APPLICATION OF CANONICAL CORRELATION ANALYSIS

"A Comparative Study on Academic Performance"

BY

KEPHER HENRY MAKAMBI.

This dissertation is submitted in partial fulfilment for the degree of Master of Science in Mathematical Statistics in the Department of Mathematics.

KENYATTA UNIVERSITY,

JUNE, 1990.

DECLARATION

This dissertation is my own work and has not been presented for a degree in any other University.

Signature .....

KEPHER HENRY MAKAMBI.

This dissertation has been submitted for examination with my approval as University Supervisor.

Signature .....

PROF. J.W. ODHIAMBO,
DEPARTMENT OF MATHEMATICS,
UNIVERSITY OF NAIROBI,
NAIROBI, KENYA.

CONTENTS

LIST OF TABLES

SUMMARY OF CONTENTS

The technique of Canonical Correlation Analysis is useful in investigating the interrelationships between two or more sets of variables. In this dissertation, some interpretive devices in canonical correlation analysis are used to study the interrelationships among different courses offered by the Department of Mathematics, Kenyatta University.

In section 1.1 of Chapter I; we give a brief introduction of Canonical correlation analysis. Section 1.2 gives an overview of canonical correlation analysis whereby the data is assumed to be from a known population. Section 1.3 outlines some of the work done early on the application of canonical correlation analysis. The aims and significance of the study are given in section 1.4

Normally the population parameters are not known and therefore their sample counterparts are used. Section 2.1 of Chapter II defines the data matrix based on a sample of size N. In section 2.2 the sample estimates of the population parameters are defined. A rigorous derivation of fundamental equations for canonical correlations and variates is given in section 2.3.

Chapter III outlines the significance tests and Interpretation of canonical correlations. In section 3.1 the tests of significance of canonical correlations, individual canonical correlation coefficients and those of particular variables are discussed. Section 3.2 outlines the interpretive devices used in canonical correlation analysis.

Finally chapter IV deals with the application of canonical

correlation analysis in studying the interrelationships among the courses offered by the Department of Mathematics, Kenyatta University. Section 4.1 gives the descriptions of the courses which were offered in first year 1984/85 and 1985/86 academic years and these courses act as the variables in our current study. Sub-section 4.2.1 of section 4.2 deals with canonical correlation analysis on the relationship between Pure Mathematics and Mathematical Statistics and the results for 1984/85 and 1985/86 are both given here. Sub-section 4.2.2 gives the comparative remarks on the results of subsection 4.2.1. Canonical Correlation analysis on the relationships between Pure and Applied Mathematics is dealt with in sub-section 4.2.3 and subsection 4.2.4 gives the comparative remarks on the results of subsection 4.2.3. The canonical correlation analysis on the relationships between Mathematical Statistics and Applied Mathematics is given in subsection 4.2.5 and subsection 4.2.6 gives the comparative remarks on the results of subsection 4.2.5. Even though this technique (canonical correlation analysis) is useful, it has a number of limitations which are outlined in section 4.3. The conclusion on the study is given in section 4.4.

ACKNOWLEDGEMENT

It gives me great pleasure to express my gratitude to those who in one way or the other assisted me in the preparation of this dissertation.

I wish to acknowledge my Supervisor Prof. J.W. Odhiambo for his deligence and excellent supervision which he showed by sacrificing a lot of his time to scrutinise my work and guide me in the best way possible.

I am especially indebted to Mr. L.B.X. Odongo and Mr. Githu of Kenyatta University for improving my interest and knowledge in Statistics. My thanks to Prof. J. Mutio and Dr. E.M. Kukuni for providing me with the data for the present study. I would also like to express my appreciation to my lecturers Dr. M. Manene and Dr. F. Njui for their encouragement during the preparation of this work.

I am thankful to Irene Karimi for typing my work speedily and with a lot of care.

Finally, thanks to my dear wife, Pamela, for her patience and cooperation while I worked through this project. I also do appreciate the active support and encouragement shown by my parents, friends and other members of my family throughout my school days.

CHAPTER I

INTRODUCTION

1.1 Introduction.

In most statistical data analyses, it is usual to examine linear interdependencies by multiple regression, which measures the relationship between a set of predictor (independent) variables and one dependent (criterion) variable. When observations are taken on a large number of correlated variables it is natural to look at various ways in which the number of variables might be reduced without sacrificing too much information. When the variables are regarded as belonging to a single set then principal components analysis is often informative.

However, in many research settings, a scientist encounters a phenomenon that is best described not in terms of a single criterion but, because of its complexity, in terms of a number of response variables. In such cases, interest may centre on the relationship between the set of criterion variables and the set of explanatory factors. In the business or economic fields, we might be interested in the relationship between a set of price indices and a set of production indices, with a view towards, say, predicting one from the other. In psychological investigations, we might be concerned with the relationship between a set of personality variables on the one hand, and various ability variables

on the other. In educational performance investigations, for example, we might be interested in the relationship between a group of science subjects and arts subjects. At university level, one can assess the degree of relationship between a group of Pure Mathematics units on the one hand, and Statistics or Applied Mathematics units on the other. Broadly, one is faced with the problem of investigating and explaining the relationships between two or more sets of variables. The study of the relationship between a set of predictor variables and a set of response variables is known as Canonical Correlation Analysis

1.2 What is Canonical Correlation Analysis?

In the most general of settings, Canonical Correlation Analysis describes a multivariate statistical technique that investigates the degree of relationship between two sets of variables. It is used in analysing predictor and criterion variables simultaneously. It is particularly appropriate when the criterion variables are themselves correlated. When one criterion variable is available, Canonical Correlation Analysis reduces to multiple regression analysis. The use of canonical correlation analysis for descriptive purposes requires no distributional assumptions. In such cases, the predictor and criterion variables can be measured at the nominal or ordinal level. To test the significance of the relationships between canonical variates, however, the data should meet the requirements of multivariate normality and homogeneity

of variance.

The Population Model of Canonical Correlation Analysis

Let p be the number of predictors and q the number of criterion variables and assume $p \geq q$. Denote by $\underline{X}' = (x_1, x_2, \ldots, x_p)$ the p dimensional vector of predictor variables, and by $\underline{Y}' = (y_1, y_2, \ldots, y_q)$ the q dimensional vector of criterion variables. Suppose that $\underline{X} \sim N_p(\underline{\mu}_x, \Sigma_{xx})$ and $\underline{Y} \sim N_q(\underline{\mu}_y, \Sigma_{yy})$ where $\underline{\mu}_x$ and $\underline{\mu}_y$ denote the respective mean vectors and Σ_{xx} and Σ_{yy} are "within-set" variance-covariance matrices defined by

$$\Sigma_{xx} = E\{(\underline{X}-\underline{\mu}_x)(\underline{X}-\underline{\mu}_x)'\}, \ \Sigma_{xx} > 0$$

$$\Sigma_{yy} = E\{(\underline{Y}-\underline{\mu}_y)(\underline{Y}-\underline{\mu}_y)'\}, \ \Sigma_{yy} > 0. \qquad (1.1)$$

We shall use Σ_{xy} to denote the "between-set" covariance matrix which is defined by

$$\Sigma_{xy} = E\{(\underline{X}-\underline{\mu}_x)(\underline{Y}-\underline{\mu}_y)'\} \qquad (1.2)$$

If we define a $(p+q)$ dimensional vector, $\underline{Z}$, by

$$\underline{Z} = (\underline{X}, \underline{Y}),$$

then we can view the problem in terms of the partitioned variance-covariance matrix Σ_{zz} given by

$$\Sigma_{zz} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \qquad (1.3)$$

The objective of canonical correlation analysis is to find a linear combination of the p predictors that maximally correlates with a linear combination of the criterion variables. Let the respective linear combinations be denoted by

$$\eta = \underline{a}' \underline{X} = a_1 x_1 + a_2 x_2 + \ldots + a_p x_p \qquad (1.4)$$

and

$$\phi = \underline{b}' \underline{y} = b_1 y_1 + b_2 y_2 + \ldots + b_q y_q \qquad (1.5)$$

The correlation between η and ϕ is given by

$$\rho(\underline{a},\underline{b}) = \frac{\underline{a}' \Sigma_{xy} \underline{b}}{\{(\underline{a}' \Sigma_{xx} \underline{a})(\underline{b}' \Sigma_{yy} \underline{b})\}^{\frac{1}{2}}} \qquad (1.6)$$

Then out of the infinite number of linear combinations between X-variables and the Y-variables, we find that set of linear combinations which maximises the correlation, $\rho(\underline{a},\underline{b})$. Since $\rho(\underline{a},\underline{b})$ is invariant to scale transformations we choose $\underline{a}$ and $\underline{b}$ so that $\underline{a}' \Sigma_{xy} \underline{b}$ is maximum subject to the conditions

$$\text{(i)} \quad \underline{a}' \Sigma_{xx} \underline{a} = 1$$

and

$$\text{(ii)} \quad \underline{b}' \Sigma_{yy} \underline{b} = 1 \qquad (1.7)$$

The value of $\rho(\underline{a},\underline{b})$ obtained in this way is called the canonical correlation between $\underline{X}$ and $\underline{Y}$. The corresponding variables η and ϕ are called canonical variables.

1.3 Brief Literature Review

The theory of canonical correlation analysis was developed by Hotelling (1935, 1936) as a means of identifying the most predictable p-variate criterion, given the availability of several predictor and criterion variables.

In finance, Waugh (1942) used canonical correlation method to measure the changes in prices relative to a stable value of currency. He did this by taking for each of the years 1921 to 1940 inclusive, the prices of beef steers and hogs and the per capita consumption of beef and pork (excluding lard) for the U.S.A.

Shephard and Tanner (1962) used canonical correlation to study proximities and growth of adolescence, respectively.

In the area of educational performance, Barnett and Lewis (1963) used the procedure to analyse qualitative and quantitative data. Their study was on the academic prediction. They managed to predict the students' average grade at the university from the grades on their GCE A-levels taken (a secondary school examination required for university entrance), and the student's average grade at the university was predicted.

In the field ofmarketing, Perry and Hamm (1969) used canonical correlation analysis to relate the degree of social risk and economic risk of 25 consumer products. They found out that the higher the risk, particularly the social risk, the greater the perceived

importance of personal influence on brand choice.

Psychology is another discipline which has really utilized the technique of canonical correlation analysis. For example, Cooley and Lohnes (1971) examined the association between a set of eleven ability-type factors (for example, verbal knowledge, mathematics, visual reasoning) and a set of eleven factors dealing with career motives (for example, interest in science, interest in business).

In Ecology, Thornton (1971) applied the method of canonical correlation to study the effect of complete removal of hippopotamus on grassland in the Queen Elizabeth National Park in Uganda. In the study of growth and development in animals, Tanner, Whitehouse, Marshall, Healy and Goldstein (1975) applied the method of canonical correlation analysis in the assessment of skeletal maturity and prediction of adult height.

Further, in psychology, Wingard, Huba, and Bentler (1979) report a canonical correlation analysis of relationships between personality variables and use of various licit and illicit drugs among junior high school students in a metropolitan area.

Cohen, Gaughran, and Cohen (1979) examined the relationships between patterns of fertility across 6 different age groups (as dependent variables) and 5 different sets of demographic characteristics as independent variables (education and occupation; income-labour force; ethnicity; marriage-life cycle; and

housing and occupancy) in 5 separate canonical analyses. The units of analysis were 338 New York City health areas.

Gittins (1979) used canonical correlation analysis to investigate the relationships between variables of two distinct but associated kinds in Ecology. He applied the method to study the connections between the occurence of plant or animal communities (and their component species) and soil or other environmental variables of several areas.

1.4 Aims and Significance of the study

Canonical correlation analysis is a technique which has not been so widely used in analysing educational performance. In contributing to the few applications of the technique available, the present study aims at:

(a) studying the interrelationships among different educational variables, in particular courses offered by the Mathematics department of Kenyatta University;

(b) showing how the results of canonical correlation analysis may be interpreted and the kind of information provided;

(c) illustrating the kind of opportunities offered by canonical correlation analysis in the analysis of academic data and so contribute towards an improved definition of the role of the method in academics;

(d) assessing the ability of canonical correlation analysis to recover known relationships between mathematics units of statistical interest;

(e) comparing the results of the analysis for two consecutive academic years that is, 1984/85 and 1985/86.

Consequently the entire study will

(a) raise many questions on interpretation of results and thus make more statisticians to venture into the technique and thus improving its applicability;

(b) act as a reference material for those involved in studying relationships between educational variables;

(c) help postgraduate students in statistics to develop their insight on the usefulness of the technique;

(d) provide guidance on how to make use of existing information (data) to arrive at useful conclusions.

CHAPTER II

SAMPLE CANONICAL CORRELATION AND CANONICAL VARIATES

2.1 Introduction

Let $\underline{X}$ be a (p+q) - component random vector. Let $\underline{X}_1, \underline{X}_2, \ldots\ldots, \underline{X}_N$ be a sample of size N from $\underline{X}$. Further, let $\underline{X}_r$, r=1,2,.....,N be partitioned into two subvectors $\underline{X}_r^{(1)}$ of p variables of X and $\underline{X}_r^{(2)}$ of q variables of Y; where p+q = n.

Then

$$\underline{X}_r = \begin{bmatrix} \underline{X}_r^{(1)} \\ \underline{X}_r^{(2)} \end{bmatrix}; \quad r = 1,2,\ldots\ldots,N$$

where

$$\left[\underline{X}_r^{(1)}\right] = \left[X_{r1}, X_{r2}, \ldots\ldots, X_{rp}\right]$$

$$\left[\underline{X}_r^{(2)}\right] = \left[X_{rp+1}, \ldots\ldots\ldots, X_{rn}\right]$$

For convenience we will assume p>q.

The observed sample values on the variates may be written collectively in matrix form as

$$X = (\underline{X}_1, \underline{X}_2, \ldots\ldots\ldots, \underline{X}_N)'$$

The Nxn matrix X is called the data matrix of the sample.

2.2 Sample Estimates

Let us define

$$\bar{X}_i = \frac{1}{N}\sum_{r=1}^{N} X_{ri}; \quad i=1,2,.....,n.$$

This is the sample mean of the N observations on the ith variable of $\underline{X}$. The vector

$$\underline{\bar{X}}' = (\bar{X}_1, \bar{X}_2, ... \bar{X}_n)$$

represents the n-sample means of the n-variables of $\underline{X}$. It is called the sample mean vector. We can write $\underline{\bar{X}}$ in the form

$$\underline{\bar{X}} = \frac{1}{N}\sum_{r=1}^{N} \underline{X}_r'$$

$$= \frac{1}{N}\underline{X}'\underline{1}$$

where $\underline{1}$ is an (Nx1) vector of ones.

