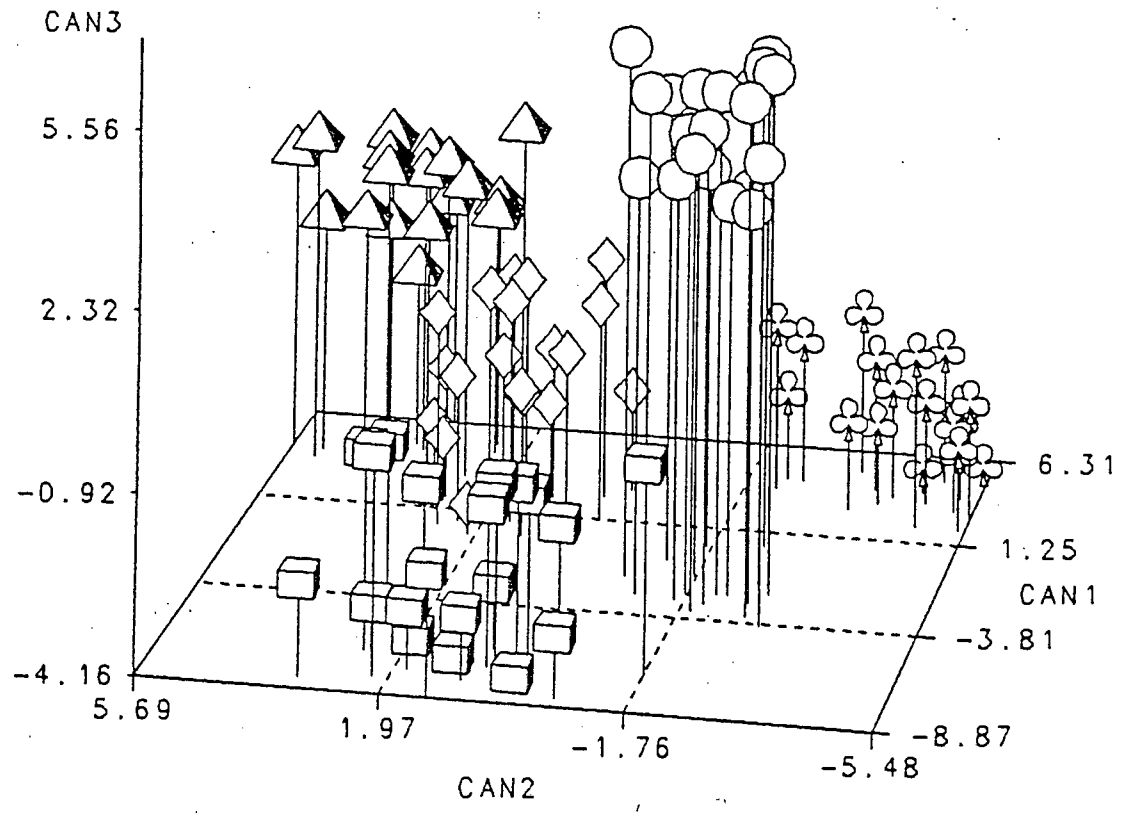
GRAPH 6.4



DISCRIMINANT ANALYSIS ON RANKED DATA

ALL VARIABLES USED FROM 5 GRAPE CULTIVARS

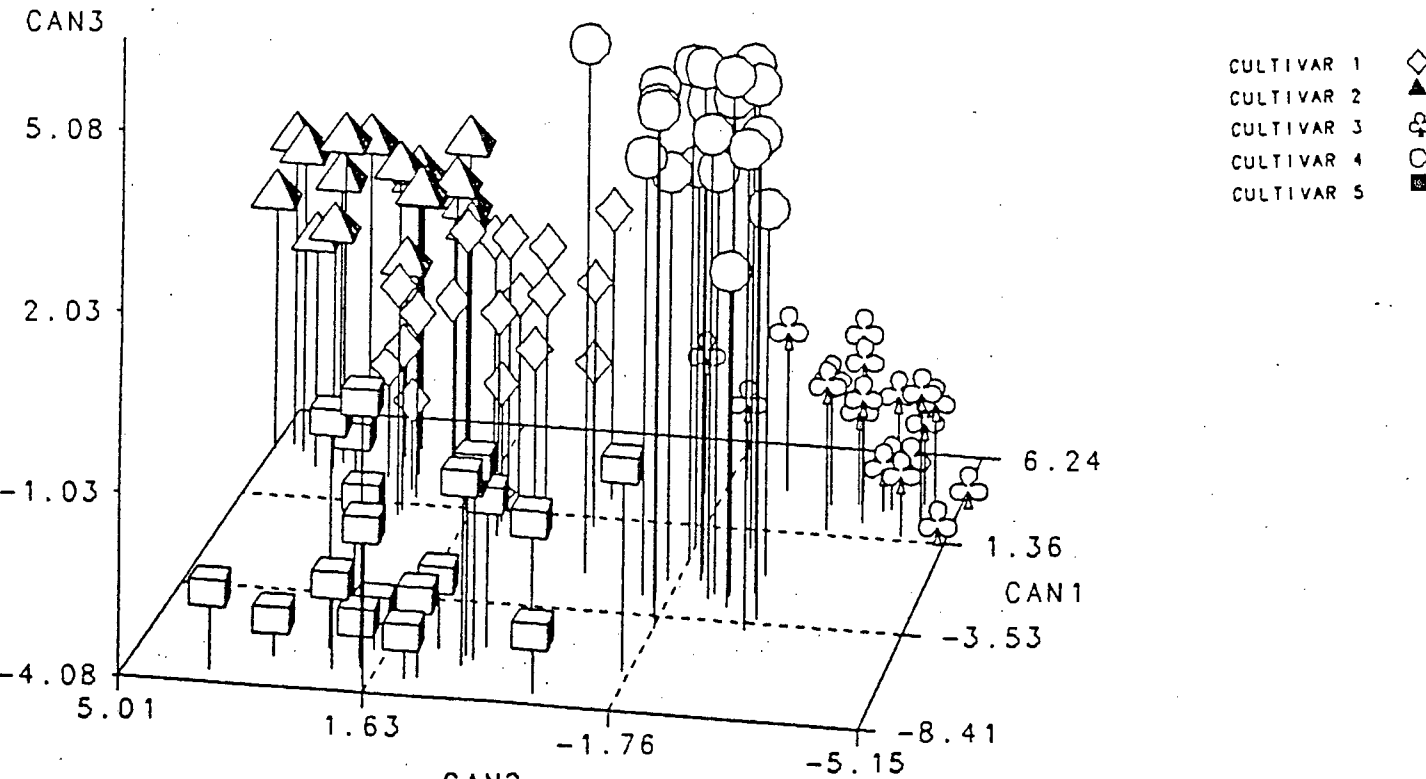
GRAPH 6.5



DISCRIMINANT ANALYSIS ON RANKED DATA

SELECTED VARIABLES USED FROM 5 GRAPE CULTIVARS

GRAPH 6.6



CHAPTER 7

SUMMARY AND CONCLUSION

The main objective of this study was to calculate a classification function for determining the cultivar of an unknown grape juice sample. The wider application of such a function is obvious and it was accordingly decided to examine the theory underlying such a function.

The practical research was undertaken to establish a solution to a real problem. The most ideal tool was found to be discriminant analysis. This became apparent only after several other methods were tested in practice. The results were so perfect as to be unbelievable and it was accordingly thought necessary to examine the principles underlying such a satisfactory technique.

Chapter 2 therefore examined the theory supporting discriminant analysis. The two approaches to discriminant analysis, namely parametrical and non-parametrical, were discussed and the situations in which each found application pointed out.

It was further concluded that of the classical approaches linear discriminant analysis proved to be optimal only when the assumptions of normality and equal covariance matrices are met although slight deviations might be tolerated. When the covariance matrices are unequal quadratic discriminant analysis is to be preferred.

When the assumptions of normality are not met, transformation of the data is attempted. If that proves unsuccessful the non-parametrical approaches are applied. The kernel method, K-nearest neighbour and ranking were discussed.

It was shown that the non-parametrical methods have the advantage of not requiring exact distributions of measurements. On the other hand, their computations are lengthy and computer programs are not always readily available.

Chapter 3 addressed the probabilities of misclassification. It was shown that when the same data used to calculate the optimal classification rule were used to estimate the probabilities of misclassification care had to be taken to avoid optimistically biased assessments of these probabilities.

It was further shown that the apparent error rate will be biased and will underestimate the true error rate. The jackknife and bootstrap estimates were discussed as methods of improving the estimate of the true error rate. It was shown that the jackknife estimated with similar bias but that the bootstrap estimate was less variable and involved extensive computations.

In chapter 4 the problem of variable selection was discussed. It was shown that by excluding variables contributing only marginally to the discrimination between the populations, the number of parameters estimated by the model can be reduced. The performance of the classification rule is improved and

the number of measurements needed is reduced.

Various selection procedures were discussed. It was concluded that no unique statistical procedure exists for determining the selection of measurements needed for an optimal classification rule. The final selection depends on the statistician's judgement and the results obtained must necessarily be interpreted with caution.

In chapter 5 matters related to discriminant analysis were discussed briefly. It was found that the methods discussed can be used in situations where it will be inadvisable to apply linear discriminant analysis. For example, the EM algorithm can be used to estimate incomplete data values, sequential discrimination can be used where the classification function is gradually built up and logistic discriminant analysis is used in a broader class of distributions than linear discriminant analysis. It was further noted that the latter marginally outperforms the linear discriminant function when the assumptions of normality and homoscedasticity are violated.

Chapter 6 and the appendices dealt with the practical application of the preceding theory. A classification function was calculated to classify grape cultivars into distinct populations by using amino acid values of grape juice samples. This derived classification function was then used to classify grape samples of unknown origin into the existing populations.

In order to obtain approximate normality the data were trans-

formed. The logarithmic transformation proved to be the most appropriate technique. The result was that equal covariance matrices could be assumed and a pooled covariance matrix could therefore be used.

The expected probabilities of misclassification were minimized by implementing variable selection techniques. Stepwise, forward selection and backward elimination yielded the same results. These selected variables were then used to calculate the classification function.

The derived discriminant function was tested for accuracy by calculating estimates of the probability of misclassification. These probabilities were extremely small and the classification function classified with great precision.

The derived function was then used to classify the grape juice samples of unknown origin and classified these samples with extreme precision.

APPENDIX A

ORIGINAL DATA

OBS	GROUP	PK1	PK2	PK3	PK4	PK5	PK6	PK7	PK8	PK9	PK10	PK11	PK12	PK13	PK14	PK15	PK16	PK17	PK18	PK19	PK20	PK21	PK22	PK23
1	1	37.4	126.8	59.8	0.0	148.0	132.7	2.9	110.3	10.7	20.0	4.1	12.8	15.1	10.5	23.8	74.2	98.1	43.6	7.2	6.0	37.6	46.7	22.3
2	1	40.5	106.3	69.0	0.0	110.1	124.0	4.3	132.4	15.5	19.0	2.6	10.1	14.9	14.0	20.4	113.9	86.1	15.1	8.3	7.4	33.9	62.2	17.4
3	1	53.4	88.6	43.3	0.0	156.8	68.3	4.3	121.3	9.4	15.8	2.5	10.6	18.3	15.4	21.1	84.4	75.7	15.0	5.7	7.5	30.7	57.7	12.5
4	1	48.6	94.8	50.1	7.7	150.9	69.2	2.8	109.3	0.0	12.9	1.7	8.2	13.6	9.0	21.5	62.3	77.2	24.3	4.4	5.4	16.8	44.1	8.5
5	1	65.3	172.7	79.2	30.2	241.1	184.0	5.3	175.6	22.7	27.7	9.0	19.5	29.8	23.3	45.5	124.6	152.4	0.0	14.7	13.3	71.0	106.6	27.8
6	1	58.9	98.8	46.3	0.0	149.4	59.5	5.7	130.3	3.2	19.4	3.5	14.1	21.5	13.6	22.7	137.4	66.4	40.0	7.2	7.7	26.5	38.7	12.4
7	1	85.9	156.0	107.0	36.2	254.7	393.7	6.3	244.5	42.0	28.6	8.3	16.0	24.2	26.2	46.5	90.8	140.0	65.5	18.0	12.2	59.8	112.4	25.9
8	1	72.3	110.9	95.4	26.6	220.5	227.0	5.1	221.2	27.3	22.0	2.3	10.8	18.8	17.7	27.1	82.8	97.2	22.8	9.1	8.9	41.8	80.3	25.4
9	1	66.4	111.7	83.0	9.7	207.9	203.3	4.7	165.4	19.5	24.5	3.9	14.2	22.3	19.1	54.6	67.1	99.7	23.0	8.6	6.7	43.8	73.4	22.6
10	1	91.7	110.6	79.7	15.8	220.9	147.4	4.1	224.0	22.4	22.1	5.3	10.9	19.8	18.0	28.9	89.1	106.6	37.8	10.4	9.1	39.3	101.0	20.9
11	1	43.2	136.2	58.5	10.9	132.5	161.7	3.8	111.6	11.7	17.4	2.9	12.0	14.4	13.7	22.3	81.4	138.3	36.5	12.2	8.1	40.0	70.0	18.1
12	1	31.3	89.6	48.5	0.0	121.5	58.7	3.8	109.2	0.0	10.8	2.7	8.2	13.8	11.6	16.4	92.7	99.4	23.9	4.2	5.3	21.2	46.4	13.3
13	1	103.7	152.2	138.8	31.7	275.4	442.6	8.3	405.0	68.0	34.2	6.6	16.4	30.1	26.8	54.0	129.2	131.2	50.5	21.3	12.2	67.6	138.0	36.6
14	1	70.8	107.6	68.2	0.0	177.4	101.9	3.9	185.4	15.4	17.4	3.9	9.2	16.3	13.1	23.5	98.8	86.0	30.6	8.2	7.3	34.6	97.4	20.7
15	1	36.9	90.3	54.7	7.3	172.7	82.3	2.9	134.9	6.8	16.2	1.3	8.8	12.9	7.4	12.4	85.2	80.3	0.0	4.4	6.0	17.8	46.5	14.0
16	1	34.5	65.6	37.3	0.0	91.3	58.7	3.6	89.5	3.1	15.0	2.2	8.6	14.1	9.7	24.5	122.4	49.5	11.5	3.0	5.2	17.4	34.4	8.6
17	1	69.8	149.7	69.6	0.0	221.8	184.7	6.1	191.9	26.9	42.4	12.6	30.0	47.0	28.3	65.6	110.4	86.3	42.5	9.9	12.4	70.8	132.8	25.3
18	1	70.3	134.5	92.8	17.5	230.7	208.1	8.1	248.2	25.1	38.0	5.9	19.2	35.5	25.4	0.0	102.5	84.9	0.0	11.2	9.5	47.3	144.3	30.2
19	1	50.8	65.3	36.5	0.0	141.8	0.0	2.7	101.2	6.8	12.7	2.0	6.8	13.0	6.7	0.0	53.3	48.7	0.0	5.3	4.1	14.0	42.4	10.8
20	3	94.5	147.3	79.2	12.3	195.0	151.4	2.9	179.3	6.8	26.2	2.3	14.9	32.8	10.6	32.2	37.8	54.0	36.3	4.8	8.4	24.4	84.7	14.6
21	3	57.9	124.5	68.8	13.1	135.4	61.4	2.9	146.3	0.0	17.4	3.1	11.8	22.3	9.1	23.0	67.9	44.6	0.0	6.5	6.5	7.9	41.3	12.4
22	3	78.1	82.3	38.5	0.0	136.4	88.7	1.7	90.6	1.7	17.1	3.4	11.5	17.3	7.0	0.0	25.0	42.6	38.8	5.6	4.5	11.2	35.1	11.4
23	3	104.5	150.9	105.5	16.1	126.8	367.2	3.5	172.3	7.3	19.9	4.1	12.6	22.9	10.2	25.0	57.4	54.7	8.2	4.1	7.5	17.8	77.9	17.1
24	3	101.9	151.2	92.9	17.4	147.3	311.0	3.3	149.1	6.2	21.7	3.6	14.7	25.3	11.5	20.8	44.0	63.4	0.0	3.9	7.9	21.7	122.5	13.1
25	3	74.7	77.5	35.5	5.4	149.9	62.9	1.5	90.8	1.9	14.0	1.8	8.3	15.5	6.8	17.7	35.9	20.6	0.0	2.3	4.2	11.6	29.7	9.4
26	3	42.7	86.9	42.4	0.0	119.5	75.3	1.9	98.6	0.0	16.3	1.8	11.7	20.3	10.8	28.3	50.8	25.6	19.3	4.0	3.8	13.9	31.6	13.6
27	3	72.4	86.6	44.8	0.0	173.6	57.7	1.8	119.2	2.1	17.9	2.0	11.8	26.1	0.0	0.0	58.8	22.6	21.4	4.5	5.4	15.1	34.0	8.6
28	3	69.2	105.5	57.9	7.7	152.9	137.6	2.4	167.1	2.9	19.0	3.7	12.6	23.4	12.1	32.5	80.1	40.8	0.0	4.2	5.3	17.6	53.4	15.2
29	3	67.8	129.4	60.9	0.0	198.3	106.2	3.4	171.7	0.0	32.9	5.8	21.9	38.8	15.3	35.7	84.8	37.8	12.2	6.1	8.0	27.1	52.0	19.0
30	3	97.7	167.9	93.4	0.0	286.9	241.5	3.2	308.6	0.0	30.2	6.1	21.3	36.4	13.7	37.1	74.8	73.4	42.4	0.5	10.3	29.6	76.7	26.1
31	3	70.1	99.2	42.5	0.0	127.0	74.0	1.4	120.5	0.0	13.1	1.8	6.8	14.2	6.9	13.9	46.1	33.1	18.0	4.1	5.1	10.4	40.5	12.6
32	3	70.8	98.2	45.6	0.0	111.0	48.1	2.3	155.1	0.0	25.7	5.2	14.0	28.3	7.4	23.3	53.0	40.7	23.8	5.0	7.1	19.5	53.8	20.0
33	3	50.7	74.8	34.1	3.0	86.7	30.1	1.9	113.0	0.0	21.0	2.1	11.9	22.9	7.2	21.0	36.6	24.6	0.0	2.9	4.5	12.3	36.8	0.0
34	3	63.6	96.7	47.5	7.8	117.3	53.8	2.4	146.2	2.4	29.6	3.2	15.4	30.8	9.9	33.9	44.1	32.9	0.0	2.6	5.6	18.7	47.4	12.1
35	3	31.8	95.4	55.9	11.5	118.1	78.1	2.4	181.8	2.4	29.9	3.0	14.7	31.6	10.5	32.8	60.0	29.9	0.0	5.2	6.2	20.2	50.2	17.9
36	3	36.2	124.8	46.2	24.7	138.0	67.1	3.7	167.4	3.0	36.8	4.5	21.8	40.4	13.3	42.3	81.0	52.1	24.5	5.3	8.8	27.1	76.7	24.6
37	3	64.5	120.1	42.9	28.6	130.0	39.1	2.3	102.1	1.0	20.2	5.1	15.8	26.1	8.1	23.5	60.0	70.3	40.2	8.7	8.3	15.5	38.9	15.0
38	3	113.8	168.5	68.6	0.0	167.8	190.0	4.1	158.6	0.0	45.3	14.4	37.3	48.9	15.0	56.8	54.3	94.7	45.4	8.2	11.5	32.8	87.4	28.5
39	2	91.2	139.0	123.6	21.4	215.0	413.8	6.8	252.7	16.2	28.3	8.5	13.8	30.3	15.9	15.1	47.0	187.7	56.8	13.2	8.7	43.5	86.6	13.2
40	2	87.9	149.7	70.7	9.4	223.6	161.6	4.9	161.5	7.9	27.0	13.1	14.7	29.0	13.6	16.5	42.2	139.1	59.6	14.8	0.0	50.8	93.4	18.0
41	2	59.5	104.1	53.9	17.2	141.9	250.0	4.2	151.5	6.6	26.2	14.1	17.3	27.6	15.3	22.4	67.4	162.5	57.1	8.3	8.5	48.6	85.1	16.0
42	2	51.2	80.4	59.6	19.2	159.0	191.7	3.9	198.2	5.2	15.6	7.9	7.4	16.5	10.5	9.7	49.6	141.0	20.0	4.6	5.3	32.2	59.3	11.0
43	2	62.1	99.8	61.8	16.1	146.4	196.3	3.4	181.2	5.3	17.5	8.0	8.7	17.3	12.9	13.1	45.0	156.0	32.6	5.1	6.3	30.7	64.2	8.5
44	2	64.8	118.4	69.1	0.0	174.2	280.3	4.9	174.6	9.0	27.7	16.5	15.7	24.9	16.0	19.5	47.5	159.5	39.8	8.0	8.5	44.2	75.3	23.4
45	2	53.4	129.1	61.1	16.9	143.5	210.2	5.0	161.4	10.5	31.8	18.4	19.9	30.1	17.0	22.7	53.0	149.7	32.9	9.4	9.7	49.5	87.4	20.5
46	2	54.4	145.2	54.5	8.2	101.0	107.2	3.9	122.7	8.4	32.6	15.1	24.1	33.0	16.0	19.0	49.7	161.2	34.5	8.3	10.3	46.0	94.6	17.4
47	2	74.1	112.1	64.7	18.3	154.4	172.1	4.0	166.6	7.2	20.1	10.0	10.9	21.3	13.7	15.6	51.0	169.0	23.3	6.5	8.4	37.8	83.4	12.3
48	2	95.2	132.3	87.0	16.0	166.9	212.8	4.8	189.6	15.3	28.9	12.6	15.0	29.7	13.7	17.6	33.8	201.4	20.3	13.6	11.0	50.7	172.6	14.0
49	2	94.4	138.1	102.3	32.9	204.4	445.3	6.3	253.8	18.2	30.8	15.0	15.0	16.2	22.9	22.9	46.5	215.0	21.3	12.7	10.8	50.7	107.2	20.8
50	2	85.3	159.4	128.5	27.0	241.2	551.1	10.6	307.6	24.1	34.9	19.6	17.4	41.3	20.0	18.2	61.7	312.8	29.0	33.2	16.7	79.0	147.9	24.9
51	2	26.7	102.3	34.6	0.0	77.7	51.0	2.4	75.8	11.8	11.8	7.8	9.2	15.5	6.9	7.7	73.3	115.5	51.5	10.3	7.9	31.6	56.2	11.2

ORIGINAL DATA CONTINUES

OBS	GROUP	PK1	PK2	PK3	PK4	PK5	PK6	PK7	PK8	PK9	PK10	PK11
52	2	51.70	161.00	61.20	0.00	127.40	180.90	4.50	135.20	10.40	33.30	19.60
53	2	75.70	117.10	68.90	22.40	141.80	245.30	3.90	203.70	9.10	19.80	9.40
54	2	48.50	146.50	52.80	0.00	142.50	63.40	3.20	92.50	5.10	28.80	15.70
55	2	102.60	140.60	106.50	35.60	262.60	481.70	6.40	296.60	28.50	31.00	20.30
56	2	67.70	104.90	48.00	5.70	125.50	108.80	2.70	124.00	5.50	19.50	10.30
57	2	85.60	125.80	55.30	0.00	184.00	130.70	3.30	145.30	8.40	23.00	12.50
58	0	76.43	146.27	82.07	25.19	93.91	375.74	6.38	154.52	13.21	28.65	16.52
59	0	81.37	137.10	106.38	33.71	103.15	681.74	6.91	241.09	19.88	33.05	16.10
60	0	76.81	107.28	76.82	25.03	88.45	354.77	4.66	191.79	8.30	21.57	9.91
61	0	58.75	117.31	66.77	24.76	94.75	404.96	5.20	214.57	18.51	33.39	18.02
62	0	60.17	135.65	89.95	38.16	98.74	405.48	7.50	282.32	21.59	37.88	20.15
63	0	66.54	117.40	97.00	37.73	130.36	650.76	8.31	343.51	19.34	37.27	17.91
64	0	55.13	66.67	66.39	27.33	110.71	431.25	5.04	256.55	4.42	25.45	8.10
65	0	52.19	114.65	68.57	13.93	81.51	265.31	5.14	224.81	7.50	28.82	10.34
66	0	59.22	120.22	72.99	24.17	81.41	354.62	6.32	216.35	12.21	41.76	16.42
67	0	62.93	101.89	101.80	32.26	151.79	431.19	6.86	279.01	7.71	27.78	11.62
68	0	57.89	120.05	82.37	23.20	112.82	241.89	5.95	216.86	10.30	29.51	10.55
69	0	72.24	135.90	102.25	28.30	118.38	488.42	6.32	310.88	4.46	36.76	11.62
70	0	68.73	125.11	52.36	10.24	58.09	128.95	2.68	132.19	1.83	20.79	7.36
71	0	56.75	148.67	96.25	16.29	81.93	271.99	7.02	293.89	14.42	41.93	15.57
72	0	53.52	140.19	92.65	38.37	106.62	265.73	8.41	292.18	13.80	40.47	15.14
73	0	53.61	148.24	92.50	30.39	104.62	391.29	6.43	275.06	19.51	34.60	12.15
74	0	78.23	144.64	92.87	28.62	97.37	316.80	5.84	260.11	20.76	31.58	11.94
75	0	46.01	121.77	62.42	7.50	59.56	109.08	5.36	145.93	7.59	28.47	10.77
76	0	37.83	88.26	50.49	8.36	59.77	86.38	4.18	155.98	1.42	22.86	10.23
77	4	45.40	40.20	23.70	0.00	66.70	38.40	2.00	65.90	0.00	15.60	3.90

