

APPLICATIONS OF CANONICAL CORRELATION IN GEOLOGY

APPLICATIONS OF CANONICAL CORRELATION IN GEOLOGY

By

PEI JEN LEE

B. Sc. M.Sc.

A Thesis

Submitted to the Faculty of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

May 1968

DOCTOR OF PHILOSOPHY (1968)
(Geology)

McMASTER UNIVERSITY
Hamilton, Ontario

TITLE: Applications of Canonical Correlation in Geology

AUTHOR: Pei Jen Lee, B. Sc. (National Taiwan University)

M. Sc. (University of Western Ontario)

SUPERVISOR: Professor Gerard V. Middleton

NUMBER OF PAGES: xi, 124

SCOPE AND CONTENTS: The purpose of this thesis is to develop a technique for constructing a multivariate trend surface through canonical correlation analysis. The surface is called a canonical trend surface and is a succinct summarization of variations of a set of variates. It has the property of maximum correlation between the variates and their geographic coordinates.

The contents of this thesis include: (1) a literature review of both the theory and application of canonical correlation in psychology, biology, economics and social sciences, (2) a practical computation scheme for canonical trend analysis, with a computer program, and examples of this application, (3) suggestions for other applications of this theory in geology with numerical examples and computer programs.

ABSTRACT

The theory of canonical correlation analysis has been combined with that of trend surface analysis in order to construct a multivariate trend surface which is called a canonical trend surface.

A canonical trend surface is a parsimonious summarization of areal variations of a set of geological variates. This trend has a property of maximum correlation between variates and geographic coordinates. It does not show the absolute value of each variate, but it shows the nature of the variation of a linear combination of the variates. The Permian system in western Kansas and eastern Colorado was studied as a numerical example to illustrate the general procedures in solving practical problems and also to demonstrate the validity of this technique. By use of this type of trend it is possible to reveal the underlying pattern of geographic variation common to a set of variates.

Other applications of canonical correlation analysis in geology have been explained with illustrative geological examples, namely: the relationships between two sets of variates, matching two factor patterns, Q-technique canonical correlation, and discriminatory analysis.

Comparison of canonical correlation analysis and principal factor solution in factor analysis suggests that factor analysis may be more appropriate for suggesting interrelationships among variables, while canonical correlation analysis may be a suitable tool for prediction problems.

FORTRAN IV programs for these computations are listed in appendices with instructions for using them.

ACKNOWLEDGEMENTS

The writer is deeply indebted to Dr. Gerard V. Middleton for his helpful suggestions all through the work and also patiently reading the manuscript and contributing valuable comments.

Special thanks are due to Dr. H. P. Schwarcz, of the Department of Geology, Dr. M. Behara, and Dr. S. K. Banerjee, of the Department of Mathematics for kindly reviewing this manuscript.

TABLE OF CONTENTS

	PAGE
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 THEORY OF CANONICAL CORRELATION	6
2.1 General Statement	6
2.2 Canonical Correlation and Variates	7
2.3 Tests of Significance for Canonical Roots	13
CHAPTER 3 TWO-DIMENSIONAL CANONICAL TREND ANALYSIS	15
3.1 General Statement	15
3.2 Computation Procedures	18
3.3 Distinguishing Characters of the Canonical Trend	25
3.31 Experiment 1	26
3.32 Experiment 2	28
3.33 Experiment 3	28
3.34 Experiment 4	29
3.4 Functions of the Canonical Trend Surface	30
3.5 Permian System in Western Kansas and Eastern Colorado	34
3.51 Geological Setting	34
3.52 Stratigraphic factor map analysis	37
3.53 Canonical trend surface analysis	38
CHAPTER 4 OTHER APPLICATIONS IN GEOLOGY	49
4.1 Relationships Between Two Sets of Variates	49
4.11 General Statement	49
4.12 Recent Sediment from Buttonwood Sound	51

TABLE OF CONTENTS, continued	PAGE
4.2 Matching Two Factor Patterns	56
4.21 General Statement	56
4.22 Factor Patterns of Sediment from Buttonwood Sound	57
4.3 Q-Technique Canonical Correlation	60
4.31 General Statement	60
4.32 Ordovician and Silurian Sediments	61
4.4 Discriminatory Analysis by Canonical Variates	63
4.41 General Statement	63
4.42 Ordovician Limestones from Colborne Quarry	67
CHAPTER 5 CONCLUSIONS	70
REFERENCES	74

	PAGE
APPENDIX I FORTRAN IV PROGRAM FOR TWO-DIMENSIONAL CANONICAL TREND	82
6.1 Instructions for Using the Program	82
6.11 Input to Program	82
6.12 Output from Program	84
6.2 Listing of the Program	85
APPENDIX II FORTRAN IV PROGRAM FOR CANONICAL CORRELATION ANALYSIS	95
7.1 Instructions for Using the Program	95
7.11 Input to Program	95
7.12 Output from Program	98
7.2 Listing of the Program	99
APPENDIX III FORTRAN IV PROGRAM FOR DISCRIMINATORY ANALYSIS BY CANONICAL VARIATES	105
8.1 Instructions for Using the Program	105
8.11 Input to Program A	105
8.12 Output from Program A	106
8.13 Input to Program B	106
8.14 Output from Program B	107
8.2 Listing of the Program	108
APPENDIX IV FORTRAN IV PROGRAM FOR PRINCIPAL COMPONENT ANALYSIS	113
9.1 Instructions for Using the Program	113
9.11 Input to Program	113
9.12 Output from Program	116
9.2 Listing of the Program	118

LIST OF FIGURES

PAGE

- Figure 1 Flow chart for generating all terms up to a sextic polynomial based on geographic coordinates X and Y. 21
- Figure 2 Canonical trend surface fitted to $Z_1 = 0.707 X + 0.707Y + E_1$, $Z_2 = E_2$, where E_1 was $N(0,4)$, and E_2 was $N(0,1)$. 31
- Figure 3 Canonical trend surface (univariate case) shows an exponential variation. 31
- Figure 4 Permian system of Western Kansas and Eastern Colorado. Stratigraphic factor maps. A. Factor 1; B. Factor 2 (Data after Imbrie, 1963; redrawn by the present writer). 39
- Figure 5 Permian system of Western Kansas and Eastern Colorado. A. Third order canonical trend associated with second canonical root showing variations in thickness of the sand and the carbonate. B. Second order canonical trend associated with the first canonical root showing variations in thickness of the evaporites, the sand, and the shale. 46
- Figure 6 Permian system of Western Kansas and Eastern Colorado. Net thickness maps of the sand, the shale, the carbonate, and the evaporite. (Data after Krumbein, 1962; redrawn by the present writer). 47

- Figure 7 Permian system of Western Kansas and Eastern Colorado
- A. Canonical trend surface for the total thickness, the sand/shale ratio, and the carbonate/evaporite ratio. Isopach map (right upper), sand/shale ratio map (left lower), and carbonate/evaporite ratio map (right lower). (Data after Krumbein, 1962; redrawn by the present writer).
- Figure 8 Ordovician limestone specimens from Colborne quarry (dot: section A; square: section B; and circle: section C) are shown on a canonical coordinates based on first two canonical variates. Mean values of all sections are shown in the upper right corner.