The sample variance of the ith variable is given by

$$S_{ii} = \frac{1}{N}\sum_{r=1}^{N}(X_{ri} - \bar{X}_i)^2$$

and the sample covariance between the ith and jth variable; ($i \neq j$) is given by:

$$S_{ij} = \frac{1}{N}\sum_{r=1}^{N}(X_{ri} - \bar{X}_i)(X_{rj} - \bar{X}_j)$$

The matrix

$$S = (S_{ij}); \quad i,j = 1,2,.....,n$$

is called the sample covariance matrix or the sample dispersion

matrix. Its diagonal elements S_{ii} are sample variances and the off diagonal elements are the covariances. Let the diagonal matrix of sample variances be given by

$$S_{\Delta} = \text{diag}(S_1^2, S_2^2, \ldots, S_n^2),$$

then the r-th vector of Standardised variables denoted by $\underline{Z}_r$ is

$$\underline{Z}_r = S_{\Delta}^{-\frac{1}{2}}(\underline{X}_r - \underline{\bar{X}}).$$

The variance covariance matrix of $\underline{Z}_r$ is the nxn matrix R, given by

$$R = \frac{1}{N}\sum_{r=1}^{N}\begin{bmatrix}\underline{Z}_r^{(1)}\\ \underline{Z}_r^{(2)}\end{bmatrix}\left[\underline{Z}_r^{(1)}, \underline{Z}_r^{(2)}\right]$$

$$= \begin{bmatrix}R_{11} & R_{12}\\ R_{21} & R_{22}\end{bmatrix}$$

where

$$R_{11} = \text{Var}(\underline{Z}_r^{(1)}), \qquad R_{22} = \text{Var}(\underline{Z}_r^{(2)})$$

$$R_{12} = R_{21}' = \text{Cov}(\underline{Z}_r^{(1)}, \underline{Z}_r^{(2)})$$

are matrices of orders pxp, qxq and pxq respectively. Note that the matrix R is the sample correlation matrix.

The inter-relationships within and between the variables composing $\underline{Z}_r^{(1)}$ and $\underline{Z}_r^{(2)}$ are specified by the correlation matrix R of $\underline{Z}_r$.

Usually the variables making up the data matrix are first standardised to have unit variance so that the variance-covariance matrix is the correlation matrix.

2.3 Derivation of Fundamental Equations for Canonical Correlations and Variates.

We now seek linear transformations of each set of variables to new variates η_k and ϕ_k, the canonical variates, for which the correlation matrix has a particular simple and appealing form. Thus, we require that:

(a) all the η_k be uncorrelated with one another;

(b) all the ϕ_k be uncorrelated with one another; and

(c) that the pairs of canonical variates η_k, ϕ_m for $k,m = 1,2,.....,q$; be maximally correlated for $k=m$ and zero otherwise.

Let the linear combinations of $\underline{z}_r^{(1)}$ and $\underline{z}_r^{(2)}$ be

$$\eta = a_1 z_{r1} + \ldots\ldots\ldots\ldots + a_p z_{rp} = \underline{a}' \underline{z}_r^{(1)}$$

$$\phi = b_{p+1} z_{rp+1} + \ldots\ldots\ldots + b_n z_{rn} = \underline{b}' \underline{z}_r^{(2)}$$

where the coefficient vectors

$$\underline{a}' = (a_1, a_2, \ldots\ldots, a_p)$$

and

$$\underline{b}' = (b_{p+1}, b_{p+2}, \ldots., b_n)$$

are chosen to maximise the simple correlation, r, between η and ϕ.

The Correlation r, between η and ϕ expressed as a function of $\underline{a}$ and $\underline{b}$ is

$$r(\underline{a},\underline{b}) = \frac{\mathrm{Cov}(\underline{a}'\underline{Z}^{(1)},\underline{b}'\underline{Z}^{(2)})}{\{\mathrm{Var}(\underline{a}'\underline{Z}^{(1)}).\mathrm{Var}(\underline{b}'\underline{Z}^{(2)}\}^{\frac{1}{2}}}$$

$$= \frac{\underline{a}'R_{12}\underline{b}}{\{(\underline{a}'R_{11}\underline{a}).(\underline{b}'R_{22}\underline{b})\}^{\frac{1}{2}}} \qquad (2.1)$$

We require that $\underline{a}$ and $\underline{b}$ be such that η and ϕ have unit variances. That is

$$\mathrm{Var}(\eta) = \underline{a}'R_{11}\underline{a} = 1$$
$$\qquad (2.2)$$
$$\mathrm{Var}(\phi) = \underline{b}'R_{22}\underline{b} = 1$$

From equations (2.1) and (2.2) we have

$$r = \underline{a}'R_{12}\underline{b} \qquad (2.3)$$

Thus the problem is to find $\underline{a}$ and $\underline{b}$ which maximise (2.3) subject to (2.2).

Using Lagranges multipliers, we define the Lagrange function:

$$F(\underline{a},\underline{b},\lambda,\mu) = \underline{a}'R_{12}\underline{b} - \frac{\lambda}{2}(\underline{a}'R_{11}\underline{a}-1) - \frac{\mu}{2}(\underline{b}'R_{22}\underline{b}-1$$

where λ and μ are Lagrange multipliers.

To maximise $F(\underline{a},\underline{b},\lambda,\mu)$ we differentiate $F(\underline{a},\underline{b},\lambda,\mu)$ first with respect to the elements of $\underline{a}$ and then those of $\underline{b}$. On setting the results equal to zero we have

$$\frac{\partial F}{\partial \underline{a}} = 0 \iff R_{12}\underline{b} - \lambda R_{11}\underline{a} = \underline{0} \qquad (2.4)$$

$$\frac{\partial F}{\partial \underline{b}} = 0 \iff R_{21}\underline{a} - \mu R_{22}\underline{b} = \underline{0} \tag{2.5}$$

Premultiplying (2.4) by $\underline{a}'$ and (2.5) by $\underline{b}'$ gives

$$\begin{aligned} \underline{a}'R_{12}\underline{b} - \lambda \underline{a}'R_{11}\underline{a} &= \underline{0} \\ \underline{b}'R_{21}\underline{a} - \mu \underline{b}'R_{22}\underline{b} &= \underline{0} \end{aligned} \tag{2.6}$$

Now, because

$$\underline{a}'R_{11}\underline{a} = \underline{b}'R_{22}\underline{b} = 1,$$

equations (2.6) reduce to

$$\underline{a}'R_{12}\underline{b} = \lambda = \mu$$

From (2.5) we obtain

$$\underline{b} = \frac{1}{\lambda}R_{22}^{-1}R_{21}\underline{a} \tag{2.7}$$

substituting for $\underline{b}$ in (2.4) we get

$$\frac{1}{\lambda}R_{12}R_{21}\underline{a} - \lambda R_{11}\underline{a} = \underline{0} \tag{2.8}$$

That is

$$R_{12}R_{22}^{-1}R_{21}\underline{a} - \lambda^2 R_{11}\underline{a} = \underline{0},$$

or equivalently

$$(R_{12}R_{22}^{-1}R_{21} - \lambda^2 R_{11})\underline{a} = \underline{0} \tag{2.9}$$

There will be a non-zero solution for $\underline{a}$ in (2.9) if and only if

$$|R_{12}R_{22}^{-1}R_{21} - \lambda^2 R_{11}| = 0$$

That is if and only if

$$|R_{11}^{-1}R_{12}R_{22}^{-1}R_{21} - \lambda^2 I| = 0 \qquad (2.10)$$

Therefore λ^2 is an eigen value of the matrix $R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$. Let $\ell_1^2 \geq \ell_2^2 \geq \ldots\ldots \geq \ell_q^2$ be the eigen values of the matrix, then,

$$\text{Max } \lambda^2 = \underset{\underline{a},\underline{b}}{\text{Max}}(\underline{a}'R_{12}\underline{b}) = \ell_1^2$$

The first canonical correlation is $r_1 = \ell_1$. Let $\underline{a}'$ be the eigen vector corresponding to ℓ_1^2. Then

$$\eta_1 = \underline{a}_1'\underline{Z}_r^{(1)}.$$

Similarly, eliminating $\underline{a}$ from (2.4) and substituting in (2.5) gives

$$|R_{21}R_{11}^{-1}R_{12} - \lambda^2 R_{22}| = 0,$$

that is

$$|R_{22}^{-1}R_{21}R_{11}^{-1}R_{12} - \lambda^2 I| = 0. \qquad (2.11)$$

Therefore, λ^2 is a charasteristic root of the matrix

$$R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$$

The number of non-zero characteristic roots of $R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ and $R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$ is the same. So again ℓ_1^2 is the maximum characteristic root of $R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$. Let $\underline{b}_1$ be the corresponding eigen vector, then

$$\phi_1 = \underline{b}_1'\underline{Z}_r^{(2)}.$$

Thus, the first pair of canonical variables is

$$\eta_1 = \underline{a}_1'\underline{Z}_r^{(1)} \quad \text{and} \quad \phi_1 = \underline{b}_1'\underline{Z}_r^{(2)}$$

and the corresponding canonical correlation is $r_1 = \ell_1$, where

$$\ell_1^2 = \text{Max. characteristic root of } R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$$

$$= \text{Max. characteristic root of } R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$$

and $\underline{a}_1$ and $\underline{b}_1$ are the eigen vectors corresponding to ℓ_1^2 in $R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ and $R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$, respectively.

The second pair of canonical variables is given by

$$\eta_2 = \underline{a}'\underline{Z}_r^{(1)} \quad \text{and} \quad \phi_2 = \underline{b}'\underline{Z}_r^{(2)}$$

where $\underline{a}$ and $\underline{b}$ are such that $\underline{a}'R_{12}\underline{b}$ is a maximum subject to

(i) $\underline{a}'R_{11}\underline{a} = 1$

(ii) $\underline{b}'R_{22}\underline{b} = 1$

(iii) (η_2, ϕ_2) are uncorrelated with (η_1, ϕ_1).

In general the (k+1)-th pair of canonical variates is given by

$$\eta_{k+1} = \underline{a}'\underline{Z}_r^{(1)} \quad \text{and} \quad \phi_{k+1} = \underline{b}'\underline{Z}_r^{(2)},$$

where $\underline{a}$ and $\underline{b}$ are such that $\underline{a}'R_{12}\underline{b}$ is a maximum subject to

(i) $\underline{a}'R_{11}\underline{a} = 1$

(ii) $\underline{b}'R_{22}\underline{b} = 1$

(iii) (η_{k+1}, ϕ_{k+1}) are uncorrelated with

(η_i, ϕ_i) for all $i = 1,2,\ldots,k$;

where

$$\eta_i = \underline{a}'_i \underline{Z}_r^{(1)} \quad \text{and} \quad \phi_i = \underline{b}'_i \underline{Z}_r^{(2)}$$

Note that the pair (η_i, ϕ_i) of canonical variates will satisfy equations (2.4) and (2.5). That is

$$R_{12}\underline{b}_i - \lambda R_{11}\underline{a}_i = \underline{0} \tag{2.12}$$

$$R_{21}\underline{a}_i - \mu R_{22}\underline{b}_i = \underline{0} \tag{2.13}$$

where $\underline{a}_i$ and $\underline{b}_i$ are the eigen vectors at the i-th step.

Now, η_{k+1} uncorrelated with η_i implies that

$$\mathrm{Cov}(\underline{a}'\underline{Z}_r^{(1)}, \underline{a}'_i \underline{Z}_r^{(1)}) = \underline{a}' R_{11}\underline{a}_i = 0.$$

It follows from (2.12) that

$$\underline{a}' R_{12}\underline{b}_i = 0.$$

That is η_{k+1} is uncorrelated with ϕ_i as well. Next ϕ_{k+1} uncorrelated with ϕ_i implies that

$$\mathrm{Cov}(\underline{b}'\underline{Z}_r^{(2)}, \underline{b}'_i \underline{Z}_r^{(2)}) = \underline{b}' R_{22}\underline{b}_i = 0.$$

It follows from (2.13) that

$$\underline{b}' R_{21}\underline{a}_i = 0.$$

That is ϕ_{k+1} is uncorrelated with η_i as well, for all $i = 1,2,.....,k$.

Therefore, the condition that (η_{k+1}, ϕ_{k+1}) be uncorrelated with (η_i, ϕ_i) can be replaced with

$$\underline{a}'R_{11}\underline{a}_i = 0 \quad \text{and} \quad \underline{b}'R_{22}\underline{b}_i = 0; \text{ for all } i = 1,2,.....,k.$$

The problem can be restated now as

$$\underset{\underline{a},\underline{b}}{\text{Maximise}} \quad \underline{a}'R_{12}\underline{b}$$

subject to

$$\text{(i)} \quad \underline{a}'R_{11}\underline{a} = 1$$

$$\text{(ii)} \quad \underline{b}'R_{22}\underline{b} = 1$$

$$\text{(iii)} \quad \underline{a}'R_{11}\underline{a}_i = 0; \quad i = 1,2,.....,k$$

$$\text{(iv)} \quad \underline{b}'R_{22}\underline{b}_i = 0; \quad i = 1,2,.....,k.$$

Using Lagranges multipliers, let us difine

$$\begin{aligned} F(\underline{a},\underline{b},\lambda,\mu,\underline{\alpha},\underline{\theta}) = {} & \underline{a}'R_{12}\underline{b} - \lambda/2(\underline{a}'R_{11}\underline{a} - 1) \\ & - \mu/2(\underline{b}'R_{22}\underline{b} - 1) - \sum_{i=1}^{k} \alpha_i(\underline{a}'R_{11}\underline{a}_i) \\ & - \sum_{i=1}^{k} \theta_i(\underline{b}'R_{22}\underline{b}_i) \end{aligned} \tag{2.14}$$

Differentiating (2.14) with respect to $\underline{a}, \underline{b}, \lambda, \mu, \alpha_i$ and θ_i respectively and equating the results to zero we obtain

$$\frac{\partial F}{\partial \underline{a}} = 0 = R_{12}\underline{b} - \lambda R_{11}\underline{a} - \sum_{i=1}^{k} \alpha_i R_{11}\underline{a}_i = \underline{0} \tag{2.15}$$

$$\frac{\partial F}{\partial \underline{b}} = 0 = R_{21}\underline{a} - \mu R_{22}\underline{b} - \sum_{i=1}^{k} \theta_i R_{22}\underline{b}_i = \underline{0} \qquad (2.16)$$

$$\frac{\partial F}{\partial \lambda} = 0 = \underline{a}'R_{11}\underline{a} = 1$$

$$\frac{\partial F}{\partial \mu} = 0 = \underline{b}'R_{22}\underline{b} = 1$$

$$\frac{\partial F}{\partial \alpha_i} = 0 = \underline{a}'R_{11}\underline{a}_i = 0, \; i = 1,2,\ldots\ldots,k. \qquad (2.17)$$

$$\frac{\partial F}{\partial \theta_i} = 0 = \underline{b}'R_{22}\underline{b}_i = 0, \; i = 1,2,\ldots\ldots,k. \qquad (.218)$$

Premultiplying (2.15) by $\underline{a}'_j$; for fixed j we get

$$\underline{a}'_j R_{12}\underline{b} - \lambda \underline{a}'_j R_{11}\underline{a} - \alpha_j \underline{a}'_j R_{11}\underline{a}_j = 0$$

which implies that

$$\alpha_j = 0, \; \forall_j = 1,2,\ldots\ldots,k.$$

Similarly, premultiplying (2.16) by $\underline{b}_j$, for fixed j we get

$$\underline{b}'_j R_{21}\underline{a} - \mu \underline{b}'_j R_{22}\underline{b} - \theta_j \underline{b}'_j R_{22}\underline{b}_j = 0$$

which implies that

$$\theta_j = 0, \quad \forall_j = 1,2,\ldots\ldots,k.$$

Thus, the system of equations reduces to

$$R_{12}\underline{b} - \lambda R_{11}\underline{a} = \underline{0}$$

$$R_{21}\underline{a} - \mu R_{22}\underline{b} = \underline{0}$$

$$\underline{a}'R_{11}\underline{a} = 1$$

$$\underline{b}'R_{22}\underline{b} = 1$$

which are the same as equations given originally. Therefore, the (k+1)-th pair of canonical variates is

$$\eta_{k+1} = \underline{a}'_{k+1}\underline{Z}_r^{(1)} \quad \text{and} \quad \phi_{k+1} = \underline{b}'_{k+1}\underline{Z}_r^{(2)}$$

where $\underline{a}_{k+1}$ is the eigen vector corresponding to ℓ^2_{k+1} in $R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ and $\underline{b}_{k+1}$ is the eigen vector corresponding to ℓ^2_{k+1} in $R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$.