OBS	PK12	PK13	PK14	PK15	PK16	PK17	PK18	PK19	PK20	PK21	PK22	PK23
52	32.90	43.60	15.70	23.20	75.20	140.80	56.60	18.10	11.60	63.90	104.10	18.60
53	9.30	20.50	12.10	14.50	61.70	182.70	2.90	12.40	9.20	40.70	96.90	17.40
54	23.00	31.10	13.00	18.70	56.80	118.20	63.80	13.10	11.40	46.80	72.90	19.70
55	15.60	31.90	19.10	32.30	53.80	262.30	31.60	22.00	13.20	55.50	126.80	28.90
56	12.10	26.40	10.40	19.80	47.10	120.30	27.20	9.20	9.10	32.30	93.70	11.20
57	13.30	27.60	12.00	16.90	32.80	111.60	14.90	11.00	10.00	33.40	90.70	14.90
58	15.65	25.77	15.14	15.09	95.95	123.32	30.20	18.06	13.22	53.06	120.04	18.08
59	14.90	29.33	16.43	16.98	104.20	151.10	27.94	19.27	11.59	50.63	132.45	17.22
60	9.09	19.12	13.07	10.73	92.27	128.15	26.26	9.93	8.53	43.91	90.60	17.75
61	16.59	28.21	13.32	21.70	128.50	126.23	39.53	13.67	12.11	47.34	111.37	26.48
62	18.28	38.24	16.69	25.47	118.04	130.60	31.51	20.77	17.51	49.04	117.03	20.21
63	14.25	29.71	15.01	21.58	140.94	146.02	26.91	18.60	13.87	49.59	138.26	30.70
64	9.18	20.62	11.00	16.63	127.69	122.78	26.55	7.89	7.68	39.40	85.29	22.38
65	13.31	26.80	8.15	15.80	138.14	149.48	11.52	9.63	12.70	39.55	131.06	16.39
66	20.43	39.64	14.62	23.58	135.64	145.25	33.11	19.51	14.58	61.56	150.85	24.80
67	11.25	26.05	11.51	13.33	136.97	147.69	18.23	19.51	12.04	43.80	106.77	18.27
68	13.03	25.68	10.09	14.15	120.18	126.04	26.55	12.46	11.13	47.19	102.33	19.62
69	17.34	28.77	10.74	25.83	133.14	135.63	41.43	11.86	10.74	47.12	102.85	18.87
70	10.87	21.93	7.89	12.97	100.15	100.60	16.85	6.74	8.53	27.84	95.71	11.11
71	18.65	38.14	16.82	19.69	133.51	120.01	23.27	19.33	19.60	49.30	161.19	17.52
72	19.87	40.30	13.71	21.93	139.18	116.10	23.34	21.37	18.49	54.76	163.62	21.24
73	17.81	34.30	11.90	18.40	133.80	115.58	19.69	17.52	15.36	51.44	135.34	20.50
74	15.37	32.99	10.74	20.28	102.06	120.75	18.45	15.77	13.87	43.80	143.75	16.59
75	14.25	26.05	9.83	16.98	111.05	71.26	16.63	10.66	13.67	37.63	101.55	21.13
76	10.68	22.02	7.37	12.62	102.43	81.43	13.57	6.86	10.16	24.08	86.00	12.70
77	9.00	20.40	12.00	14.60	94.30	28.00	17.30	4.10	6.90	16.70	48.50	12.80

ORIGINAL DATA CONTINUES

OBS	GROUP	PK1	PK2	PK3	PK4	PK5	PK6	PK7	PK8	PK9	PK10	PK11	PK12	PK13	PK14	PK15	PK16	PK17	PK18	PK19	PK20	PK21	PK22	PK23
78	4	31.9	27.7	19.2	0.0	60.7	37.0	2.10	56.9	1.1	12.2	3.0	7.8	15.1	7.2	13.4	86.2	19.0	2.9	2.9	3.9	12.1	24.50	11.30
79	4	58.9	72.0	45.3	0.0	164.1	88.8	4.50	126.8	4.5	23.2	6.5	12.0	26.9	17.0	23.5	92.0	53.8	20.2	3.3	6.6	23.2	66.80	25.90
80	4	39.7	32.4	21.7	3.4	70.6	33.3	2.30	47.2	0.0	17.7	3.3	11.0	18.0	8.9	10.7	97.9	21.2	16.2	4.0	7.3	12.7	21.40	26.40
81	4	33.4	43.2	26.0	5.0	83.6	35.9	3.50	56.7	0.0	19.0	3.4	11.4	22.2	11.7	20.1	124.4	27.0	6.5	2.2	5.2	15.1	35.80	22.70
82	4	77.2	96.8	98.8	6.5	166.1	127.8	5.90	238.1	7.9	39.4	10.0	18.1	43.8	29.4	40.9	109.1	85.3	26.7	4.4	7.8	35.4	103.10	60.00
83	4	51.9	101.7	57.9	0.0	133.2	81.6	6.20	173.7	10.2	38.3	11.5	19.7	47.0	29.2	27.1	173.1	67.8	10.3	5.0	11.4	39.9	124.00	74.30
84	4	32.6	77.1	39.5	6.4	70.6	45.3	4.50	102.8	0.0	28.4	6.1	15.3	31.9	16.2	30.4	113.4	44.0	16.1	4.2	8.5	26.9	80.30	38.90
85	4	29.8	34.4	20.6	3.2	81.9	25.2	3.80	51.3	0.0	21.8	3.8	11.4	23.8	13.6	12.6	141.0	29.4	20.8	4.4	8.4	19.7	43.10	56.10
86	4	39.1	76.3	42.5	0.0	99.6	51.3	7.40	105.5	3.8	38.3	4.5	18.7	38.2	28.0	34.5	198.2	38.5	12.9	2.9	9.2	26.5	65.40	91.70
87	4	54.3	103.0	57.3	0.0	108.3	59.5	5.50	131.7	6.6	35.8	5.5	18.2	33.2	24.0	20.0	117.2	49.5	24.3	4.4	9.8	27.4	80.60	45.30
88	4	39.8	54.9	30.2	0.0	68.2	52.4	5.00	106.7	0.0	22.1	6.8	13.7	25.0	14.1	18.3	134.2	39.4	9.3	2.1	5.9	19.0	49.60	39.60
89	4	42.7	64.5	40.2	6.2	82.9	84.9	4.40	156.7	5.2	28.3	6.5	14.9	35.8	20.4	25.7	138.2	54.9	25.1	5.1	10.4	31.0	94.30	30.40
90	4	45.7	73.6	44.2	0.0	82.6	92.4	5.00	154.6	6.7	20.0	7.7	14.8	33.7	19.1	28.4	115.4	75.0	19.1	4.0	7.1	27.4	94.50	61.40
91	4	46.4	49.3	29.8	0.0	95.4	28.0	3.60	83.9	1.9	23.4	5.7	15.4	27.9	17.0	18.4	88.0	32.7	20.9	3.2	7.3	21.3	46.50	30.40
92	4	33.2	62.7	31.1	0.0	65.0	25.0	4.10	89.3	2.3	25.9	5.1	18.2	34.5	16.2	23.7	114.0	36.4	22.7	3.5	9.4	23.4	70.10	26.90
93	4	27.4	47.1	24.7	0.0	53.7	16.6	4.20	58.3	0.0	33.6	4.7	28.5	44.5	15.5	17.6	119.3	23.2	13.7	1.6	8.7	22.0	52.40	32.90
94	4	46.9	85.3	54.1	0.0	168.4	66.7	6.40	139.9	6.0	38.8	6.3	20.8	39.8	24.4	34.9	85.3	45.8	27.6	5.1	11.6	32.7	95.60	35.60
95	4	33.5	92.8	50.5	7.9	60.8	132.2	4.60	116.3	6.0	37.1	9.3	19.6	43.2	24.0	34.6	106.4	84.9	17.7	3.6	9.3	38.6	121.10	42.10
96	4	62.9	93.5	80.3	11.4	122.6	205.0	5.80	306.7	12.0	42.5	10.5	18.6	46.1	31.0	36.4	192.8	76.1	18.8	3.7	9.7	37.9	133.40	51.50
97	5	25.5	21.5	19.6	0.0	58.2	29.4	1.90	34.6	0.0	14.9	2.0	10.2	14.3	2.7	7.0	119.4	29.2	8.7	2.8	2.4	7.4	10.89	73.14
98	5	26.7	37.8	32.2	0.0	133.8	37.9	2.90	135.1	0.0	19.0	2.2	9.6	15.1	4.6	8.7	93.4	74.8	1.5	3.5	3.8	15.9	22.15	80.92
99	5	30.2	51.1	37.7	0.0	107.5	41.2	3.60	114.9	2.0	36.9	3.8	17.7	26.7	10.4	15.6	106.7	48.1	8.7	3.3	4.6	19.6	21.73	175.68
00	5	25.6	50.8	35.3	0.0	99.0	39.6	3.30	101.4	0.0	39.1	4.0	18.8	28.7	13.6	18.7	133.4	56.1	18.7	4.4	5.6	24.1	23.48	200.19
01	5	15.5	23.5	17.0	0.0	35.7	28.7	1.70	52.7	2.3	16.2	2.2	10.9	17.4	3.4	7.0	103.5	55.8	15.8	2.4	2.4	14.6	14.51	55.28
02	5	29.5	32.8	29.0	2.5	122.1	22.8	4.40	95.9	1.4	26.7	2.9	12.9	19.7	7.2	10.2	124.1	49.2	11.0	3.4	4.8	22.3	19.38	131.29
03	5	31.4	63.9	51.9	1.5	95.1	90.7	1.90	182.5	6.9	42.0	4.5	19.9	37.3	10.9	26.5	133.3	98.2	8.9	4.8	4.4	43.7	57.61	77.25
04	5	20.5	61.0	40.8	0.0	53.3	55.3	3.10	107.8	8.2	41.2	4.6	20.3	37.7	16.3	17.6	199.1	100.0	19.5	6.7	6.5	38.1	66.62	175.19
05	5	28.9	105.3	80.8	0.0	154.5	118.3	7.30	309.8	14.3	76.2	8.2	32.7	57.8	18.5	30.7	328.6	130.2	9.3	10.8	10.9	66.5	97.99	262.16
06	5	30.8	46.9	35.3	0.0	78.8	47.1	3.20	125.3	3.8	37.0	4.0	19.4	29.6	6.6	14.6	128.3	83.4	19.7	6.5	7.3	27.3	31.35	138.93
07	5	30.6	38.2	32.1	0.0	71.3	39.4	2.60	86.5	1.0	21.8	3.2	10.8	17.6	5.3	7.7	143.7	59.7	14.7	3.6	3.4	12.7	21.12	86.57
08	5	34.8	44.1	32.2	5.9	79.3	41.5	3.20	88.2	2.3	27.8	3.4	15.0	23.1	6.2	10.0	82.6	53.4	7.5	3.0	3.7	16.6	22.93	164.65
09	5	35.2	41.1	33.7	0.0	91.5	46.7	3.30	102.3	4.1	29.3	2.6	15.4	23.4	6.3	13.8	135.0	65.5	22.3	7.6	6.8	23.9	35.81	119.73
10	5	28.7	58.7	39.4	0.0	79.8	56.5	3.80	121.8	3.7	36.7	3.5	20.2	32.1	7.5	12.5	104.0	66.3	8.6	5.7	5.7	22.9	39.19	148.38
11	5	37.1	44.2	32.1	0.0	83.7	45.1	3.60	93.5	0.0	31.3	2.0	16.9	26.3	5.0	14.0	112.3	58.4	7.7	5.7	5.8	17.4	38.96	138.20
12	5	45.9	96.7	77.4	0.0	147.2	123.3	5.72	270.5	9.7	50.6	5.8	22.3	40.4	14.2	20.0	129.5	90.4	16.0	12.8	9.2	52.2	86.45	181.94
13	5	37.0	52.4	38.4	0.0	106.7	60.9	2.80	84.8	2.5	44.7	3.7	22.5	33.6	7.1	17.4	140.5	78.1	14.5	5.4	6.2	20.5	36.20	151.93
14	5	28.2	54.0	39.4	0.0	70.4	59.1	3.80	87.0	4.2	40.8	3.5	22.0	33.3	6.9	13.7	158.5	94.1	16.3	7.4	7.3	31.3	35.11	228.64
15	5	24.2	30.9	26.6	0.0	46.9	36.4	4.20	65.3	1.5	23.9	2.1	14.5	21.9	4.2	11.2	143.9	57.0	10.7	2.9	3.5	13.9	16.73	169.78
16	5	19.5	34.0	27.7	0.0	47.2	27.4	2.10	56.1	0.0	37.2	2.9	22.9	25.6	5.4	11.9	149.8	41.6	14.8	3.6	5.3	13.9	10.34	110.49

APPENDIX B

TABLE B.

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
				GROUP=0				
PK1	19	61.80789474	11.61607008	37.83000000	81.37000000	18.794	-0.01126282	-0.40264488
PK2	19	123.01421053	21.54461168	66.67000000	148.67000000	17.514	-1.04007336	1.22979757
PK3	19	81.73157895	17.07360317	50.49000000	106.38000000	20.890	-0.37746792	-0.98612803
PK4	19	24.92315789	9.73070744	7.50000000	38.37000000	39.043	-0.45725576	-0.67810941
PK5	19	96.52315789	23.91776538	58.09000000	151.79000000	24.779	0.27058663	0.46360483
PK6	19	350.33421053	158.27541984	86.38000000	681.74000000	45.178	0.33393814	0.33513614
PK7	19	6.02684211	1.40222067	2.68000000	8.41000000	23.266	-0.40152020	0.62409283
PK8	19	236.18947368	60.29811343	132.19000000	343.51000000	25.530	-0.21440050	-0.81986275
PK9	19	11.93473684	6.65972586	1.42000000	21.59000000	55.801	-0.02002267	-1.31792062
PK10	19	31.71526316	6.57323814	20.79000000	41.93000000	20.726	-0.02887215	-0.94895512
PK11	19	13.18000000	3.66244606	7.36000000	20.15000000	27.788	0.27882455	-1.01568270
PK12	19	14.77894737	3.49020677	9.09000000	20.43000000	23.616	-0.11128575	-0.99993044
PK13	19	29.14052632	6.51914316	19.12000000	40.30000000	22.371	0.38344325	-0.86127598
PK14	19	12.31736842	2.98579013	7.37000000	16.82000000	24.240	-0.03410865	-1.05827813
PK15	19	18.09157895	4.46746294	10.73000000	25.83000000	24.694	0.19482304	-0.95975310
PK16	19	120.72842105	16.73398646	92.27000000	140.94000000	13.861	-0.40536893	-1.48648613
PK17	19	124.10631579	21.58866838	71.26000000	151.10000000	17.395	-1.04015952	1.12019120
PK18	19	24.81789474	8.18476537	11.52000000	41.43000000	32.979	0.37734793	-0.27463054
PK19	19	14.70578947	5.03129574	6.74000000	21.37000000	34.213	-0.28834784	-1.48784790
PK20	19	12.91473684	3.27197930	7.68000000	19.60000000	25.335	0.42206939	-0.16102881
PK21	19	45.31789474	8.90664707	24.08000000	61.56000000	19.654	-0.84350390	1.19648192
PK22	19	119.79263158	24.69884407	85.29000000	163.62000000	20.618	0.30807336	-1.02768565
PK23	19	19.55578947	4.52504428	11.11000000	30.70000000	23.139	0.58793623	1.25775202
				GROUP=1				
PK1	19	59.56315789	20.49531682	31.30000000	103.70000000	34.409	0.50025381	-0.41711724
PK2	19	114.11578947	30.05822615	65.30000000	172.70000000	26.340	0.24943950	-0.57603124
PK3	19	69.35263158	26.25343085	36.50000000	138.80000000	37.855	1.04193092	1.23733009
PK4	19	10.18947368	12.56896065	0.00000000	36.20000000	123.352	0.96904243	-0.44608406
PK5	19	180.28421053	52.76732220	91.30000000	275.40000000	29.269	0.13899722	-1.06171768
PK6	19	153.04210526	112.52535673	0.00000000	442.60000000	73.526	1.35857982	1.92050463
PK7	19	4.66842105	1.64622553	2.70000000	8.30000000	35.263	0.95382789	0.38839023
PK8	19	169.01052632	76.50041318	89.50000000	405.00000000	45.264	1.73444849	3.92772487
PK9	19	17.71052632	16.42169617	0.00000000	68.00000000	92.723	1.72132061	3.99047175
PK10	19	21.90000000	8.75918565	10.80000000	42.40000000	39.996	1.04567236	0.44053158
PK11	19	4.38421053	2.95150868	1.30000000	12.60000000	67.321	1.50934120	2.08015789
PK12	19	12.96842105	5.54427158	6.80000000	30.00000000	42.752	1.75424538	3.91684573
PK13	19	20.81052632	9.10756064	12.90000000	47.00000000	43.764	1.63868229	2.63472743
PK14	19	16.28947368	6.87627925	6.70000000	28.30000000	42.213	0.45086875	-1.02937750
PK15	19	27.93684211	17.71945605	0.00000000	65.60000000	63.427	0.57443561	-0.03313714
PK16	19	94.86842105	23.42626856	53.30000000	137.40000000	24.693	0.15726711	-0.69373202
PK17	19	94.94736842	28.83933210	48.70000000	152.40000000	30.374	0.45831886	-0.19513835
PK18	19	25.40000000	18.83100281	0.00000000	65.50000000	74.138	0.25443105	-0.49470340
PK19	19	9.12105263	4.80781528	3.00000000	21.30000000	52.711	1.14557396	1.17052207
PK20	19	8.12105263	2.73568577	4.10000000	13.30000000	33.686	0.61042677	-0.64823773
PK21	19	38.52105263	18.26577184	14.00000000	71.00000000	47.418	0.53080265	-0.63942702
PK22	19	77.64736842	36.12305156	34.40000000	144.30000000	46.522	0.58800948	-0.99773868
PK23	19	19.64736842	7.80173545	8.50000000	36.60000000	39.709	0.34379500	-0.43948207

TABLE B. (CONTINUES)

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
GROUP=2								
PK1	19	70.10526316	20.19203566	26.70000000	102.60000000	28.802	-0.21355178	-0.56831093
PK2	19	126.62105263	22.16630523	80.40000000	161.00000000	17.506	-0.28405293	-0.62895418
PK3	19	71.79473684	25.83037561	34.60000000	128.50000000	35.978	1.08902721	0.36733741
PK4	19	14.01578947	11.29002820	0.00000000	35.60000000	80.552	0.24272194	-0.79735047
PK5	19	164.89473684	47.43037557	77.70000000	262.60000000	28.764	0.39787871	-0.11060188
PK6	19	234.43157895	141.92021253	51.00000000	551.10000000	60.538	0.97420896	0.15402600
PK7	19	4.68947368	1.86693294	2.40000000	10.60000000	39.811	1.86242844	4.80588772
PK8	19	178.65789474	62.98802634	75.80000000	307.60000000	35.256	0.60166848	-0.00505240
PK9	19	11.19473684	6.54679512	5.10000000	28.50000000	58.481	1.49989982	1.77764097
PK10	19	25.71578947	6.61742340	11.80000000	34.90000000	25.733	-0.59963847	-0.65780469
PK11	19	13.38947368	4.24144822	7.80000000	20.30000000	31.677	0.19548969	-1.22312053
PK12	19	15.54210526	6.19482901	7.40000000	32.90000000	39.858	1.26657145	2.22268354
PK13	19	27.04210526	7.89544987	15.50000000	43.60000000	29.197	0.31954812	-0.08412567
PK14	19	14.56315789	3.69807916	6.90000000	22.90000000	25.393	0.27815198	0.69838828
PK15	19	18.17894737	5.43870685	7.70000000	32.30000000	29.918	0.47398810	1.73648763
PK16	19	52.37368421	11.54310001	32.80000000	75.20000000	22.040	0.40928939	-0.02065485
PK17	19	168.75263158	51.31749072	111.60000000	312.80000000	30.410	1.51177269	2.53589214
PK18	19	35.56315789	17.43716371	2.90000000	63.80000000	49.032	0.14855933	-0.92617035
PK19	19	12.30526316	6.63772666	4.60000000	33.20000000	53.942	1.89954132	4.78320176
PK20	19	9.29473684	3.35335217	0.00000000	16.70000000	36.078	-0.66065400	3.24368170
PK21	19	45.67894737	12.17404247	30.70000000	79.00000000	26.651	1.06646912	1.85543634
PK22	19	94.64736842	28.91862853	56.20000000	172.60000000	30.554	1.31136056	2.09400900
PK23	19	16.94210526	5.31897189	8.50000000	28.90000000	31.395	0.50964320	-0.09249699
GROUP=3								
PK1	19	71.73157895	22.94140005	31.80000000	113.80000000	31.982	0.10669285	-0.51238343
PK2	19	115.14210526	30.50628926	74.80000000	168.50000000	26.494	0.44819460	-1.07782192
PK3	19	58.05789474	21.29873016	34.10000000	105.50000000	36.685	0.99089507	-0.05887298
PK4	19	7.76842105	8.99642229	0.00000000	28.60000000	115.808	1.00540153	0.15913330
PK5	19	148.31052632	43.77288474	86.70000000	286.90000000	29.514	1.86148626	4.87953791
PK6	19	117.95789474	95.17829652	30.10000000	367.20000000	80.688	1.60284051	1.81525513
PK7	19	2.57894737	0.79273013	1.40000000	4.10000000	30.739	0.27552743	-0.97162295
PK8	19	149.38421053	49.36598315	90.60000000	308.60000000	33.046	1.75052031	5.36407515
PK9	19	1.98421053	2.40376167	0.00000000	7.30000000	121.144	1.23079489	0.56222132
PK10	19	23.90526316	8.45224978	13.10000000	45.30000000	35.357	0.98546952	0.69401729
PK11	19	4.05263158	2.85528135	1.80000000	14.40000000	70.455	2.84264277	9.99004417
PK12	19	15.30526316	6.69962467	6.80000000	37.30000000	43.773	2.10357468	5.96369532
PK13	19	27.59473684	9.00379387	14.20000000	48.90000000	32.629	0.68043977	0.27121811
PK14	19	9.75789474	3.59309767	0.00000000	15.30000000	36.822	-0.79462157	1.77571828
PK15	19	26.30526316	13.47295537	0.00000000	56.80000000	51.218	-0.09280329	1.01545787
PK16	19	55.38947368	16.76666318	25.00000000	84.80000000	30.270	0.20376123	-0.64955428
PK17	19	45.17894737	19.69172356	20.60000000	94.70000000	43.586	0.95424078	0.67301044
PK18	19	17.39473684	16.80996641	0.00000000	45.40000000	96.638	0.37851608	-1.36949725
PK19	19	4.65789474	1.94545504	0.50000000	8.70000000	41.767	0.17481779	0.75721144
PK20	19	6.78421053	2.13053545	3.80000000	11.50000000	31.404	0.56214512	-0.27003484
PK21	19	18.65263158	7.00986356	7.90000000	32.80000000	37.581	0.46145651	-0.62418356
PK22	19	56.34736842	24.75058627	29.70000000	122.50000000	43.925	1.20002809	1.17920693
PK23	19	15.32631579	6.60974393	0.00000000	28.50000000			