48

69

LIST OF TABLES

	PAGE
Table 1 Rotated and Normalized Eigenvectors of Sediment Thickness of Permian system.	40
Table 2 Record of the Successive Evaluation of the order of Polynomial for Four Variates.	41
Table 3 Record of the Successive Evaluation of the order of Polynomial for Three Variates.	44
Table 4 Simple Correlation Coefficients for Sediments from Buttonwood Sound.	52
Table 5 Canonical Variates for Sediments from Buttonwood Sound.	53
Table 6 Rotated and Normalized Five Principal Components for Sediments from Buttonwood Sound.	55
Table 7 Factor Patterns for Sediments from Buttonwood Sound.	58,59
Table 8 Canonical Variates for Two Factor Patterns	59,60
Table 9 Canonical Correlation Matrix for the Four Formations from Ontario	62
Table 10 Canonical Coordinates of Limestones from Colborne Quarry, Ontario.	68

CHAPTER 1

INTRODUCTION

Suppose we have a number of rock specimens taken from a rock body, and that the abundance of various kinds of minerals, major elements, trace elements, fossils and so on, are estimated from each specimen. It may be asked whether or not there are some relationships between the major and trace elements, or between the minerals and size parameters, or between trace elements and clay minerals, or between sediment types and fossils.

Let us take the trace elements and clay minerals as an example. One simple correlation coefficient indicates a possible relation between one of the trace elements and one of the clay minerals. If we had five clay minerals and ten trace elements, we will have 50 simple correlation coefficients to show all possible relationships. The relationships indicated by 50 coefficients are not easily understood: what is needed is some technique which will express the relationship between the set of trace elements and the set of clay minerals in a more concise form.

Canonical correlation analysis was introduced by Hotelling in 1936 (the paper was read in 1935). It is a technique to understand the relationships between two sets of variables (for example, trace elements and clay minerals) and was an outgrowth of simple correlation analysis. The relationship between the two sets of variables are summarized and expressed by a simple index. For example, the clay mineral and

trace element abundances are weighted and combined into two linear combinations such as

$$U = a_1 (\text{kaolinite}) + a_2 (\text{illite}) + a_3 (\text{chlorite}) + \dots$$

$$V = b_1 (\text{vanadium}) + b_2 (\text{zirconium}) + b_3 (\text{titanium}) + \dots$$

where $a_1, a_2, a_3, \dots, b_1, b_2, b_3, \dots$ are weights to be determined such that U and V have a maximum simple correlation coefficient which is called the canonical correlation coefficient or canonical root.

The strength of the link between the sets is measured by the canonical root, whereas the nature of the link is indicated by those trace elements and clay minerals having larger weights.

The principle of canonical correlation analysis was immediately used as a procedure for discriminating between two populations and also as a technique for testing difference between groups. The technique has the properties of maximizing the difference between two populations and also displaying each individual on a set of canonical coordinates (Bartlett, 1938). Canonical correlation was generalized for three or more sets of variables and partial canonical correlation was also proposed by Roy (1957, p. 26). Canonical factor analysis (Rao, 1955) was developed by making use of the canonical correlation principle. Horst (1961, 1961a) has applied the same principle to the problem of matching two or more factor patterns. The probability function of the canonical roots has been studied by a number of mathematical statisticians (Hsu, 1941; Bartlett, 1941, 1947; Constantine and James 1958). Uses of this technique has been suggested in various fields: psychology (Bartlett, 1947, 1948; Thomson, 1947; Burt, 1948; Hotelling,

1957; Beech and Maxwell, 1958; Maxwell, 1961; Kendall, 1961, p. 75-85; Meredith, 1964; Das, 1965; Duntzman, 1967), biology (Pearce and Holland, 1960; Kshirsagar, 1962; Bartlett, 1965, p. 201-224; Seal, 1964, p. 123-152), economics (Waugh, 1942; Tintner, 1946; Morrison, 1967, p. 207-220), sociology (Cooley and Lohnes, 1962, p. 31-49; Koons, 1962), and geology (Buzas, 1966; Reyment, 1966).

Suppose that a number of rock specimens have been collected from an area, so that each specimen has its own geographic locality. After estimation of trace elements, for instance, we would like to know the areal variation of each trace element over the area studied. The simplest technique is to plot each individual value of a particular trace element on the map according to its locality, and then to draw contour lines showing equal abundance of that trace element. This type of map shows not only regional variation but also local fluctuations. If we are able to filter out the local fluctuations, then the regional variation will be accentuated. This could be done by fitting a polynomial surface to this set of data over the area. Deviations from the polynomial are considered to be residuals or local fluctuations and the computed surface is considered to be the regional 'trend'. If we had ten trace elements, then we will have ten polynomial surfaces or trend maps indicating variations of trace elements. It may be asked whether or not is it possible to evaluate a single trend which is common to all or some of the trace elements. Again, we encounter the same problem as in the previous example, i.e. we need a summarization of the ten maps.

After a period of discussion about the possible application of the canonical correlation analysis in geology at McMaster University in 1965, Dr. G. V. Middleton suggested the canonical trend idea. Since that time, extensive studies were carried out on both real geological and hypothetical data in order to understand the geological meanings of this technique.

A canonical trend is a polynomial surface which shows the variations of many variables over an area simultaneously by making use of the principle of canonical correlation analysis. In this case, one set of variables consists of geological variables such as trace elements, whereas the second set of variables is composed of location of specimens, i.e. (X, Y) coordinates. The canonical trend is a succinct summarization of areal variations of many geological variables and their geographical coordinates.

Application of canonical trend analysis was introduced by Lee and Middleton (1967) before a colloquium on trend analysis held in Lawrence, Kansas, and also by Middleton and Lee (1967) before the seventh International Sedimentological Congress held in England.

A quadratic surface which is an homogeneous expression of a second degree is reduced to a linear combination of squares only, the cross-product terms being eliminated. A form of this type is said to be a canonical form. This reduction process is also called a canonical analysis in chemical engineering (Hill and Hunter, 1966). This canonical reduction of chemical engineers is algebraically equivalent to the canonical analysis of Hotelling's method, but the underlying purposes, assumptions and implications of the canonical reduction are completely

different from that of the canonical trend analysis discussed in this paper.

In Section 2.1 the physical meaning of the canonical correlation coefficient is introduced. A proof of the theory of the analysis is presented in Section 2.2. Section 2.3 states the probability distribution of canonical roots. Detailed computation procedures of the two-dimensional canonical trend are formulated in Section 3.2. Sections 3.3 and 3.4 are devoted to a full discussion of the physical meaning and uses of canonical trends, using hypothetical data. A numerical example, taken from the Permian of Kansas and Colorado is given in Section 3.5. Other applications in geology are demonstrated with illustrative examples in Chapter 4.

CHAPTER 2

THEORY OF CANONICAL CORRELATION

2.1 General Statement

Suppose we have a sample from a p -dimensional space, then the purpose of canonical correlation analysis (Hotelling, 1936) is to find a linear function of the first p_1 -variates and a linear function of the last p_2 -variates ($p_1 + p_2 = p$), so that these two linear functions have the highest possible correlation coefficient. In a practical situation, interrelationships between two sets of measurements made on the same samples would be analyzed through this technique. Under the assumption of normality, if the canonical correlation is zero, these two sets are completely independent, and it is useless to predict the dependent variates by means of the independent variates. If the canonical correlation is unity, this means that the dependent variates would be predicted perfectly by means of the independent variates based on the particular linear functions. For the special case in which the number of dependent variates is equal to one, the problem is multiple regression. In this case, it is usual to make use of a regression model rather than a correlation model.

The geometrical meaning of the canonical correlation can be stated as follows: In a p -dimensional space, a sample of $p_1 + p_2$ variates determines one hyperplane of p_1 and one of p_2 dimensions, intersecting at the origin, and containing a swarm of points representing

the two sets. Linear transformations are developed for the first p_1 coordinate axes and also for the p_2 coordinate axes such that these two hyperplanes are as parallel as possible in a new p -dimensional space. The cosine of the angle between these two hyperplanes is defined as the canonical correlation coefficient.

The assumption is that the observed variates are linear functions of the canonical variates. Furthermore we will assume that the observed variates are normally distributed in order to make a statistical inference on the dependence between two sets and to derive the probability distribution of the canonical correlation coefficients.