The corresponding canonical correlation is $r_{k+1} = \ell_{k+1}$.

CHAPTER III

TESTS OF SIGNIFICANCE AND INTERPRETIVE DEVICES

3.1 Tests of Significance

For purposes of testing hypotheses we shall assume that the observations are obtained from a normal population.

(a) Joint Nullity of all the q Canonical Correlations

We know that $-1 \leq r_k \leq 1$, where r_k is the k^{th} canonical correlation coefficient. It is often necessary to test the simultaneous departure of the canonical correlation coefficients from zero. Joint nullity of all the q canonical correlations would indicate the absence of any linear relationship between the two or more sets of variables. Suppose that we wish to test the null hypothesis that, say, the X-variables set is uncorrelated with the Y-variables sets that is,

$$H_o: \rho_1 = \rho_2 = \ldots = \rho_q = 0$$

against

$$H_1: \rho_k \neq 0 \text{ for some } k,\ k = 1,2,\ldots,q,$$

where ρ_k is the k^{th} population canonical correlation; or equivalently

$$H_o: \Sigma_{12} = \underline{0}$$

against

$$H_1: \Sigma_{12} \neq \underline{0}$$

where Σ_{12} is the matrix of intercorrelations. Then the likelihood ratio test statistic is given by

$$\Lambda = \left| I - R_{22}^{-1} R_{21} R_{11}^{-1} R_{12} \right|$$

$$= \prod_{k=1}^{q} (1-r_k^2)$$

where Λ is the Wilks' lambda statistic. The range of Λ is 0 to 1. It is apparent that Λ may be regarded as the product of the proportion of variance left unexplained by the q canonical correlations. We see that if there is little correlation between the two sets of varables, Λ will be close to unity, while if they are closely correlated Λ will approach zero.

Bartlett's chi-square approximation for the distribution of Λ is

$$\chi^2 = -\left[N-\tfrac{1}{2}(p+q+3)\right]\log_e \Lambda.$$

Under the null hypothesis χ^2 is distributed approximately as a chi-squared variable with pq degrees of freedom asymptotically as $N \to \infty$. The hypothesis is rejected if

$$\chi^2 > \chi_\alpha^2(pq),$$

where α is the size of the test. If the null hypothesis (no relationship) can be rejected, the contribution of the first pair of canonical variates can be removed from Λ and the statistical significance of the remaining pairs of canonical variates assessed. It is generally of interest to remove the contribution of the largest root, the first two roots, and so on, from Λ and then to assess the significance of the remaining canonical correlations.

(b) Joint Nullity of the Smallest q-k Canonical Correlations

A general criterion for testing the joint nullity of the canonical correlations $r_k, r_{k+1}, \ldots, r_q$ is given by

$$\Lambda_k = \prod_{i=k}^{q} (1-r_i^2) .$$

The significance of Λ_k may be assessed by the chi-square approximation given by

$$\chi^2 = -\left[N-\tfrac{1}{2}(p+q+3)\right] \log_e \Lambda_k ,$$

where χ^2 is distributed approximately as a chi-squared variable with $(p-k+1)(q-k+1)$ degrees of freedom. Λ_k provides a test of the residuals after the effects of the preceding correlations have been removed.

This test will eventually give us the number of useful and interpretable canonical variates. It is generally accepted that if the overall test is significant then r_1 at least must be significant. The significance of r_1 itself is usually never directly tested.

(c) The Significance of Individual Canonical Correlation Coefficients.

An alternative to the likelihood ratio tests of the hypotheses considered above is arrived at through the application of Roy's union intersection procedure. In this context we write the multivariate hypothesis.

$$H_0: \rho_1 = \rho_2 = \ldots = \rho_q = 0$$

against

$$H_1: \rho_k \neq 0 \quad \text{for some} \quad k, \quad k = 1,2,..q.$$

as an intersection of composite univariate hypotheses and as a union of corresponding alternative hypotheses respectively, that is

$$H_0: \bigcap_{\underline{a},\ \underline{b}} (\rho(\underline{a},\underline{b}) = 0)$$

against

$$H_1: \bigcup_{\underline{a},\ \underline{b}} (\rho(\underline{a},\underline{b}) \neq 0)$$

where $\rho(\underline{a}, \underline{b})$ is the simple correlation between η and ϕ (the linear composites $\underline{a}'\underline{z}_r^{(1)}$ and $\underline{b}'\underline{z}_r^{(2)}$ respectively).

The test criterion for the hypothesis above is the square of the maximum attainable correlation between the linear composites $\underline{a}'\underline{z}_r^{(1)}$ and $\underline{b}'\underline{z}_r^{(2)}$ for all choices of $\underline{a}$ and $\underline{b}$ subject to the normalisation condition $\underline{a}'R_{11}\underline{a} = \underline{b}'R_{22}\underline{b} = 1$. This quantity is the largest root, r_1^2, of

$$(R_{22}^{-1} R_{21} R_{11}^{-1} R_{12} - \lambda^2 I)\underline{b} = \underline{0}.$$

The significance of r_1^2 may be tested by referring it to critical points of the greatest characteristic root (gcr) distribution defined as follows:

Let

$$V_1 \sim W_p(I, m_1) \quad \text{and} \quad V_2 \sim W_p(I, m_2);\ m_1 \geq p,$$

be two independent Wishart matrices. Then the largest eigenvalue θ of $(V_1+V_2)^{-1}V_2$ is called the greatest root statistic and its distribution denoted by

$$\theta(p, m_1, m_2).$$

Note that θ can also be defined as the largest root of the determinantal equation

$$|V_2-\theta(V_1+V_2)| = 0 .$$

If λ is an eigenvalue of $V_1^{-1} V_2$, then $\lambda/(1+\lambda)$ is an eigenvalue of $(V_1+V_2)^{-1}V_2$. Since this is a monotonic function of λ, θ is given by

$$\theta = \frac{\lambda_1}{1+\lambda_1} ,$$

where λ_1 is the largest eigenvalue of $V_1^{-1} V_2$. Clearly, since $\lambda_1 > 0$, we see that $0 < \theta < 1$.

Therefore, the null hypothesis is accepted at the level α if

$$r_1^2 \leq \theta_\alpha(q,m_1,m_2)$$

and rejected otherwise. Here $\theta_\alpha(q,m_1,m_2)$ is the upper 100α percentage point of the $_{qcr}$ distribution with parameters q, $m_1 = (|p-q|-1)/_2$ and $m_2 = (N-p-q-2)/_2$.

If the overall hypothesis of independence given above is rejected, the significance of the k^{th} root may be tested $(k=2,\ldots,q)$. The procedure is identical to the overall test using r_1^2, but an adjustment to the first degree-of-freedom parameter however is required. The parameter q_k is given by $q_k = \min(p-k+1, q-k+1)$.

(d) The Contribution of Particular Variables

The significance of the contribution of particular variables to the canonical relationship can be tested by a modification of the likelihood ratio test procedure described above. The test is accomplished by comparing the χ^2 values which result when the variables of

interest are first included and then omitted. The value of Λ is first calculated from

$$\Lambda = \prod_{k=1}^{q} (1-r_k^2)$$

with all the variables included. Now, suppose K_1 X's and K_2 Y's are omitted. The value of Λ is then recalculated for the remaining $(p-k_1)$ plus $(q-k_2)$ variables. Denoting this second value by Λ^*, the quantity

$$\chi^2 = -\left[N-\tfrac{1}{2}(p+q+3)\right] \log_e \Lambda^*$$

is distributed approximately as a chi-squared variate with $(p-k_1)(q-k_2)$ degrees of freedom. The test statistic is then given by

$$\chi^2 = -\left[N-\tfrac{1}{2}(p+q+3)\right] \log_e (\Lambda/\Lambda^*),$$

which has an approximate chi-square distribution with $pk_2+qk_1-k_1k_2$ degrees of freedom.

3.2 INTERPRETIVE DEVICES

(a) Canonical Correlation Coefficients (r_k)

These are product-moment correlation coefficients between the k^{th} pair of canonical variates, η_k and ϕ_k. They can also be interpreted as multiple correlation coefficients between a particular canonical variate of one set and the complete set of variables of the other. The magnitude of r_k expresses the degree of linear correlation between η_k and ϕ_k. They are dimensionless quantities and hence are invariant under non-singular linear transformations of the variables of either

or both sets.

A squared canonical correlation coefficient, r_k^2, represents the ratio of two determinants or generalized variances - namely the ratio of the generalized total variance. The square of a canonical correlation is also interpretable as the overlapping variance between the k^{th} pair of canonical variates. It should be noted that canonical correlation coefficients calculated from the sums of squares and product matrix, the variance-covariance matrix and the correlation matrix for a particular set of data are the same.

The magnitude of the sample correlation, r_k, does depend to a large extent on the relative number of variables and sample involved. As the number of variables, $(p+q)$, approaches the sample size N, the value of r_1 (the first canonical correlation) tends rapidly to unity. When $(p+q) \geq N$ one or more canonical correlations of unity will inevitably arise.

Apart from measuring the relationship between canonical variates, canonical correlations have two other uses. First they provide an indication of the dimensionality of linear relationship between the measurement domains. That is, they help in determining the number of useful and interpretable canonical variates. Secondly, they are used in the construction of further interpretive devices.

(b) Canonical Weights

These are elements of the vectors $\underline{a}_k$ and $\underline{b}_k$. They serve to transform the original variables so that the correlation between the two sets of variables is maximal. The magnitude of the weight tells us the importance of a variable from one set with regard to the other set in obtaining a maximum correlation between the sets.

The numerical values of canonical weights depend on the selection of variables as well as on their scale. Addition or deletion of variables in either set is likely to produce major alterations in the remaining coefficients. Standardising the observed variables to zero mean and unit variance will remove the scaling effects but the inter-dependencies still remain.

As in multiple regression, the canonical weight coefficients may be highly unstable due to multicollinearity. Thus, some variable may have a small or even negative weight because the variance in the variable has already been accounted for by some other variable(s). The adverse effects of collinearity on the algebraic sign and magnitude of the canonical weight coefficient give a rather warped view of the relevance of the variables.

(c) Canonical Loadings (Intraset Correlation Coefficients)

A canonical loading (an intraset correlation coefficient) gives the ordinary product-moment correlation

of the original variable and its respective canonical variate. Thus, it reflects the degree to which a variate is represented by a canonical variate. More precisely, a canonical loading gives the correlation between a canonical variate and an observed variable of the same set.

There are two sets of intraset correlations corresponding to the two measurement domains. The vector of canonical loadings associated with the k^{th} canonical variate can be obtained by premultiplying the vector of canonical weights by the appropriate matrix of within-set correlations. For the variables and canonical variates of $\underline{z}_r^{(1)}$ the canonical loadings are given by

$$\begin{aligned} \text{Corr}(\underline{z}_r^{(1)}, \eta_k) &= \text{Corr}(\underline{z}_r^{(1)}, \underline{a}'_k \underline{z}_r^{(1)}) \\ &= \frac{1}{N} \sum_{r=1}^{N} \underline{z}_r^{(1)} \left[\underline{z}_r^{(1)}\right]' \underline{a}_k \\ &= R_{11}\, \underline{a}_k \\ &= \underline{\varepsilon}_k^{(1)} \quad \text{(say)}, \end{aligned} \tag{3.1}$$

where $\underline{\varepsilon}_k^{(1)}$ is the pxl vector of correlations between the k-th canonical variate of $\underline{z}_r^{(1)}$ and the observed variables of $\underline{z}_r^{(1)}$.

Similarly, for the Y-variables set we have

$$\text{Corr}(\underline{z}_r^{(2)}, \phi_k) = R_{22}\, \underline{b}_k = \underline{\varepsilon}_k^{(2)} \quad \text{(say)} \tag{3.2}$$

(d) Cross Loadings (Interset Correlation Coefficients)

A cross-loading gives the relationship between an observed variable from one set with a canonical variate from the other set.

There are two sets of cross-loadings corresponding to the two measurement domains. For the correlations of the variables of $\underline{z}^{(1)}$ with the canonical variates of $\underline{z}^{(2)}$ we have

$$\begin{aligned} \text{Corr}(\underline{z}_r^{(1)}, \phi_k) &= \text{Corr}(\underline{z}_r^{(1)}, \underline{b}_k' \underline{z}_r^{(2)}) \\ &= \frac{1}{N} \sum_{r=1}^{N} \underline{z}_r^{(1)} \left[\underline{z}_r^{(2)}\right]' \underline{b}_k \\ &= R_{12} \underline{b}_k = \underline{\varepsilon}_k^{(x\phi)} \quad \text{(say)} \end{aligned} \tag{3.3}$$

where $\underline{\varepsilon}_k^{(x\phi)}$ is the pxl vector of interset correlations between the k-th canonical variate, ϕ_k, of $\underline{z}^{(2)}$ and the observed variables of $\underline{z}^{(1)}$.

From equation (2.4) it is easily seen that

$$R_{12} \underline{b}_k = r_k \; R_{11} \underline{a}_k. \tag{3.4}$$

Thus, cross-loadings can be obtained by taking the product of the canonical correlation coefficient and the canonical loading.