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
				GROUP=4				
PK1	20	43.63500000	12.60944821	27.40000000	77.20000000	28.898	1.07987039	1.18500841
PK2	20	66.42500000	24.39898348	27.70000000	103.00000000	36.732	-0.02501603	-1.29459348
PK3	20	41.88000000	20.66121716	19.20000000	98.80000000	49.334	1.34059103	1.92140730
PK4	20	2.50000000	3.52300709	0.00000000	11.40000000	140.920	1.15766150	0.35676333
PK5	20	95.25000000	36.92021982	53.70000000	168.40000000	38.761	1.03236628	-0.13136958
PK6	20	66.36500000	46.46318605	16.60000000	205.00000000	70.012	1.61663164	2.98809456
PK7	20	4.54000000	1.44090615	2.00000000	7.40000000	31.738	-0.09076099	-0.22282136
PK8	20	118.45000000	65.87365498	47.20000000	306.70000000	55.613	1.45238468	2.49236784
PK9	20	3.71000000	3.75189777	0.00000000	12.00000000	101.129	0.67694118	-0.47596689
PK10	20	28.07000000	9.23534971	12.20000000	42.50000000	32.901	0.01434970	-1.35244985
PK11	20	6.20500000	2.49282655	3.00000000	11.50000000	40.174	0.75401434	-0.27224049
PK12	20	15.85500000	4.82324143	7.80000000	28.50000000	30.421	0.58755221	1.10512534
PK13	20	32.55000000	9.82550387	15.10000000	47.00000000	30.186	-0.13744371	-1.12170465
PK14	20	18.94500000	7.07445181	7.20000000	31.00000000	37.342	0.23568501	-0.97278015
PK15	20	24.29000000	8.87295714	10.70000000	40.90000000	36.529	0.23900766	-1.03330106
PK16	20	122.02000000	33.13653732	85.30000000	198.20000000	27.157	1.18629097	0.80173380
PK17	20	46.59500000	21.30296434	19.00000000	85.30000000	45.719	0.60127733	-0.83114248
PK18	20	17.45500000	6.66573919	2.90000000	27.60000000	38.188	-0.51469681	-0.18887179
PK19	20	3.68500000	0.99009569	1.60000000	5.10000000	26.868	-0.49158415	-0.32009791
PK20	20	8.22000000	1.99330458	3.90000000	11.60000000	24.249	-0.28014092	-0.13438979
PK21	20	25.44500000	8.48351187	12.10000000	39.90000000	33.341	0.18926142	-0.89009834
PK22	20	72.55000000	32.98657622	21.40000000	133.40000000	45.467	0.27076191	-0.86174526
PK23	20	40.81000000	20.25084534	11.30000000	91.70000000	49.622	0.87688161	0.75102153
				GROUP=5				
PK1	20	29.29000000	6.86906873	15.50000000	45.90000000	23.452	0.25467803	0.93753876
PK2	20	49.44500000	21.17947057	21.50000000	105.30000000	42.834	1.39976670	2.25694368
PK3	20	37.93000000	15.98667537	17.00000000	80.80000000	42.148	1.76886627	3.30257998
PK4	20	0.49500000	1.42181315	0.00000000	5.90000000	287.235	3.35534245	11.86897321
PK5	20	88.10000000	33.33776111	35.70000000	154.50000000	37.841	0.45109371	-0.42232322
PK6	20	52.36500000	27.84156822	22.80000000	123.30000000	53.168	1.67847107	2.27109099
PK7	20	3.42100000	1.32075696	1.70000000	7.30000000	38.607	1.43705658	3.08145534
PK8	20	115.80000000	68.04159718	34.60000000	309.80000000	58.758	1.86677895	3.50854579
PK9	20	3.39500000	3.78034738	0.00000000	14.30000000	111.350	1.59361217	2.52334205
PK10	20	34.66500000	13.88616407	14.90000000	76.20000000	40.058	1.21196625	3.13363540
PK11	20	3.55500000	1.48021158	2.00000000	8.20000000	41.637	1.74286976	4.24260761
PK12	20	17.74500000	5.61412457	9.60000000	32.70000000	31.638	0.66108175	1.18881722
PK13	20	28.08000000	10.35626431	14.30000000	57.80000000	36.881	1.15446785	2.22585038
PK14	20	8.11500000	4.42567568	2.70000000	18.50000000	54.537	1.10120694	0.31525064
PK15	20	14.44000000	6.18847826	7.00000000	30.70000000	42.856	1.18933766	1.51783245
PK16	20	138.48000000	51.55095180	82.60000000	328.60000000	37.226	2.85686147	10.15378913
PK17	20	69.47500000	24.14473627	29.20000000	130.20000000	34.753	0.76917445	0.61537211
PK18	20	12.74500000	5.26982223	1.50000000	22.30000000	41.348	-0.06079202	-0.44566190
PK19	20	5.31500000	2.74653370	2.40000000	12.80000000	51.675	1.44080463	1.95724457
PK20	20	5.48000000	2.16371999	2.40000000	10.90000000	39.484	0.80068630	0.79508013
PK21	20	25.24000000	14.72039044	7.40000000	66.50000000	58.322	1.54816186	2.24824866
PK22	20	35.42750000	24.24673887	10.34000000	97.99000000	68.440	1.46504183	1.59777274

TABLE B. (CONTINUED)

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
				GROUP=1				
LPK1	19	4.040	0.346	3.459	4.646	8.566	-0.059	-0.998
LPK2	19	4.708	0.271	4.187	5.154	5.756	-0.313	-0.288
LPK3	19	4.184	0.360	3.611	4.937	8.609	0.219	-0.457
LPK4	19	1.166	1.860	-0.693	3.603	159.466	0.045	-2.020
LPK5	19	5.155	0.306	4.520	5.620	5.936	-0.334	-0.680
LPK6	19	4.604	1.420	-0.693	6.094	30.836	-3.066	11.663
LPK7	19	1.598	0.302	1.163	2.175	18.925	0.379	-0.518
LPK8	19	5.054	0.393	4.500	6.005	7.776	0.720	0.113
LPK9	19	2.371	1.316	-0.693	4.227	55.518	-1.293	1.534
LPK10	19	3.043	0.368	2.425	3.759	12.085	0.372	-0.466
LPK11	19	1.441	0.536	0.588	2.573	37.181	0.530	-0.410
LPK12	19	2.535	0.358	1.988	3.418	14.130	0.771	0.484
LPK13	19	2.989	0.367	2.595	3.861	12.262	0.955	0.155
LPK14	19	2.738	0.424	1.974	3.360	15.480	-0.134	-0.925
LPK15	19	2.928	1.347	-0.693	4.191	46.004	-2.255	4.692
LPK16	19	4.528	0.255	3.985	4.927	5.626	-0.355	-0.297
LPK17	19	4.514	0.311	3.896	5.030	6.891	-0.311	0.104
LPK18	19	2.525	1.760	-0.693	4.190	69.699	-1.319	0.082
LPK19	19	2.153	0.484	1.253	3.082	22.501	0.073	-0.393
LPK20	19	2.108	0.313	1.526	2.625	14.871	0.105	-0.742
LPK21	19	3.553	0.497	2.674	4.270	13.980	-0.239	-0.901
LPK22	19	4.257	0.464	3.552	4.975	10.899	0.130	-1.398
LPK23	19	2.926	0.413	2.197	3.614	14.111	-0.328	-0.862
				GROUP=2				
LPK1	19	4.212	0.325	3.303	4.636	7.725	-1.086	1.884
LPK2	19	4.830	0.184	4.393	5.085	3.800	-0.653	0.045
LPK3	19	4.226	0.334	3.558	4.860	7.893	0.411	0.075
LPK4	19	1.932	1.665	-0.693	3.586	86.227	-0.952	-0.917
LPK5	19	5.068	0.298	4.359	5.573	5.876	-0.425	0.585
LPK6	19	5.280	0.638	3.942	6.313	12.086	-0.345	0.006
LPK7	19	1.596	0.315	1.065	2.407	19.735	0.746	1.230
LPK8	19	5.128	0.362	4.335	5.730	7.059	-0.310	0.202
LPK9	19	2.337	0.489	1.723	3.367	20.923	0.663	-0.315
LPK10	19	3.231	0.288	2.510	3.567	8.925	-1.066	0.547
LPK11	19	2.585	0.316	2.116	3.035	12.217	-0.157	-1.319
LPK12	19	2.711	0.365	2.067	3.509	13.446	0.227	0.039
LPK13	19	3.275	0.297	2.773	3.786	9.073	-0.295	-0.638
LPK14	19	2.682	0.259	2.001	3.153	9.656	-0.682	1.621
LPK15	19	2.884	0.312	2.104	3.490	10.822	-0.777	1.596
LPK16	19	3.945	0.220	3.506	4.327	5.586	-0.176	0.137
LPK17	19	5.094	0.272	4.719	5.747	5.344	0.808	0.602
LPK18	19	3.424	0.682	1.224	4.164	19.926	-1.868	5.301
LPK19	19	2.446	0.457	1.629	3.517	18.705	0.362	0.647
LPK20	19	2.147	0.728	-0.693	2.845	33.905	-3.603	14.509
LPK21	19	3.802	0.251	3.440	4.376	6.611	0.337	0.020
LPK22	19	4.516	0.282	4.038	5.154	6.243	0.468	-0.550

TABLE B.2.

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
				GROUP=3				
LPK1	19	4.227	0.347	3.475	4.739	8.199	-0.645	0.091
LPK2	19	4.718	0.261	4.321	5.130	5.538	0.151	-1.252
LPK3	19	4.013	0.338	3.544	4.663	8.431	0.552	-0.801
LPK4	19	1.139	1.672	-0.693	3.371	146.779	-0.113	-1.935
LPK5	19	4.968	0.260	4.468	5.661	5.234	0.844	1.911
LPK6	19	4.528	0.692	3.421	5.907	15.288	0.599	-0.403
LPK7	19	1.093	0.261	0.642	1.526	23.929	-0.094	-1.045
LPK8	19	4.966	0.299	4.512	5.734	6.025	0.473	1.056
LPK9	19	0.423	1.055	-0.693	2.054	249.661	0.117	-1.592
LPK10	19	3.142	0.329	2.610	3.824	10.474	0.341	-0.555
LPK11	19	1.394	0.470	0.833	2.701	33.763	1.051	1.902
LPK12	19	2.693	0.363	1.988	3.632	13.467	0.655	1.749
LPK13	19	3.287	0.320	2.688	3.900	9.743	-0.058	-0.313
LPK14	19	2.188	0.742	-0.693	2.760	33.897	-3.545	14.183
LPK15	19	2.919	1.312	-0.693	4.048	44.955	-2.490	5.421
LPK16	19	3.978	0.318	3.239	4.446	7.991	-0.483	0.072
LPK17	19	3.738	0.420	3.049	4.556	11.237	0.128	-0.684
LPK18	19	1.780	1.983	-0.693	3.826	111.402	-0.461	-1.828
LPK19	19	1.551	0.486	0.000	2.219	31.304	-1.800	5.208
LPK20	19	1.946	0.291	1.459	2.485	14.948	0.060	-0.890
LPK21	19	2.887	0.378	2.128	3.506	13.085	-0.193	-0.648
LPK22	19	3.961	0.400	3.408	4.812	10.102	0.518	-0.685
LPK23	19	2.582	0.853	-0.693	3.367	33.031	-3.394	13.438
				GROUP=4				
LPK1	20	3.752	0.269	3.329	4.353	7.164	0.470	-0.252
LPK2	20	4.132	0.404	3.339	4.640	9.774	-0.503	-0.913
LPK3	20	3.647	0.450	2.981	4.598	12.347	0.366	-0.505
LPK4	20	0.322	1.297	-0.693	2.477	402.742	0.548	-1.737
LPK5	20	4.499	0.355	3.993	5.129	7.891	0.605	-0.762
LPK6	20	4.003	0.642	2.839	5.325	16.027	0.228	-0.468
LPK7	20	1.573	0.317	0.916	2.067	20.131	-0.799	0.218
LPK8	20	4.649	0.516	3.865	5.727	11.097	0.241	-0.496
LPK9	20	0.852	1.255	-0.693	2.526	147.351	-0.276	-1.732
LPK10	20	3.298	0.348	2.542	3.761	10.567	-0.456	-0.655
LPK11	20	1.840	0.364	1.253	2.485	19.782	0.154	-0.816
LPK12	20	2.752	0.302	2.116	3.367	10.980	-0.303	0.063
LPK13	20	3.451	0.326	2.747	3.861	9.435	-0.610	-0.508
LPK14	20	2.899	0.392	2.041	3.450	13.508	-0.458	-0.280
LPK15	20	3.145	0.379	2.416	3.723	12.050	-0.301	-0.882
LPK16	20	4.777	0.249	4.452	5.292	5.218	0.721	-0.073
LPK17	20	3.754	0.460	2.970	4.452	12.244	-0.005	-1.047
LPK18	20	2.790	0.513	1.224	3.336	18.395	-1.766	3.633
LPK19	20	1.401	0.266	0.742	1.723	18.962	-1.034	0.700
LPK20	20	2.138	0.251	1.482	2.493	11.734	-0.[illegible]	1.117
LPK21	20	3.201	0.347	2.534	3.699	10.852	-0.394	-0.603
LPK22	20	4.179	0.512	3.086	4.897	12.262	-0.599	-0.245
LPK23	20	3.601	0.521	2.468	4.524	14.469	-0.469	0.336

TABLE B.2.1 (CONTINUES)

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
				GROUP=5				
LPK1	20	3.368	0.242	2.773	3.837	7.176	-0.606	1.064
LPK2	20	3.835	0.395	3.091	4.662	10.299	0.216	0.409
LPK3	20	3.581	0.365	2.862	4.398	10.190	0.545	1.420
LPK4	20	-0.407	0.725	-0.693	1.856	-178.275	2.443	5.066
LPK5	20	4.413	0.395	3.589	5.043	8.961	-0.327	-0.411
LPK6	20	3.863	0.447	3.148	4.819	11.581	0.763	0.379
LPK7	20	1.318	0.312	0.788	2.054	23.664	0.364	0.608
LPK8	20	4.626	0.509	3.558	5.738	11.002	0.326	0.983
LPK9	20	0.875	1.090	-0.693	2.695	124.570	-0.262	-0.948
LPK10	20	3.489	0.391	2.734	4.340	11.209	-0.164	0.232
LPK11	20	1.347	0.323	0.916	2.163	23.991	0.680	0.714
LPK12	20	2.859	0.312	2.313	3.503	10.914	-0.182	-0.360
LPK13	20	3.295	0.348	2.695	4.066	10.564	0.136	-0.094
LPK14	20	2.040	0.483	1.163	2.944	23.701	0.291	-0.485
LPK15	20	2.630	0.392	2.015	3.440	14.907	0.232	-0.249
LPK16	20	4.888	0.290	4.420	5.796	5.938	1.528	4.415
LPK17	20	4.192	0.347	3.391	4.873	8.277	-0.181	0.322
LPK18	20	2.478	0.537	0.693	3.127	21.676	-1.910	5.691
LPK19	20	1.671	0.421	1.065	2.588	25.210	0.608	-0.331
LPK20	20	1.727	0.363	1.065	2.434	21.021	-0.117	-0.220
LPK21	20	3.117	0.514	2.067	4.205	16.500	0.335	0.225
LPK22	20	3.395	0.616	2.383	4.590	18.148	0.313	-0.393
LPK23	20	4.896	0.409	4.021	5.571	8.360	-0.530	-0.350
				GROUP=0				
LPK1	19	4.115	0.194	3.646	4.405	4.705	-0.512	0.459
LPK2	19	4.799	0.199	4.207	5.005	4.147	-1.632	3.370
LPK3	19	4.387	0.223	3.932	4.672	5.077	-0.690	-0.480
LPK4	19	3.141	0.487	2.079	3.660	15.514	-1.131	0.256
LPK5	19	4.545	0.256	4.071	5.026	5.643	-0.438	0.064
LPK6	19	5.739	0.551	4.465	6.525	9.598	-1.033	0.789
LPK7	19	1.851	0.240	1.157	2.187	12.982	-1.260	2.764
LPK8	19	5.433	0.276	4.888	5.841	5.079	-0.632	-0.599
LPK9	19	2.323	0.726	0.652	3.095	31.270	-1.059	0.492
LPK10	19	3.452	0.211	3.058	3.748	6.126	-0.368	-0.727
LPK11	19	2.581	0.273	2.062	3.028	10.565	-0.128	-0.867
LPK12	19	2.700	0.241	2.261	3.041	8.913	-0.463	-0.792
LPK13	19	3.366	0.220	2.977	3.709	6.522	0.032	-0.824
LPK14	19	2.524	0.244	2.063	2.852	9.660	-0.399	-0.771
LPK15	19	2.895	0.246	2.419	3.271	8.489	-0.184	-0.866
LPK16	19	4.788	0.143	4.530	4.952	2.989	-0.508	-1.351
LPK17	19	4.808	0.197	4.273	5.021	4.089	-1.540	2.490
LPK18	19	3.180	0.337	2.487	3.736	10.604	-0.328	-0.370
LPK19	19	2.661	0.372	1.980	3.085	13.997	-0.621	-0.992
LPK20	19	2.568	0.246	2.102	3.001	9.593	-0.137	-0.326
LPK21	19	3.803	0.222	3.202	4.128	5.845	-1.488	2.515
LPK22	19	4.770	0.205	4.452	5.101	4.305	0.037	-1.146

TABLE B.2. (CONTINUES)

B_0

TABLE B.2.2

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
GROUP=0								
SQ_PK1	19	7.82796033	0.74861743	6.15060973	9.02053214	9.563	1.24888382	-0.05647899
SQ_PK2	19	11.04575998	1.03017109	8.16516993	12.19303080	9.326	-1.32675517	2.19484742
SQ_PK3	19	8.99078087	0.97329127	7.10563157	10.31406806	10.825	-0.53050133	-0.76916441
SQ_PK4	19	4.88101524	1.07698432	2.73861279	6.19435227	22.065	-0.81220397	-0.30218370
SQ_PK5	19	9.75151393	1.22908153	7.62167960	12.32030844	12.604	-0.09941950	0.13879821
SQ_PK6	19	18.20374608	4.47337148	9.29408414	26.11015128	24.574	-0.37168050	0.17124360
SQ_PK7	19	2.43747623	0.30050723	1.63707055	2.90000000	12.329	-0.85424276	1.57849090
SQ_PK8	19	15.24149687	2.02537638	11.49739101	18.53402277	13.289	-0.42773628	-0.76001767
SQ_PK9	19	3.29152187	1.07785259	1.19163753	4.64650406	32.746	-0.50737503	-0.70818803
SQ_PK10	19	5.60213235	0.59142716	4.55960525	6.47533783	10.557	-0.19875840	-0.86525239
SQ_PK11	19	3.59692201	0.50557390	2.71293199	4.48887514	14.056	0.07670093	-1.00973499
SQ_PK12	19	3.81775135	0.46372397	3.01496269	4.51995575	12.147	-0.29306348	-0.92594385
SQ_PK13	19	5.36649452	0.60018495	4.37264222	6.34822810	11.184	0.21061748	-0.87032833
SQ_PK14	19	3.48421134	0.43302271	2.71477439	4.10121933	12.428	-0.22039575	-0.94373596
SQ_PK15	19	4.22226191	0.52797214	3.27566787	5.08232230	12.504	0.00388509	-0.96268664
SQ_PK16	19	10.96177574	0.77423733	9.60572746	11.87181536	7.063	-0.45606308	-1.42392801
SQ_PK17	19	11.09545952	1.02590829	8.44156384	12.29227400	9.246	-1.29196011	1.76358544
SQ_PK18	19	4.91621513	0.82750452	3.39411255	6.43661402	16.832	0.01939895	-0.47339822
SQ_PK19	19	3.77549893	0.69027165	2.59615100	4.62276973	18.283	-0.45264822	-1.27688759
SQ_PK20	19	3.56643361	0.45402366	2.77128129	4.42718872	12.730	0.13322269	-0.31418990
SQ_PK21	19	6.69709966	0.70191276	4.90713766	7.84601810	10.481	-1.17749761	1.77355666
SQ_PK22	19	10.89017826	1.12389038	9.23525852	12.79140336	10.320	0.17212143	-1.11080059
SQ_PK23	19	4.39435240	0.50901175	3.33316666	5.54075807	11.583	0.14805914	1.00175454
GROUP=1								
SQ_PK1	19	7.60996630	1.32035021	5.59464029	10.18331969	17.350	0.20887843	-0.81192348
SQ_PK2	19	10.59312665	1.41671933	8.08084154	13.14153720	13.374	-0.02213357	-0.52143485
SQ_PK3	19	8.19602055	1.51620313	6.04152299	11.78134118	18.499	0.61081941	0.15203478
SQ_PK4	19	2.22899198	2.34757914	0.00000000	6.01664358	105.320	0.33099007	-1.59265029
SQ_PK5	19	13.28626953	1.99200902	9.55510335	16.59518002	14.993	-0.08493423	-0.96885278
SQ_PK6	19	11.47810191	4.74113415	0.00000000	21.03806075	41.306	-0.10795702	1.31340316
SQ_PK7	19	2.13095533	0.36678467	1.64316767	2.88097206	17.212	0.63687448	-0.18270137
SQ_PK8	19	12.73563393	2.68192448	9.46044396	20.12461180	21.058	1.18535983	1.62633145
SQ_PK9	19	3.70764937	2.04550166	0.00000000	8.24621125	55.170	0.05057523	0.45210618
SQ_PK10	19	4.59836770	0.89272601	3.28633535	6.51152824	19.414	0.71364581	-0.13714521
SQ_PK11	19	1.99862788	0.64136336	1.14017543	3.54964787	32.090	0.96395963	0.39020613
SQ_PK12	19	3.53570275	0.70227074	2.60768096	5.47722558	19.862	1.22471677	1.82679855
SQ_PK13	19	4.47546733	0.90779501	3.59165700	6.85565460	20.284	1.27283094	1.18547329
SQ_PK14	19	3.94979906	0.85253415	2.58843582	5.31977443	21.584	0.16879853	-1.06309373
SQ_PK15	19	4.86651555	2.11900792	0.00000000	8.09938269	43.543	-1.00618766	1.58590979
SQ_PK16	19	9.66801210	1.21475416	7.30068490	11.72177461	12.565	-0.08919928	-0.57405626
SQ_PK17	19	9.63694957	1.48051902	6.97853853	12.34503949	15.363	0.09188801	-0.16958655
SQ_PK18	19	4.36351587	2.59095494	0.00000000	8.09320703	59.378	-0.78426123	-0.50280111
SQ_PK19	19	2.92785909	0.76103664	1.73205081	4.61519230	25.993	0.58422307	0.07010521
SQ_PK20	19	2.81244021	0.47219469	2.02484567	3.64691651	16.790	0.35523963	-0.76141985
SQ_PK21	19	6.03736785	1.47861798	3.74165739	8.42614977	24.491	0.14573594	-0.87743422
SQ_PK22	19	8.59000485	2.01831225	5.86515132	12.01249350	23.496	0.35939199	-1.26461849
SQ_PK23	19	4.34631481	0.89384949	2.91547595	6.04979338	20.566	-0.01308477	-0.84703411

TABLE B.2.2 (CONTINUES)