2.2 Canonical Correlation and Variates

Suppose $(Z_{ij}, i = 1, 2, \dots, p; j = 1, 2, \dots, N, N > p)$ is a random sample of size N from a p -dimensional distribution and has covariance matrix R which is known to be a positive definite real symmetric matrix. Without loss of generality we may suppose that Z_i has zero mean, i.e. $E[Z_i] = 0$.

We partition Z into two subvectors of p_1 and p_2 components ($p = p_1 + p_2$) respectively,

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}, \text{ where } Z_1 = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{p_1} \end{bmatrix}, \quad Z_2 = \begin{bmatrix} z_{p_1+1} \\ z_{p_1+2} \\ \vdots \\ z_p \end{bmatrix} \quad (1)$$

For convenience we shall assume $p_1 \leq p_2$. The covariance matrix is partitioned into matrices as follows:

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array} \right] \quad (2)$$

where R_{11} is the correlation matrix for Z_1 , R_{22} for Z_2 , and $R_{12} = R'_{21}$ is the correlation matrix between Z_1 and Z_2 . The canonical variates U and V are defined as:

$$U = A' Z_1, \quad V = B' Z_2 \quad (3)$$

where

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{p_1} \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{p_2} \end{bmatrix} \quad (4)$$

We require A and B to be such that U and V have unit variance, that is

$$1 = E[U^2] = E[A' Z_1 Z_1' A] = A' R_{11} A \quad (5)$$

$$1 = E[V^2] = E[B' Z_2 Z_2' B] = B' R_{22} B \quad (6)$$

$$\begin{aligned} E[UV] &= \text{COV}[UV] + E[U] E[V] \\ &= E[A' Z_1 Z_2' B] + E[A' Z_1] \cdot E[B' Z_2] \\ &= A' E[Z_1 Z_2'] B + A' E[Z_1] B' E[Z_2] \\ &= A' R_{12} B \end{aligned} \quad (7)$$

Thus the problem is to find A and B to maximize (7) or $E[UV]$ subject to (5) and (6).

$$\text{Let } \psi = A' R_{12} B - \frac{1}{2} \lambda (A' R_{11} A - 1) - \frac{1}{2} \mu (B' R_{22} B - 1) \quad (8)$$

where λ and μ are Lagrange multipliers. We differentiate ψ with respect to the variables of A and B. The vectors of derivatives set equal to zero are

$$R_{12} B - \lambda R_{11} A = 0 \quad (9)$$

$$R_{12}' A - \mu R_{22} B = 0 \quad (10)$$

Multiplication of (9) on the left by A' and (10) on the left by B' gives

$$A' R_{12} B - \lambda A' R_{11} A = 0 \quad (11)$$

$$B' R_{12}' A - \mu B' R_{22} B = 0 \quad (12)$$

Since $A' R_{11} A = 1$, $B' R_{22} B = 1$, and $(A' R_{12} B)' = B' R_{12}' A$. Thus, we have

$$\lambda = \mu = A' R_{12} B \quad (13)$$

Thus (9) and (10) can be written as

$$-\lambda R_{11} A + R_{12} B = 0 \quad (14)$$

$$R_{21} A - \lambda R_{22} B = 0 \quad (15)$$

In one matrix equation this is

$$\begin{bmatrix} -\lambda R_{11} & R_{12} \\ R_{21} & -\lambda R_{22} \end{bmatrix} \cdot \begin{bmatrix} A \\ B \end{bmatrix} = 0 \quad (16)$$

In order that there be a nontrivial solution, the matrix on the left must be singular. Multiplication of (14) by λ and (15) on the

left by R_{22}^{-1} , gives

$$\lambda R_{12} B = \lambda^2 R_{11} A \quad (17)$$

$$R_{22}^{-1} R_{21} A = \lambda B \quad (18)$$

Substitution from (18) into (17) gives

$$R_{12} R_{22}^{-1} R_{21} A = \lambda^2 R_{11} A \quad (19)$$

or

$$(R_{12} R_{22}^{-1} R_{21} - \lambda^2 R_{11}) A = 0 \quad (20)$$

or

$$(R_{11}^{-1} R_{12} R_{22}^{-1} R_{21} - \lambda^2 I) A = 0 \quad (21)$$

The solution involves finding latent roots λ^2 of the equation

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

The matrix, $R_{11}^{-1} R_{12} R_{22}^{-1} R_{21}$, is $(p_1 \times p_1)$ in dimension. Thus λ is the $(p_1 \times p_1)$ diagonal matrix.

From (13) we see that $\lambda = A' R_{12} B$ is the correlation between U and V . Thus, the elements, λ_i , of λ were called the canonical roots or canonical correlation coefficients by Hotelling (1936). Values of A_i in equation (3) are the eigenvectors associated with λ_i^2 . Solving for B from equation (15), we have B_i for a particular λ_i ,

$$B_i = R_{22}^{-1} R_{21} A_i / \lambda_i \quad (\lambda_i \neq 0) \quad (23)$$

The A_i and B_i are normalized, so that we have

$$A_i' A_i = 1 \quad \text{and} \quad B_i' B_i = 1 \quad (24)$$

Then U_i and V_i are normalized linear functions of Z_1 and Z_2 , respectively, with maximum correlation.

We now consider finding the second linear functions of Z_1 and Z_2 , respectively, such that each of these two linear functions are uncorrelated with the first linear functions. This procedure is continued. At the r -th step we have obtained linear combinations

$U_1 = A_1' Z_1$, $V_1 = B_1' Z_2$, ..., $U_r = A_r' Z_1$, $V_r = B_r' Z_2$ with corresponding roots λ_1 , ..., λ_r . Let the $(r+1)$ th linear functions of Z_1 and Z_2 be $U_{r+1} = A_{r+1}' Z_1$, and $V_{r+1} = B_{r+1}' Z_2$, respectively. The condition that U_{r+1} be uncorrelated with U_i ($1 \leq i \leq r$) is

$$\begin{aligned} 0 &= E[U_i U_{r+1}] = E[A_i' Z_1 Z_1' A_{r+1}] \\ &= A_i' R_{11} A_{r+1}. \end{aligned} \quad (25)$$

$$R_{11} A_i = R_{12} B_i / \lambda_i \quad (\lambda_i \neq 0) \quad (26)$$

$$\text{Therefore, } E[U_{r+1} V_i] = A_{r+1}' R_{12} B_i = \lambda_i A_{r+1}' R_{11} A_i = 0 \quad (27)$$

If $\lambda_i = 0$, $R_{12} B_i = 0$ and (27) holds.

The condition that V_{r+1} be uncorrelated with V_i is

$$0 = E[V_{r+1} V_i] = B_{r+1}' R_{22} B_i \quad (1 \leq i \leq r) \quad (28)$$

By the same argument we have

$$0 = B_{r+1}' R_{21} A_i = E[V_{r+1} U_i] \quad (29)$$

We now maximize $E[U_{r+1} V_{r+1}]$, choosing A and B to satisfy (5), (6), (25), and (28) for $i = 1, 2, \dots, r$. Let

$$\begin{aligned} \psi_{r+1} &= A' R_{12} B - \frac{1}{2} \lambda (A' R_{11} A - 1) - \frac{1}{2} \mu (B' R_{22} B - 1) \\ &\quad + \sum_{i=1}^r v_i A' R_{11} A_i + \sum_{i=1}^r \theta_i B' R_{22} B_i \end{aligned} \quad (30)$$

where $\lambda, \mu, v_1, \dots, v_r, \theta_1, \dots, \theta_r$ are Lagrange multipliers.