In a similar way,

$$\begin{aligned} \text{Corr}(\underline{z}_r^{(1)}, \eta_k) &= R_{21} \underline{a}_k \\ &= r_k \; R_{22} \underline{b}_k = \underline{\varepsilon}_k^{(y\eta)} \quad \text{(say)} \end{aligned} \tag{3.5}$$

where $\underline{\varepsilon}_k^{(y\eta)}$ is the qxl vector of correlations between the k-th canonical variate, η_k, of $\underline{z}^{(1)}$ and the observed variables in $\underline{z}^{(2)}$.

The attractive feature of cross-loadings is that they isolate the relationship of each variable

separately with the canonical variate from the other set. The cross-loadings are more conservative, less inflated than within-set loadings.

(e) Proportion of Explained Variance (Variance Extracted by a Canonical Variate)

The proportion of the total variance of a measurement domain which is associated with a canonical variate is referred to as the variance extracted or accounted for by the variate. It is an index of the extent to which a canonical variate explains the total variance of the domain of which it is a linear composite.

The variance extracted by a canonical variate is assessed as the sum of squared loadings on a variate divided by the number of variables in the set.

The proportion of explained variance in the X-set variables that is accounted for by a particular canonical variate is given by

$$R^2_{(k)x} = \sum_{i=1}^{p} \varepsilon^2_{ik/p} , \qquad (3.6)$$

where $R^2_{(k)x}$ denotes the proportion of variance in the x-set variables accounted for by a particular canonical variate ε^2_{ik} is the ith squared intraset correlation with η_k, that is, the kth canonical variate. Similarly, the proportion of variance in the Y-set accounted for by a particular canonical variate, denoted by $R^2_{(k)y}$, can be computed by

$$R^2_{(k)y} = \sum_{h=1}^{q} \varepsilon^2_{hk} /q . \qquad (3.7)$$

A pair of canonical variates may extract very different amounts of variance from their respective sets of variates. Because canonical variates are independent of one another (orthogonal), percentages of variance can be summed across variates to arrive at the total variance extracted from the variables by all significant canonical variates.

(f) Redundancy

Redundancy is the proportion of the total variance of a measurement domain predictable from a linear composite of the other domain, given the availability of the second domain.

The redundancy, $R^2_{(k)x/y}$, of $\underline{z}^{(1)}$ with respect to the k-th canonical variate $\phi_k = \underline{b}'_k \underline{z}^{(2)}$ of $\underline{z}^{(2)}$, given the availability of $\underline{z}^{(2)}$, is given by the mean of the squared interset correlations (cross-loadings) between the elements of $\underline{z}^{(1)}$ and ϕ_k. That is

$$R^2_{(k)x/y} = \sum_{i=1}^{p} \varepsilon^2_{ik/p} \quad , \tag{3.8}$$

where ε_{ik} is the correlation between the ith variable of $\underline{z}^{(1)}$ and the k-th canonical variate of $\underline{z}^{(2)}$.

Similary, the redundancy, $R^2_{(k)y/x}$, of $\underline{z}^{(2)}$ given the availability of $\underline{z}^{(1)}$, is given by

$$R^2_{(k)y/x} = \sum_{h=1}^{q} \varepsilon^2_{hk/q} \quad , \tag{3.9}$$

where ε_{hk} is the correlation between the h-th variable of $\underline{z}^{(2)}$ and the k-th canonical variate of $\underline{z}^{(1)}$.

Alternatively, redundancy may be regarded as the product of the within-set variance times the between-set variance accounted for by a canonical variate. In other words, redundancy is the product of the variance extracted by a canonical variate from its own domain times the the variance which the canonical variate shares with its counterpart of the domain. Thus, we write

$$R^2_{(k)x/y} = R^2_{(k)x}\, r^2_k$$

for the redundancy of $\underline{z}^{(1)}$, given the availability of the set $\underline{z}^{(2)}$, and

$$R^2_{(k)y/x} = R^2_{(k)y}\, r^2_k$$

for the redundancy of $\underline{z}^{(2)}$ given the availability of $\underline{z}^{(1)}$.

In more general terms, redundancy answers the question: If I knew the score on a canonical variate from one set, how much would my uncertainty regarding the other set be reduced? It is possible, for instance, for a canonical variate from one set to be an important dimension in its own set of variables, but correlated with an unimportant dimension among the other set (and vice versa). Therefore the redundancies for a pair of canonical variates are not usually equal. That is

$$R^2_{(k)x/y} \neq R^2_{(k)y/x} .$$

Because canonical variates are orthogonal, redundancies for a group of variables can be added

across canonical variates to get a total for the variables and, thus, a total redundancy measure for one set relative to the other, and vice versa.

For the total redundancy in the set $\underline{z}^{(1)}$, given the canonical variates $\phi_1, \phi_2, \ldots, \phi_m$

$$R^2_{(T)x/y} = \sum_{k=1}^{m} \left(\sum_{i=1}^{p} \varepsilon^2_{ik/p} \right)$$

$$= \sum_{k=1}^{m} R^2_{(k)x/y} \quad , \tag{3.10}$$

where m is the number of retained canonical variates.

Similarly, the total redundancy in $\underline{z}^{(2)}$, given the canonical variates $\eta_1, \eta_2, \ldots, \eta_m$ of $\underline{z}^{(1)}$ is given by

$$R^2_{(T)y/x} = \sum_{k=1}^{m} \left(\sum_{h=1}^{q} \varepsilon_{hk/q} \right)$$

$$= \sum_{k=1}^{m} R^2_{(k)y/x} \tag{3.11}$$

This index provides an overall measure of the variance explained in one variable set by the variables of the other. Total redundancy is asymmetric between domains, to that in general

$$R^2_{(T)x/y} \neq R^2_{(T)y/x} \, .$$

(g) Variable Communalities

(i) Intraset Variable Communalities

These are the proportions of variance accounted for by the retained canonical variates of the variables'

own set.

These are obtained as the sum of squared intraset correlations (loadings) between a variable and the retained canonical variates. For the i-th variable of $\underline{z}^{(1)}$, the intraset or within-set communality, h^2_{wi}, is therefore

$$h^2_{wi} = \sum_{k=1}^{m} \varepsilon^2_{ik} , \qquad (3.12)$$

where ε_{ik} is the intraset correlation of the i-th variable with the k-th canonical variate of $\underline{z}^{(1)}$.

Similarly, for the within-set communality of the h-th variable of $\underline{z}^{(2)}$, we have

$$h^2_{wh} = \sum_{k=1}^{m} \varepsilon^2_{hk} \qquad (3.13)$$

(ii) Interset Variable Communalities

These are the proportions of variance accounted for by the retained canonical variates of the other set.

They are obtained as the sum of squared cross loadings (interset correlations) between a variable and the retained canonical variates. The interset or between-set communality, h^2_{bi}, of the i-th variable of $\underline{z}^{(1)}$ with the retained canonical variates of $\underline{z}^{(2)}$ is therefore given by

$$h^2_{bi} = \sum_{k=1}^{m} \varepsilon^2_{ik} , \qquad (3.14)$$

where ε_{ik} is the interset correlation of the i-th variable of $\underline{z}^{(1)}$ with the k-th canonical variate of $\underline{z}^{(2)}$.

Similarly, the interset communality of the h-th variable of $\underline{z}^{(2)}$ with the retained canonical variates of $\underline{z}^{(1)}$ is

$$h_{bh}^2 = \sum_{k=1}^{m} \varepsilon_{hk}^2 \quad , \tag{3.15}$$

where ε_{hk} denotes an interset correlation coefficient.

CHAPTER IV

APPLICATIONS

4.1 INTRODUCTION

This chapter is aimed at applying the technique of Canonical Correlation Analysis to data on scores of students in Mathematics in first year for two consecutive academic years, that is, 1984/85 and 1985/86 academic years. Six cases on the applications are discussed. The various fields of Mathematics, namely: Pure Mathematics, Statistics and Applied Mathematics, provided the three sets of variables for the analysis, whereas the units within each field formed the specific variables. The three sets are each characterized by N observations, where N is the same for each academic year (for 1984/85 academic year, N = 66 and for 1985/86, N = 44). The number N represents the actual number of students who sat for the examinations in a particular academic year. The three sets of variables are denoted by X,Y and Z, where the X-variables set represents scores in Pure Mathematics, the Y-variables set represents scores in Mathematical Statistics and the Z-variables set represents scores in Applied Mathematics. The variable sets are as given below:

(a) X-Variables

X_1 - Calculus I

X_2 - Calculus II

X_3 - Linear Algebra I

X_4 - Linear Algebra II

(b) Y-Variables

Y_1 - Probability and Statistics I

Y_2 - Prabability and Statistics II

(c) Z-Variables

Z_1 - Vector Analysis

Z_2 - Classical Mechanics.

4.2. RESULTS.

4.2.1. CANONICAL CORRELATION ANALYSIS ON THE RELATIONSHIP BETWEEN PURE MATHEMATIC AND MATHEMATICAL STATISTICS UNITS(VARIABLES).

This will be split into two analyses:

(I) 1984/85 academic year

(II) 1985/86 academic year

(I) 1984/85

Table 1 below gives the sample means, standard deviations and the correlation matrix, R, for scores in Pure Mathematics and Mathematical Statistics for first year, 1984/85 academic year.

Table 1: Sample means, standard deviations and the correlation matrix, R, for scores in Pure Mathematics and Mathematical Statistics for first year, 1984/85

Variable	Mean	Standard deviation
X_1	58.3	11.6
X_2	65.3	10.3
X_3	68.5	13.9
X_4	67.0	8.0
Y_1	62.4	12.7
Y_2	62.3	8.6

$$R = \begin{bmatrix} 1.0000 & 0.4222 & 0.5652 & 0.3148 & 0.6047 & 0.3102 \\ 0.4222 & 1.0000 & 0.5589 & 0.3022 & 0.3291 & 0.3534 \\ 0.5652 & 0.5589 & 1.0000 & 0.3914 & 0.5798 & 0.3429 \\ 0.3148 & 0.3022 & 0.3914 & 1.0000 & 0.4922 & 0.4143 \\ 0.6047 & 0.3291 & 0.5798 & 0.4922 & 1.0000 & 0.4005 \\ 0.3102 & 0.3534 & 0.3429 & 0.4143 & 0.4005 & 1.0000 \end{bmatrix}$$

Canonical Correlations

Table 2 below shows the canonical correlation coefficients (r_k) and other indices, for first year 1984/85. The first canonical correlation, r_1, is 0.737 representing about 54% ($r_1^2 = 0.543$) overlapping variance between the first pair of canonical variates. The second canonical correlation, r_2, is 0.260 which represents about 7% ($r_2^2 = 0.068$) overlapping variance between the second pair of canonical variates.

Table 2: Canonical Correlation Coefficients, weights and loadings for scores in Pure Mathematics and Mathematical Statistics for first year, 1984/85.

K	eigen-value, r_k^2	Canonical correlation, r_k	Canonical Weights and Loadings			
			Pure Mathematics (X-Variables)		Statistics (Y-Variables)	
			Weights	Loadings	Weights	Loadings
1	0.543	0.737	0.485	0.821	0.866	0.971
			-0.036	0.512	0.261	0.608
			0.379	0.803		
			0.436	0.726		
2	0.068	0.260	-0.564	-0.279	-0.664	-0.239
			0.964	0.600	1.060	0.794
			-0.504	-0.082		
			0.515	0.432		

Number of Significant Canonical Variates (Dimensionality)

In order to find the number of significant canonical variates, we test the overall null hypothesis

$$H_o: \Sigma_{12} = 0$$

against

$$H_1: \Sigma_{12} \neq 0,$$

where Σ_{12} is the intercorrelation matrix between X- and Y-variables.

The computed chi-square, X_c^2, and the table chi-square, X_α^2, values for various degrees of freedom appear in table (3) below.

Table 3: Test Statistics for scores in Pure Mathematics and Mathematical Statistics for first year, 1984/85.

K	X_c^2	$X^2_{(0.01)}$	$X^2_{(0.05)}$	d.f.
1	53.306	1.65	2.73	8
2	4.378	0.115	0.352	3

Bartlett's X^2 statistic given in the table provides tests of the joint nullity of the residuals after the larger roots are successively removed. From the table we see that

$$X_c^2 > X_\alpha^2; \quad \text{for all } r_k (K = 1,2).$$

Thus, the null hypothesis of independence is rejected (for $\alpha = 0.01$ and 0.05). Consequently, the two canonical variates are required to fully account for the relationships which are present between the X- and Y-variables. Therefore, a dimensionality of 2 is required to study this relationship.

Canonical Weights.

These are also given in Table 2 above. From the table we see that the first canonical variate is associated with X_1 and X_4 from the X-variables set and Y_1 from the Y-variables set. This

implies that X_1 and Y_1 contribute more towards the first canonical correlation than all the other variables.

The second canonical variate is associated with X_2 in the X-variables set and Y_2 in the Y-variables set. Thus, X_2 and Y_2 contribute more significantly towards the second canonical correlation.

Canonical Loadings (Intraset Correlations)

The canonical loadings also appear in Table 2 above and are helpful in establishing the nature of the canonical variates defined on each set of variables.

The correlation of the X-variables with $\eta_1 = \underline{a}_1', \underline{z}^{(1)}$ are alike in sign. The variable X_1 shows the strongest correlation (0.821) with η_1 closely followed by X_3(0.803). The canonical variate η_1 shows that the performance in all the X-variables is correlated though to varying extents. Thus, it is possible to get a linear combination of these variables.

The corresponding correlations of the Y-variables with $\phi_1 = \underline{b}_1', \underline{z}^{(2)}$ have the same sign, indicating also that the performance in the two variables comprising the Y-variables set is correlated. The variate ϕ_1 is mainly represented by Y_1 with a correlation of 0.971.

The first pair of canonical variates, (η_1, ϕ_1), indicate that X_1 and Y_1 contribute most significantly to the first canonical correlation. This implies that the performance in X_1 and Y_1 should be most highly interrelated, a fact that is established from the correlation matrix, R, which gives the correlation of X_1 and Y_1

as 0.605 which is the highest correlation between the variables. Further, these units (variables) account for most of the variance in their respective variates.

Turning to the second pair of canonical variates we see that the correlations tend to be smaller than those of the variables with η_1 and ϕ_1. The loadings of the X-variables with η_2 vary in sign. The canonical variate η_2 appears to represent $X_2(0.600)$ which is least correlated with η_1. However, we notice that X_1 and X_3 (the variables which are highly correlated with η_1) have negative correlations with η_2. Similarly, the loadings of the Y-variables with ϕ_2 vary in sign. More so, ϕ_2 is largely represented by $Y_2(0.794)$ which has the least correlation with ϕ_2.

It should be clear that the first canonical variates, η_1 and ϕ_1, tend to represent those variables whose performance is most highly correlated.

The fact that X_1 and Y_1 contribute more to the first canonical correlation is consistent with the interpretation given by the canonical weights above. Further, X_2 and Y_2 contribute more to the second canonical correlation as also given by the canonical weights. Therefore, we can assume that the canonical weights given in this analysis are quite stable (their interpretation is valid).