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
GROUP=2								
SQ_PK1	19	8.28190344	1.26472291	5.16720427	10.12916581	15.271	-0.59751504	0.31709464
SQ_PK2	19	11.20990824	1.00612537	8.96660471	12.68857754	8.975	-0.46179163	-0.35533403
SQ_PK3	19	8.35465549	1.45095563	5.88217647	11.33578405	17.367	0.78854255	0.07970766
SQ_PK4	19	3.13183100	2.10740833	0.00000000	5.96657356	67.290	-0.60084172	-1.09871635
SQ_PK5	19	12.71350910	1.85542810	8.81476035	16.20493752	14.594	0.01557002	0.03111179
SQ_PK6	19	14.66456522	4.52315039	7.14142843	23.47551916	30.844	0.40341439	-0.30289801
SQ_PK7	19	2.13128129	0.39406438	1.54919334	3.25576412	18.490	1.24522903	2.54343648
SQ_PK8	19	13.16998957	2.34492504	8.70631954	17.53852901	17.805	0.17473027	-0.10839099
SQ_PK9	19	3.23225963	0.88811462	2.25831796	5.33853913	27.477	1.06958206	0.51545174
SQ_PK10	19	5.02618057	0.69172366	3.43511281	5.90762220	13.762	-0.82344500	-0.16101088
SQ_PK11	19	3.61459185	0.58498762	2.79284801	4.50555213	16.184	0.01005947	-1.30117907
SQ_PK12	19	3.87467743	0.74724016	2.72029410	5.73585216	19.285	0.71929485	0.80937902
SQ_PK13	19	5.14663432	0.76488734	3.93700394	6.60302961	14.862	-0.01064757	-0.45918573
SQ_PK14	19	3.78603731	0.49173772	2.62678511	4.78539445	12.988	-0.19886076	0.90212779
SQ_PK15	19	4.21674053	0.64819781	2.77488739	5.68330890	15.372	-0.18187005	1.26149929
SQ_PK16	19	7.19544443	0.79533391	5.72712843	8.67179336	11.053	0.12374133	-0.01851560
SQ_PK17	19	12.86512797	1.84964207	10.56409012	17.68615278	14.377	1.15834356	1.45162140
SQ_PK18	19	5.75436593	1.60828035	1.70293864	7.98749022	27.949	-0.61213696	0.65352872
SQ_PK19	19	3.40775495	0.85495020	2.14476106	5.76194412	25.088	1.10727930	2.10942182
SQ_PK20	19	2.94434275	0.81261136	0.00000000	4.08656335	27.599	-2.74838572	10.20492938
SQ_PK21	19	6.70532462	0.87030687	5.54075807	8.88819442	12.979	0.67910570	0.77711221
SQ_PK22	19	9.63115051	1.41181244	7.49666593	13.13773192	14.659	0.89660391	1.18319036
SQ_PK23	19	4.06818732	0.64322046	2.91547595	5.37587202	15.811	0.17577158	-0.44577469
GROUP=3								
SQ_PK1	19	8.36026445	1.39271094	5.63914887	10.66770828	16.659	-0.26148590	-0.33202080
SQ_PK2	19	10.64279273	1.40610366	8.64869932	12.98075499	13.212	0.30113451	-1.19331698
SQ_PK3	19	7.50834648	1.33270674	5.83952053	10.27131929	17.750	0.77266798	-0.48963348
SQ_PK4	19	2.03383613	1.95798509	0.00000000	5.34789678	96.271	0.18564997	-1.55042662
SQ_PK5	19	12.06999093	1.66484997	9.31128348	16.93812268	13.793	1.36090946	3.13173616
SQ_PK6	19	10.18743923	3.86799946	5.48634669	19.16246331	37.968	1.14432092	0.45486743
SQ_PK7	19	1.58765398	0.24807505	1.18321596	2.02484567	15.625	0.05403710	-1.05519200
SQ_PK8	19	12.08353586	1.88672355	9.51840323	17.56701454	15.614	1.07388326	2.82599996
SQ_PK9	19	1.02443533	0.99331412	0.00000000	2.70185122	96.962	0.27455765	-1.31784288
SQ_PK10	19	4.82188755	0.83128442	3.61939221	6.73052747	17.240	0.64959411	-0.10535536
SQ_PK11	19	1.93205022	0.58101716	1.34164079	3.79473319	30.073	1.85303774	5.09236306
SQ_PK12	19	3.83999413	0.76863717	2.60768096	6.10737259	20.017	1.38477961	3.33652286
SQ_PK13	19	5.18828259	0.84500989	3.76828874	6.99285349	16.287	0.30631053	-0.18565183
SQ_PK14	19	3.01429193	0.84218101	0.00000000	3.91152144	27.940	-2.64955589	9.43219518
SQ_PK15	19	4.78422134	1.89902457	0.00000000	7.53657747	39.693	-1.69495513	3.13185356
SQ_PK16	19	7.35860292	1.14426832	5.00000000	9.20869155	15.550	-0.11879845	-0.45040045
SQ_PK17	19	6.57884196	1.41535099	4.53872229	9.73139250	21.514	0.52666520	-0.23116123
SQ_PK18	19	3.22580298	2.71610124	0.00000000	6.73795221	84.199	-0.19196269	-1.73868518
SQ_PK19	19	2.10224724	0.50169570	0.70710678	2.94957624	23.865	-0.92337245	2.43836963
SQ_PK20	19	2.57474534	0.40435420	1.94935887	3.39116499	15.705	0.28513772	-0.69075412
SQ_PK21	19	4.24592197	0.81208874	2.81069386	5.72712843	19.126	0.13785196	-0.77343654
SQ_PK22	19	7.35292987	1.55195259	5.44977064	11.06797181	21.107	0.83670552	0.01541844
SQ_PK23	19	3.75940058	1.12228037	0.00000000	5.33853913	29.853	-1.98856545	6.82533483

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
GROUP=4								
SQ_PK1	20	6.54502319	0.91632645	5.23450093	8.78635305	14.000	0.76158725	0.32966240
SQ_PK2	20	8.00837422	1.55290637	5.26307895	10.14889157	19.391	-0.25751537	-1.17999802
SQ_PK3	20	6.30558848	1.49368984	4.38178046	9.93981891	23.688	0.82502311	0.42088654
SQ_PK4	20	0.98126128	1.27201717	0.00000000	3.37638860	129.631	0.64424411	-1.45111162
SQ_PK5	20	9.60190780	1.79278277	7.32802838	12.97690256	18.671	0.82640188	-0.48209134
SQ_PK6	20	7.74218505	2.60031793	4.07430976	14.31782106	33.586	0.90376487	0.60827242
SQ_PK7	20	2.10257431	0.35419481	1.41421356	2.72029410	16.846	-0.49665558	-0.07082781
SQ_PK8	20	10.53160312	2.81636864	6.87022561	17.51285242	26.742	0.81460597	0.60784577
SQ_PK9	20	1.49099100	1.25108199	0.00000000	3.46410162	83.909	-0.09551500	-1.58398701
SQ_PK10	20	5.22581487	0.89493247	3.49284984	6.51920241	17.125	-0.20201884	-1.10604277
SQ_PK11	20	2.44509006	0.48832107	1.73205081	3.39116499	19.971	0.43466121	-0.63666446
SQ_PK12	20	3.93796782	0.60472652	2.79284801	5.33853913	15.356	0.10636756	0.28995350
SQ_PK13	20	5.63861903	0.89205584	3.88587185	6.85565460	15.820	-0.36645684	-0.89381884
SQ_PK14	20	4.27684009	0.82948228	2.68328157	5.56776436	19.395	-0.09060558	-0.77629849
SQ_PK15	20	4.84732036	0.91391886	3.27108545	6.39531078	18.854	-0.02810338	-1.05073560
SQ_PK16	20	10.95816879	1.42848343	9.23579991	14.07835218	13.036	0.95752634	0.33138358
SQ_PK17	20	6.65858254	1.54179613	4.35889894	9.23579991	23.155	0.31040012	-1.02843454
SQ_PK18	20	4.08330198	0.90707451	1.70293864	5.25357021	22.214	-1.10577897	1.18951481
SQ_PK19	20	1.90100684	0.27371324	1.26491106	2.25831796	14.398	-0.80238668	0.19003225
SQ_PK20	20	2.84526525	0.36196199	1.97484177	3.40587727	12.722	-0.62191812	0.38478565
SQ_PK21	20	4.97483650	0.85594013	3.47850543	6.31664468	17.205	-0.09865852	-0.82833566
SQ_PK22	20	8.29068339	2.00383025	4.62601340	11.54989177	24.170	-0.13638299	-0.79806109
SQ_PK23	20	6.20199454	1.57120931	3.36154726	9.57601170	25.334	0.23310290	0.03435017
GROUP=5								
SQ_PK1	20	5.37563620	0.64280258	3.93700394	6.77495387	11.958	-0.18554104	0.79246663
SQ_PK2	20	6.89511923	1.41508071	4.63680925	10.26157883	20.523	0.82589779	1.07483205
SQ_PK3	20	6.04966604	1.18390114	4.12310563	8.98888202	19.570	1.21942141	2.17737185
SQ_PK4	20	0.26174376	0.67002765	0.00000000	2.42899156	255.986	2.53991205	5.79129300
SQ_PK5	20	9.22329968	1.78612871	5.97494770	12.42980289	19.365	0.07163332	-0.57891995
SQ_PK6	20	7.03850622	1.72426342	4.77493455	11.10405331	24.498	1.26028074	1.22666464
SQ_PK7	20	1.82020156	0.33696210	1.30384048	2.70185122	18.512	0.81817719	1.44178212
SQ_PK8	20	10.40497399	2.81658927	5.88217647	17.60113633	27.070	1.20050727	1.89461451
SQ_PK9	20	1.48849811	1.11420180	0.00000000	3.78153408	74.854	0.18131834	-0.50333796
SQ_PK10	20	5.78159813	1.14161613	3.86005181	8.72926114	19.746	0.46540706	1.12109262
SQ_PK11	20	1.85238277	0.36081495	1.41421356	2.86356421	19.478	1.12385746	1.98562082
SQ_PK12	20	4.16280845	0.66175670	3.09838668	5.71839138	15.897	0.19142130	0.14164734
SQ_PK13	20	5.21938998	0.93918678	3.78153408	7.60263112	17.994	0.60804491	0.71010585
SQ_PK14	20	2.75732700	0.73423630	1.64316767	4.30116263	26.629	0.70729538	-0.28646433
SQ_PK15	20	3.72419731	0.77483745	2.64575131	5.54075807	20.805	0.68936350	0.39193676
SQ_PK16	20	11.62212374	1.89354574	9.08845421	18.12732744	16.293	2.22676098	7.13035277
SQ_PK17	20	8.21858743	1.42526848	5.40370243	11.41052146	17.342	0.31538699	0.12927788
SQ_PK18	20	3.47964074	0.81892121	1.22474487	4.72228758	23.535	-0.86355355	1.52501949
SQ_PK19	20	2.24249449	0.54889214	1.54919334	3.57770876	24.477	0.97418633	0.52474716
SQ_PK20	20	2.29846098	0.45546637	1.54919334	3.30151480	19.816	0.29099652	0.01049505
SQ_PK21	20	4.85187825	1.33742778	2.72029410	8.15475322	27.565	0.98515048	0.80825891
SQ_PK22	20	5.66956663	1.85912096	3.21558704	9.89898985	32.791	0.92443527	0.32244294
SQ_PK23	20	11.76696785	2.30684737	7.43505212	16.19135572	19.604	-0.12782311	-0.47680806

TABLE B.2.2 (CONTINUES)

TABLE B.2.3

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
GROUP=0								
RES_PK1	19	0.01662947	0.00339178	0.01221449	0.02608923	20.396	1.12390879	2.12005001
RES_PK2	19	0.00840807	0.00195787	0.00670376	0.01488760	23.286	2.27560819	6.29827639
RES_PK3	19	0.01274447	0.00303267	0.00935629	0.01961169	23.796	1.03490360	0.32585658
RES_PK4	19	0.04921705	0.02974401	0.02572678	0.12500000	60.434	1.69165003	1.84967311
RES_PK5	19	0.01096860	0.00297496	0.00656642	0.01706776	27.123	1.00456089	0.41472580
RES_PK6	19	0.00380399	0.00268934	0.00146576	0.01151013	70.698	1.98459907	3.33099694
RES_PK7	19	0.16189737	0.04554775	0.11223345	0.31446541	28.134	2.21504789	6.62690972
RES_PK8	19	0.00453940	0.00134585	0.00290689	0.00753636	29.648	1.00872132	-0.02607644
RES_PK9	19	0.13233956	0.13065193	0.04526935	0.52083333	98.725	2.25648545	4.66750714
RES_PK10	19	0.03238013	0.00707879	0.02356823	0.04697041	21.862	0.71781085	-0.27853499
RES_PK11	19	0.07840530	0.02168653	0.04842615	0.12722646	27.660	0.61854399	-0.07777227
RES_PK12	19	0.06912551	0.01741575	0.04777831	0.10427529	25.194	0.80965572	-0.31870415
RES_PK13	19	0.03530840	0.00772437	0.02450980	0.05096840	21.877	0.35960344	-0.52815921
RES_PK14	19	0.08251833	0.02095164	0.05773672	0.12706480	25.390	0.78587219	-0.19937419
RES_PK15	19	0.05693964	0.01425029	0.03797949	0.08904720	25.027	0.59943218	-0.30256903
RES_PK16	19	0.00841012	0.00123846	0.00707014	0.01077935	14.726	0.61550499	-1.16681912
RES_PK17	19	0.00832604	0.00188533	0.00659631	0.01393534	22.644	2.00996062	4.14393262
RES_PK18	19	0.04396508	0.01564720	0.02384927	0.08319468	35.590	1.03819962	0.82182085
RES_PK19	19	0.07489002	0.03030223	0.04572474	0.13812155	40.462	1.01708170	-0.10666271
RES_PK20	19	0.07894467	0.01983151	0.04975124	0.12224939	25.121	0.68309375	0.05857201
RES_PK21	19	0.02287809	0.00594787	0.01611344	0.04068348	25.998	2.04995608	4.35380432
RES_PK22	19	0.00865062	0.00176431	0.00609310	0.01165637	20.395	0.23428940	-1.07338812
RES_PK23	19	0.05238111	0.01251399	0.03205128	0.08613264	23.890	1.15407569	2.32289253
GROUP=1								
RES_PK1	19	0.01862817	0.00645244	0.00959693	0.03144654	34.638	0.53562263	-0.83785568
RES_PK2	19	0.00935145	0.00264579	0.00577367	0.01519757	28.293	0.93301764	0.71063984
RES_PK3	19	0.01617521	0.00565195	0.00717875	0.02702703	34.942	0.45881363	-0.55483288
RES_PK4	19	0.98318935	0.99143639	0.02724796	2.00000000	100.839	0.11223090	-2.23200369
RES_PK5	19	0.00604136	0.00193428	0.00362450	0.01089325	32.017	0.92551893	0.62899667
RES_PK6	19	0.11365006	0.45682707	0.00225683	2.00000000	401.959	4.35804930	18.99479738
RES_PK7	19	0.21090461	0.06071807	0.11363636	0.31250000	28.789	0.16285565	-0.80754999
RES_PK8	19	0.00682606	0.00238697	0.00246609	0.01111111	34.968	-0.03406214	-1.00454517
RES_PK9	19	0.28979507	0.60718079	0.01459854	2.00000000	209.521	2.72330307	6.22202501
RES_PK10	19	0.05069956	0.01765374	0.02331002	0.08849558	34.820	0.36721601	-0.27499208
RES_PK11	19	0.26772773	0.12786459	0.07633588	0.55555556	47.759	0.46078823	-0.17490142
RES_PK12	19	0.08384526	0.02700300	0.03278689	0.13698630	32.206	0.01863123	-0.52778498
RES_PK13	19	0.05327901	0.01674612	0.02105263	0.07462687	31.431	-0.42984349	-1.02569719
RES_PK14	19	0.07044798	0.03042492	0.03472222	0.13888889	43.188	0.83719572	0.06731216
RES_PK15	19	0.24478726	0.61871799	0.01512859	2.00000000	252.757	2.79522031	6.49696403
RES_PK16	19	0.01114871	0.00297000	0.00725163	0.01858736	26.640	0.93796089	0.81237248
RES_PK17	19	0.01148109	0.00378240	0.00654022	0.02032520	32.945	1.14521060	1.39362932
RES_PK18	19	0.45102122	0.82198663	0.01515152	2.00000000	182.250	1.54310411	0.41654139
RES_PK19	19	0.12950241	0.06262806	0.04587156	0.28571429	48.361	0.94002359	0.57326007
RES_PK20	19	0.12722880	0.03934791	0.07246377	0.21739130	30.927	0.50346394	-0.10704175
RES_PK21	19	0.03223623	0.01659743	0.01398601	0.06896552	51.487	0.92399492	-0.23701452
RES_PK22	19	0.01562068	0.00686438	0.00690608	0.02865330	43.944	0.33576452	-1.19875773
RES_PK23	19	0.05821884	0.02524238	0.02695418	0.11111111	43.358	0.92481704	-0.06353580

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
				GROUP=2				
RES_PK1	19	0.01567152	0.00616993	0.00969932	0.03676471	39.370	2.32147432	7.30343821
RES_PK2	19	0.00812135	0.00157662	0.00619195	0.01236094	19.413	1.08691554	1.30612056
RES_PK3	19	0.01536624	0.00492749	0.00775194	0.02849003	32.067	0.63313910	1.63139745
RES_PK4	19	0.57364980	0.87633082	0.02770083	2.00000000	152.764	1.16550292	-0.72244095
RES_PK5	19	0.00657654	0.00211806	0.00380084	0.01278772	32.206	1.43507010	3.17108016
RES_PK6	19	0.00623591	0.00455909	0.00181291	0.01941748	73.110	1.86833842	3.51689916
RES_PK7	19	0.21178436	0.06181254	0.09009009	0.34482759	29.187	0.24821806	0.30378843
RES_PK8	19	0.00631826	0.00245870	0.00324570	0.01310616	38.914	1.35159784	2.30455289
RES_PK9	19	0.10698609	0.04554380	0.03448276	0.17857143	42.570	0.12766146	-0.95491429
RES_PK10	19	0.04127211	0.01367622	0.02824859	0.08130081	33.137	1.65876504	2.88013829
RES_PK11	19	0.07905632	0.02518691	0.04807692	0.12048193	31.859	0.48005744	-1.19361378
RES_PK12	19	0.07067162	0.02525297	0.02994012	0.12658228	35.733	0.63651441	-0.03465183
RES_PK13	19	0.03945848	0.01216029	0.02267574	0.06250000	30.818	0.75314060	-0.57931098
RES_PK14	19	0.07076350	0.02035643	0.04273504	0.13513514	28.767	1.75638522	4.84226575
RES_PK15	19	0.05874922	0.02104127	0.03048780	0.12195122	35.815	1.84043262	4.06935668
RES_PK16	19	0.01980276	0.00447872	0.01321004	0.03003003	22.617	0.80931484	0.92001639
RES_PK17	19	0.00633845	0.00158227	0.00319183	0.00892061	24.963	-0.15609921	-0.37015146
RES_PK18	19	0.04538082	0.06174259	0.01555210	0.29411765	136.054	4.01966429	16.89358396
RES_PK19	19	0.09526530	0.04247414	0.02967359	0.19607843	44.585	0.95113840	0.88797842
RES_PK20	19	0.20253629	0.43601637	0.05813953	2.00000000	215.278	4.33433824	18.84831632
RES_PK21	19	0.02298866	0.00560152	0.01257862	0.03205128	24.366	0.21119117	-0.76061189
RES_PK22	19	0.01133259	0.00306691	0.00577701	0.01763668	27.063	0.36336569	0.22153309
RES_PK23	19	0.06274058	0.01984461	0.03401361	0.11111111	31.630	0.77926298	0.30887031
				GROUP=3				
RES_PK1	19	0.01550840	0.00597851	0.00874891	0.03095975	38.550	1.39146969	1.63004126
RES_PK2	19	0.00922112	0.00235042	0.00591716	0.01328021	25.490	0.15864709	-1.18867877
RES_PK3	19	0.01901689	0.00588616	0.00943396	0.02890173	30.952	-0.09539239	-1.03459606
RES_PK4	19	0.90119897	0.96426933	0.03436426	2.00000000	106.998	0.33781834	-2.10715604
RES_PK5	19	0.00716733	0.00173903	0.00347947	0.01146789	24.263	0.21607114	1.43957147
RES_PK6	19	0.01315181	0.00779348	0.00271961	0.03267974	59.258	0.78601901	0.77089769
RES_PK7	19	0.34638452	0.09114664	0.21739130	0.52631579	26.314	0.47233332	-0.74590760
RES_PK8	19	0.00726426	0.00208858	0.00323520	0.01097695	28.751	0.38590357	-0.37719160
RES_PK9	19	1.03191718	0.85647262	0.12820513	2.00000000	82.998	0.28705561	-2.06739302
RES_PK10	19	0.04537441	0.01423645	0.02183406	0.07352941	31.376	0.26062511	-0.59277455
RES_PK11	19	0.27194932	0.10897516	0.06711409	0.43478261	40.072	0.14385158	-0.86180715
RES_PK12	19	0.07175666	0.02477050	0.02645503	0.13698630	34.520	0.83307869	1.98893618
RES_PK13	19	0.03921023	0.01276726	0.02024291	0.06802721	32.561	0.78023048	0.28982600
RES_PK14	19	0.19867639	0.43690231	0.06329114	2.00000000	219.907	4.33584369	18.85870487
RES_PK15	19	0.24384445	0.61897962	0.01745201	2.00000000	253.842	2.79643247	6.50135538
RES_PK16	19	0.01968581	0.00678818	0.01172333	0.03921569	34.483	1.35445664	2.50418423
RES_PK17	19	0.02582579	0.01058934	0.01050420	0.04739336	41.003	0.58381678	-0.53783135
RES_PK18	19	0.76564957	0.96882243	0.02178649	2.00000000	126.536	0.59182359	-1.85558387
RES_PK19	19	0.24585155	0.19364195	0.10869565	1.00000000	78.764	3.61001939	14.34634357
RES_PK20	19	0.14866150	0.04276420	0.08333333	0.23255814	28.766	0.38714068	-0.83556364
RES_PK21	19	0.05969924	0.02334256	0.03003003	0.11904762	39.100	0.94828549	0.71338908
RES_PK22	19	0.02044040	0.00741807	0.00813008	0.03311258	36.291	0.03618077	-1.12308519
RES_PK23	19	0.16790458	0.44411281	0.03448276	2.00000000	264.503	4.34412236	18.90938064

TABLE B.2.3 (CONTINUES)