We differentiate ψ_{r+1} with respect to the variables of A and B and set equal to zero. Thus, we have

$$R_{12} B - \lambda R_{11} A + \sum_{i=1}^r v_i R_{11} A_i = 0 \quad (31)$$

$$R_{21} A - \mu R_{22} B + \sum_{i=1}^r \theta_i R_{22} B_i = 0 \quad (32)$$

Multiplication of (31) on the left by A'_j ($j = 1, 2, \dots, r$) and (32) on the left by B'_j gives

$$0 = v_j A'_j R_{11} A_j = v_j \quad (33)$$

$$0 = \theta_j B'_j R_{22} B_j = \theta_j \quad (34)$$

Thus equations (31) and (32) are simply (16). Therefore any λ_i from the p_1 roots satisfies the conditions (5), (6), (25), and (28) for $i = 1, 2, \dots, r$.

Finally, the canonical transformation can be summarized into the following matrix equation:

$$\begin{bmatrix} A' & 0 \\ 0 & B'_1 \\ 0 & B'_2 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \begin{bmatrix} A & 0 & 0 \\ 0 & B_1 & B_2 \end{bmatrix} = \begin{bmatrix} I & \Lambda & 0 \\ \Lambda & I & 0 \\ 0 & 0 & I \end{bmatrix} = E \begin{bmatrix} U \\ V \end{bmatrix} \begin{bmatrix} U' & V' \end{bmatrix} \quad (35)$$

where $U = A' Z_1$, $V = [V_1 \ V_2]'$, $[B_1 \ B_2]'$, $[Z_1 \ Z_2]'$,

$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{p_1})$, and λ_i are the population canonical roots.

Consider approximating U by a multiple of V , say $K V$; then the mean square error of approximation is

$$\begin{aligned} E \left[(U - K V)^2 \right] &= \sigma_u^2 - 2 K \sigma_u \sigma_v \rho + K^2 \sigma_v^2 \\ &= \sigma_u^2 (1 - \rho^2) + (K \sigma_v - \rho \sigma_u)^2 \end{aligned} \quad (36)$$

where the σ_u^2 and σ_v^2 are variances of U and V , respectively, ρ is simple correlation coefficient between U and V . This is minimized by taking $K = \sigma_u \rho / \sigma_v$. We can consider $K V$ as a linear prediction of U from V ; then $\sigma_u^2 (1 - \rho^2)$ is the mean square error of prediction.

Detailed discussion of canonical correlation is given by Anderson (1958, p. 288-306) and Wilks (1963, p. 587-590).

2.3 Tests of Significance for Canonical Roots

The distribution of λ_i^2 of equation (22) has been discussed by Hsu (1941), Bartlett (1947), Constantine and James (1958), Anderson (1958, p. 323-324), and Wilks (1963, p. 590-592). For the case $p_1 \leq p_2$, the distribution of λ_i^2 , when these roots are arranged in descending order of magnitude, is given by

$$\begin{aligned} \pi^{-\frac{1}{2} p_1} \prod_{i=1}^{p_1} \frac{\Gamma \left[\frac{1}{2} (N-1) \right]}{\Gamma \left[\frac{1}{2} (N-p_2-i) \right] \Gamma \left[\frac{1}{2} (p_1+1-i) \right] \Gamma \left[\frac{1}{2} (p_2+1-i) \right]} \\ \prod_{i=1}^{p_1} \left[(\lambda_i^2)^{\frac{1}{2} (p_2-p_1-1)} (1-\lambda_i^2)^{\frac{1}{2} (N-p_2-p_1-2)} \right] \prod_{i < j}^{p_1} (\lambda_i^2 - \lambda_j^2) \end{aligned} \quad (37)$$

This distribution holds when Z_1 and Z_2 are independently distributed and Z_1 has a multivariate normal distribution, whereas

Z_2 has any distribution.

A criterion which is useful in detecting the simultaneous departure of several roots λ_i^2 from zero was suggested by Bartlett (1938, 1941).

Bartlett's statistic

$$X^2 = - \left[N - \frac{1}{2} (p_1 + p_2 + 1) \right] \log_e \Lambda \quad (38)$$

where $\Lambda = \prod_{i=1+r}^{p_1} (1 - \lambda_i^2)$, follows approximately a chi-square distribution with $(p_1 - r)(p_2 - r)$ degrees of freedom. The assumption is made that Z_1 and Z_2 follow a multivariate normal distribution with zero means.

A slight improvement which was suggested by Lawley (1959) for the test of significance of residual canonical roots is that the multiplying factor $N - \frac{1}{2}(p_1 + p_2 + 1)$ should be replaced by taking the factor as

$$N - r - \frac{1}{2}(p_1 + p_2 + 1) + \sum_{i=1}^r (1 / \lambda_i^2) \quad (39)$$

If λ_i^2 's are equal to 1, Bartlett's criterion and Lawley's criterion are identical. If the sample size N is large, these two criteria are approximately equal.

If canonical roots are much less than 1.000, then Lawley's criterion will yield higher chi-square value than that of Bartlett's criterion. This means that type I error will be high if Bartlett's criterion is adopted, on the other hand, type II error must be high if Lawley's criterion is used. As will be illustrated later, the use of statistical significant tests presents a problem of reconciling geological significance and statistical significance. In the present study, Bartlett's criterion is used for reference but not as the sole basis for making critical decisions.

CHAPTER 3

TWO-DIMENSIONAL CANONICAL TREND ANALYSIS

3.1 General Statement

Investigations may be performed to analyze a set of geological variates measured on a stratigraphic unit or a rock body spreading over a large or small area. What geologists would like to do is to filter the error variance from the systematic areal variance in order to evaluate the trend of a certain variate. The technique (trend analysis) which has been frequently used in geology is basically the fitting of a polynomial surface to the observed data of a single variate, by the principle of least squares. In geology, where most problems are beset with a highly multiple determination of events, conventional trend analysis is essentially a univariate technique which is not adequate to handle the inherently multivariate data.

The concept of a composite end member in facies analysis was introduced by Krumbein (1955). Three predetermined end members were used to construct a facies triangle. Selection of a particular point within this facies triangle was considered as an optimum combination of end members. Contour lines around this point were used to measure the deviation of facies from this optimum facies. The deviation was called the distance function. This technique, as Forgotson (1960) pointed out, does not distinguish end members or give any information on the absolute values of end members. It does not indicate the nature

of the change in composition from the optimum facies.

An entropy function derived from information theory was applied in multicomponent system mapping (Pelto, 1954) to express the degree of intermixing of the end members and to define facies quantitatively with one set of contours. This technique does not distinguish between end members and show the change of each individual variate.

The D-function (Pelto, 1954) divides a system into classes based on the difference in amounts of components. It provides information on the relative proportion of a specific end member within its own class. It does not distinguish between end members and provide information on the absolute value of components.

One elegant approach to the problem of dealing with spatially distributed multivariate data is to process the data first through principal component analysis, and then use trend analysis to map each individual principal component. It provides information on the areal distribution of the principal component which is composed of a set of geological variates with different weights. There are two ways to approach this problem.

(1) An R-mode factor or principal component analysis is first performed on the well (and/or outcrop) data, and then factors or principal components are expressed in terms of observed variates by using the principle of multivariate regression. After calculating the factors or principal components for each well, trend surfaces may be fitted to each factor or principal component over all wells. The result is a factor or principal component map. When a factor model is used, the factor map implies lithologic associations resulting from

underlying deposition environmental factors. When based on the concept of principal component analysis, the first principal component map indicates the spatial variation of the first principal component, which is the component which explains the most variance of a set of geological variates. In the both models, the factor or principal component may or may not have a clear geological interpretation.

(2) A factor-vector map is obtained from a Q-mode factor analysis as suggested by Imbrie (1963) and Krumbein and Imbrie (1963). Reference wells are chosen for each factor-vector map and all other wells are expressed in terms of similarity to the reference well for that particular map. The geological meaning of each map can only be obtained by looking at the particular lithologic association of each reference well. The concept of a reference well is analogous to the concept of an optimum facies in the distance function technique (Krumbein, 1955). This kind of factor-vector map expresses the intergradation between the reference wells in which respect it is similar to the concept of an entropy function map (Pelto, 1954).