Cross Loadings (Interset Correlations)

These appear in Table 4 below. These loadings are helpful in clarifying the nature of interrelationships among the two sets

of variables. From the table we notice that all the X-variables are directly related to ϕ_1. However, X_1 seems to be more correlated with ϕ_1. Further, Y_1 is more correlated with η_1(0.715) than Y_2(0.448).

Table 4: Cross Loadings for scores in Pure Mathematics and Mathematical Statistics for first year, 1984/85.

	ϕ_2	ϕ_2		η_1	η_2
X_1	0.605	-0.073	Y_1	0.715	-0.062
X_2	0.377	0.156	Y_2	0.448	0.206
X_3	0.592	-0.021			
X_4	0.535	0.112			

Thus, it is clear that X_1 and Y_1 contribute highly to the first canonical correlation (composed with the other variables). This confirms the fact that the performance in X_1 in more related to that of Y_1. That is, students who performed well in X_1 tended to do the same in Y_1.

Two of the X-variables vary inversely with ϕ_2. Nonetheless, ϕ_2 and η_2 tend to identify the interrelationship between X_2 and Y_2.

Variance Extracted by Canonical Variates

The proportion of explained variance in the X-set variables that is accounted for by the first canonical variate, η_1, is given by

$$R^2_{(1)X} = \left[(0.821)^2 + (0.512)^2 + (0.803)^2 + (0.726)^2\right]/4$$

$$\simeq 0.527$$

Thus, the first canonical variate, η_1, explains about 53% of the total variance of the X-variables set.

Next, the proportion of variance extracted by the second canonical variate, η_2, in the X-variables set is

$$R^2_{(2)X} = \left[(0.279)^2 + (0.600)^2 + (-0.082)^2 + (0.432)^2\right]/4$$

$$\simeq 0.158$$

which is about 16% of the total variance. Therefore, the two canonical variates account for about 69% of the total variance in the X-variables set.

Considering the Y-variables set, the proportion of variance accounted for by the canonical variate, ϕ_1, is

$$R^2_{(1)Y} = \left[(0.971)^2 + (0.608)^2\right]/2$$

$$\simeq 0.344.$$

Thus, the two variates account for almost 100% of the variance in the Y-variables set.

Redundancy:

The redundancy in the X-variables set generated by each of the canonical variates $\phi_k (k=1,2)$ is

$$R^2_{(1)x/y} = R^2_{(1)x}\, r_1^2$$

$$= 0.527 \times 0.543$$

$$= 0.286$$

and

$$R^2_{(2)x/y} = R^2_{(2)x}\, r_2^2$$

$$= 0.158 \times 0.068$$

$$\simeq 0.011$$

respectively.

Also the redundancy in the Y-variable set attributable to the canonical variates η_k (k = 1,2) is

$$R^2_{(1)y/x} = R^2_{(1)y}\, r_1^2$$

$$= 0.656 \times 0.543$$

$$\simeq 0.356$$

and

$$R^2_{(2)y/x} = R^2_{(2)y}\, r_2^2$$

$$= 0.344 \times 0.068$$

$$\simeq 0.023$$

respectively.

From the above calculations, we notice that a higher proportion of variance in the Y-variables set is predictable from the X-variables set than that of X-variables set predictable from Y-variables set. We can thus claim that, given the X-variables we can predict the Y-variables with more certainty than we can predict the X-variables

given the Y-variables.

From the redundancies given above, it is clear that ϕ_1 (which mainly represent Y_1) accounts for a greater part of the explained variance of the X-variables. Also η_1 (represented mainly by X_1) accounts for a higher proportion of the explained variance of the Y-variables.

The two canonical variates η_1 and η_2 account collectively for about 29% of the variance in the Y-variables set, whereas $\phi_k (K = 1,2)$ account for about 38% of the variance in the X-variables set; that is

$$R^2_{(T)y/x} \simeq 0.29$$

and

$$R^2_{(T)x/y} \simeq 0.38$$

respectively.

(II) 1985/86

Table 5 below shows the sample means, standard deviations and the correlation matrix, R, for scores in Pure Mathematics and Mathematical statistics for first year, 1985/86.

Table 5: Sample means, standard deviations and the correlation matrix, R, for scores in Pure Mathematics and Mathematical Statistics for first year, 1985/86.

Variable	Mean	Standard deviation
X_1	58.8	9.9
X_2	53.6	14.0
X_3	69.8	11.7
X_4	77.4	13.7
Y_1	60.8	8.9
Y_2	64.3	6.9

$$R = \begin{bmatrix} 1.0000 & 0.4554 & 0.4467 & 0.5603 & 0.5133 & 0.3044 \\ 0.4554 & 1.0000 & 0.2862 & 0.6244 & 0.5569 & 0.4755 \\ 0.4467 & 0.2862 & 1.0000 & 0.4131 & 0.3163 & 0.3756 \\ 0.5603 & 0.6244 & 0.4131 & 1.0000 & 0.4870 & 0.6988 \\ 0.5133 & 0.5569 & 0.3163 & 0.4870 & 1.0000 & 0.4559 \\ 0.3044 & 0.4755 & 0.3756 & 0.6988 & 0.4559 & 1.0000 \end{bmatrix}$$

Canonical Correlations

Table 6 below gives the canonical correlation coefficients (r_k, k = 1,2), the canonical weights and loadings for scores in Pure Mathematics and Mathematical statistics for first year, 1985/86. The first canonical correlation, r_1, is given as 0.749 which represents 56% (r_1^2 = 0.561) of the variance common to the first

pair of canonical variates. The second canonical correlation, r_2, is 0.468 representing about 22% (r_2^2 = 0.219) overlapping variance between the second pair of canonical variates.

Table 6: Canonical Correlation Coefficients, Weights and Loadings for scores in Pure Mathematics and Mathematical Statistics for first year, 1985/86.

K	Eigen-value, r_k^2	Canonical Correla-tion, r_k	Canonical Weights and Loadings			
			Pure Mathematics (X-variables)		Statistics (Y-variables)	
			Weights	Loadings	Weights	Loadings
1	0.561	0.749	-0.051	0.573	0.383	0.732
			0.278	0.771	0.766	0.940
			0.182	0.546		
			0.743	0.963		
2	0.219	0.468	-0.957	-0.624	-1.056	-0.682
			-0.664	-0.422	0.822	0.341
			0.142	-0.055		
			1.020	0.128		

Dimensionality

To determine the number of significant canonical variates, we test the null hypothesis

$$H_o: \Sigma_{12} = 0$$

against

$$H_1: \Sigma_{12} \neq 0,$$

where Σ_{12} is the intercorrelation matrix between the X- and Y-variables. Table 7 below reports the computed chi-square, χ^2_c, and the table chi-square, χ^2_α, values for various degrees of freedom. These are necessary for testing the above hypothesis.

From the table we see that

$$\chi^2_c > \chi^2_\alpha; \text{ for all } k = 1,2$$

Table 7: Test Statistics for scores in Pure Mathematics and Mathematical Statistics for first year, 1985/86.

k	χ^2_c	$\chi^2_{(0.01)}$	$\chi^2_{(0.05)}$	d.f.
1	42.29	1.65	2.73	8
2	9.77	0.115	0.352	3

Therefore, the null hypothesis of independence is rejected, at the two levels of significance, α = 0.01 and 0.05. This suggests that the two canonical variates are required to account for the relationships which are present between the two variable sets.

Canonical Weights

The canonical weights are given in Table 6 above. We notice that the first canonical variate, η_1, is associated with X_4 which has a weight of 0.743. From the Y-variables set, Y_2 seems to

represent ϕ_1 more than Y_1. Therefore, X_4 and Y_2 contribute more towards obtaining the first canonical correlation than all the other variables making up the two sets.

The second canonical variate, η_2, is associated with X_4 with a weight of 1.020, and closely followed by X_1 which has a weight of -0.957. The variate ϕ_2 is associated with Y_1. This implies that X_4, X_1 and Y_1 contribute most significantly towards the second canonical correlation.

Canonical loadings

All the X-variables contribute in the same direction to the first canonical variate, η_1, (see Table 6 above) and all have sizeable correlations with this variate. Above all, η_1 seems to be an expression of X_4(0.963).

The correlations of the variables with η_2 are noticeably weaker than their correlations with η_1; the strongest correlations being those of X_1(-0.624) and X_2(-0.422). Accordingly, η_2 may be regarded as an expression of X_1. It can also be seen from the table that three variables, X_1, X_2 and X_3 are inversely related with η_2.

Considering the canonical variates $\phi_k = \underline{b}_k^1 \underline{z}^{(2)}$ (k=1,2) of Y-variables, we see that the two variables making up this set contribute in the same direction to ϕ_1 and the correlations are quite high. The strongest correlation with ϕ_1 is that of Y_2(0.940). The second canonical variate, ϕ_2, is largely characterized by Y_1(-0.682).

The first pair of canonical variates, (η_1, ϕ_1), indicate that X^4 and Y_2 contribute most heavily to the first canonical correlation and accordingly the performance of students in the two variables is more correlated than to the rest of the variables. We note, from the correlation matrix, R, that X_4 and Y_2 have the highest correlation of about 0.699.

The second canonical variates, η_2 and ϕ_2, indicate X_1 and Y_1 contribute more to the second canonical correlation. From the correlation matrix, R, we notice that the correlation of X_1 with Y_1 is 0.513 which is the second largest after that of X_2 with Y_1(0.557). However, the difference between the two correlations is not very significant.

Whereas the interpretation that X_4 and Y_2 contribute more significantly to thefirst canonical correlation is consistent in both the canonical weights and loadings, the canonical weights and loadings do not tally in their interpretation of the second pair canonical variates. That is, according to the canonical weights, X_4 and Y_1 contribute more to the second canonical correlation whereas the canonical loadings identify X_1 and Y_1 as the most important variables. However, we normally tend to prefer the interpretation given by the canonical loadings.

Cross Loadings

From Table 8 below, we see that all the X-variables are directly related to ϕ_1. The strength of the relationships does not vary so much; that of X_4 is highest (0.722) and X_3 the lowest (0.409). The variable Y_2 is more correlated with η_1(0.704). Thus, we confirm

Table 8: Cross Loadings for scores in Pure Mathematics and Mathematical Statistics for first year, 1985/86.

	ϕ_1	ϕ_2		η_1	η_2
X_1	0.429	-0.292	Y_1	0.548	-0.319
X_2	0.577	-0.197	Y_2	0.704	0.159
X_3	0.409	-0.026			
X_4	0.722	0.060			

our early claim that X_4 and Y_2 are more correlated with each other in terms of performance than in the other variables.

Three of the X-variables vary inversely with ϕ_2. However, ϕ_2 and η_2 tend to identify the interrelationships between X_1 and Y_1.

Variance Extracted by Canonical variates.

The variance extraced by η_1 from the X-variables set is

$$R^2_{(1)x} = \left[(0.573)^2 + (0.771)^2 + (0.546)^2 + (0.963)^2\right]/4$$

$$\simeq 0.537$$

This suggests that η_1 explains about 54% of the total variance in the X-set of variables. Next, the proportion of variance extracted by the second canonical variate, η_2, in the X-variables is

$$R^2_{(2)x} = \left[(-0.634)^2 + (-0.422)^2 + (-0.055)^2 + (0.128)^2\right]/4$$

$\simeq 0.147$

which is about 15% of the total variance. Therefore, the two canonical variates account for about 69%, (54+15)%, of the total variance in the X-variables set.

Considering the Y-variables set, the proportion of variance accounted for by the first canonical variate, ϕ_1, is

$$R^2_{(1)y} = \left[(0.732)^2 + (0.940)^2\right]/2$$

$$\simeq 0.710,$$

which is 71% of the total variance. Further, the second canonical the second canonical variate accounts for about 29% of the total variance, that is

$$R^2_{(2)y} = \left[(-0.682)^2 + (0.341)^2\right]/2$$

$$\simeq 0.290$$

Thus, the two variates account for about 100% of the variance in the Y-variables set.

Redundancy:

The redundancy in the X-variables set generated by each of the canonical variates ϕ_1 and ϕ_2 is

$$R^2_{(1)x/y} = R^2_{(1)x} r_1^2$$

$$= 0.537 \times 0.561$$

$$\simeq 0.301,$$

and

$$R^2_{(2)x/y} = R^2_{(2)x} r_2^2$$

$$= 0.147 \times 0.219$$

$$\simeq 0.032,$$

respectively.

Further, the redundancy in the Y-variables set attributable to the canonical variates $\eta_k (k = 1,2)$ is

$$R^2_{(1)y/x} = R^2_{(1)y} r_1^2$$

$$= 0.710 \times 0.561$$

$$\simeq 0.398,$$

and

$$R^2_{(2)y/x} = R^2_{(2)y} r_2^2$$

$$= 0.290 \times 0.219$$

$$\simeq 0.064$$

respectively.

This shows that a slightly higher proportion of variance in the Y-variables set is predictable from the X-variables set than that of the X-variables set predictable from Y-variables set. From these redundancies it is clear that ϕ_1 (which represents Y_2) accounts for much the greater part of the explained variance of the X-variables. Also η_1 (representing mainly X_4) accounts for a higher proportion, of the explained variance of the Y-variables.

The variates η_k (k=1,2) account collectively for about 46% of the variance in Mathematical Statistics, whereas ϕ_k(k=1,2) account for about 33% of the variance in Pure Mathematics; that is

$$R^2_{(T)y/x} = 0.46$$

and

$$R^2_{(T)x/y} = 0.33$$

respectively.

4.2.2 Comparative Remarks

From the two analyses above, it is valid to compare the results obtained in the two academic years because the variables concerned are the same. However, we note that the number of students who actually sat for the examinations is different for 1984/85, N = 66 and for 1985/86, N = 44. It might be instructive to note this difference in trying to explain the differences in the results.

The squared canonical correlations (Table 2 and Table 6) express the proportion of the variance of one member of a pair of canonical variates which can be accounted for or predicted by the other; or the overlapping variance between the pairs of canonical variates. In both analyses, $0.50 < r_1^2 < 0.60$ which implies that about half of the overlapping variance is represented in the first pair of canonical variates. However, ther is a significant difference in the two analyses as concerns the second squared canonical correlation. For 1984/85, $r_2^2 = 0.068$ which represents only about 7% overlapping variance between the second pair of canonical variates;

whereas for 1985/86, $r_2^2 = 0.219$ which is about 22% of the overlapping variance. However, the number of significant canonical variates in both cases is the same.

As for 1984/85, X_1 and Y_1 are the variables which contribute most significantly to the first canonical correlation, and accordingly their performance is more related. This is not the case with 1985/86 because in this year, X_4 and Y_2 are the two variables whose performance is more correlated. This tells us that the performance in the two years was not consistent. The interest correlations put more light on this claim.