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	C.V.	SKEWNESS	KURTOSIS
				GROUP=4				
RES_PK1	20	0.02426280	0.00618677	0.01287001	0.03584229	25.499	0.05179678	-0.74568536
RES_PK2	20	0.01741204	0.00757468	0.00966184	0.03546099	43.503	1.03796283	0.19697644
RES_PK3	20	0.02854939	0.01196972	0.01007049	0.05076142	41.926	0.35887008	-0.95783259
RES_PK4	20	1.26743092	0.92135913	0.08403361	2.00000000	72.695	-0.44809477	-2.00255260
RES_PK5	20	0.01175987	0.00376865	0.00592066	0.01845018	32.047	-0.12574142	-0.99304205
RES_PK6	20	0.02193821	0.01348128	0.00486618	0.05847953	61.451	1.09596909	1.29465874
RES_PK7	20	0.21822914	0.07763052	0.12658228	0.40000000	35.573	1.39306013	1.21938356
RES_PK8	20	0.01078836	0.00528075	0.00325521	0.02096436	48.949	0.57389546	-0.87708794
RES_PK9	20	0.84433690	0.87882471	0.08000000	2.00000000	104.085	0.61990236	-1.72579159
RES_PK10	20	0.03924841	0.01466497	0.02325581	0.07874016	37.364	1.13280403	1.24072157
RES_PK11	20	0.16893483	0.05981476	0.08333333	0.28571429	35.407	0.43948300	-0.71750830
RES_PK12	20	0.06667561	0.02122575	0.03448276	0.12048193	31.834	1.04329363	0.92799485
RES_PK13	20	0.03344368	0.01185891	0.02105263	0.06410256	35.459	1.17090818	0.92456266
RES_PK14	20	0.05943954	0.02544223	0.03174603	0.12987013	42.804	1.35859615	2.00575980
RES_PK15	20	0.04616909	0.01827829	0.02415459	0.08928571	39.590	0.90033874	0.10543587
RES_PK16	20	0.00865737	0.00199747	0.00503271	0.01165501	23.072	-0.24714967	-0.62882560
RES_PK17	20	0.02586112	0.01175808	0.01165501	0.05128205	45.466	0.68653677	-0.37812415
RES_PK18	20	0.07237822	0.05834661	0.03558719	0.29411765	80.613	3.24385847	11.74080358
RES_PK19	20	0.25553225	0.07711483	0.17857143	0.47619048	30.178	1.60580455	2.49824998
RES_PK20	20	0.12176353	0.03440791	0.08264463	0.22727273	28.258	1.70062412	3.68632257
RES_PK21	20	0.04319060	0.01592565	0.02475248	0.07936508	36.873	1.00943129	0.38424245
RES_PK22	20	0.01752281	0.01031677	0.00746826	0.04566210	58.876	1.61614072	2.40460755
RES_PK23	20	0.03131886	0.01890291	0.01084599	0.08474576	60.356	1.83496607	3.43464950
				GROUP=5				
RES_PK1	20	0.03548802	0.00931089	0.02155172	0.06250000	26.237	1.42380912	2.76252393
RES_PK2	20	0.02322034	0.00913609	0.00945180	0.04545455	39.345	0.94528003	1.04162439
RES_PK3	20	0.02956078	0.01043433	0.01230012	0.05714286	35.298	0.95191926	2.13122285
RES_PK4	20	1.74947917	0.61438793	0.15625000	2.00000000	35.118	-2.16041754	3.03629270
RES_PK5	20	0.01308023	0.00551272	0.00645161	0.02762431	42.145	1.15340634	1.12149852
RES_PK6	20	0.02287498	0.00906276	0.00807754	0.04291845	39.619	0.33495757	0.05660324
RES_PK7	20	0.27978012	0.08482741	0.12820513	0.45454545	30.319	0.52668081	0.02468421
RES_PK8	20	0.01102011	0.00569810	0.00322269	0.02849003	51.706	1.58253178	3.79130180
RES_PK9	20	0.72285483	0.77106335	0.06756757	2.00000000	106.669	1.11744044	-0.63417325
RES_PK10	20	0.03287733	0.01350786	0.01303781	0.06493506	41.086	1.09665283	0.70513934
RES_PK11	20	0.27238508	0.08132378	0.11494253	0.40000000	29.856	0.11110424	-0.66535925
RES_PK12	20	0.06009550	0.01928766	0.03012048	0.09900990	32.095	0.75171568	-0.41476169
RES_PK13	20	0.03924148	0.01353710	0.01715266	0.06756757	34.497	0.63350943	-0.21040840
RES_PK14	20	0.14453935	0.06676203	0.05263158	0.31250000	46.190	0.80574945	0.78713532
RES_PK15	20	0.07738848	0.02947573	0.03205128	0.13333333	38.088	0.58033344	-0.36580618
RES_PK16	20	0.00780748	0.00195685	0.00303859	0.01203369	25.064	-0.17617214	1.37909467
RES_PK17	20	0.01602157	0.00588351	0.00765111	0.03367003	36.722	1.35575284	3.14119884
RES_PK18	20	0.10168209	0.09741180	0.04385965	0.50000000	95.800	3.93961773	16.69274919
RES_PK19	20	0.20326823	0.07658509	0.07518797	0.34482759	37.677	0.06402129	-1.01546376
RES_PK20	20	0.18941232	0.07087544	0.08771930	0.34482759	37.419	0.93531347	0.51731764
RES_PK21	20	0.04991118	0.02523883	0.01492537	0.12658228	50.567	1.35213040	3.37880828
RES_PK22	20	0.03966730	0.02301534	0.01015332	0.09225092	58.021	0.95007093	0.53940848
RES_PK2[illegible]	20	0.00813205	0.00366390	0.00380720	0.01792757	45.055	1.27764606	1.20659167

TABLE B.2.3 (CONTINUES)

SUMMARY OF THE SHAPIRO-WILK STATISTICS

ORIGINAL DATABEFORE TRANSFORMATIONS

	P - VALUES				
VARIABLE	GROUP1	GROUP2	GROUP3	GROUP4	GROUP5
1	0.43	0.55	0.60	0.09	0.82
2	0.61	0.72	0.15	0.36	0.01
3	0.21	0.01	0.02	0.02	0.01
4	0.01	0.13	0.01	0.01	0.01
5	0.54	0.67	0.01	0.01	0.60
6	0.02	0.04	0.01	0.01	0.01
7	0.06	0.01	0.48	0.76	0.02
8	0.01	0.44	0.01	0.01	0.01
9	0.01	0.01	0.01	0.02	0.01
10	0.07	0.30	0.14	0.25	0.05
11	0.01	0.21	0.01	0.14	0.01
12	0.01	0.07	0.01	0.39	0.26
13	0.01	0.32	0.57	0.47	0.14
14	0.21	0.88	0.26	0.44	0.02
15	0.09	0.49	0.48	0.48	0.05
16	0.86	0.45	0.69	0.01	0.01
17	0.34	0.01	0.22	0.09	0.41
18	0.42	0.32	0.01	0.72	0.47
19	0.08	0.01	0.74	0.47	0.01
20	0.16	0.13	0.46	0.95	0.41
21	0.15	0.06	0.60	0.53	0.01
22	0.05	0.03	0.01	0.55	0.01
23	0.58	0.76	0.51	0.36	0.79

SUMMARY OF THE SHAPIRO-WILK STATISTICS

LOG TRANSFORMATIONS ON THE ORIGINAL DATA

	P - VALUES				
VARIABLE	GROUP1	GROUP2	GROUP3	GROUP4	GROUP5
1	0.61	0.08	0.34	0.65	0.55
2	0.53	0.43	0.34	0.23	0.74
3	0.93	0.29	0.24	0.58	0.09
4	0.01	0.01	0.01	0.01	0.01
5	0.55	0.67	0.37	0.09	0.81
6	0.01	0.62	0.40	0.95	0.24
7	0.45	0.50	0.57	0.15	0.50
8	0.29	0.80	0.16	0.58	0.42
9	0.02	0.27	0.01	0.01	0.07
10	0.81	0.04	0.77	0.27	0.60
11	0.63	0.19	0.08	0.62	0.32
12	0.47	0.90	0.23	0.58	0.41
13	0.04	0.26	0.97	0.30	0.94
14	0.50	0.52	0.01	0.47	0.62
15	0.01	0.33	0.01	0.53	0.80
16	0.82	0.52	0.63	0.17	0.02
17	0.41	0.36	0.92	0.51	0.91
18	0.01	0.01	0.01	0.01	0.01
19	0.96	0.85	0.01	0.07	0.36
20	0.56	0.01	0.77	0.34	0.89
21	0.37	0.36	0.82	0.49	0.80
22	0.25	0.53	0.31	0.40	0.62
23	0.48	0.97	0.01	0.69	0.48

SUMMARY OF THE SHAPIRO-WILK STATISTICS

ORIGINAL TRIMMED DATA (7 DELETEDOBS)

	P - VALUES				
VARIABLE	GROUP1	GROUP2	GROUP3	GROUP4	GROUP5
1	0.43	0.57	0.51	0.05	0.62
2	0.63	0.81	0.17	0.43	0.86
3	0.56	0.03	0.01	0.01	0.51
4	0.01	0.08	0.03	0.01	0.01
5	0.43	0.69	0.33	0.01	0.93
6	0.08	0.07	0.01	0.06	0.06
7	0.23	0.44	0.38	0.76	0.49
8	0.05	0.70	0.08	0.20	0.62
9	0.40	0.01	0.01	0.01	0.03
10	0.03	0.20	0.31	0.15	0.26
11	0.01	0.31	0.25	0.19	0.30
12	0.01	0.05	0.08	0.43	0.27
13	0.01	0.26	0.69	0.67	0.62
14	0.31	0.67	0.24	0.47	0.03
15	0.07	0.58	0.22	0.58	0.22
16	0.95	0.32	0.58	0.02	0.51
17	0.23	0.21	0.54	0.09	0.47
18	0.34	0.33	0.01	0.76	0.53
19	0.39	0.47	0.42	0.44	0.09
20	0.20	0.02	0.32	0.97	0.50
21	0.20	0.34	0.65	0.72	0.23
22	0.07	0.01	0.01	0.59	0.06
23	0.48	0.63	0.47	0.24	0.80

TABLE B.2.

STEPWISE SELECTION: SUMMARY

STEP	VARIABLE ENTERED	VARIABLE REMOVED	NUMBER IN	PARTIAL R**2	F STATISTIC	PROB > F	WILKS' LAMBDA	PROB < LAMBDA	AVERAGE SQUARED CANONICAL CORRELATION	PROB > ASCC
1	LPK23		1	0.7247	60.543	0.0001	0.27530673	0.0001	0.18117332	0.0001
2	LPK17		2	0.7603	72.150	0.0001	0.06599843	0.0001	0.35002743	0.0001
3	LPK2		3	0.7329	61.753	0.0001	0.01762510	0.0001	0.43048769	0.0001
4	LPK11		4	0.6670	44.572	0.0001	0.00586867	0.0	0.59173095	0.0001
5	LPK16		5	0.4880	20.971	0.0001	0.00300457	0.0	0.69500447	0.0
6	LPK22		6	0.3977	14.359	0.0001	0.00180979	0.0	0.74074109	0.0
7	LPK13		7	0.3573	11.953	0.0001	0.00116313	0.0	0.77603654	0.0
8	LPK21		8	0.3302	10.475	0.0001	0.00077908	0.0	0.80211327	0.0
9	LPK7		9	0.1911	4.962	0.0012	0.00063019	0.0	0.81584376	0.0
10	LPK8		10	0.2622	7.375	0.0001	0.00046495	0.0	0.82993454	0.0
11	LPK12		11	0.2376	6.390	0.0002	0.00035446	0.0	0.84099208	0.0
12	LPK6		12	0.1396	3.285	0.0151	0.00030498	0.0	0.84755627	0.0
13	LPK14		13	0.1218	2.774	0.0326	0.00026784	0.0	0.85410886	0.0
14	LPK1		14	0.1350	3.082	0.0206	0.00023169	0.0	0.86184334	0.0
15	LPK10		15	0.1352	3.049	0.0217	0.00020036	0.0	0.86633348	0.0
16	LPK4		16	0.0889	1.878	0.1228	0.00018255	0.0	0.86792370	0.0
17	LPK18		17	0.1107	2.364	0.0604	0.00016235	0.0	0.87199850	0.0
18	LPK19		18	0.0915	1.889	0.1212	0.00014749	0.0	0.87540002	0.0

TABLE B.3.1 B_14

FORWARD SELECTION: SUMMARY

STEP	VARIABLE ENTERED	NUMBER IN	PARTIAL R**2	F STATISTIC	PROB > F	WILKS' LAMBDA	PROB < LAMBDA	AVERAGE SQUARED CANONICAL CORRELATION	PROB > ASCC
1	LPK23	1	0.7247	60.543	0.0001	0.27530673	0.0001	0.18117332	0.0001
2	LPK17	2	0.7603	72.150	0.0001	0.06599843	0.0001	0.35002743	0.0001
3	LPK2	3	0.7329	61.753	0.0001	0.01762510	0.0001	0.43048769	0.0001
4	LPK11	4	0.6670	44.572	0.0001	0.00586867	0.0	0.59173095	0.0001
5	LPK16	5	0.4880	20.971	0.0001	0.00300457	0.0	0.69500447	0.0
6	LPK22	6	0.3977	14.359	0.0001	0.00180979	0.0	0.74074109	0.0
7	LPK13	7	0.3573	11.953	0.0001	0.00116313	0.0	0.77603654	0.0
8	LPK21	8	0.3302	10.475	0.0001	0.00077908	0.0	0.80211327	0.0
9	LPK7	9	0.1911	4.962	0.0012	0.00063019	0.0	0.81584376	0.0
10	LPK8	10	0.2622	7.375	0.0001	0.00046495	0.0	0.82993454	0.0
11	LPK12	11	0.2376	6.390	0.0002	0.00035446	0.0	0.84099208	0.0
12	LPK6	12	0.1396	3.285	0.0151	0.00030498	0.0	0.84755627	0.0
13	LPK14	13	0.1218	2.774	0.0326	0.00026784	0.0	0.85410886	0.0
14	LPK1	14	0.1350	3.082	0.0206	0.00023169	0.0	0.86184334	0.0
15	LPK10	15	0.1352	3.049	0.0217	0.00020036	0.0	0.86633348	0.0
16	LPK4	16	0.0889	1.878	0.1228	0.00018255	0.0	0.86792370	0.0
17	LPK18	17	0.1107	2.364	0.0604	0.00016235	0.0	0.87199850	0.0
18	LPK19	18	0.0915	1.889	0.1212	0.00014749	0.0	0.87540002	0.0

TABLE B.3.2

KWARD ELIMINATION: SUMMARY

STEP	VARIABLE REMOVED	NUMBER IN	PARTIAL R**2	F STATISTIC	PROB > F	WILKS' LAMBDA	PROB < LAMBDA	AVERAGE SQUARED CANONICAL CORRELATION	PROB > ASCC
0		23	.	.	.	0.00011624	0.0	0.88221720	0.0
1	LPK9	22	0.0256	0.459	0.7655	0.00011929	0.0	0.88125349	0.0
2	LPK15	21	0.0328	0.603	0.6618	0.00012334	0.0	0.88031450	0.0
3	LPK5	20	0.0481	0.910	0.4629	0.00012958	0.0	0.87847815	0.0
4	LPK3	19	0.0490	0.940	0.4460	0.00013625	0.0	0.87675950	0.0
5	LPK20	18	0.0762	1.526	0.2034	0.00014749	0.0	0.87540002	0.0

TABLE B.3.3

B_15

PWISE SELECTION: SUMMARY

STEP	VARIABLE ENTERED	VARIABLE REMOVED	NUMBER IN	PARTIAL R**2	F STATISTIC	PROB > F	WILKS' LAMBDA	PROB < LAMBDA	AVERAGE SQUARED CANONICAL CORRELATION	PROB > ASCC
1	LPK23		1	0.7757	38.902	0.0001	0.22431848	0.0001	0.19392038	0.0001
2	LPK17		2	0.8013	44.360	0.0001	0.04457191	0.0001	0.37471649	0.0001
3	LPK2		3	0.8377	55.477	0.0001	0.00723491	0.0001	0.46419617	0.0001
4	LPK11		4	0.7234	27.467	0.0001	0.00200086	0.0001	0.63067865	0.0001
5	LPK22		5	0.5584	12.961	0.0001	0.00088358	0.0001	0.70263555	0.0001
6	LPK7		6	0.4801	9.236	0.0001	0.00045933	0.0001	0.77322853	0.0001
7	LPK21		7	0.4057	6.656	0.0003	0.00027299	0.0001	0.81034772	0.0001
8	LPK13		8	0.4337	7.275	0.0002	0.00015460	0.0001	0.83990647	0.0001
9	LPK16		9	0.3546	5.081	0.0023	0.00009979	0.0001	0.86902077	0.0001
10	LPK8		10	0.2944	3.755	0.0118	0.00007041	0.0001	0.87705645	0.0001
11	LPK12		11	0.3026	3.796	0.0115	0.00004911	0.0001	0.88787971	0.0001
12	LPK10		12	0.3080	3.784	0.0119	0.00003398	0.0001	0.89459989	0.0001
13	LPK1		13	0.2451	2.679	0.0487	0.00002565	0.0001	0.89965339	0.0001

TABLE B.3.4

STEPWISE SELECTION: SUMMARY

STEP	VARIABLE ENTERED	VARIABLE REMOVED	NUMBER IN	PARTIAL R**2	F STATISTIC	PROB > F	WILKS' LAMBDA	PROB < LAMBDA	AVERAGE SQUARED CANONICAL CORRELATION	PROB > ASCC
1	LPK16		1	0.7053	28.123	0.0001	0.29468645	0.0001	0.17632839	0.0001
2	LPK17		2	0.6514	21.493	0.0001	0.10271638	0.0001	0.32096966	0.0001
3	LPK2		3	0.7285	30.187	0.0001	0.02788739	0.0001	0.44769036	0.0001
4	LPK14		4	0.7054	26.336	0.0001	0.00821625	0.0001	0.59986785	0.0001
5	LPK10		5	0.7450	31.399	0.0001	0.00209554	0.0001	0.70104700	0.0001
6	LPK11		6	0.5090	10.886	0.0001	0.00102887	0.0001	0.78098838	0.0001
7	LPK4		7	0.3189	4.799	0.0029	0.00070079	0.0001	0.79370048	0.0001
8	LPK9		8	0.3433	5.228	0.0017	0.00046020	0.0001	0.81632145	0.0001
9	LPK8		9	0.2698	3.603	0.0136	0.00033603	0.0001	0.82653539	0.0001
10	LPK22		10	0.2645	3.416	0.0176	0.00024715	0.0001	0.83545478	0.0001
11	LPK23		11	0.2579	3.214	0.0232	0.00018342	0.0001	0.84542589	0.0001
12		LPK16	10	0.1103	1.147	0.3497	0.00020617	0.0001	0.83958045	0.0001
13	LPK5		11	0.1972	2.272	0.0799	0.00016551	0.0001	0.85008966	0.0001
14	LPK21		12	0.2610	3.178	0.0246	0.00012232	0.0001	0.86260924	0.0001
15	LPK20		13	0.4005	5.845	0.0010	0.00007333	0.0001	0.87843976	0.0001
16	LPK18		14	0.3243	4.079	0.0083	0.00004955	0.0001	0.88669026	0.0001
17		LPK9	13	0.1569	1.581	0.2016	0.00005877	0.0001	0.88271752	0.0001
18	LPK12		14	0.2472	2.791	0.0417	0.00004425	0.0001	0.89484671	0.0001

TABLE B.3.5

STEPWISE SELECTION: SUMMARY

STEP	VARIABLE ENTERED	VARIABLE REMOVED	NUMBER IN	PARTIAL R**2	F STATISTIC	PROB > F	WILKS' LAMBDA	PROB < LAMBDA	AVERAGE SQUARED CANONICAL CORRELATION	PROB > ASCC
1	LPK23		1	0.8190	55.437	0.0001	0.18098095	0.0001	0.20475476	0.0001
2	LPK2		2	0.8157	53.105	0.0001	0.03335819	0.0001	0.27907869	0.0001
3	LPK17		3	0.7534	35.903	0.0001	0.00822528	0.0001	0.46309732	0.0001
4	LPK14		4	0.7349	31.882	0.0001	0.00218040	0.0001	0.61959388	0.0001
5	LPK11		5	0.6089	17.518	0.0001	0.00085265	0.0001	0.74895770	0.0001
6	LPK4		6	0.3739	6.570	0.0003	0.00053381	0.0001	0.78606645	0.0001
7	LPK16		7	0.2978	4.560	0.0037	0.00037482	0.0001	0.81086258	0.0001
8	LPK5		8	0.1973	2.581	0.0509	0.00030087	0.0001	0.82523086	0.0001
9	LPK8		9	0.2169	2.839	0.0363	0.00023561	0.0001	0.83630519	0.0001
10	LPK22		10	0.2236	2.880	0.0347	0.00018293	0.0001	0.84771935	0.0001
11	LPK21		11	0.2326	2.956	0.0317	0.00014038	0.0001	0.86096149	0.0001
12	LPK15		12	0.2388	2.980	0.0310	0.00010685	0.0001	0.87189679	0.0001
13	LPK12		13	0.1814	2.050	0.1073	0.00008747	0.0001	0.87978257	0.0001

TABLE B.3.6

APPENDIX C

TABLE C.1.1 CANONICAL DISCRIMINANT ANALYSIS

ALL VARIABLES LOG TRANSFORMED STANDARDIZED CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
LPK1	-0.0186	-0.4700	-0.5393
LPK2	3.0488	-1.3747	-1.3623
LPK3	-0.3211	0.5826	0.6900
LPK4	0.5145	-0.2859	0.1191
LPK5	0.1815	0.4301	0.0578
LPK6	0.0583	-0.3773	-0.0510
LPK7	0.5482	0.7168	0.7302
LPK8	-0.3964	-1.3461	-1.3756
LPK9	-0.0588	0.2139	0.0025
LPK10	-1.2824	0.5434	0.2569
LPK11	1.1250	0.8091	1.9239
LPK12	-0.4273	-0.6742	-1.6088
LPK13	0.1348	-0.8052	0.6573
LPK14	0.1790	-0.1227	0.3248
LPK15	-0.1958	-0.0605	0.0313
LPK16	-0.4297	-0.0645	0.1750
LPK17	-1.6418	2.0381	-1.4260
LPK18	0.2302	-0.0852	0.4092
LPK19	-0.2109	0.0657	-0.6007
LPK20	0.3464	-0.1170	0.1654
LPK21	0.4069	1.4974	-0.5056
LPK22	-0.1837	-0.2751	1.8822
LPK23	-1.5137	-0.7496	-0.0884

RAW CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
LPK1	-0.041796911	-1.055175525	-1.210889307
LPK2	6.077756505	-2.740398043	-2.715779589
LPK3	-0.706486537	1.281725874	1.517897518
LPK4	0.308426220	-0.171371432	0.071417758
LPK5	0.408514991	0.968160978	0.130190029
LPK6	0.060708917	-0.393128979	-0.053116695
LPK7	1.531573543	2.002582157	2.040203104
LPK8	-0.848317493	-2.880417945	-2.943537278
LPK9	-0.043897252	0.159761964	0.001882760
LPK10	-3.442862591	1.458955803	0.689666034
LPK11	1.831183672	1.316883886	3.131474832
LPK12	-1.223065797	-1.929643629	-4.604390482
LPK13	0.376159310	-2.246396598	1.833645784
LPK14	0.306817711	-0.210298998	0.556787838
LPK15	-0.223388617	-0.069003929	0.035752196
LPK16	-0.901942908	-0.135360876	0.367336758
LPK17	-2.630667672	3.265678762	-2.284870754
LPK18	0.171842196	-0.063634179	0.305512304
LPK19	-0.366562641	0.114218221	-1.043898236
LPK20	0.777006054	-0.262559176	0.371072243
LPK21	0.788431516	2.901304446	-0.979598890
LPK22	-0.306960033	-0.459573277	3.144331097
LPK23	-1.519590163	-0.752479648	-0.088734382

TABLE C.1.1

TABLE C.1.2 - SELECTED VARIABLES LOG TRANSFORMED

STANDARDIZED CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
LPK1	0.0288	-0.1952	-0.4180
LPK2	2.7955	-0.9079	-1.0243
LPK4	0.4580	-0.2860	0.1412
LPK6	-0.0791	-0.2771	0.0773
LPK7	0.5805	0.7860	0.8164
LPK8	-0.3655	-0.9913	-1.1619
LPK10	-1.2879	0.7688	0.4001
LPK11	1.1654	0.7541	1.8935
LPK12	-0.2072	-0.9696	-1.6386
LPK13	-0.0399	-0.7089	0.6106
LPK14	0.0351	-0.2766	0.3200
LPK16	-0.4224	-0.2899	0.1725
LPK17	-1.5921	2.1191	-1.2629
LPK18	0.2056	-0.1789	0.3943
LPK19	-0.1291	0.1236	-0.5610
LPK21	0.4058	1.6941	-0.5611
LPK22	-0.0519	-0.6111	1.8788
LPK23	-1.4984	-0.5369	-0.0863