Another technique will be discussed here, namely the construction of a trend surface for a set of multivariate data through a canonical transformation. The basic difference between this canonical trend analysis and the method of least squares is that in canonical trend analysis the covariance or correlation between geological variates and geographic coordinates is maximized, while in conventional trend analysis the variance is maximized (i.e. the error variance is minimized). The canonical trend surface is a polynomial surface which

has a maximum correlation with a set of geological variates combined together as a linear function. The coefficient of each variate indicates the degree of contribution of the particular variate to the trend surface, because the sum of squares of each coefficient is equal to 1, the total variance of the linear function.

3.2 Computation Procedures

The geographic coordinates, X and Y , and their various powers and cross-products constitute one set of variates, and the geological variates, Z_i , constitute the second set. What the canonical correlation analysis can do is to find one linear function for the (X,Y) coordinates, and one for Z_i such that these two linear functions have a maximum correlation. The trend surface evaluated by this principle is considered to be a most predictable surface for the particular linear function of a set of geological variates. An additional assumption is that X and Y are uncorrelated. This means that samples are taken by using a systematic or stratified sampling method from a two-dimensional area but not from an one-dimensional line. The calculation procedures are fully explained below:

STEP 1. Transformation of geological variates if necessary.

When the distribution of raw data is highly skewed or if the error variance is not stable, transformations may be carried out in order to symmetrize the distribution or stabilize the variance, or both. Bartlett (1947b) has summarized those transformations appropriate to a particular situation, and also suggested a general transformation as follows:

$$g(m) = \int \frac{c}{\sqrt{f(m)}} dm \quad (1)$$

where $f(m)$ is a function between variance, σ^2 , and mean, m , of a particular variate, c is a constant.

There are three alternatives allowed in the computer program prepared for the present study, (1) arc sine square root transformation; (2) logarithm transformation; and (3) no transformation.

STEP 2. Standardization of geological variates.

A correlation coefficient is a dimensionless value, therefore a trend surface equation derived from a correlation matrix is also scale independent. In order to have a scale dependent trend surface equation, the covariance matrix for the (X, Y) coordinates and geological variates z_i , should be used. On the other hand, geological variates may be measured in noncomparable units. Thus, we should deal with the correlation matrix rather than the covariance matrix. In order to satisfy both of the two requirements, the geological variates are standardized according to equation (2), and then the covariance matrix is computed for z_i and the (X, Y) coordinates.

The equation for standardization of the geological variates is as follows:

$$z_{ij} = \frac{z_{ij} - \bar{z}_j}{\sqrt{s_j^2}} \quad (2)$$

where s_j^2 = sample variance of variate z_j , \bar{z}_j = sample mean of variate z_j , z_{ij} is the i -th observation on the j -th variate.

STEP 3. Generation of high power and cross-product terms of the (X, Y) coordinates.

The high power and cross-product terms of the (X, Y) coordinates are generated within the computer based on the input raw data (X,Y) as follows:

Cumulative

Polynomial Term

2	X	Y					
5	X ²	XY	Y ²				
9	X ³	X ² Y	XY ²	Y ³			
14	X ⁴	X ³ Y	X ² Y ²	XY ³	Y ⁴		
20	X ⁵	X ⁴ Y	X ³ Y ²	X ² Y ³	XY ⁴	Y ⁵	
27	X ⁶	X ⁵ Y	X ⁴ Y ²	X ³ Y ³	X ² Y ⁴	XY ⁵	Y ⁶

The polynomial terms of each observation will follow the geological variates in an array D(I), where I = 1, 2, ..., p₁ + p₂, where p₁ = the number of geological variates, p₂ = the number of polynomial terms. The process of generation is listed in figure 1.

STEP 4. Computation of covariance matrix.

The covariance between i-th and j-th variates is calculated by standard formula which is defined as

$$S_{ij}^2 = \frac{\sum_{k=1}^N (Z_{ik} Z_{jk}) - \sum_{k=1}^N Z_{ik} \sum_{k=1}^N Z_{jk}}{N(N-1)} \quad i \neq j \quad (3)$$

If i = j, S_{ii}² is the variance of i-th variate, which is defined as

$$S_{ii}^2 = \frac{\sum_{k=1}^N Z_{ik}^2 - \left(\sum_{k=1}^N Z_{ik} \right)^2}{N(N-1)} \quad (4)$$

The covariance matrix is partitioned into four parts as

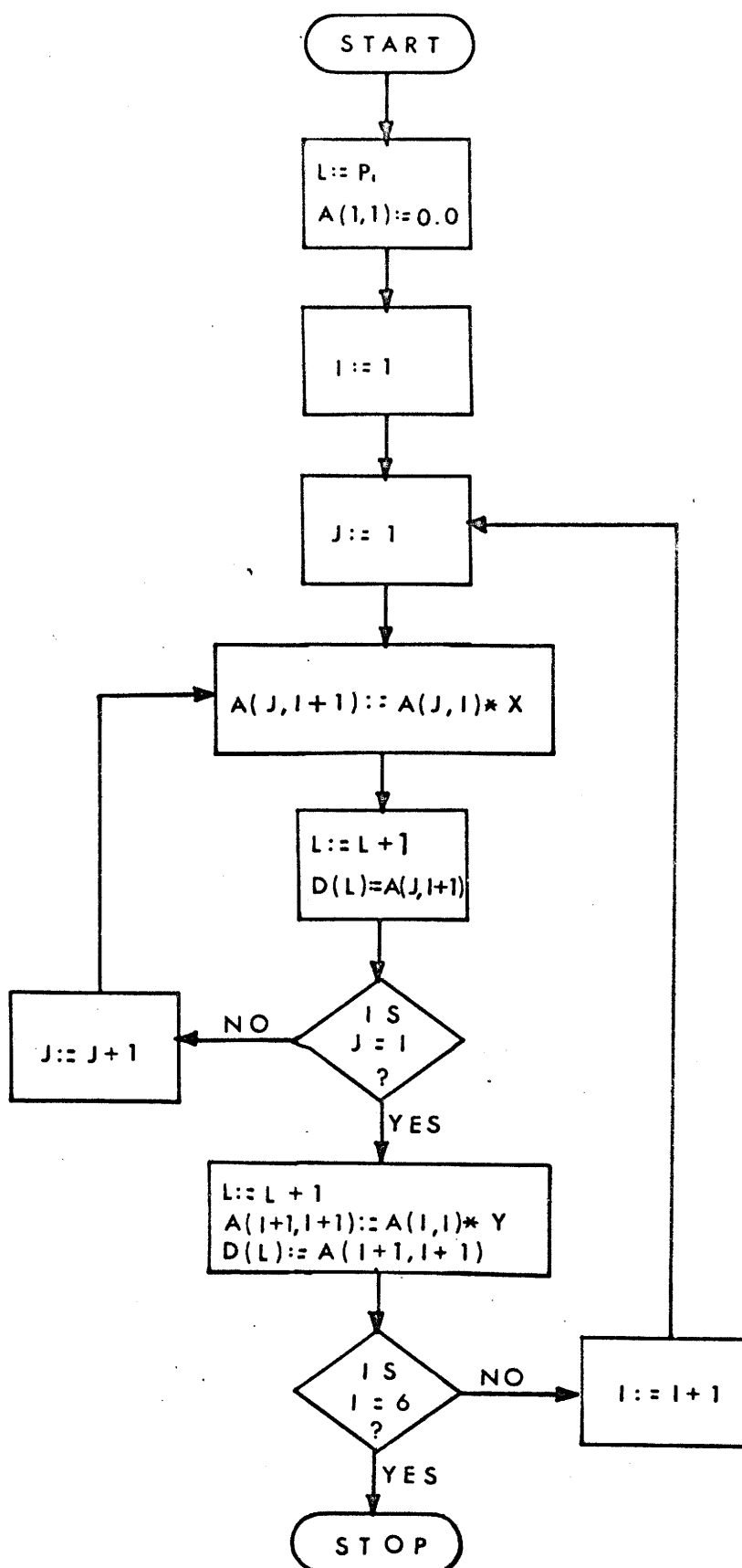


FIGURE 1 - Flow chart for generating all terms up to a sextic polynomial based on geographic coordinates X and Y.