The variance extracted from the X-variables set by η_1 and η_2 is 69% of the total variance in the set, that is, for the year 1984/85. This fortunately tallies with the variance extracted by the same variables in the following year, 1985/86. Likewise, the proportion of explained variance in the Y-variables set that is accounted for by ϕ_1 and ϕ_2 is 100% for the two years. This gives us a rather consistent interpretation even though each canonical variate extracts a different amount of variance from the respective sets.

In the 1984/85 academic year, the two canonical variates η_1 and η_2 account for only 29% of the variance in the Y-variables set. In the following year 1985/86, these same variables accounted for 46% of the variance in the Y-variables set. This tells us that in the 1985/86 academic year, the redundancy in the Y-variables set generated by the canonical variates η_1 and η_2 of the X-variables set is more than that of 1984/85. In the year 1984/85 the total redundancy in the X-variables explained by ϕ_1 and ϕ_2 is 38% whereas that of 1985/86 is 33%. Therefore, we note that in 1984/85,

if we are given the X-variables, we could be more certain about the Y-variables in terms of performance. In 1985/86, the reverse in true.

4.2.3. CANONICAL CORRELATION ANALYSIS ON THE RELATIONSHIPS BETWEEN PURE AND APPLIED MATHEMATICS

This will also be divided into two cases

(I) 1984/85 academic year

(II) 1985/86 academic year

(I). 1984/85

The sample means, standard deviations and the correlation matrix, R, for scores in Pure and Applied Mathematics for first year, 1984/85 appear in Table 9 below:

Table 9: Sample means, Standard deviations and the Correlation matrix, R, for scores in Pure and Applied Mathematics for first year, 1984/85.

Variable	Mean	Standard deviation
X_1	58.3	11.6
X_2	65.3	10.3
X_3	68.5	13.9
X_4	67.0	8.0
Z_1	52.7	11.9
Z_2	61.1	9.1

$$R = \begin{bmatrix} 1.0000 & 0.4222 & 0.5652 & 0.3148 & 0.5653 & 0.4055 \\ 0.4222 & 1.0000 & 0.5589 & 0.3022 & 0.3874 & 0.3482 \\ 0.5652 & 0.5589 & 1.0000 & 0.3914 & 0.4665 & 0.2916 \\ 0.3148 & 0.3022 & 0.3914 & 1.0000 & 0.4106 & 0.4511 \\ 0.5653 & 0.3874 & 0.4665 & 0.4106 & 1.0000 & 0.3476 \\ 0.4055 & 0.3482 & 0.2916 & 0.4511 & 0.3476 & 1.0000 \end{bmatrix}$$

Canonical Correlations

The first canonical correlation, r_1, is 0.711 (see Table 10 below). This is about 51% ($r_1^2 = 0.506$) overlapping variance between the first pair of canonical variates. The second Canonical Correlation, r_2, is 0.196 which implies that about 4% ($r_1^2 = 0.038$) of the variance is common between the second pair of canonical variates.

Table 10: Canonical Correlation Coefficients, Weights and loadings of scores in Pure and Applied Mathematics for first year, 1984/85.

K	Eigen value, r_k^2	Canonical Correlation r_k^2	Canonical Weights and Loadings			
			Pure Mathematics (X-Variables)		Applied Mathematics(Z-Variables)	
			Weights	Loadings	Weights	Loadings
1	0.506	0.711	0.591	0.848	0.702	0.879
			0.227	0.632	0.509	0.753
			0.027	0.669		
			0.463	0.728		
2	0.038	0.196	-0.308	-0.377	-0.803	-0.477
			0.485	0.077	0.937	0.658
			-0.920	-0.516		
			0.784	0.473		

Dimensionality

To determine whether both canonical variates are necessary to account for the differences, we require a test of the residual association which remains after r_1^2 is eliminated. The Table 11 below gives the computed Chi-square, χ_c^2, and the table Chi-square, χ_α^2, values for various degrees of freedom.

Table 11: Test Statistics for scores in Pure and Applied Mathematics for first year, 1984/85.

K	χ_c^2	$\chi^2_{(0.01)}$	$\chi^2_{(0.05)}$	df.
1	46.50	1.65	2.73	8
2	2.46	0.115	0.352	3

From the table we see that

$$\chi_c^2 > \chi_\alpha^2; \quad \text{for all k.}$$

Thus, the null hypothesis of independence is rejected at levels of significance α = 0.01 and 0.05. Thus, it appears that both r_1^2 and r_2^2 will be required in order to explain the relationships between the two sets of data.

Canonical Weights

These appear in Table 10 above. The first canonical variate, η_1, is mainly associated with X_1 and X_4 and ϕ_1 is associated with Z_1. Thus, X_1 and Z_1 contribute more towards the first canonical

correlation.

The variable X_3 and X_4 with weights -0.920 and 0.784 respectively are the variables which represent η_2 most significantly whereas ϕ_2 is associated with mainly Z_2. Hence X_3, X_4 and Z_2 contribute more towards the second canonical correlation.

Canonical loadings

The first canonical variate, η_1, of the X-variables (Table 10) is dominated by X_1 and X_4 with loadings 0.848 and 0.728 respectively. However, all the variables have sizeable loadings, that is their correlations with η_1 are quite high. On the Z-variables set, ϕ_1 is dominated by Z_1(0.879).

The second canonical variate, η_2, expresses mainly X_3(-0.516) which has a negative correlation with this canonical variate. It should be noted that X_1 and X_3 have inverse relation with η_2. The variable Z_1 which is highly correlated with ϕ_1 has a negative correlation with ϕ_2(-0.477). However, ϕ_2 is mostly associated with Z_2(0.658).

The first pair of canonical variates, (η_1,ϕ_1), suggest that X_1 and Z_1 contribute most heavily to the first canonical correlation. Thus, the performance in X_1 and Z_1 is more correlated relative to the other variables. This is confirmed from the correlation matrix, R, where the correlation between X_1 and Y_1 is given as 0.565 (the highest correlation between the sets).

The variates η_2 and ϕ_2 indicate that X_3 and Z_2 contribute more to the second canonical correlation. The correlation matrix,

R, gives the correlation between X_3 and Z_2 as 0.292 which is the lowest correlation for both between the sets and within the sets.

Interset Correlations

The cross-loadings are given in Table 12 below. We find, from the table, that the first canonical variate, ϕ_1, of the Z-variables set is characterized in the X-variables set by X_1 and X_4. The other two variables X_2 and X_3 have fair correlations with this

Table 12: Cross loadings for scores in Pure and Applied Mathematics for first year, 1984/85.

	ϕ_1	ϕ_2		η_1	η_2
X_1	0.603	-0.074	Z_1	0.625	-0.094
X_2	0.449	0.015	Z_2	0.536	0.129
X_3	0.476	-0.101			
X_4	0.518	0.093			

canonical variate (0.449 and 0.476 respectively). The canonical variate η_1 is mainly associated with Z_1 on the Z-variables set. Thus, X_1, X_4 and Z_1 contribute more to the first canonical correlation.

The correlations of the X-variables and Z-variables with ϕ_2 and η_2, respectively, are not very significant and thus we can't say much about them.

Variance Extracted by Canonical Variates

The proportion of variance accounted for by η_1 from the X-variables is

$$R^2_{(1)x} = \left[(0.848)^2 + (0.632)^2 + (0.669)^2 + (0.728)^2\right]/4$$

$$\simeq 0.524$$

This implies that about 52% of the total variance associated with the X-variables is absorbed by η_1. The proportion of variance extracted by the second canonical variate, η_2, in the X-variables set is

$$R^2_{(2)x} = \left[(-0.377)^2 + (0.077)^2 + (-0.516)^2 + (0.473)^2\right]/4$$

$$\simeq 0.160,$$

which is 16% of the total variance. The two canonical variates account for 68% of the total variance in the X-variables set.

Considering the Z-variables set, the proportion of variance accounted for by the first canonical variate, ϕ_1, is

$$R^2_{(1)z} = \left[(0.879)^2 + (0.753)^2\right]/2$$

$$\simeq 0.670.$$

This is 67% of the total variance in the Z-variables set. The second canonical variate, ϕ_2, accounts for approximately 33% of the total variance, that is

$$R^2_{(2)z} = \left[(-0.477)^2 + (0.658)^2\right]/2$$

$\simeq$ 0.330.

This indicates that the two canonical variates account for about 100% of the variance in the Y-variables set, that is, all the variance in the set is accounted for.

Redundancy:

The redundancy in the X-variables set attributable to the canonical variates ϕ_k of the Z-variables set is

$$R^2_{(1)x/z} = R^2_{(1)x} r_1^2$$

$$= 0.524 \times 0.506$$

$$\simeq 0.265$$

and

$$R^2_{(2)x/2} = R^2_{(2)x} r_2^2$$

$$= 0.160 \times 0.038$$

$$\simeq 0.006.$$

Collectively, ϕ_1 and ϕ_2 account for only 27% of the total variance of the X-variables set. That is

$$R^2_{(T)x/z} = 0.27$$

Further, the redundancy in the Z-variables set attributable to the canonical variates $\eta_k (k=1,2)$ is

$$R^2_{(1)z/x} = R^2_{(1)z} r_1^2$$

$$= 0.670 \times 0.506$$

$$\simeq 0.339$$

and

$$R^2_{(2)z/x} = R^2_{(2)z} r_2^2$$

$$= 0.330 \times 0.038$$

$$\simeq 0.013$$

Thus, the two canonical variates together account for about 35% of the variance in the Z-variables set.

(II). 1985/86

The sample means, standard deviations and the correlation matrix, R, for scores in Pure and Applied Mathematics for first year, 1985/86 are given in Table 13 below.

Table 13: Sample means, standard deviations and the correlation matrix, R, for scores in Pure and Applied Mathematics for first year, 1985/86.

Variable	Mean	Standard deviation
X_1	58.8	9.9
X_2	53.6	14.0
X_3	69.8	11.7
X_4	77.4	13.7
Z_1	62.9	10.8
Z_2	57.1	12.2

$$R = \begin{bmatrix} 1.0000 & 0.4554 & 0.4467 & 0.5603 & 0.4708 & 0.5302 \\ 0.4554 & 1.0000 & 0.2862 & 0.6244 & 0.3915 & 0.4032 \\ 0.4467 & 0.2862 & 1.0000 & 0.4131 & 0.3678 & 0.2368 \\ 0.5603 & 0.6244 & 0.4131 & 1.0000 & 0.6901 & 0.5505 \\ 0.4708 & 0.3915 & 0.3678 & 0.6901 & 1.0000 & 0.5820 \\ 0.5302 & 0.4032 & 0.2368 & 0.5505 & 0.5820 & 1.0000 \end{bmatrix}$$

Canonical Correlations

The Canonical Correlation Coefficients and other indices are given in Table 14 below. The first canonical correlation coefficient, r, is 0.735. This represent about 54% (r_1^2 = 0.540) overlapping variance between the first pair of canonical variates, (η_1,ϕ_1). The second canonical correlation coefficient, r_2, is 0.280 representing approximately 8% (r_2^2 = 0.078) of the variance common to the second pair of canonical variates, (η_2,ϕ_2).

Table 14: Canonical Correlation Coefficients, Weights and Loadings for scores in Pure and Applied Mathematics for first year, 1985/86

K	Eigen Value, r_k^2	Canonical Correlation r_k	Canonical Weights and Loadings			
			X-Variables		Z-Variables	
			Weights	Loadings	Weights	Loadings
1	0.540	0.735	0.291	0.742	0.724	0.949
			-0.064	0.597	0.386	0.808
			0.031	0.976		
			0.833	0.969		
2	0.078	0.280	1.021	0.539	-0.994	-0.314
			0.486	0.292	1.167	0.589
			-0.588	-0.318		
			-0.787	-0.154		

Dimensionality:

This is determined by testing the null hypothesis

$$H_o: \Sigma_{12} = 0$$

against

$$H_i: \Sigma_{12} \neq 0,$$

where R_{12} is the intercorrelation matrix between the X- and Z-Variables. The computed Chi-square, X_c^2, and the table Chi-square, X_α^2, values for various degrees of freedom are given in table (15) below.

Table 15: Test statistics for scores in Pure and Applied Mathematics for first year, 1985/86

K	X_c^2	$X^2_{(0.01)}$	$X^2_{(0.05)}$	d.f.
1	34.85	1.65	2.73	8
2	3.32	0.115	0.352	3

From the table, it is clear that

$$X_c^2 > X_\alpha^2; \quad \text{for all } k=1,2$$

This implies that the null hypothesis of independence is rejected at the two levels of significance, $\alpha = 0.01$ and 0.05. Thus, the two canonical variates are necessary in accounting for the inter-relationships between the variables.

Canonical Weights

The Canonical weights(Coefficients) are given in Table 14 together with other indices. The variables X_1, X_2 and X_3 don't seem to contribute significantly towards η_1; that is, by considering their weights (0.291, -0.064 and 0.031 respectively). However, X_4 tends to contribute so much to the first canonical variate, η_1. On the Z-variables set, Z_1 with a canonical weight of 0.724 seems to contribute highly towards ϕ_1. Thus, the variables X_4 and Z_1 dominate in getting the first canonical correlation.

The second canonical variate is associated with X_1 and X_4 with Coefficients 1.021 and -0.787 respectively (that is, on the X-variables set). On the Z-variables set, ϕ_2 is associated with Z_2 even though there is an indication that both variables, Z_1 and Z_2, contribute significantly towards this variate. Therefore, X_1, X_4, Z_1 and Z_2 contribute sizeably towards the second canonical correlation.

Canonical Loadings

These appear in Table 14 and they actually give us a completely different picture from the canonical weight, for at least the first canonical variate, η_1.

We see from the correlations of X-variables with η_k that η_1 is characterized principally by X_3 and X_4 (with correlations 0.976 and 0.969 respectively), although X_1 and X_2 are to some extent also related in a direct sense to this canonical variate. The second canonical variate, η_2, is essentially represented by X_1(0.539); we note also the inverse correlation between X_3, X_4 (with correlations

-0.318 and -0.154 respectively) and η_2.

Turning now to the correlations of Z-variables with ϕ_k, we see that ϕ_1 is mainly represented by Z_1 with a correlation of 0.949. However, Z_2 also shows a high correlation with this canonical variate. The second canonical variate, ϕ_2, is represented by Z_2(0.589); we note the negative correlation of Z_1 (-0.314) with this variate.

It is of interest to see that those variables which have the highest correlation with the first canonical variates, η_1 and ϕ_1 are assigned or rather have negative correlations with the second canonical variates, η_2 and ϕ_2. This is also evident from all the preceding analyses.

The pair (η_1,ϕ_1) show that the performance in X_3 and Z_1 is most highly correlated. However, this is not supported from the correlation matrix, R, since the matrix shows that the correlation between X_3 and Z_1 is 0.368. The correlation between X_4 and Z_1 is 0.690, which is the highest between the variables. This is expected because X_4 also has a very high correlation with η_1 (a correlation of 0.969 which is not significantly different from that of X_3 with this variate).