CANONICAL DISCRIMINANT ANALYSIS

RAW CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
LPK1	0.064731451	-0.438332861	-0.938522410
LPK2	5.572766746	-1.809909813	-2.042018896
LPK4	0.274522792	-0.171414393	0.084646979
LPK6	-0.082396855	-0.288766451	0.080545123
LPK7	1.621830806	2.196065484	2.280860211
LPK8	-0.782133872	-2.121180025	-2.486290546
LPK10	-3.457601794	2.063927312	1.074268447
LPK11	1.896942737	1.227449878	3.081984630
LPK12	-0.593089345	-2.774990189	-4.689894653
LPK13	-0.111198097	-1.977736132	-1.703552205
LPK14	0.060221967	-0.474105387	0.548599635
LPK16	-0.886640618	-0.608620295	0.362065100
LPK17	-2.551072927	3.395551740	-2.023574245
LPK18	0.153487700	-0.133575788	0.294419676
LPK19	-0.224362607	0.214741175	-0.974931622
LPK21	0.786347143	3.282423652	-1.087171261
LPK22	-0.086735152	-1.020860864	3.138752071
LPK23	-1.504201779	-0.539037199	-0.086639873

TABLE C.1.2

TABLE C.2.1

ALL VARIABLES RANKED

CANONICAL DISCRIMINANT ANALYSIS

STANDARDIZED CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
RPK1	-0.1349	-0.3941	-0.2997
RPK2	2.6693	-0.8272	-2.2257
RPK3	-1.2799	-0.2014	1.5696
RPK4	0.2345	-0.0864	0.1866
RPK5	-0.0719	0.5399	-0.1714
RPK6	0.7131	-0.3882	-0.0527
RPK7	0.2582	0.6491	0.5501
RPK8	0.3418	-0.5074	-1.6291
RPK9	-0.1709	0.3409	-0.1013
RPK10	0.1484	0.4614	-0.5141
RPK11	0.6205	0.4936	1.9517
RPK12	-0.8068	0.2490	-1.0901
RPK13	0.2748	-1.1912	0.6625
RPK14	0.4028	-0.5515	0.7263
RPK15	0.0215	-0.5112	-0.2015
RPK16	-0.9137	0.0336	-0.2339
RPK17	-0.6937	2.0113	-1.0437
RPK18	0.1237	-0.2016	0.2279
RPK19	0.1213	0.0379	-0.4760
RPK20	0.3718	-0.0710	0.7167
RPK21	-0.0177	1.3669	0.2642
RPK22	-0.7994	-0.1460	0.7915
RPK23	-2.1165	-0.5993	-0.3403

CANONICAL DISCRIMINANT ANALYSIS

RAW CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
RPK1	-.0047920256	-.0140030871	-.0106479219
RPK2	0.0948396018	-.0293920782	-.0790807176
RPK3	-.0454771818	-.0071568797	0.0557698495
RPK4	0.0090108539	-.0033188294	0.0071695219
RPK5	-.0025553151	0.0191818391	-.0060909695
RPK6	0.0253351955	-.0137921732	-.0018734901
RPK7	0.0091767243	0.0230706635	0.0195531419
RPK8	0.0121445506	-.0180281256	-.0578811045
RPK9	-.0061090835	0.0121851962	-.0036197045
RPK10	0.0052724313	0.0163934972	-.0182657914
RPK11	0.0220506173	0.0175404413	0.0693563721
RPK12	-.0286668236	0.0088461637	-.0387360218
RPK13	0.0097650236	-.0423245179	0.0235389374
RPK14	0.0143123566	-.0195990023	0.0258108707
RPK15	0.0007638788	-.0181652818	-.0071587006
RPK16	-.0324631715	0.0011944010	-.0083114092
RPK17	-.0246473101	0.0714608709	-.0370813411
RPK18	0.0043972896	-.0071665275	0.0081029220
RPK19	0.0043097204	0.0013455785	-.0169186284
RPK20	0.0132137663	-.0025231765	0.0254700630
RPK21	-.0006301055	0.0485691838	0.0093883512
RPK22	-.0284020623	-.0051859723	0.0281233141
RPK23	-.0752014099	-.0212944546	-.0120909099

TABLE C.2.1

TABLE C.2.2 CANONICAL DISCRIMINANT ANALYSIS

SELECTED VARIABLES RANKED STANDARDIZED CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
RPK1	-0.1943	-0.0545	-0.1907
RPK2	1.9629	-1.0236	-1.4311
RPK4	0.2564	-0.1705	0.1590
RPK6	0.2716	-0.4199	0.0959
RPK7	0.2801	0.7740	0.6863
RPK8	-0.1585	-0.3907	-1.2823
RPK10	-0.1876	0.4000	-0.1627
RPK11	0.8882	0.3295	1.5457
RPK12	-0.4169	0.1818	-1.0095
RPK13	0.1895	-1.1827	0.3098
RPK14	0.3225	-0.8671	0.8913
RPK16	-0.7830	-0.1748	0.0784
RPK17	-0.8239	2.0785	-0.7096
RPK18	0.1856	-0.3255	0.2227
RPK19	0.1519	0.2421	-0.2884
RPK21	0.0504	1.6722	0.1009
RPK22	-0.5311	-0.3784	1.2520
RPK23	-2.1078	-0.4222	-0.4103

RAW CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3
RPK1	-.0069031515	-.0019355564	-.0067746932
RPK2	0.0697417917	-.0363697871	-.0508474231
RPK4	0.0098523795	-.0065544983	0.0061098093
RPK6	0.0096501876	-.0149188942	0.0034069257
RPK7	0.0099557097	0.0275123686	0.0243957195
RPK8	-.0056328769	-.0138806773	-.0455589601
RPK10	-.0066667245	0.0142137034	-.0057823883
RPK11	0.0315640911	0.0117080970	0.0549294621
RPK12	-.0148130735	0.0064581536	-.0358700446
RPK13	0.0067335252	-.0420222305	0.0110086486
RPK14	0.0114585244	-.0308131088	0.0316734592
RPK16	-.0278202012	-.0062115459	0.0027844027
RPK17	-.0292741169	0.0738496755	-.0252137750
RPK18	0.0066003391	-.0115725933	0.0079166247
RPK19	0.0053979254	0.0086052422	-.0102499648
RPK21	0.0017922976	0.0594150420	0.0035851284
RPK22	-.0188686997	-.0134452643	0.0444836613
RPK23	-.0748911310	-.0150008192	-.0145788988

APPENDIX D

OTHER MULTIVARIATE APPROACHES APPLIED

D.1 Cluster analysis

Items of similar nature are grouped into clusters by using cluster analysis. Originally each item is a single cluster, the two closest clusters are then merged into a new cluster. This process continues until the required number of clusters is reached.

Let X_{ij}, $i=1,..,G$ and $j=1,..,n_i$, be the j-th item from the i-th population with no initial group membership defined. The aim of cluster analysis is to group or classify the n_i items into G homogeneous clusters where G is unknown and $G \leq n$.

Numerous clustering methods were developed in several different fields. The data structures of items to be clustered also take many forms. The most common forms are :

- a similarity or square matrix such as the correlation matrix.

- a coordinate matrix, in which the rows are observations and the columns are variables, as in the usual multivariate data set. The observations, or the variables or both may be clustered.

In applying cluster analysis various clustering methods were used. The two methods that proved the most satisfactory for the problem at hand were the two stage density linkage method and the average linkage method.

D.1.1 Results when using the two stage density linkage method:

The percentage of cultivar items clustered into similar populations were:

TABLE D.1

Cultivar 1	68.42 %	into cluster 3
Cultivar 2	73.68 %	into cluster 4
Cultivar 3	89.5 %	into cluster 3
Cultivar 4	80.00 %	into cluster 2
Cultivar 5	70.00 %	into cluster 1

From these results it appears that the majority of items of both cultivar 1 and cultivar 3 were clustered into the same cluster. The final result was that most of the items were clustered into only 4 clusters. The fifth cluster contained only 1 item from cultivar 5.

The graphical representation obtained by using the two stage density linkage method can be seen in GRAPH D.2.

D.1.2 Results when using the average linkage method:

The percentage of cultivar items clustered into similar populations were:

TABLE D.2

Cultivar 1	52.63 %	into cluster 1
Cultivar 2	52.63 %	into cluster 2
Cultivar 3	63.16 %	into cluster 1
Cultivar 4	90.00 %	into cluster 1
Cultivar 5	90.00 %	into cluster 1

These results show that the majority of items were clustered into the same cluster. The final result was that most of the items were clustered into only 2 clusters. This result is unsatisfactory for distinction between the cultivars.

The graphical representation obtained by using the average linkage method can be seen in GRAPH D.1.

Cluster analysis was unsuccessful in clustering the items into the correct cultivar populations. It would thus be impossible to classify an item of unknown origin into the correct cultivar population.

D.2 Principal component analysis

Principal component analysis was first formulated by Pearson (1901) as finding "lines and planes of closest fit to the system of the points in space". Principal component analysis is an explanatory technique for examining relationships among several quantitative measurements. This technique is used for summarizing data by finding linear relationships among the items.

Given a data set with m measurements, m principal components can be computed. Each component is a linear combination of the original variables, with coefficients equal to the eigenvectors of the correlation or covariance matrix. The eigenvectors are taken with unit-norm. The principal components are sorted in descending order of the eigenvalues, which are equal to the variances of the components.

The first j components give a least-squares solution to the model $Y = XB + E$,

where Y is an n x m matrix for the centered observed measurements; X is the n x j matrix of scores on the first j principal components; B is the j x m matrix-of eigenvectors; E is an n x m matrix of residuals.

Principal component analysis can be used with a well selected set of items and measurements to build a model for predicting the population origin of newly measured data items.

Tables D.1 and D.2 reveal the principal components computed on the original data and on the selected variables (using the subset selected in chapter 6). By examining the cumulative proportion of the variance explained by the first three principal components it appears that the selected variables' proportion is marginally higher. Thus when using the selected variables slightly more variation is explained by the first three dimensions.

The graphical representations are given in GRAPHS D.3 and D.4. The graphical displays show that principal component analysis does not calculate clear distinctive populations. This makes the classification of items of unknown origin difficult.

D.3 Correspondence analysis

Correspondence analysis is an exploratory multivariate technique whereby a simultaneous graphical image is made of the points representing the rows and the points representing the columns of a matrix. The main purpose is to provide a visual method for comparing row and column proportions.

The original concept of correspondence analysis was motivated by contingency data and the metrics involved are methods used in the calculation of contingency data. Nevertheless, correspondence analysis is a more fundamental way of showing the structure of a data matrix. In a paper by Greenacre (1981) he clearly pointed out on the first page that , " ... although correspondence analysis is primarily a technique of displaying rows and columns of a contingency table..., the technique ... may be extended with suitable care, to the display of a wide range of data matrices."

Correspondence analysis was applied to the problem at hand, as the examples in the above mentioned paper are of similar nature.

The initial data table is transformed into a table of row profiles by calculating the sum of each row and then dividing each element in the row by this sum for the

row. The sum of the elements in a row profile, which are the co-ordinates of the row profile, is now equal to 1.0 and the initial sum for the row is kept as the weight of the row. No information is lost during this transformation. An average row profile is calculated from the data table, which is the profile of the bottom margin of the table; it defines the coordinates of the center of gravity of the row profiles space. The distance between rows is calculated using the formula of the distributional distance between row profiles, which is the chi-square distance between two profiles using the average row profile as reference. The same transformation can be applied to the columns.

In order to calculate the factor space for the data table a symmetric matrix is calculated from the frequency table deduced from the data table and from the two diagonal matrices formed respectively by the row weights and the column weights. Eigenvalues and eigenvectors are calculated by diagonalizing this symmetric matrix.

Correspondence analysis summarizes the row and column proportions by replacing them with a smaller set of co-ordinates. These co-ordinates are computed so that each successive co-ordinate axis accounts for a decreasing portion of the total association between the rows and columns as represented by the Pearson´s chi-square statistic. The first co-ordinate accounts for the largest part of

the association, the second for the second largest part, and so on.

TABLE D.5 shows the spread of the cloud of points representing the rows which is quantified by the moment of inertia of the points in relation to the row center of gravity. The principal inertia's and the percentage of the total inertia is given. The histogram of the eigenvalues is useful since it provides a visual representation of the relative importance of each principal axis. Thus the first three principal axes account for 75.51 percent of the total inertia when all variables are used.

TABLE D.8 shows that when using selected variables, the first three principal axes account for 79.11 percent of the total inertia.

TABLES D.6 & D.7 represent the row and column contributions (when using all variables) respectively. TABLES D.9 & D.10 represent similar contributions calculated when only selected variables are used.

The "I" represents the number of rows (different items). The "NAME" is the cultivar origin. As an example the result obtained from the first row(item) of TABLE D.6 is examined.

For each element the so-called quality of the representation of the element in the subspace of the factorial axis is given by "QLT". This quantity is either the squared correlation of this element with the subspace or the squared cosine of the angle that it makes with the subspace. It is calculated by summing the correlations for all the factors. Thus, the first row has a quality of display in three dimensions of 0.359 and mass "MAS" of 0.010 which is scaled so as to sum to 1000.

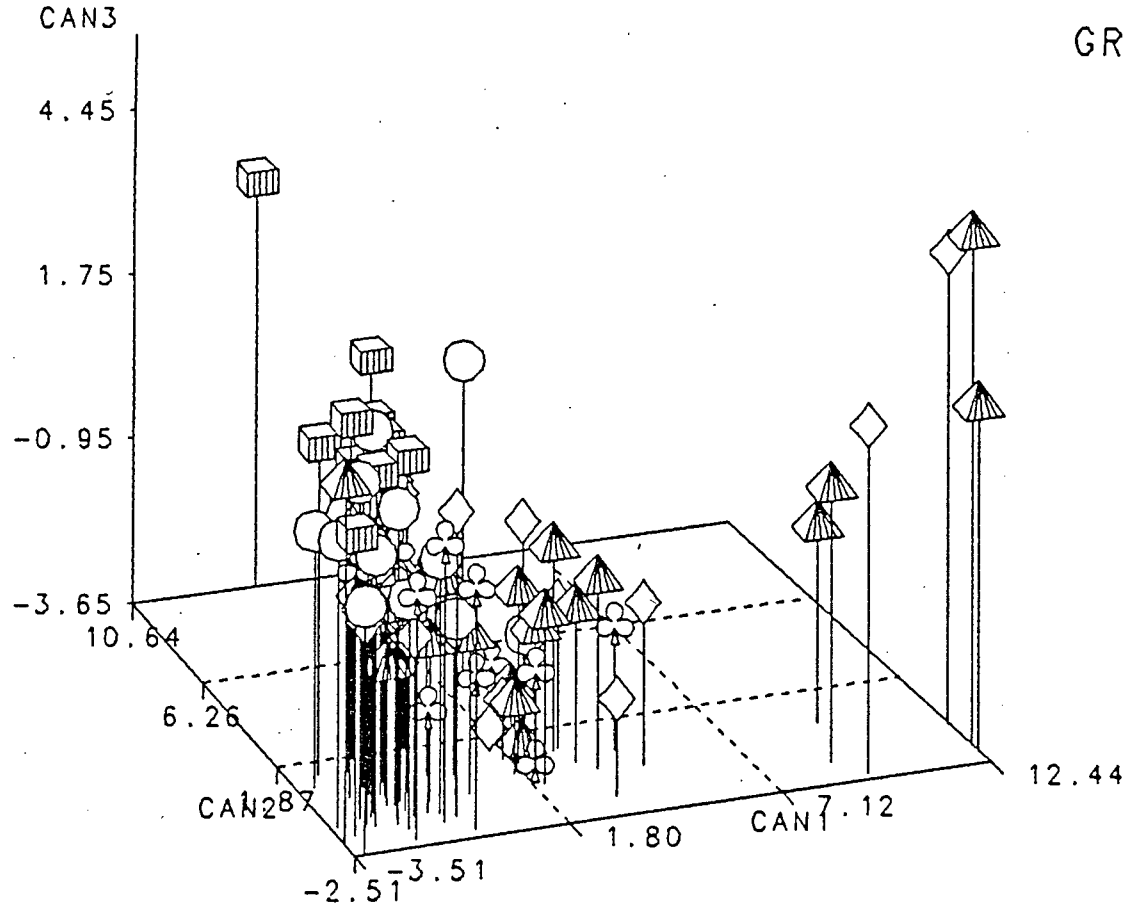
The inertia relative to the total inertia of the cloud for each item is given by "INR" = 0.004 * total inertia.

The principal co-ordinate on axis one is -0.138 with a squared cosine (or correlation) of 0.204. The contribution of the item to the first axis is 0.2 percent. The principal co-ordinates of the second and third axes follow.

Graphical displays of this technique is given by GRAPHS D.5 and D.6. It can be seen that this technique did extremely well when grouping the items into distinct cultivars and is extremely useful as an exploratory technique.

CLUSTER ANALYSIS BY THE AVERAGE LINKAGE METHOD
ALL VARIABLES USED FROM 5 GRAPE CULTIVARS

GRAPH D.1



CLUSTER ANALYSIS BY THE TWO-STAGE DENSITY LINKAGE
ALL VARIABLES USED FROM 5 GRAPE CULTIVARS

GRAPH D.2

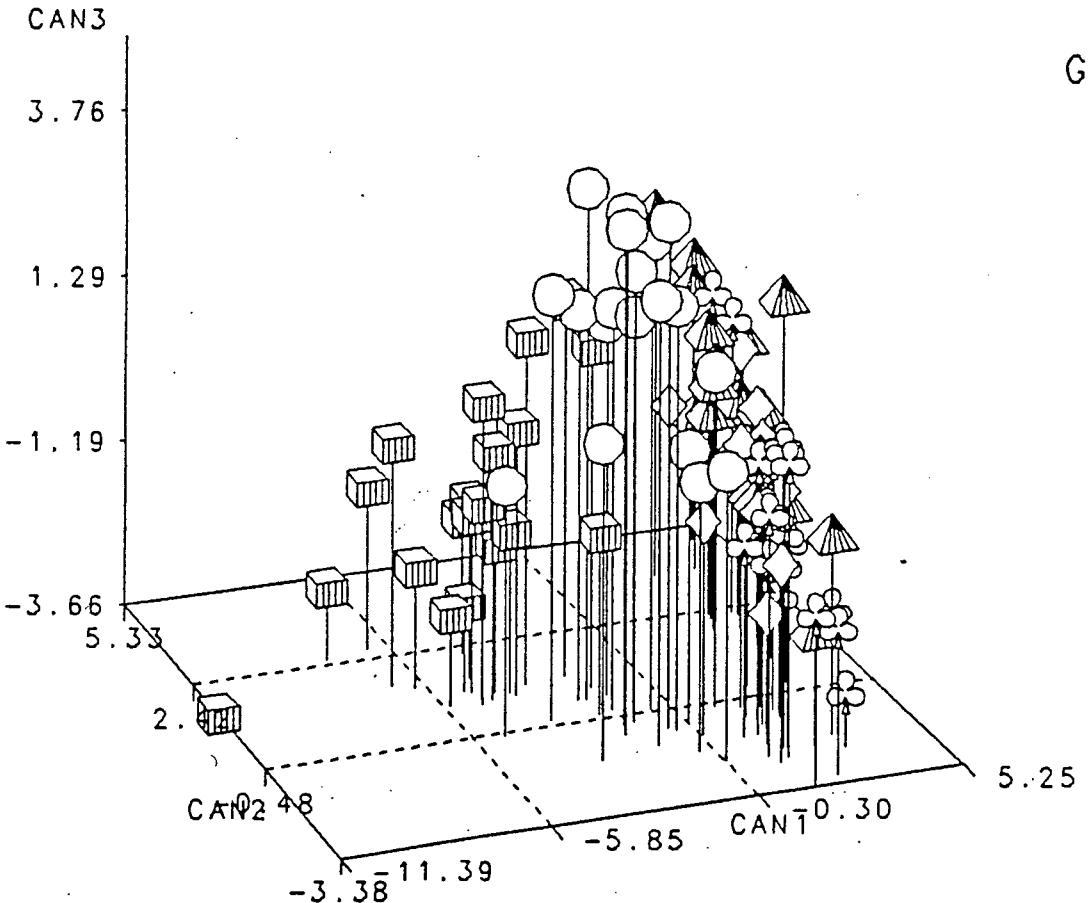


TABLE D.3

RESULTS FROM PRINCIPAL COMPONENT ANALYSIS USING ALL VARIALBES

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	11.2728	7.19460	0.490122	0.490122
PRIN2	4.0782	2.50744	0.177313	0.667435
PRIN3	1.5708	.	0.068294	0.735729

PRINCIPAL COMPONENT ANALYSIS ON ALL VARIABLES
EIGENVECTORS USED FOR GRAPHICAL DISPLAY

EIGENVECTORS

	PRIN1	PRIN2	PRIN3
PK1	0.205038	-.228784	-.222734
PK2	0.245208	-.129324	-.243886
PK3	0.267069	-.089559	0.056578
PK4	0.196899	-.188358	0.182067
PK5	0.232512	-.162648	-.046646
PK6	0.254778	-.140349	0.148154
PK7	0.228090	0.154952	0.196858
PK8	0.251089	0.004638	0.085540
PK9	0.232267	-.042140	0.236466
PK10	0.120774	0.416823	0.008080
PK11	0.216252	0.033410	-.056102
PK12	0.115291	0.369391	-.240535
PK13	0.138251	0.374027	-.256445
PK14	0.213690	0.160561	-.141563
PK15	0.167375	0.113081	-.394423
PK16	-.031387	0.400993	0.247176
PK17	0.225848	-.084185	0.312856
PK18	0.147816	-.064446	-.131809
PK19	0.236982	-.055693	0.277901
PK20	0.241883	0.104919	-.099057
PK21	0.265020	0.096355	0.136364
PK22	0.263799	0.028340	-.080172
PK23	-.063046	0.359820	0.370935

TABLE D.4

RESULTS FROM PRINCIPAL COMPONENT ANALYSIS USING SELECTED VARIABLES

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	8.39508	4.53772	0.466393	0.466393
PRIN2	3.85736	2.55692	0.214298	0.680691
PRIN3	1.30044	.	0.072247	0.752938