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array} \right] \quad (5)$$

where the order of R_{11} is less than that of R_{22} , R_{11} and R_{22} are covariance matrices of geological variates and polynomial terms, respectively if $p_1 \leq p_2$. The contents of R_{11} and R_{22} should be exchanged if $p_1 > p_2$. This covariance matrix is a symmetrical matrix: each element is stored into an one-dimensional array at location L , where $L = j(j-1)/2 + i$, $i=1, 2, \dots, p_1 + 27$ for a sextic polynomial, $j \geq i$.

STEP 5. Calculation of canonical roots and variates.

The canonical roots λ 's and their associated canonical variates (A_i) for variates in matrix R_{11} , are respectively the eigenvalues and eigenvectors of the matrix equation:

$$\left| \begin{array}{cc} R_{11}^{-1} & R_{12} \\ R_{21} & R_{22}^{-1} - \lambda^2 I \end{array} \right| A = 0 \quad (6)$$

The matrix on the left side of (6) is nonsymmetric ($p_1 \times p_1$). A Jacobi-like method (Eberlein, 1962) was used to solve for the p_1 roots, λ^2 's. The eigenvectors are A_i each of which is the canonical variate for one set of variates, either the canonical variate for the geological variates or for the (X, Y) coordinates. The A_i are normalized, i.e. $A_i^T A_i = 1$.

STEP 6. Determination of degree of polynomial surface.

We begin with a linear polynomial and evaluate the degree through canonical correlation analysis. When the canonical root is

greater than 0.950 or the difference between two successive roots is less than 0.05 (the initial canonical root is set to zero), then the canonical root corresponding to the degree of the polynomial is used as the canonical trend surface. If not, the degree of the polynomial is increased 1, and steps 5 and 6 are repeated until the iterative process reaches a suitable degree.

It should be kept in mind that the sample size should always be greater than the sum of the number of the geological variates and the number of polynomial terms. Suppose we have a sample of size 30 with 4 variates, the highest order of polynomial that may be fitted to this data is 5, because a sextic polynomial must have 27 terms. In this example, for a sextic, the sum (= 31) of the number of variates (= 4) and number of polynomial terms (= 27) is greater than the sample size.

STEP 7. Calculation of the canonical variate of the second set of Variates.

The canonical variate of the second set is computed as follows:

$$B_i = R_{22}^{-1} R_{21} A_i / \lambda_i \quad (\lambda_i \neq 0) \quad (7)$$

The B_i is normalized, i.e. $B_i^t B_i = 1$.

STEP 8. Calculation of residuals.

Consider approximating U (a linear function of geological variates) by a multiple of V (a linear function of the X-, Y-coordinates) plus a constant, say $c + kV$, then k and c are defined as follows:

$$\underline{U} = c + k V \quad (8)$$

$$k = \frac{N \sum_{i=1}^N (V_i U_i) - \sum_{i=1}^N V_i \sum_{i=1}^N U_i}{N \sum_{i=1}^N V_i^2 - \left(\sum_{i=1}^N V_i \right)^2} \quad (9)$$

$$c = \frac{\sum_{i=1}^N V_i^2 \sum_{i=1}^N U_i - \sum_{i=1}^N (V_i U_i)}{N \sum_{i=1}^N V_i^2 - \left(\sum_{i=1}^N V_i \right)^2} \quad (10)$$

The residual of canonical trend surface for the j-th sample and i-th root is defined as:

$$\text{Residual}_j = U_j - (k V_j + c) = A'_j Z_j - (k V_j + c) \quad (11)$$

where Z_j is column matrix for standardized geological variates, thus, $A'_j Z_j$ is called observed value, whereas $C + kV_j$ is called calculated value for U_j . The U and V should be exchanged, i.e. A_i is replaced by B_i if U is a linear function of the (X, Y) coordinates while V is a linear function of geological variates.

STEP 9. Contouring of canonical trend surface.

In the canonical trend surface map, the X-axis is the abscisa, whereas the Y-axis is the ordinate. In the computer printout the length along X-axis is greater than or equal to that of Y-axis. The length of X-axis is divided into 50 units, while the length of Y-axis will be assigned to a certain units proportional to their relative lengths. Each point (having integer coordinates) of the new coordinate system is substituted into the polynomial equation $V = f(X, Y)$, so we have a set of V_{ij} , where $i = 1, 2, \dots, 50$, $j \leq i$. The values, V_{ij} ,

are scaled according to equation $c + k V_{ij}$.

Suppose U is a linear function of geological variates, then the calculated value for U is $k V + c$ as mentioned above. The difference between U_{\max} and U_{\min} , ΔU , is divided into ten parts, and the values so defined are used for contours. The purpose of using ΔU rather than $\Delta V = V_{\max} - V_{\min}$ is to avoid extreme values introduced from high order polynomial.

Each V_{ij} is then replaced by one of following ten integers in order to indicate its relative value: 0(low), 1, 2, 3, 4, 5, 6, 7, 8, 9 (high). Finally a map composed of these integers is printed by the computer. The scale is printed in feet per unit in the (X,Y) coordinates.

STEP 10. Residual plot.

Residuals of all observations are calculated according to equation (11). The difference between the maximum and minimum values is divided into ten parts, each residual is represented by one of the following integers according to its value: 0 (low), 1, 2, 3, 4, 5, 6, 7, 8, 9 (high). These integer numbers are plotted as a residual map according to their relative geographic coordinates.

3.3 Canonical Trend Surface of Hypothetical Data

An illustration of some of the physical meaning of the canonical trend which results from the computations described in the previous sections can be given by making use of hypothetical examples. One hundred and fifty sampling points were spread over a rectangular area, the points being located by a stratified sampling method (Cochran,

1963, p. 87-88). Geographic coordinates of each sampling point within a grid were defined by a pair of random numbers. The sampling points, defined by a set of 150 pairs of random numbers from a table (Krumbein and Graybill, 1965, p. 445-450), were used for all the examples discussed later. Values for the hypothetical variates at the sampling points were computed from various equations incorporating an independent, random, normally distributed, error term with zero mean and any desired variance.

3.31 Experiment 1: The canonical trend surface achieves a parsimonious summarization of a set of trend surface showing distributions of variates from a single population.

(a) Let two hypothetical variates be

$$Z_1 = 0.707 X + 0.707 Y + E_1 \quad (12)$$

$$Z_2 = -0.707 X - 0.707 Y + E_2 \quad (13)$$

where E_1 and E_2 are independent variates with normal distribution $N(0,1)$. The canonical trend was

$$U = -0.652 Z_1 + 0.758 Z_2 \quad (14)$$

$$V = -0.693 X - 0.721 Y \quad (15)$$

The canonical root was 0.9834. Equation (15) is the polynomial of the canonical trend surface. Equation (14) represents the linear combination of variates Z_1 and Z_2 . The absolute values, -0.652 and 0.758 ($-0.652^2 + 0.758^2 = 1.000$), indicate the contributions to the trend. The variates, Z_1 and Z_2 have approximately equal loadings but are in opposite signs, so that these two variates vary approximately in opposite directions. Equations (14) and (15) show that variate

Z_1 increases toward north-east, while variate Z_2 increases toward south-west. This interpretation agrees with the equations (12) and (13).