The other pair of canonical variates, (η_2,ϕ_2) seems to isolate at least one variable which is lowly correlated with (η_1,ϕ_1) and give it a negative sign.

Cross Loadings

The interset correlations are given in Table 16 below. From these correlations of X-variables with ϕ_1, it can be seen that all

the variables are closely related to ϕ_1; the positive sign indicates that the X-variables are directly related to ϕ_1. The variables X_3 and X_4 correlate most highly with ϕ_1.

Table 16: Cross Loadings for scores in Pure and Applied Mathematics for first year, 1985/86.

	ϕ_1	ϕ_2		η_1	η_2
X_1	0.546	0.151	Z_1	0.698	-0.088
X_2	0.439	0.082	Z_2	0.594	0.165
X_3	0.718	-0.089			
X_4	0.713	-0.043			

The X-variables show very low correlations with ϕ_2. However, X_1 has the highest correlation with ϕ_2(0.151). The variables X_3 and X_4 which have negative correlations with η_2 also have negative correlations with ϕ_2. The variable Z_1 has a negative correlation with η_2; we also note that Z_1 has a negative correlation with ϕ_2.

Thus, the cross loadings tend to give more light on the correlations of variables across sets.

Variance Extracted by Canonical Variates:

This is an index of the extent to which a canonical variate explains the total variance of the domain of which it is a linear composite. The proportion of variance in the X-variables set extracted by η_1 is given as

$$R^2_{(1)X} = \left[(0.742)^2 + (0.597)^2 + (0.976)^2 + (0.969)^2\right]/4$$

$$\simeq 0.713;$$

a very reasonable percentage (71%) of variance which is explained by η_1 on X-variables set. The second canonical variate, η_2, explains about 13% of the total variance in the X-variables set. That is

$$R^2_{(2)X} = \left[(0.539)^2 + (0.292)^2 + (-0.318)^2 + (-0.154)^2\right]/4$$

$$\simeq 0.125.$$

The two canonical variates explain about 84% of the total variance in the X- set of variables.

Next, the proportion of variance accounted for the canonical variate ϕ_1 is

$$R^2_{(1)Z} = \left[(0.949)^2 + (0.808)^2\right]/2$$

$$\simeq 0.777$$

which is about 78% of the total variance. The second canonical variate, ϕ_2, accounts for about 22% of the total variance in the Z-variables set. In other words

$$R^2_{(2)Z} = \left[(-0.314)^2 + (0.589)^2\right]/2$$

$$\simeq 0.223.$$

The two variates, ϕ_1 and ϕ_2, account for almost 100% of the total variance in the Z-variables set.

Redundancy:

The redundancy, $R^2_{(k)x/y}$, of the X-variables with respect to

the k-th canonical variate $\phi_k = \underline{b}_k' \underline{Z}^{(2)}$ of $\underline{Z}^{(2)}$, given the availability of $\underline{Z}^{(2)}$, can be calculated using the formula

$$R^2_{(k)x/z} = \sum_{i=1}^{p} \varepsilon^2_{ik}/p,$$

where ε_{ik} is the correlation between the ith variable of $\underline{Z}^{(1)}$ and the K-th canonical variate of $\underline{Z}^{(2)}$. Therefore, the redundancy in the X-variables set generated by the canonical variates ϕ_k (K=1,2) is

$$R^2_{(1)x/z} = \sum_{i=1}^{4} \varepsilon^2_{i1}/4$$

$$= \left[(0.546)^2 + (0.439)^2 + (0.718)^2 + (0.713)^2\right]/4$$

$$\simeq 0.379$$

and

$$R^2_{(2)x/z} = \sum_{i=1}^{4} \varepsilon^2_{iz}/4$$

$$= \left[(0.151)^2 + (0.082)^2 + (-0.089)^2 + (-0.043)^2\right]/4$$

$$\simeq 0.010$$

respectively. Thus, the two canonical variates, ϕ_1 and ϕ_2, account for about 39% of the variance in X-variables set, that is

$$R^2_{(T)x/z} = 0.390.$$

Further, the redundancy in the Z-variables set attributable to the canonical variates η_k (k=1,2) is

$$R^2_{(1)z/x} = \sum_{h=1}^{2} \varepsilon^2_{h1}/2$$

$$= \left[(0.698)^2 + (0.594)^2\right]/2$$

$$\simeq 0.420$$

and

$$R^2_{(2)z/x} = \sum_{h=1}^{2} \varepsilon^2_{h2}/2$$

$$= \left[(-0.088)^2 + (0.165)^2\right]/2$$

$$\simeq 0.018,$$

respectively. The two canonical variates, η_1 and η_2, account for about 44% of the variance in the Z-variables set, that is

$$R^2_{(T)z/x} \simeq 0.44.$$

4.2.4 Comparative Remarks

We see from the analysis of pure and Applied Mathematics, for 1984/85 and 1985/86, that the first canonical correlations are 0.711 and 0.735 respectively. Therefore, these represents, respectively, 51% and 54% of the overlapping variance between the first pair of canonical variates, (η_1, ϕ_1). The second canonical correlation for 1984/85 is 0.196 and that of 1985/86 is 0.280. This means that there is no significant difference in the two analyses as concerns the first and second squared canonical correlation. However, we note that both r_1^2 and r_2^2 are required in order to establish the relationships between the X- and Z - variables set.

In the academic year 1984/85, X_1 and Z_1 contribute more towards the first canonical correlation and thus their performance is more correlated. In the following year, 1985/86, X_3 and Z_1 are the variables which contribute more towards the first canonical correlation. This does not establish consistency in the performance for the two academic years. From the tables of intraset correlations for the two academic years, we note that the variable which is most highly correlated with η_1 or ϕ_1 has a negative correlation with η_2 or ϕ_2. The interset correlations have similar trends of correlations.

The total variance extracted by η_1 and η_2 from the variables on which they are defined (X-variables) is 52% and 84% of the total variance in the set for 1984/85 and 1985/86 respectively. The proportion of explained variance in the Z-variables set that is accounted for by ϕ_1 and ϕ_2 is 100% for both 1984/85 and 1985/86. Thus, the redundancy in the X-variables set attributable to the canonical variates ϕ_k (k=1,2) is 27% for 1984/85 and 39% for the year 1985/86. On the other hand, the canonical variates η_1 and η_2 account for 35% and 44% of the total variance in the Z-variables set for the years 1984/85 and 1985/86 respectively. This implies that given the X-variables, in both years, we could predict the performance in the Z-variables with more confidence than the reverse case.

4.2.5 CANONICAL CORRELATION ANALYSIS ON THE RELATIONSHIPS BETWEEN MATHEMATICAL STATISTICS AND APPLIED MATHEMATICS.

Like all the preceding analyses, this will be divided into two cases:

(I) 1984/85 academic year

(II) 1985/86 academic year

(I) 1984/85

Table 17 below reports the sample means, standard deviations and the correlation matrix, R, for scores in Mathematical Statistics and Applied Mathematics for first year, 1984/85.

Table 17: Sample means, standard deviations and the correlation matrix, R, for scores in Mathematical Statistics and Applied Mathematics for first year, 1984/85.

Variable	Mean	Standard deviation
Y_1	62.4	12.7
Y_2	62.3	8.6
Z_1	52.7	11.9
Z_2	61.1	9.1

$$R = \begin{bmatrix} 1.0000 & 0.4005 & 0.3880 & 0.4900 \\ 0.4005 & 1.0000 & 0.3380 & 0.4978 \\ 0.3880 & 0.3380 & 1.0000 & 0.3476 \\ 0.4900 & 0.4978 & 0.3476 & 1.0000 \end{bmatrix}$$

Canonical Correlations

These are given in Table 18 below. The first canonical correlation coefficient, r_1, is 0.639 which represents 41% ($r_1^2 = 0.408$) of the shared variance between the first pair of canonical variates, (η_1, ϕ_1). The second canonical correlation coefficient, r_2, is 0.050 which is about 0.3% of the variance common between the second pair of canonical variates; this is an extremely small percentage of variance which might not expose much about the variables on which the canonical variates are defined.

Table 18: Canonical correlation coefficients, Weights and Loadings of scores in Mathematical Statistics and Applied Mathematics for first year, 1984/85.

k	Eigen value r_k^2	Canonical Correlation r_k	Canonical Weights and Loadings			
			Y-Variables		Z-Variables	
			Weights	Loadings	Weights	Loadings
1	0.408	0.639	0.616	0.848	0.409	0.680
			0.579	0.825	0.782	0.924
2	0.003	0.050	-0.901	-0.530	-0.985	-0.733
			0.925	0.565	0.726	0.383

Dimensionality:

It is necessary to determine whether the two canonical variates are important in accounting for the relationships that exist between variables. We require a test of the residual association which remains after r_1^2

is eliminated. Table 19 below gives the computed chi-square, χ_c^2 and the table chi-square, χ_α^2, values for various degrees of freedom.

From the table we see that

$$\chi_c^2 > \chi_\alpha^2; \quad \text{for all} \quad k.$$

Table 19: Test Statistics for scores in Mathematical Statistics and Applied Mathematics, for first year, 1984/85.

k	χ_c^2	$\chi^2(0.01)$	$\chi^2(0.05)$	d.f
1	33.44	0.297	0.711	4
2	0.16	0.0002	0.004	1

This implies that the null hypothesis of independence, that is

$$H_o: \Sigma_{12} = 0$$

against

$$H_1: \Sigma_{12} \neq 0,$$

is rejected at levels of significance $\alpha = 0.01$ and 0.05. This suggests that both r_1^2 and r_2^2 are required to explain the relationships between the two sets of data.

Canonical Weights

These are given together with other indices in Table 18 above. The variate η_1 is associated mainly with Y_1 whereas ϕ_1 is represented by z_2 with a

weight of 0.782. This Y_1 and z_2 contribute relatively more to the first canonical correlation than Y_2 and z_1.

Both Y_1 and Y_2 seem to represent η_2 quite well but Y_2 has a slightly higher canonical weight (0.925) than Y_1 (-0.901). The variate ϕ_2 is mainly represented by z_1 (-0.985). Thus Y_2 and z_1 contribute more to the second canonical correlation, r_2. It is clear here that those variables which contribute most highly towards the first canonical correlation contribute least towards the second canonical correlation.

Canonical loadings

The canonical loadings are reported in Table 18 above. The variables Y_1 and Y_2 have fairly high correlations with η_1, that is 0.848 and 0.825 respectively. Thus, the two variables, Y_1 and Y_2 represent η_1 almost equally even though Y_1 is more pronounced. The second canonical variate, η_2, seem to identify Y_1 with an inverse correlation of -0.530. However, this variate is an expression of mainly Y_2(0.565).

On the z-variables set, we see that ϕ_1 is mainly identified with z_2 (0.924). The variate ϕ_2 represents mainly z_1(-0.733). We notice that z_1 which is lowly correlated with ϕ_1 has a higher but negative correlation with ϕ_2.

From above, η_1 and ϕ_1 suggest that Y_1 and z_2 are the two variables whose performance is more correlated and thus the variables contribute more towards the first canonical correlation. The variable Y_2 and z_1 contribute more towards the second canonical correlation. From the correlation matrix, R, the correlation between Y_1 and z_2 is 0.490 which is the highest correlation between the variables. That is the first canonical relationship tends to isolate those variables whose performance is most highly correlated.

Interset Correlations

The interset correlations appear in Table 20 below. The variables Y_1 and Y_2 have correlations of 0.542 and 0.527 with ϕ_1 respectively. These variable are thus related in a direct way to ϕ_1. However

Table 20. Interset Correlations for scores in Mathematics Statistics and Applied Mathematics for first year, 1984/85.

	ϕ_1	ϕ_2		η_1	η_2
Y_1	0.542	-0.027	z_1	0.434	-0.037
Y_2	0.527	0.028	z_2	0.590	0.019

Y_1(0.542) tends to have a higher correlation with this variate, ϕ_1. The variable Y_1 shows a negative

correlation of -0.027 with ϕ_2 and Y_2 gives a positive correlation (0.028) with this variate that is ϕ_2 (compare this correlations, the sign, with those of Y_1 and Y_2 with η_2).

The correlation of the z-variables with $\eta_k (k=1,2)$ follow almost a similar trend to that of X-variables with $\phi_k (k=1,2)$. The variate η_1 has a higher correlation with $z_2(0.590)$ and η_2 with $z_1(-0.037)$.

Variance Extracted by Canonical Variates

The proportion of variance in the Y-variables set extracted by η_1 is given by

$$R^2_{(1)y} = \left[(0.848)^2 + (0.825)^2\right]/_2$$

$$\simeq 0.700,$$

that is, 70% of the variance in Y_1 and Y_2 can be explained by η_1. The second canonical variate of the Y-variables extracts

$$R^2_{(2)y} = \left[(-0.530)^2 + (0.565)^2\right]/_2$$

$$\simeq 0.300$$

of the total proportion of variance in the Y-variables set. Therefore the two canonical variates, η_1 and η_2, extract 100% of the variance in Y-set of variables.

Looking at the z-variables set we see that the proportion of variance accounted for by ϕ_1 is

$$R^2_{(1)z} = \left[(0.680)^2 + (0.924)^2\right]/_2$$

$$\simeq 0.658,$$

which is about 66% of the total variance in z domain. The canonical variate ϕ_2 extracts the remaining 34% of the variance in the z-domain so that the two canonical variates, ϕ_1, ϕ_2 extract 100% of the variance in this measurement domain. In other words

$$R^2_{(2)z} = \left[(-0.733)^2 + (0.383)^2\right]/2$$

$$\simeq 0.342,$$

so that

$$R^2_{(1)z} + R^2_{(2)z} = 0.658 + 0.342 = 1.000$$

Redundancy:

The redundancy in Y-variables set generated by $\phi_k = \underline{b}_k' \underline{z}^{(2)}$ $(k=1,2,)$ is

$$R^2_{(1)y/z} = \sum_{i=1}^{2} \varepsilon^2_{i1}/2$$

$$= \left[(0.542)^2 + (0.527)^2\right]/2$$

$$\simeq 0.286$$

and

$$R^2_{(2)y/z} = \sum_{i=1}^{2} \varepsilon^2_{i2}/2$$

$$= \left[(-0.027)^2 + (0.028)^2\right]/2$$

$$\simeq 0.0008$$

respectively. Note that ε_{ik} is as defined in

Chapter III. It seems as if the second canonical variate, ϕ_2, does not explain any significant proportion of variance in the Y-variables set. Thus,only ϕ_1 explains about 29% of the variance in the Y measurement domain.