PRINCIPAL COMPONENT ANALYSIS ON SELECTED VARIABLE
EIGENVECTORS USED FOR GRAPHICAL DISPLAY

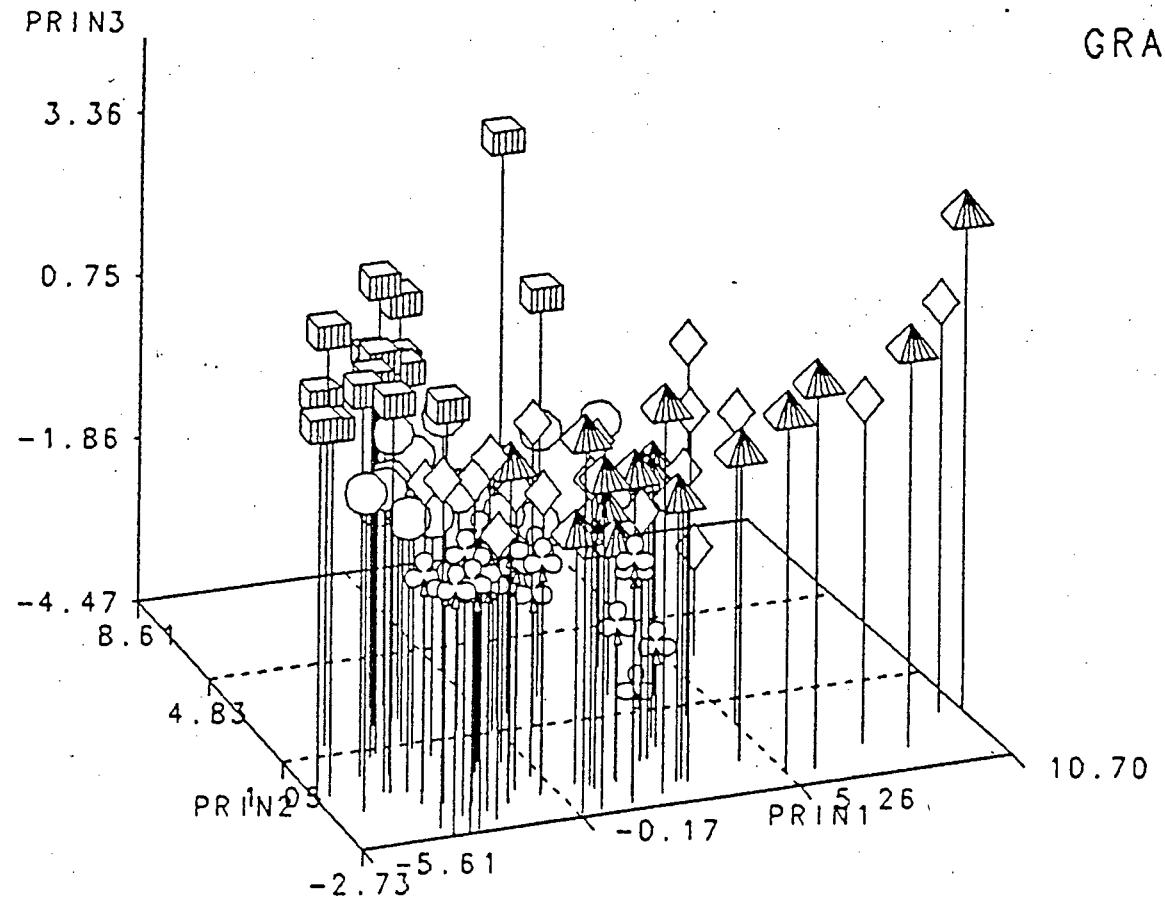
EIGENVECTORS

	PRIN1	PRIN2	PRIN3
PK1	0.225876	-.244836	-.139810
PK2	0.277474	-.152762	-.235981
PK4	0.220325	-.203435	0.325791
PK6	0.291297	-.154087	0.194968
PK7	0.266594	0.151522	0.265080
PK8	0.281414	0.005042	0.217027
PK10	0.151442	0.425902	-.045334
PK11	0.271676	0.004496	-.237801
PK12	0.146920	0.362076	-.385209
PK13	0.172982	0.365876	-.285748
PK14	0.244659	0.142833	0.012744
PK16	-.032957	0.422977	0.305588
PK17	0.273877	-.098599	0.185833
PK18	0.176509	-.085119	-.375237
PK19	0.277919	-.067898	0.152568
PK21	0.312150	0.083027	0.058820
PK22	0.307242	0.006080	-.021988
PK23	-.062387	0.392951	0.280178

PRINCIPAL COMPONENT ANALYSIS

ALL VARIABLES USED FROM 5 GRAPE CULTIVARS

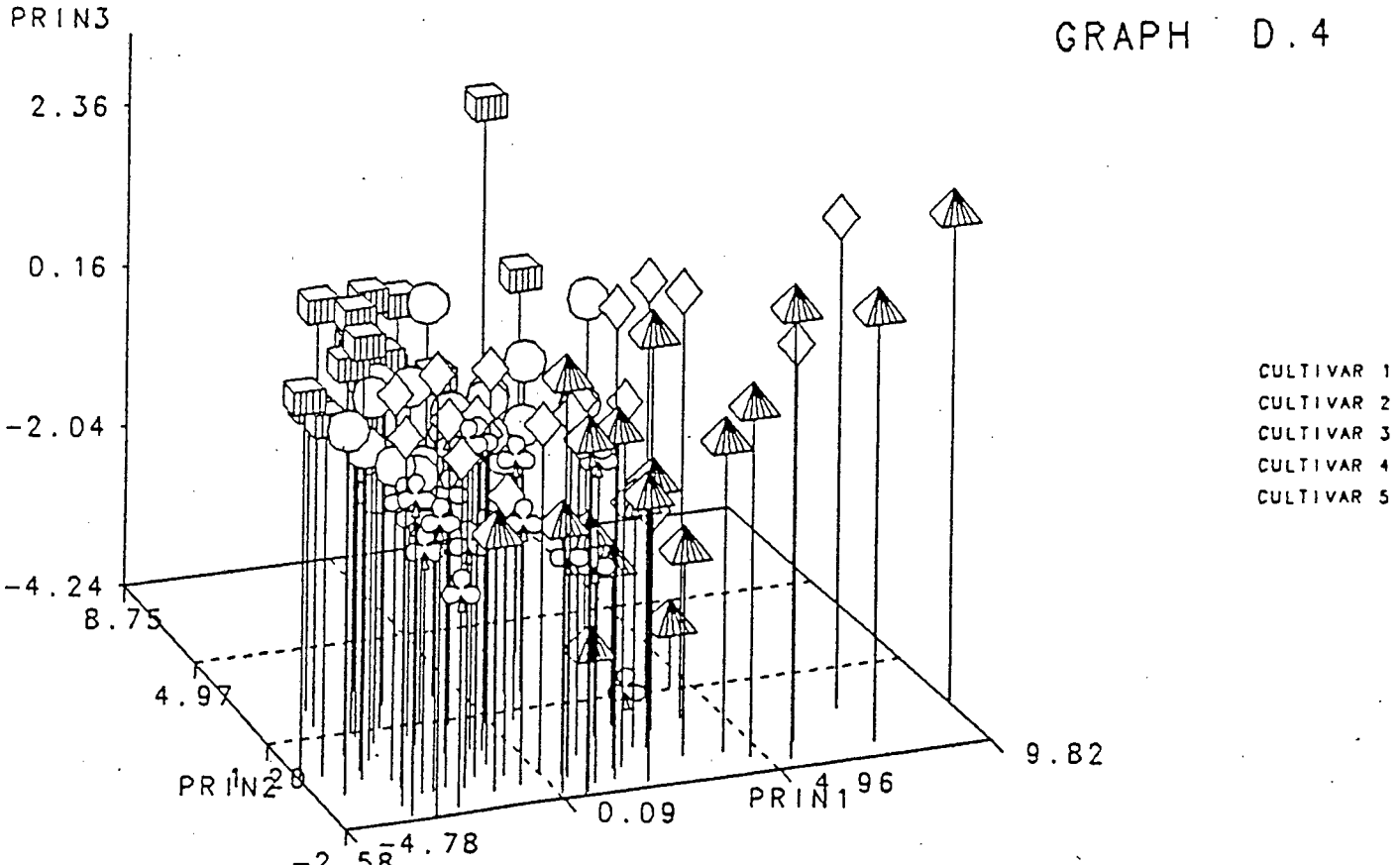
GRAPH D.3



PRINCIPAL COMPONENT ANALYSIS

SELECTED VARIABLES FROM 5 GRAPE CULTIVARS

GRAPH D.4



ABLE D.5

ORRESPONDENCE ANALYSIS ON ALL VARIABLES

ino Acids of five grape cultivars

ERTIAS AND PERCENTAGES OF INERTIA

0.115694	52.71%	**********
0.033184	15.12%	**********
0.016356	7.45%	*******
0.013665	6.23%	******
0.008118	3.70%	****
0.007289	3.32%	***
0.005549	2.53%	**
0.004873	2.22%	**
0.003594	1.64%	**
0.003044	1.39%	*
0.002438	1.11%	*
0.001620	0.74%	*
0.001309	0.60%	*
0.000629	0.29%	
0.000579	0.26%	
0.000447	0.20%	
0.000320	0.15%	
0.000269	0.12%	
0.000171	0.08%	
0.000162	0.07%	
0.000109	0.05%	
0.000076	0.03%	
0.219494		

TABLE D.6

CORRESPONDENCE ANALYSIS ON ALL VARIABLES

ROW CONTRIBUTIONS

I	NAME	QLT	MAS	INR	k=1 COR	CTR	k=2 COR	CTR	k=3 COR	CTR
1	A	359	10	4	-138 204	2	14 2	0	-119 153	9
2	A	146	10	3	-75 85	0	32 15	0	-55 46	2
3	A	527	9	3	-71 69	0	183 455	9	-13 2	0
4	A	476	8	4	-148 213	2	164 263	6	8 1	0
5	A	228	16	7	-153 226	3	6 0	0	-13 2	0
6	A	600	9	6	35 9	0	273 538	21	-86 53	4
7	A	694	19	16	-312 540	16	-165 152	16	20 2	0
8	A	693	14	7	-235 486	7	-77 52	2	133 155	15
9	A	553	13	5	-214 499	5	-14 2	0	69 52	4
10	A	512	14	5	-179 438	4	61 51	2	42 24	1
11	A	706	11	5	-196 373	4	-74 53	2	-169 280	19
12	A	354	8	4	-32 9	0	135 169	4	-138 175	9
13	A	551	23	20	-267 370	14	-150 116	15	113 66	18
14	A	413	11	3	-101 152	1	132 261	6	-1 0	0
15	A	301	9	4	-97 85	1	100 90	3	117 125	7
16	A	407	7	5	75 37	0	231 356	11	-46 14	1
17	A	355	15	7	-112 132	2	128 173	8	-69 50	4
18	A	357	15	8	-161 234	3	1 0	0	116 123	12
19	A	405	6	8	-17 1	0	325 348	19	130 56	6
20	C	737	12	7	-242 457	6	136 144	7	132 136	13
21	C	621	8	7	-109 68	1	233 310	14	206 242	22
22	C	286	7	8	-217 193	3	141 81	4	53 12	1
23	C	669	13	20	-369 400	16	-191 106	14	236 163	45
24	C	655	13	16	-365 476	15	-119 51	6	189 128	28
25	C	763	6	8	-171 111	2	264 264	13	320 388	40
26	C	631	7	4	-91 71	0	229 445	11	116 115	6
27	C	556	7	9	-105 44	1	301 361	20	195 151	17
28	C	847	10	5	-144 202	2	122 145	4	227 500	30
29	C	791	11	6	-64 36	0	252 544	20	157 211	16
30	C	637	16	10	-198 290	5	95 67	4	195 280	37
31	C	650	7	5	-150 147	1	220 317	10	168 185	12
32	C	599	8	5	-38 11	0	268 551	18	68 36	2
33	C	764	6	6	-112 55	1	359 567	22	180 142	11
34	C	695	8	6	-90 46	1	273 424	18	199 224	19
35	C	516	8	6	-64 27	0	178 209	8	205 279	22
36	C	299	10	7	-32 7	0	207 289	13	19 2	0
37	C	223	8	10	-115 49	1	209 161	11	-60 13	2
38	C	283	14	8	-155 181	3	113 96	5	-27 6	1
39	B	894	17	18	-378 626	22	-247 266	32	20 2	0
40	B	538	13	8	-244 451	7	24 5	0	-104 82	9
41	B	853	13	9	-265 453	8	-171 187	11	-182 213	26
42	B	720	11	8	-290 540	8	-165 176	9	25 4	0
43	B	799	11	8	-306 632	9	-149 149	8	-53 19	2
44	B	763	13	9	-275 504	9	-183 222	14	-75 37	5
45	B	861	13	6	-240 553	6	-122 144	6	-130 163	13
46	B	696	11	8	-164 171	3	2 0	0	-287 525	56
47	B	732	12	7	-268 593	7	-102 85	4	-81 53	5
48	B	647	15	12	-300 506	11	-107 64	5	-118 78	12
49	B	951	18	21	-386 571	23	-313 375	53	38 5	2
50	B	930	23	28	-368 492	26	-343 426	80	-58 12	5

TABLE D.6

CORRESPONDENCE ANALYSIS ON ALL VARIABLES

ROW CONTRIBUTIONS

I	NAME	QLT	MAS	INR	k=1	COR	CTR	k=2	COR	CTR	k=3	COR	CTR
51	B	737	8	10	-86	26	0	56	11	1	-452	701	94
52	B	735	13	9	-166	188	3	-4	0	0	-283	547	64
53	B	761	13	10	-292	502	10	-209	258	17	-12	1	0
54	B	633	10	11	-97	41	1	152	100	7	-337	492	72
55	B	954	21	22	-363	567	24	-300	386	56	10	0	0
56	B	691	10	4	-203	437	4	38	15	0	-151	239	14
57	B	490	11	6	-229	459	5	59	30	1	-9	1	0
58	D	649	6	4	130	96	1	291	481	14	-113	72	4
59	D	486	4	4	181	157	1	261	327	9	25	3	0
60	D	599	9	2	-8	1	0	174	566	8	42	32	1
61	D	576	5	5	285	339	4	234	227	8	-48	10	1
62	D	587	6	7	270	279	4	282	304	14	-31	4	0
63	D	245	14	3	23	12	0	84	159	3	58	75	3
64	D	590	12	7	222	407	5	141	164	7	-49	20	2
65	D	647	8	5	196	266	3	204	290	10	-115	91	6
66	D	783	6	11	495	643	13	188	92	7	-135	48	7
67	D	818	10	14	479	722	19	171	91	9	-38	4	1
68	D	742	10	4	143	234	2	203	473	12	-55	35	2
69	D	650	7	6	284	459	5	178	181	7	-42	10	1
70	D	398	10	5	104	98	1	149	202	6	-103	98	6
71	D	405	10	3	154	320	2	37	18	0	-70	67	3
72	D	840	7	4	196	269	2	277	539	15	-67	32	2
73	D	839	7	7	234	247	3	308	428	20	-190	163	16
74	D	716	6	13	412	372	9	345	261	22	-196	84	14
75	D	570	10	5	50	24	0	239	542	18	-20	4	0
76	D	251	10	6	35	10	0	7	0	0	-174	241	19
77	D	39	15	7	30	9	0	30	8	0	47	21	2
78	E	822	5	15	749	820	23	-35	2	0	16	0	0
79	E	632	7	9	380	538	9	-49	9	1	150	84	10
80	E	939	8	26	758	835	42	-197	57	10	181	47	17
81	E	946	9	34	856	879	57	-220	58	13	88	9	4
82	E	847	5	9	567	739	13	-75	13	1	-204	96	12
83	E	922	7	19	705	881	32	-91	15	2	121	26	7
84	E	555	10	5	249	538	6	-37	12	0	-25	5	0
85	E	977	10	27	712	886	46	-176	54	10	-145	37	13
86	E	916	19	33	587	881	55	-113	33	7	26	2	1
87	E	961	9	16	607	898	28	-159	62	7	-26	2	0
88	E	854	7	12	577	847	19	-45	5	0	-26	2	0
89	E	909	7	26	759	751	36	-305	121	20	167	36	12
90	E	957	8	13	573	929	23	-97	26	2	-21	1	0
91	E	942	9	17	607	845	27	-193	86	10	71	12	3
92	E	961	8	17	664	898	30	-156	49	6	80	13	3
93	E	778	14	13	358	622	16	-143	99	9	109	58	10
94	E	943	9	19	637	894	32	-148	48	6	14	0	0
95	E	970	10	41	876	841	66	-342	128	35	-14	0	0
96	E	975	7	36	1005	887	60	-314	87	21	40	1	1
97	E	916	6	24	861	906	41	-66	5	1	-60	4	1

TABLE D.7

CORRESPONDENCE ANALYSIS ON ALL VARIABLES

COLUMN CONTRIBUTIONS

J	NAME	QLT	MAS	INR	k=1	COR	CTR	k=2	COR	CTR	k=3	COR	CTR
1	1	522	50	26	-147	188	9	156	211	36	119	123	43
2	2	565	86	31	-153	299	17	143	261	53	-19	5	2
3	3	616	51	7	-112	393	5	8	2	0	84	221	22
4	4	423	6	35	-606	303	20	-347	99	23	159	21	10
5	5	534	123	40	-98	132	10	126	221	59	114	181	98
6	6	899	113	140	-386	546	145	-304	339	314	61	14	26
7	7	249	4	1	117	168	0	51	32	0	-63	49	1
8	8	402	133	24	-44	49	2	28	19	3	116	335	110
9	9	272	7	24	-345	156	7	-294	114	18	-36	2	1
10	10	718	24	15	303	680	19	70	37	4	-14	2	0
11	11	508	6	7	-131	62	1	-46	7	0	-350	439	43
12	12	579	14	10	250	397	8	142	129	9	-91	53	7
13	13	614	25	16	222	356	11	181	236	24	-55	22	5
14	14	298	12	8	38	10	0	180	220	12	-100	68	8
15	15	303	20	19	16	1	0	250	297	38	30	4	1
16	16	814	84	141	523	743	199	145	57	54	-69	13	25
17	17	739	77	55	-113	82	9	-225	324	118	-228	332	245
18	18	422	20	38	-101	24	2	97	22	6	-402	377	194
19	19	357	6	5	-123	84	1	-130	94	3	-179	179	12
20	20	438	7	2	36	16	0	140	247	4	-118	174	6
21	21	521	28	8	-11	2	0	-35	18	1	-183	502	57
22	22	340	61	26	-112	137	7	80	70	12	-110	133	46
23	23	980	43	319	1189	872	527	-401	99	209	121	9	39

TABLE D.8

CORRESPONDENCE ANALYSIS ON SELECTED VARIABLES

Amino Acids of five grape cultivars

INERTIAS AND PERCENTAGES OF INERTIA

1	0.136839	56.23%	**********
2	0.036863	15.15%	**********
3	0.018818	7.73%	*******
4	0.013460	5.53%	*****
5	0.009181	3.77%	***
6	0.007806	3.21%	***
7	0.005863	2.41%	**
8	0.005037	2.07%	**
9	0.003255	1.34%	*
10	0.002541	1.04%	*
11	0.001209	0.50%	
12	0.000938	0.39%	
13	0.000590	0.24%	
14	0.000456	0.19%	
15	0.000207	0.09%	
16	0.000161	0.07%	
17	0.000115	0.05%	
	0.243340		

TABLE D.9

CORRESPONDENCE ANALYSIS ON SELECTED VARIABLES

ROW CONTRIBUTIONS

I	NAME	QLT	MAS	INR	k=1	COR	CTR	k=2	COR	CTR	k=3	COR	CTR
1	a	565	10	5	-148	187	2	12	1	0	-210	377	22
2	a	181	10	3	-74	86	0	74	87	1	-21	7	0
3	a	598	8	2	-61	52	0	190	514	8	-48	32	1
4	a	564	7	3	-161	279	1	133	192	4	-93	93	3
5	a	226	15	7	-165	224	3	-9	1	0	-14	2	0
6	a	607	9	6	51	15	0	299	520	22	-111	72	6
7	a	762	18	15	-347	600	16	-179	159	16	20	2	0
8	a	788	13	6	-252	548	6	-102	90	4	132	151	12
9	a	872	12	3	-237	787	5	-71	72	2	31	14	1
10	a	544	13	4	-187	482	3	65	59	1	15	3	0
11	a	751	11	6	-222	395	4	-56	25	1	-203	332	24
12	a	478	8	4	-31	8	0	143	161	4	-199	309	16
13	a	781	22	15	-286	501	13	-139	119	11	162	161	31
14	a	462	11	3	-97	125	1	158	332	7	-20	5	0
15	a	189	8	3	-83	74	0	70	52	1	77	63	2
16	a	389	7	5	91	45	0	251	343	11	11	1	0
17	a	449	15	5	-118	172	2	141	247	8	-49	30	2
18	a	628	15	6	-163	275	3	35	13	0	181	340	26
19	a	473	5	6	41	6	0	368	462	19	39	5	0
20	c	611	11	7	-274	468	6	120	89	4	93	54	5
21	c	502	8	7	-113	61	1	225	246	11	200	194	16
22	c	309	7	7	-238	215	3	141	76	4	-70	19	2
23	c	792	13	21	-420	451	17	-201	104	15	304	237	66
24	c	760	13	18	-416	513	16	-125	46	5	260	200	46
25	c	551	6	6	-178	114	1	211	161	7	277	276	23
26	c	494	6	3	-96	87	0	194	357	6	72	50	2
27	c	532	7	6	-87	34	0	312	440	18	113	58	5
28	c	901	9	4	-161	218	2	88	65	2	271	617	36
29	c	684	10	5	-60	32	0	226	453	14	150	198	12
30	c	522	15	9	-225	332	5	47	14	1	163	175	21
31	c	480	7	5	-164	151	1	210	249	8	119	80	5
32	c	548	8	5	-45	13	0	283	516	17	55	19	1
33	c	721	5	6	-124	54	1	375	499	21	218	168	14
34	c	635	7	6	-97	46	1	264	340	14	227	249	20
35	c	567	8	6	-68	26	0	161	144	6	267	397	30
36	c	277	10	7	-41	10	0	203	251	11	51	16	1
37	c	238	8	12	-131	50	1	208	125	10	-148	63	9
38	c	352	14	8	-190	263	4	98	70	4	-51	19	2
39	b	914	18	19	-426	674	23	-254	239	31	-2	0	0
40	b	826	13	8	-271	505	7	28	6	0	-214	315	32
41	b	880	14	9	-306	554	9	-152	137	8	-179	189	23
42	b	773	11	8	-328	590	9	-183	183	10	5	0	0
43	b	861	12	8	-348	695	10	-148	126	7	-84	40	4
44	b	801	14	10	-314	542	10	-190	198	13	-105	60	8
45	b	859	13	6	-275	635	7	-103	90	4	-127	135	11
46	b	770	12	8	-193	230	3	62	24	1	-289	515	52
47	b	780	12	7	-305	637	8	-95	61	3	-109	81	8
48	b	634	15	13	-337	550	12	-70	24	2	-112	60	10
49	b	965	18	23	-435	619	26	-320	336	51	56	10	3

TABLE D.9

CORRESPONDENCE ANALYSIS ON SELECTED VARIABLES

ROW CONTRIBUTIONS (CONTINUE)

I	NAME	QLT	MAS	INR	k=1	COR	CTR	k=2	COR	CTR	k=3	COR	CTR
51	b	907	8	10	-98	31	1	146	68	5	-502	807	105
52	b	742	14	9	-195	247	4	52	18	1	-271	477	54
53	b	789	14	10	-331	589	11	-192	198	14	16	1	0
54	b	865	10	12	-111	42	1	194	129	10	-449	693	109
55	b	960	21	24	-409	599	26	-317	360	57	8	0	0
56	b	763	10	5	-234	476	4	63	34	1	-171	253	15
57	b	533	11	6	-256	479	5	56	23	1	-65	31	2
58	d	658	6	5	132	89	1	333	568	17	-15	1	0
59	d	542	4	5	206	166	1	279	306	9	134	70	4
60	d	624	9	2	8	1	0	163	595	6	35	28	1
61	d	573	5	6	319	361	4	245	212	8	-2	0	0
62	d	587	6	7	309	302	4	293	272	13	66	14	1
63	d	344	13	2	32	23	0	88	175	3	80	146	5
64	d	675	13	7	239	411	5	184	245	12	51	19	2
65	d	720	8	4	193	289	2	235	430	12	-11	1	0
66	d	765	6	11	535	660	13	202	94	7	-70	11	2
67	d	847	10	15	513	741	19	184	96	9	59	10	2
68	d	867	10	4	159	246	2	252	621	17	-2	0	0
69	d	751	8	5	283	461	4	212	257	9	76	33	2
70	d	503	10	4	92	85	1	201	406	11	35	13	1
71	d	448	10	3	146	359	2	72	87	1	11	2	0
72	d	828	6	5	226	297	2	300	521	16	-42	10	1
73	d	856	7	7	238	248	3	366	585	26	-73	23	2
74	d	711	7	13	418	367	8	403	341	29	-37	3	0
75	d	547	10	5	76	48	0	244	497	16	-13	1	0
76	d	63	11	4	8	1	0	62	46	1	-37	17	1
77	d	461	16	6	17	3	0	76	61	3	194	397	32
78	e	830	5	16	804	826	23	-59	4	0	6	0	0
79	e	730	7	9	457	666	11	-129	53	3	57	11	1
80	e	936	9	29	830	839	43	-272	90	17	75	7	3
81	e	957	9	36	905	873	56	-282	85	20	-13	0	0
82	e	821	5	9	558	750	12	-31	2	0	-169	69	8
83	e	976	7	21	808	935	35	-169	41	6	20	1	0
84	e	568	11	5	255	561	5	-17	3	0	22	4	0
85	e	974	12	25	691	924	41	-126	31	5	-100	19	6
86	e	934	20	33	601	899	53	-94	22	5	71	13	5
87	e	969	9	17	620	898	26	-162	61	7	-63	9	2
88	e	863	7	12	603	857	19	-38	3	0	-28	2	0
89	e	902	8	27	801	746	36	-362	153	27	49	3	1
90	e	968	9	14	609	929	23	-110	30	3	-60	9	2
91	e	937	9	17	627	844	26	-206	91	10	28	2	0
92	e	962	8	18	695	892	29	-192	68	8	29	2	0
93	e	764	15	14	380	630	16	-149	97	9	92	37	7
94	e	958	10	20	681	889	32	-185	66	9	-46	4	1
95	e	978	11	41	874	838	61	-347	132	35	-84	8	4
96	e	977	8	35	996	884	56	-323	93	22	-14	0	0
97	e	913	7	24	868	907	38	-60	4	1	-44	2	1