(b) Let two hypothetical variates be

$$Z_1 = 0.707 X + 0.707 Y + E_1 \quad (16)$$

$$Z_2 = -X + E_2 \quad (17)$$

where E_1 and E_2 were $N(0,1)$. The canonical trend was

$$U = -0.549 Z_1 + 0.836 Z_2 \quad (18)$$

$$V = -0.941 X - 0.339 Y \quad (19)$$

The canonical root was 0.9837. Equation (19) indicates that the canonical trend (NNW-SSE) is intermediate between the trends of the two variates, but slightly closer to Z_2 . Because the variance in X was larger than that in Y (the area studied was rectangular, with the larger dimension in the E-W direction), the variance of Z_2 was larger than that of Z_1 . After standardization, therefore, the contribution of the variance of the variate E_2 to the variance of Z_2 was proportionally less than the contribution of E_1 to Z_1 . Thus the weighting of Z_2 is greater than that of Z_1 in the equation (18).

The second canonical root is equal to 0.8509, its associated canonical variates are

$$U = 0.724 Z_1 + 0.690 Z_2 \quad (20)$$

$$V = -0.163 X + 0.987 Y \quad (21)$$

The combination of equation (20) which is orthogonal to equation (18) and contradictory to that of equations (16), (17) and (18) shows that both the variates Z_1 and Z_2 increase toward northwest. This root should be rejected even though it is statistically

significant at 1 percent level.

3.32 Experiment 2: The coefficients of geological variates are a function of the strength of the trend or the magnitude of the random variance error.

(a) Let two hypothetical variates be

$$Z_1 = 0.707 X + 0.707 Y + E_1 \quad (22)$$

$$Z_2 = -X + E_2 \quad (23)$$

where E_1 was $N(0,1)$, E_2 was $N(0,4)$. The canonical trend was

$$U = 0.996 Z_1 - 0.093 Z_2 \quad (24)$$

$$V = 0.736 X + 0.677 Y \quad (25)$$

The canonical root is 0.9666. It is clear that Z_1 is so strongly weighted that Z_2 has a negligible effect on the trend.

(b) Let two hypothetical variates be

$$Z_1 = 0.707 X + 0.707 Y + E_1 \quad (26)$$

$$Z_2 = E_2 \quad (27)$$

where E_1 was $N(0,4)$, E_2 was $N(0,1)$. The canonical trend was

$$U = 0.986 Z_1 - 0.196 Z_2 \quad (28)$$

$$V = 0.954 X - 0.284 Y - 0.014 X^2 + 0.025 XY + 0.088 Y^2 \quad (29)$$

The canonical root is equal to 0.6967. It is obvious that variate Z_2 should contribute nothing to the trend and in fact contribute very little. Figure 2 shows the variation of Z_1 which approximately agrees with equation (26).

3.33 Experiment 3: If the information (areal variations of the variates) cannot be obtained from the first canonical root alone, the second canonical root will supply part of the remaining information.

Let five constructed variates be

$$Z_1 = X + E_1 \quad (30)$$

$$Z_2 = Y + E_2 \quad (31)$$

$$Z_3 = -X + E_3 \quad (32)$$

$$Z_4 = -Y + E_4 \quad (33)$$

$$Z_5 = E_5 \quad (34)$$

where E_i ($i = 1, 2, 3, 4, 5$) where $N(0,1)$. The first canonical root was 0.9887, its associated canonical variates were

$$U = 0.622 Z_1 - 0.163 Z_2 - 0.763 Z_3 + 0.048 Z_4 + 0.034 Z_5 \quad (35)$$

$$V = 0.974 X - 0.228 Y \quad (36)$$

The second canonical root is 0.9701, its associated canonical variates were

$$U = -0.098 Z_1 + 0.527 Z_2 - 0.272 Z_3 - 0.799 Z_4 - 0.005 Z_5 \quad (37)$$

$$V = 0.101 X + 0.995 Y \quad (38)$$

The first canonical root explains the variations of variates Z_1 and Z_3 whereas the second canonical root explains the variations of variates Z_2 and Z_4 . The variate Z_5 does not contribute to the trend.

3.34 Experiment 4: The univariate case of the canonical trend surface is also a most predictable surface for a single geological variate.

(a) Let the constructed variate be

$$Z_1 = 0.707 X + 0.707 Y + E_1 \quad (39)$$

where E_1 was $N(0,1)$. The canonical root for the linear polynomials is 0.9649. The linear trend was accepted as follows:

$$V = 0.709 X + 0.705 Y \quad (40)$$

(b) Let the constructed variate be

$$Z_1 = 0.13 X - 0.13 Y + 0.60 X^2 - 0.50 XY + 0.60 Y^2 + E_1 \quad (41)$$

where E_1 was $N(0,1)$. The canonical trend surface for variate Z_1 is that

$$V = 0.098 X + 0.092 Y + 0.611 X^2 - 0.513 XY + 0.587 Y^2 \quad (42)$$

The canonical root is equal to 0.9996.

(c) Let the constructed variate be

$$Z_1 = e^{X+Y} + E_1 \quad (43)$$

where E_1 was $N(0,1)$. The canonical trend surface for variate Z_1 is that

$$\begin{aligned} V = & 0.6613 X + 0.6246 Y - 0.1812 X^2 - 0.3195 XY - 0.1754 Y^2 \\ & + 0.0220 X^3 + 0.0521 X^2 Y + 0.0555 XY^2 + 0.0224 Y^3 \\ & - 0.0012 X^4 - 0.0035 X^3 Y - 0.0050 X^2 Y^2 - 0.0042 XY^3 \\ & - 0.0013 Y^4 + 0.0001 X^4 Y + 0.0001 X^3 Y^2 + 0.0002 X^2 Y^3 \\ & + 0.0001 XY^4 \end{aligned} \quad (44)$$

The canonical root is equal to 0.9700. The expression for V was a quintic polynomial. Figure 3 illustrates this use of a canonical trend to approximate an exponential variation.

3.4 Functions of Canonical Trend Surface

Stratigraphers or petrologists usually handle a great number of maps, and like to sort them into groups showing analogous features, and other groups showing quite different features. Canonical trend analysis achieves a parsimonious summarization of a set of maps showing distributions of variates from a single population. It is thus a useful technique for screening maps, at least for exploratory studies.

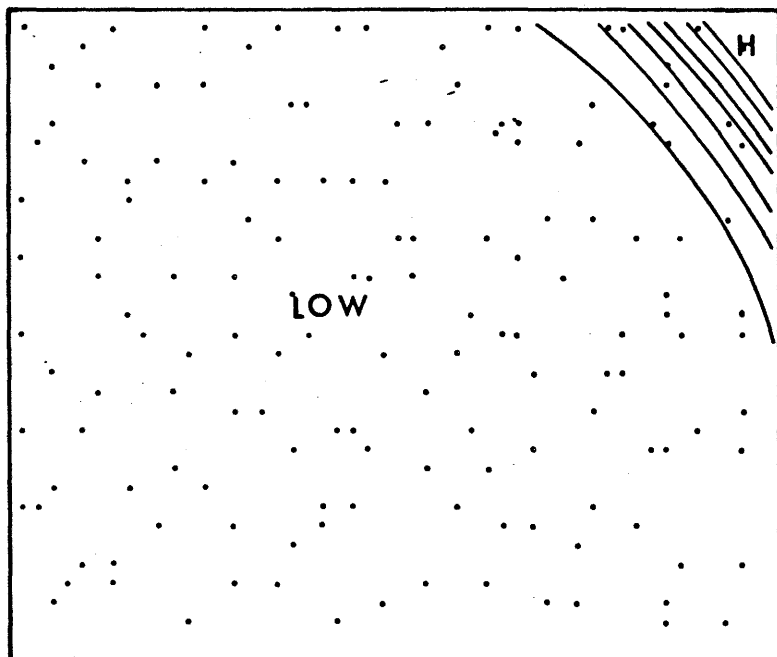


Figure 3 - Canonical trend surface (univariate case) shows an exponential variation.