Taking the Z-measurement domain, we see that the redundancy in this set attributable to the canonical variates $\eta_k (k=1,2)$ of the Y domain is

$$R^2_{(1)z/y} = \sum_{h=1}^{2} \varepsilon^2_{h1/2}$$

$$= \left[(0.435)^2 + (0.590)^2\right]/2$$

$$\simeq 0.268$$

and

$$R^2_{(2)z/y} = \sum_{h=1}^{2} \varepsilon^2_{h2/2}$$

$$= \left[(-0.037)^2 + (0.019)^2\right]/2$$

$$\simeq 0.0009$$

respectively. That is, only the η_1 explains about 27% of the variance in the z-variables set. From both domains, we notice that $R^2_{(T)y/z}$ and $R^2_{(T)z/y}$ are attributed by only the first canonical variates from the other set.

(II) 1985/86

Given below is a table of sample means, standard deviations and the correlation matrix, R, for scores

in Mathematical Statistics and Applied Mathematics for first year, 1985/86.

Table 21. Sample means, standard deviations and the correlation matrix, R, for scores in Mathematical Statistics and Applied Mathematics for first year, 1985/86.

Variable	Mean	Standard deviation
Y_1	60.8	8.9
Y_2	64.3	6.9
Z_1	62.9	10.8
Z_2	57.1	12.2

$$R = \begin{bmatrix} 1.0000 & 0.4559 & 0.5012 & 0.3596 \\ 0.4559 & 1.0000 & 0.4294 & 0.3531 \\ 0.5012 & 0.4294 & 1.0000 & 0.5820 \\ 0.3596 & 0.3531 & 0.5820 & 1.0000 \end{bmatrix}$$

Canonical Correlations

These are given in Table 22 below together with other variable. The first canonical correlation, r_1, is 0.526 representing about 32% ($r_1^2 = 0.316$) overlapping variance between η_1 and ϕ_1. The second canonical correlation is 0.055 representing a mere 0.3% overlapping variance between the pair (η_2, ϕ_2).

Dimensionality:

To determine the number of significant canonical

variates, we have to test the hypothesis

$$H_o: \quad \Sigma_{12} = 0$$

against

$$H_1: \quad \Sigma_{12} \neq 0.$$

Table 22: Canonical correlation coefficients, weights and loadings of scores in Mathematical Statistics and Applied Mathematics for first year, 1985/86.

k	Eigenvalue r_k^2	Canonical Correlation r_k	Canonical Weights and Loadings			
			Y-variables		Z-variables	
			Weights	Loadings	Weights	Loadings
1	0.316	0.562	0.683	0.904	0.827	0.978
			0.483	0.794	0.259	0.740
2	0.003	0.055	-0.892	-0.430	-0.910	-0.210
			1.014	0.608	1.202	0.673

Table 23 below gives the computed chi-square, χ_c^2, and the table chi-square, χ_α^2, values for various degrees of freedom. We see from the table that

$$\chi_c^2 > \chi_\alpha^2; \quad \text{for all} \quad k = 1,2 \quad .$$

Table 23: Test statistics for scores in Mathematical Statistics and Applied Mathematics for first year, 1985/86.

k	χ^2_c	$\chi^2_{(0.01)}$	$\chi^2_{(0.05)}$	d.f
1	15.51	0.297	0.711	4
2	0.12	0.0002	0.004	1

Thus, the null hypothesis of independence is rejected at levels of significance $\alpha = 0.01$ and 0.05. Although highly significant, only the first canonical correlation represents substantial relationship. Thus, the interpretation of the second pair of canonical variates is marginal.

Canonical Weights

We see that the first canonical variate, η_1, (Table 22) is associated with $Y_1(0.683)$. The variate ϕ_1 is mainly represented by $z_1(0.827)$. Therefore, Y_1 and Z_1 contribute more towards the first canonical correlation.

The second canonical variate, η_2, is mainly dominated by $Y_2(1.014)$ and Y_2 by $Z_2(1.202)$. Thus Y_2 and Z_2 contribute more significantly towards the second canonical correlation.

Canonical Loadings

These are also reported in Table 22, above. All the two Y-variables contribute in the same direction to the first canonical variate, η_1, and

all have appreciably high correlations with this variate. Above all, η_1 seems to be an expression of $Y_1(0.904)$. The correlations of the two Y-variables with η_2 are weaker than their correlations with η_1; the strongest correlation being that of $Y_2(0.608)$. We notice the negative correlation of Y_1 with η_2.

Considering the canonical variates ϕ_k (k=1,2) of the Z-variables, we see that the two variables making up this set, z_1 and z_2, have high correlations with ϕ_1 and the correlation are both positive. The strongest correlation with ϕ_1 is that of $z_1(0.978)$. The second canonical variate, ϕ_2, is largely characterized by $z_2(0.673)$; also notice the negative correlation of z_1 with ϕ_2.

The pair $(\eta_1;\phi_1)$ indicate that Y_1 and z_1 contribute most heavily to the first canonical correlation and accordingly the performance of students in the two variables is more correlated than in the other variables. We note, from the correlation matrix, R, that Y_1 and Z_1 have the highest correlation of 0.501, that is, between variables of the two measurement domains.

The second canonical variates, η_2 and ϕ_2, indicate that Y_2 and Z_2 contribute more to the second canonical correlation. We should note that the interpretation given by canonical loadings is consistent with that given by the canonical weights.

Cross Loadings

In Table 24 below, we notice that all the Y-variables are directly related to ϕ_1. The

Table 24: Cross Loadings for scores in Mathematical Statistics and Applied Mathematics for first year, 1985/86.

	ϕ_1	ϕ_2		η_1	η_2
Y_1	0.508	-0.024	Z_1	0.550	-0.012
Y_2	0.447	0.034	Z_2	0.416	0.037

correlation of Y_1 with ϕ_1 is about 0.508 which is the highest. We also recall from above that the variable Y_1 is more correlated with η_1 than Y_2. Thus, we can say that once a variable is highly correlated with a given canonical variate from one set it should be relatively highly correlated with the corresponding canonical variate from the other set. The variable correlations with ϕ_2 are small and not very important results can be got from them.

The variable $z_1(0.550)$ is more correlated with η_1 than $Z_2(0.416)$. Also, the correlations of variables with η_2 are not very important.

Variance Extracted by Canonical Variates

The first canonical variate, η_1, extracts 72% of variance from its own set of variables. That is

$$R^2_{(1)y} = \left[(0.904)^2 + (0.794)^2\right]/2$$

$$\simeq \quad 0.724$$

The second canonical variate, η_2, extracts

$$R^2_{(1)y} = \left[(0.430)^2 + (0.608)^2\right]/2$$

$$\simeq 0.277$$

of variance from its own set of variables (Y-set). This is about 28% of the total variance in this domain. Together, the two canonical variates extract 100% of the variance in Y domain.

Looking at the z-variables set, the proportion of variance accounted for by the first canonical variate, ϕ_1, is

$$R^2_{(1)z} = \left[(0.978)^2 + (0.740)^2\right]/2$$

$$\simeq 0.752,$$

which is 75% of the variance in the z-variables set. Further, the second canonical variate accounts for about 25% of the variance, that is

$$R^2_{(2)z} = \left[(-0.210)^2 + (0.673)^2\right]/2$$

$$\simeq 0.248 .$$

Therefore, the two canonical variates account for approximately 100% of the variance in the z-measurement domain.

Redundancy.

The redundancy in the Y-variables set generated by each of the canonical variates ϕ_1 and ϕ_2 is

$$R^2_{(1)y/z} = R^2_{(1)y}\, r^2_1$$

$$= 0.724 \times 0.316$$

$$\simeq 0.229,$$

and

$$R^2_{(2)y/z} = R^2_{(2)y}\, r^2_2$$

$$= 0.277 \times 0.003$$

$$\simeq 0.0008 ,$$

respectively.

Further, the redundancy in the z-variables set attributable to the canonical variates $\eta_k (k=1,2)$ is

$$R^2_{(1)z/y} = R^2_{(1)z}\, r^2_1$$

$$= 0.752 \times 0.316$$

$$\simeq 0.238 ,$$

and

$$R^2_{(2)z/y} = R^2_{(2)z}\, r^2_2$$

$$= 0.248 \times 0.003$$

$$\simeq 0.0007$$

respectively.

We see clearly that the redundancy in each variables set generated by the second canonical variates, η_2 and ϕ_2 are negligible. In other words, the total redundancies are only given by the first canonical variates, η_1 and ϕ_1.

4.2.6: Comparative Remarks

The squared canonical correlations expressing the overlapping variance between the pairs of canonical variates are given in Tables 18 and 22 for 1984/85

and 1985/86 respectively, In both cases, less than 50% of the overlapping variance is represented in first pair of canonical variates (r_1^2 = 0.408 and 0.316 for 1984/85 and 1985/86 respectively). However, for the two academic years, $r_2^2 \simeq 0.003$ which represents an almost neglible proportion of variance common to the second pair of canonical variates, (η_2, ϕ_2). Despite r_2^2 being so small, the number of significant canonical variates is two in both years.

Considering the canonical loadings for the two years, we see that in 1984/85 Y_1 and Z_2 are the variables which contribute most significantly to the first canonical correlation and therefore their performance is more correlated than in any other variables. In 1985/86, the same case does not arise because in this academic year, Y_1 and Z_1 are the variables whose performance is more correlated. We don't have much to say about r_2^2 because it is so small that its interpretation is marginal and thus may not be very clear.

The variance extracted from the Y-variables set by η_1 and η_2 is 100% of the total variance in the for for the year 1984/85. In the following year, 1985/86, η_1 and η_2 extract the same proportion of variance from Y-variables. Likewise, the proportion of explained, variance in the Z-variables set that is accounted for by ϕ_1 and ϕ_2 is 100% for the two academic years.

The redundancy in the Y-variables set generated by ϕ_1 and ϕ_2 is 29% for 1984/85 and 23% for 1985/86. On the other hand, the variates η_1 and η_2 extract 27% and 24% of the total variance in the Z-variables set for 1984/85 and 1985/86 respectively. However, we note that in both years, η_2 and ϕ_2 do not contribute significantly to the total redundancy. That is, the total redundancy may be attributable to only the first canonical variates, η_1 and ϕ_1.

4.3 LIMITATIONS OF THE TECHNIQUE

Canonical correlation has several important theoretical limitations that may explain its rarity in the literature. Perhaps the most critical limitation involves the interpretation and evaluation of the results obtained from a canonical correlation analysis. Although in principal component analysis and factor analysis linear combinations are usually rotated to facilitate interpretation, rotation of canonical variates is not common or even available through SPSS or BMD.

Not only are pairs of canonical variates restricted to linear relationships and limited in number, but also they are calculated to be independent of all other pairs. In this technique, only the orthogonal solution is available.

Occassionally, there arises the problem of missing data which can create inconsistencies and they should be estimated carefully. Cases that are unusual can

have untoward effects on canonical analysis just as they can on other multivariate techniques. The search for outliers should be conducted separately with each set of variables. Because calculation of canonical variates requires inversion of correlation matrices, it is important that the matrices be of full rank. That is, none of the variables should be too highly correlated with, or near linear combinations of, others in the set. Therefore, the problem of multicollinearity and singularity in correlation matrices should be identified and eliminated.

4.4 CONCLUSIONS:

Techniques for studying interrelationships among variables are quite few. Of the ones available, canonical correlation analysis appears to be well-suited to the analysis of one class of problem encountered in Statistics, namely that in which the relationship between the variables is essentially linear and continuous.

In the preceding discussions on the application of canonical correlation analysis, it cannot be claimed that the results are fully reliable. Nevertheless, it is true that much has been done in an attempt to interpret the results and thus indicating how the results of canonical correlation analysis can be interpreted. Furthermore, we have been able to indicate that as long as data can be grouped into two or more sets of variables, canonical correlation analysis can be used

to assess the interrelationships among the variables of the various sets (of course taking two sets at a time).

When we look at section 4.2.1, we see from the correlation matrices, R, that in 1984/85 academic year, Calculus I and Probability and Statistics I have the highest interrelationship; whereas in 1985/86 Linear Algebra II and Probability and Statistics II have the highest interrelationship. On using the interpretive devices, particularly the canonical loadings and cross loadings, these high interrelationships are confirmed. Further, this is true when we look at Sections 4.2.3 and 4.2.5 both case (I) and (II). Thus, canonical correlation analysis is able to recover known relationships among variables and in this case mathematics units.

Case (I) of Section 4.2.1 indicates that Calculus I and Probability and Statistics I contribute most significantly towards the first canonical correlation whereas of Case (II), Linear Algebra II and Probability and Statistics II contribute most towards the first canonical correlation. On looking at Case (I) of Section 4.2.3, we see that Calculus I and Vector Analysis contributed most to the first canonical correlation unlike in Case (II) where Linear Algebra I and Vector Analysis contributed more towards the first canonical correlation. Section 4.2.5 Case (I) indicates that Probability and Statistics I and Classical Mechanics contributed heavily towards the first canonical correlation whereas in

Case (II) it was Probability and Statistics I and Vector Analysis. This broad view on the analysis reveals that there is very little consistency in performance over the two academic years, 1984/85 and 1985/86. Of course we do not expect much consistency given that the number, N, of students who sat for the examinations was different, the students were themselves different and thus their interests in various units different, the lecturers who taught, set and marked the papers may have been different and without forgetting the conditions that might have surrounded the students when doing the examinations (short time for revision among others). Consistency is further disqualified when we look at the other interpretive devices like canonical weights, variance extracted from the canonical variates and the redundancy in Sections 4.2.1, 4.2.3 and 4.2.5.

It is my belief that in further most researchers in Statistics will see the necessity to venture into canonical correlation analysis so that we can improve its interpretability and thus propose it to other researchers in other areas as a useful technique in analysing complex sets of data.

REFERENCES

1. Anderson, T.W. (1984): An Introduction to Multivariate Statistical Analysis; Wiley, New York.

2. Gittins, R. (1979): "Ecological applications of canonical analysis", Multivariate Methods in Ecological Work; International Co-operative Publishing House, Fairland (U.S.A.)

3. Green, P.E. and Carroll, J.D. (1976): Mathematical Tools for Applied Multivariate Analysis; Academic Press, New York.

4. Harris, R.J. (1975): A Primer to Multivariate Statistics; Academic Press, New York.

5. Kendall, S.M.G. (1975): Multivariate Analysis; Griffins, London.

6. Mardia, K.W., Kent, J.T. and Bibby, J.M. (1979): Multivariate Analysis; Academic Press, London.

7. Morrison, D.F. (1979): Multivariate Statistical Methods; McGraw-Hill, New York.

8. Nie, N.H. (1975): Statistical Package for the Social Sciences; McGraw-Hill, New York.

9. Tabachnick, G.B., and Fidell, L.S. (1983): Using Multivariate Statistics; Harper and Row, New York.

10. Tanner, J.M. (1962): "Growth at Adolescence," 2nd edition; Oxford: Blackwell.

11. Waugh, F.W. (1942): "Regression between sets of variates", Econometrika; 10, 290.

12. William, R.D. and Mathew, G. (1984): "Multivariate Analysis," Methods and Applications; Wiley, New York.