TABLE D.10

CORRESPONDENCE ANALYSIS ON SELECTED VARIABLES

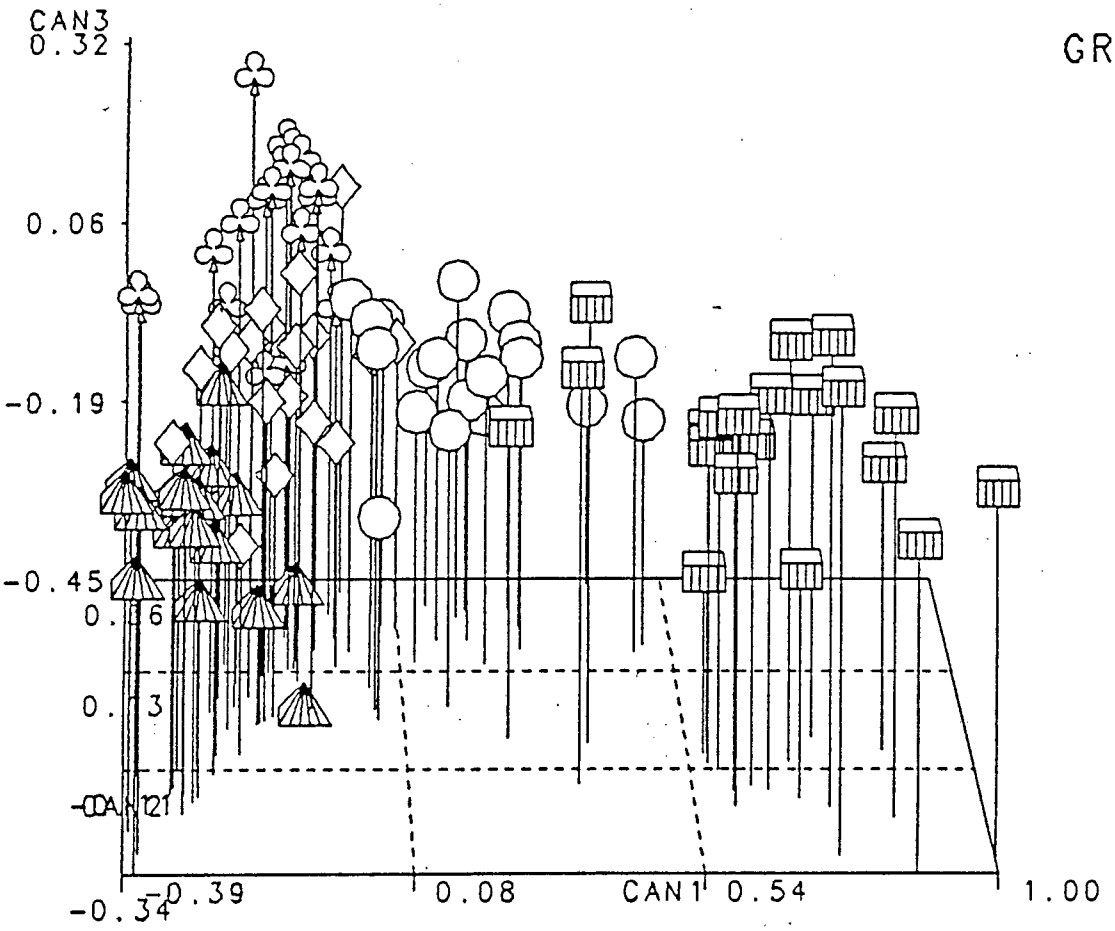
COLUMN CONTRIBUTIONS

J	NAME	QLT	MAS	INR	k=1	COR	CTR	k=2	COR	CTR	k=3	COR	CTR
1	1	448	63	36	-163	191	12	161	187	44	98	70	32
2	2	591	108	42	-171	306	23	162	276	77	-29	9	5
3	4	436	8	41	-626	315	23	-332	89	24	201	32	17
4	6	927	143	160	-410	614	175	-277	280	297	95	33	69
5	7	203	5	1	96	118	0	81	84	1	-1	0	0
6	8	505	168	34	-63	82	5	42	37	8	138	387	169
7	10	703	31	15	275	630	17	90	68	7	24	5	1
8	11	420	7	8	-157	95	1	14	1	0	-292	325	33
9	12	547	18	10	221	339	6	165	188	13	-53	19	3
10	13	632	31	17	195	295	9	208	334	37	18	3	1
11	14	325	16	9	18	2	0	217	319	20	23	4	0
12	16	811	106	148	493	721	189	173	89	86	23	2	3
13	17	798	97	58	-140	135	14	-181	226	87	-252	437	329
14	18	567	25	44	-125	36	3	140	45	13	-460	486	279
15	19	396	8	6	-145	117	1	-87	42	2	-205	237	18
16	21	359	35	9	-35	19	0	8	1	0	-146	339	40
17	22	380	77	29	-133	196	10	128	182	34	-13	2	1
18	23	979	54	332	1133	865	511	-411	114	250	23	0	2

CORRESPONDENCE ANALYSIS ON ORIGINAL DATA

ALL VARIABLES USED FROM 5 GRAPE CULTIVARS

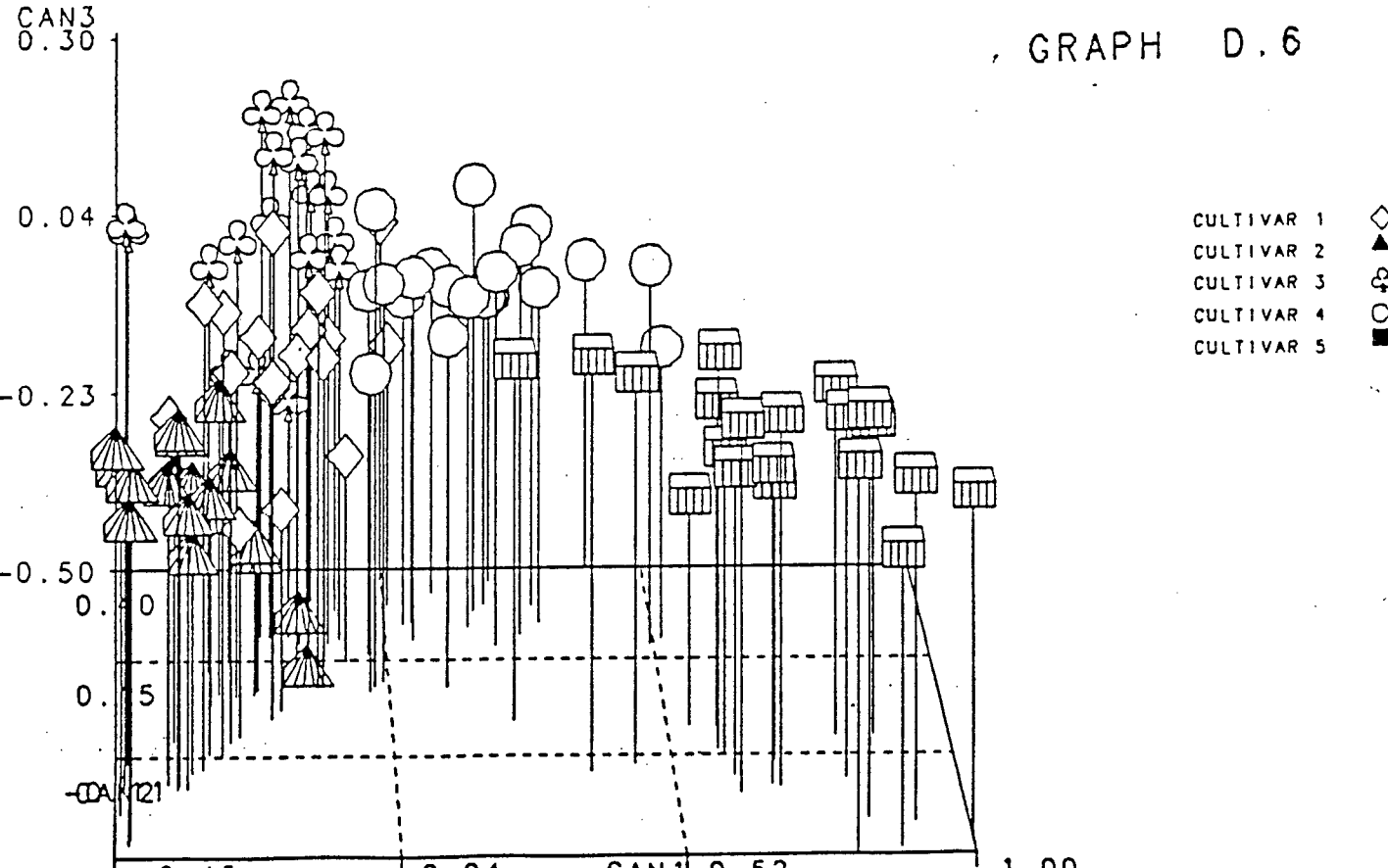
GRAPH D.5



CORRESPONDENCE ANALYSIS ON ORIGINAL DATA

SELECTED VARIABLES USED FROM 5 GRAPE CULTIVARS

, GRAPH D.6



REFERENCES

Aitchison, J. Aitken, C.G.G. (1976). Multivariate discrimination by the Kernel method. Biometrika, 63, pp. 413-420.

Amoh, R.K. (1985). Estimates of a Discriminant Function from a mixture of two inverse Gaussian distributions. Journal Statist. Comput. Simul., Vol. 20, pp. 275-286.

Anderberg, M.R. (1973). Probability and Mathematical Statistics. Academic Press.

Anderson, T.W. (1958). An Introduction to Multivariate Statistical analysis. John Wiley & Sons, Inc.

Balakrishnan, N., Kocherlakota, S., Kocherlakota, K. (1988). Robustness of the Double Discriminant function in non-normal situations. S.A. Statist. Journ., 22, pp. 15-43.

Bayne, A., Tan, W.Y., (1981). QDF misclassification probabilities for known population parameters. Commun. Statist.-Theor. Meth. A10(22), pp. 2315-2326 .

Bayne, C.K., Beachamp, J.J., Kane, V.E. (1984). Misclassification probabilities for second-order discriminant functions used to classify bivariate normal populations. Commun.Statist. -Simula. Computa. Vol. 13(5), pp. 683-704 .

Beauchamp, J.J., Folkert, J.E., Robson, D.S. (1980). A note on the effect of logarithmic transformation on the probability of misclassification. Commun. Statist.-Theor. Meth., Vol. A9(8), pp. 777-794 .

Beauchamp, J.J., Robson, D.S. (1986). Transformation considerations in discriminant analysis. Commun Statist -Simula. Vol. 15(1), pp. 147-147 .

Berk, K.N. (1980). Forward and Backward stepping in variable selection. Journal Statist. Comput. Simul., pp. 177-185.

Birks, H.J.B. (1987). Multivariate Analysis in Geology and Geochemistry: an Introduction. Chem. and Intelligent Lab. Syst. , 2, pp. 15-28 .

Broffitt, B., Clarke, W.R., Lachenbruch, P.A. (1980). The effect of Huberizing and Trimming on the Quadratic Discriminant Function. Commun. Statist., A9(1), pp. 13-25.

Caccoulas, T. (1973). Discriminant analysis and Applications, New York: Academic Press,Inc.

Clarke, M.R.B. (1971). Computer developments in research and diagnosis. Proc. Roy. Soc. Med., 64, pp. 819-822.

Conover, W.J., Iman, R.L. (1980). The rank transformation as a method of discrimination with some examples. Commun. Statist. Theor. Math. , A9(5), pp. 465-487.

Constanza, M.C. Variable selection in forward stepwise discriminant analysis: Some result for small samples when variables are independent. Amer. Statist. Assoc. Proceedings of the Statist. Comput. Section 5-8 August 1985 .

Constanza, M.C., Afifi, A.A.(1979) Comparisons of stopping rules in Forward Stepwise Discriminant analysis. Journ. Amer. Statist. Assoc., 74, pp.777-785.

Craft, D. (1988). Evaluation of Chemical test for Engineering Soils using Discriminant Analysis. Chemometrics and Intelligent Systems. Vol. 3, pp. 111-118 .

Cureton, E.E., D'Agostino, R.B. (1982). Factor Analysis an Applied Approach. Lawrence Eurlbaum Ass Publ. London pp.223.

Dixon, W.J.(1983). BMDP Statistical Software. University of California Press, Ltd.

Dunn, C.L., Smith, W.B. (1980). Combinatoric Classification of Multivariate Normal variates. Commun. Statist. Theor. Meth. A9(13) pp. 1317-1340 .

Efron, B. (1975). The efficiency of Logistic-regression compared to Normal Discriminant Analysis. Journ. Amer. Statist. Assoc., Vol 70, pp. 892-898.

Efron, B. (1979). Bootstrap methods: another look ar the jackknife. Ann. Statist., Vol 7, pp. 1-26.

Efron, B. (1982). The Jackknife, the Bootstrap and other Resampling plans. Society for Industrial and Applied Mathematics CBMS 38.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-validation. Journ. Amer. Stat. Assoc. Vol. 78(382), pp. 316-331 .

Enis, P., Geiser, S. (1970). Sample discriminants which minimize posterior squared error. S.African Stat. Journ. , 4, pp.85-93.

Fisher, R. A. (1936). The use of multiple measurement in Taxonomic problems. (See Hawkins (1982))

Ganesalingam, S., McLachlan, G.J. (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. Journal Statist. Comput Simul., Vol. 9, pp. 151-164 .

Geisser, S. (1964). Posterior odds for multivariate normal classification. Journ. Roy. Statist. Soc., B26, pp. 69-76.

Goldstein , M.(1975). Comparisons of the density estimate in classification procedures. Journ. Amer. Statist. Assoc., 70, pp. 666-669.

Greenacre, M.J. (1981) . Practical Correspondence Analysis. appearing in Interpreting Multivariate data. Barnett, W. Wiley, pp. 119-146.

Greenacre, M.J. (1984) . Theory and applications of Correspondence Analysis. Academic Press.

Habbema, J.D.F., Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics, 19, pp. 487-493.

Habbema, J.D.F., Hermans, J., Remme, J. (1978). Variable kernel density estimation in Discriminant Analysis. In Compstat 1978: Proceedings in computational statistics. Physica-Verlag, Vienna, pp. 178-185.

Haff, L.R. (1986). On Linear Log-odds and estimation of discriminant coefficients. Commun. Statist. -Theor. Meth. Vol. 15(7), pp. 2131-2144 .

Hand, D.J. (1981). Discrimination and Classification. John Wiley & Sons, Ltd.

Hangos, K.M., Leitzner, L. (1987). The systematic error caused by random errors through data reduction. Journal of Automatic Chem. 9(1), pp. 23-29 .

Hawkins, D.M. (1982) Topics in Applied Multivariate Analysis. Cambridge University Press, Cambridge, pp. 1-71 .

Hawkins, D.M., Fatti, L.P. (1984). Exploring Multivariate Data using the Minor Principal Components. The Statistician. Vol. 33, pp. 325-338 .

Hills, M. (1966). Allocation rules and their error rates. Journ. Roy. Statist. Soc. , B28. pp.1.

Hora, S.C. (1980). Sequential Discrimination. Commmun. Statist.-Theor. Meth. A9(9), pp. 905-916 .

Johnson, R.A., Wichen, D.W. (1982). Applied Statistical Analysis. Prentice Hall .

Kelley, P.R. (1981). Bayesian adjustment of the matching discriminant function. Amer. Statist. Assoc. 1985 Proceedings of the stat. Comput. SECTION 10-13 .

Kendall, M.G., Stuart, A. (1961). The advanced theory of statistics, Vol 3. London: Charles Griffin and Company, Ltd.

Khatri, C.G., Rao, C.R., Sun, Y.N. (1986). Tables for obtaining confidence bounds for realized signal to noise ration with an estimated discriminant function. Commun. Statist. -Simula., Vol. 15(1), pp. 1-14 .

Koffler, S.L., Penfield, D.A. (1979). Non-parametric discriminant procedures for non-normal distributions. Journal Statist. Comput. Simul. Vol. 8, pp. 281-299 .

Kwan, W.O., Kowalski, B.R. (1978). Classification of wines by Applying Pattern Recognition to Chemical Composition data. Journal of Food Science. Vol. 43, pp. 1320-1323 .

Lachenbruch, P.A. (1967). An almost unbaised method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics. Vol 23, pp.639-645.

Lachenbruch, P.A. (1975). Discriminant Analysis. New York: Macmillon Publishers Co.,Inc.

Lavine, B.K., Jurs, P.C., Henry, D.R., Vander Meer, R.K., Pino, J.A., McMurray, J.E. (1988). Pattern recognition studies of Complex Chromatographic Data sets: Design and analysis of Pattern recognition experiments. Chemometrics and Intel. Lab. Systems. Vol. 3, pp. 79-89 .

Lindeman, R.H., Marenda, P.F., Gold, R.Z. (1980). Introduction to bivariate and Multivariate Analysis. Scott, Foresman and Company, US.

Little, R.J.A., Rubin, D.B. (1987). Statistical Analysis with missing data John Wiley & Sons, Inc.

Mallows, C.L. (1953). Sequential discrimination. See Hawkins (1982) .

Mardia, K.V., Kent, J.T., Bibby, J.M., (1979). Multivariate Analysis. Academic Press, New York, NY . pp. 213-393 .

McCulloch, E. (1986). Some remarks on allocatory and separatory linear discrimination. Journal of Stat. Planning and Inference. Vol. 14, pp. 323-330 .

Mckay, R.J. (1976). Simultaneous procedures in Discriminant analysis involving two groups. Technometrics, 18, pp. 47-53.

McLachlan, G.J., Ganesalingam, S. (1982). Updating a discriminant function on the basis of unclassified data. Commun. Statist.-Simula. Computa. Vol. 11(6), pp. 753-767.

Mellinger, M. (1987). Correspondence analysis: The method and its applications. Chem. and Intelligent Lab. Syst. Vol 2, pp. 61-77.

Mellinger, M. (1987). Interpretation of Lithogeochemistry using Correspondence analysis. Chem. and Intelligent Lab. Syst. Vol 2, pp. 93-108.

Menzefricke, U. (1981). A decision-theoretic approach to variable selection in discriminant analysis. Commun. Statist.-Theor. Meth. A10(7), pp. 669-696.

Moore, K.K., Smith, W.B. (1975). A rank order approach to discriminant analysis. Proceedings of the business and economic statistics section of the American Statistical Association , pp. 451-455.

Mueller, R.O., Cozad, J.B. (1988). Standardized discriminant coefficients : which variance estimate is appropriate? Journ. Educ. Statist. Vol 13(4), pp.313-318.

Nakaniski, H., Sato, Y. (1985). Performance of the LDF and QDF for three types of non-normal distributions. Commun Statist.-Theor. Meth. Vol. 14(5), pp. 1181-1200 .

Novotny, T.J., McDonald, L.L. (1986). Model selection using discriminant analysis. Journal of Applied Statist. Vol. 13(2). pp. 159-165 .

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Stat., 33, pp. 1065-1076.

Radhakrishnan, R. (1985). Influence functions for certain parameters in discriminant analysis when a single discriminant function is not adequate. Commun. Statist. -Theor. Meth. Vol. 14(3), pp. 535-549 .

Rao, C.R. (1972). Recent trends of research work in multivariate statistics. Biometrika, 28, pp. 3-22.

Remme, J.,Habbema, J.D.F., Hermans, J. (1980). A simulative Comparison of Linear, Quadratic and Kernel discrimination. Journal Statist. Comput. Simul. Vol. 11. pp. 87-106 .

Robert, P., Bertrand, D., Devaux M.F., Grappin, R. (1987). Multivariate analysis applied to Near-Infrared Spectra of milk. Anal. Chem. Vol. 59, pp. 2187-2191 .

Rolin, J. (1983). Selection of variables in Discriminant analysis. Proceedings of the 4-th Frans-Belgian Meeting of Statisticians.

Romesberg, H.C. (1984). Cluster analysis for researchers. Lifetime Learning Publ.

Saint, Y., Testa, A. (1986). Application of Data Analysis Methods to the Chemical Analysis of Tobacco. Coresta Simposium October, Giardini Naxos - Taormina .

SAS Institute Inc. SAS User's Guide: Basics, Version 5 Edition. Cary,NC: SAS Institute Inc.,(1985).

SAS Institute Inc. SAS User's Guide: Statistics, Version 5 Edition. Cary,NC: SAS Institute Inc.,(1985).

SAS Institute Inc. SAS/GRAPH User's Guide, Version 5 Edition. Cary,NC: SAS Institute Inc.,(1985).

Saxberg, E.H., Duewer, D.L., Booker, J.L. (1978). Pattern Recognition and Blind Assay Techniques Applied for Forensic separation of Whiskies. Analytical Chimica Acta. Vol. 103, pp. 201-212 .

Schervish, M.J. (1984). Linear Discrimination for Three known Normal populations. Journal of Stat Planning and Inference. Vol. 10, pp. 167-175 .

Schmitz, P.I.M., Habbema, J.D.F., Hermans, J., Raatgever, J.W. (1984). Comparative performance of four Discriminant Analysis methods for Mixtures of continuous and discrete data variables. Commun. Statist.-Simula. Computa., Vol. 12(6), pp. 727-751 .

Schmitz, P.I.M., Habbema, J.D.F., Hermans, J. (1985). A Simulation study of the Performance of five Discriminant analysis methods for mixtures of continuous and binary variables. Journal Statist. Comput. Simul. Vol. 23, pp. 69-95.

Schwemer, G.T., Mickey, M.R. (1980). A note of the Linear discriminant function when group means are equal. Commun .Statist.-Simula. Computa., B9(6), pp. 633-638 .

Schwemer, G.T., Dunn, O.J. (1980). Posterior Probability estimators in Classification simulations. Commun. Statist. -Simula. Computa., B9(2), pp. 133-140 .

Snapinn, S.M., Knoke, J.D. (1985). An Evaluation of Smoothing Classification Error-rate Estimators. Technometrics. Vol. 27(2). pp. 327-352 .

Van der Merwe, C.A. (1982). Statistiese klassifikasietegnieke: Aannames en metodes. HSRC Report WS 26 .

Van Ness, J.W.,Simpson, C. (1976). On the effects of dimension in discriminant analysis. Technometrics, 18, pp. 175-187.

Wald, S. (1987) . Principal Component analysis. Chem. and intelligent Lab Syst. Vol 2, pp. 37-52 .

Schwemer, G.T., Mickey, M.R. (1980). A note of the Linear discriminant function when group means are equal. Commun .Statist.-Simula. Computa., B9(6), pp. 633-638 .

Schwemer, G.T., Dunn, O.J. (1980). Posterior Probability estimators in Classification simulations. Commun. Statist. -Simula. Computa., B9(2), pp. 133-140 .

Snapinn, S.M., Knoke, J.D. (1985). An Evaluation of Smoothing Classification Error-rate Estimators. Technometrics. Vol. 27(2). pp. 327-352 .

Van der Merwe, C.A. (1982). Statistiese klassifikasietegnieke: Aannames en metodes. HSRC Report WS 26 .

Van Ness, J.W.,Simpson, C. (1976). On the effects of dimension in discriminant analysis. Technometrics, 18, pp. 175-187.

Wald, S. (1987) . Principal Component analysis. Chem. and intelligent Lab Syst. Vol 2, pp. 37-52 .