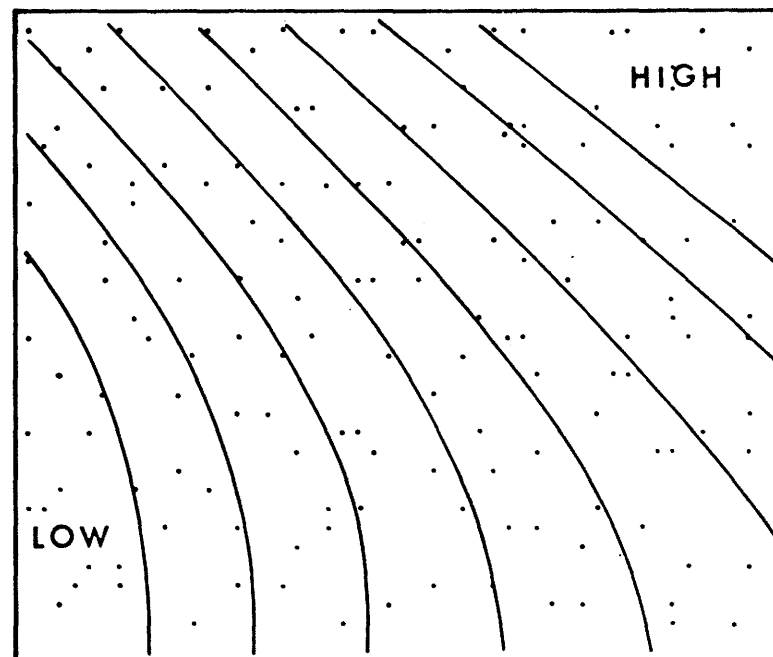


Figure 2 - Canonical trend surface fitted to $Z_1 = 0.707 X + 0.707 Y + E_1$, $Z_2 = E_2$, where E_1 was $N(0,4)$, and E_2 was $N(0,1)$.

The fundamental principle of canonical trend analysis is to maximize the covariance between a set of geological variates and the (X, Y) coordinates. This implies that the canonical trend surface is a most predictable surface to the particular set of variates. It weighs each variate according to its error variance. No other trend surface possesses these characters.

Canonical trend surface will show the nature of variation of any number or any kind of geological variates if they can be amalgamated in a linear function. We can interpret the variations of each variate even though the geological meaning of the particular combination is not yet known. This type of trend surface does not show the absolute values at each contour. On the other hand it does show the general trends of the variates at least for the first approximation. The coefficients of a particular set of variates would not drastically change from low to high order polynomials.

Theoretically, we may map as many variates as we wish at a time. It is generally true that if many variates are analyzed simultaneously, the interpretation of the result will not be as clear as the interpretation for fewer variates. There is no way to decide the appropriate number of variates in advance. Based on the writer's experience of practical problems, a reasonable number is four or fewer but not more than five. It is not necessary that the geological variates be uncorrelated variates. But, if they are highly correlated, the correlation matrix will be nearly singular, and matrix inversion may yield a meaningless solution as a result of the limited number of

digits in the computer. With a sufficiently large computer, this is not always a serious problem. However, we should keep this problem in our mind when we are dealing with highly intercorrelated variates. If a number of variates is available for canonical trend analysis, the type of variates which should be chosen for the analysis is dependent on what geologically meaningful hypothesis we expect to establish. Poor selection of variates would produce geologically ambiguous and trivial result.

Real linear combinations other than the first one cannot be obtained unless the equations (5), (6), (25), and (28) of Chapter 2 are satisfied. This means that variations obtained from canonical correlation analysis exist in nature if they are orthogonal to the previous linear combinations, otherwise, the variations obtained are mere artefacts brought in by mathematical operations. We should realize, however, that the coefficients in the linear function which give maximum correlation may not be suitable as a solution to a practical problem. The linear combination associated with the largest canonical root is most likely to have some physical interpretations. If the subsequent combinations associated with smaller canonical roots give contradictory combinations, these subsequent canonical roots should be rejected without hesitation, even though these canonical roots are statistically significant at a high probability level. On the other hand it can be demonstrated by using geological example that even where the largest canonical root is not significant statistically, the information resulting from the canonical root may be correct. This

illustrates a controversial problem regarding the difference between geological and statistical significance.

Coefficients of canonical variates will yield ambiguous signs in the linear function, if the sample size is insufficient, if the order of polynomial is too high, if the number of variates is too large, and if geological variates and/or (X, Y) coordinates are highly correlated. The investigator should be extremely careful that results have a real geological base. Does this destroy the value of the technique? The answer is not at all. In dealing with multivariate statistical analysis such as multiple regression analysis, principal component analysis, and so on, results which cannot be interpreted may be obtained in attempting to solve practical problems. This is a general difficulty in applying multivariate techniques in geology and results more from the real complexity of nature than from problems caused by the mathematical analysis.

After working a number of practical problems, the writer is convinced that the canonical trend analysis is an adequate technique to evaluate a trend common to a set of variates in preliminary and concept-formation stages of research as part of data-reduction scheme.

3.5 Permian System in Western Kansas and Eastern Colorado

3.51 Geological Setting: The upper part of the Permian rock in western Kansas and eastern Colorado, which is above the Stone Corral dolomite, is composed of Guadalupean series and Upper Leonardian (Nippewalla) series. The rock sequence consists of alternating sandstones and shale with thin dolomites and appreciable amounts of

evaporites. The stratigraphic sequence studied is equivalent to the upper part of intervals B and C-D (Mudge, 1967) in the central Midcontinent region.

A total of 31 wells were selected from this area and analyzed by W. T. Fox (see Krumbein, 1962) by using conventional stratigraphic maps. Integration of stratigraphic maps shows that four components of this System, namely sandstone, shale, carbonate, and evaporites thicken toward a center, i.e. the area shows the characteristics of a sedimentary basin.

The Permian basin was a broad, shallow, fairly stable restricted marine basin. The basin was encircled on the west by the Front Range and Wet Mountains from which feldspathic and quartz grains were mainly derived. On the north and east the Permian basin was probably bounded by low-lying land areas. There was a restricted connection with open sea on the south of the basin.

The petrography of the Permian rocks (Swineford, 1955) and the stratigraphic correlations between western Kansas and southeastern Colorado (Maher, 1953) are listed below:

Western Kansas

Southeastern Colorado

Big Basin (Taloga) formation (O'Connor, 1963)

Red (at the top) and greenish-gray (at the base) montmorillonitic shale, silty shale, siltstones and some very fine-grained sandstone

Day Creek dolomite Upper "crinkly" limestone

Pale-gray to pink fine-grained dolomite bed

(2 to 3 feet)

Whitehorse sandstone

Reddish-brown, fine-grained, well-sorted
sandstone, siltstone, and shale, and
minor quantities of white to buff
sandstone and dolomite

Dog Creek shale Lower "crinkly" limestone

Dark-red silty shale, brownish-red bed

and greenish-gray siltstone, and
very fine-grained sandstone, dolomite,
dolomitic and gypsiferous sandstone
and gypsum

Blaine formation

Anhydrite or anhydritic gypsum,
thin dolomite, and brownish red shale

Flowerpot shale

Reddish-brown gypsiferous shale with
a few thin beds of sandstone and
siltstone

Cedar Hills sandstone

Brownish-red, massive very fine-grained

Lyons sandstone

Fine-grained, quartz,

sandstones and siltstone separated by subangular, littoral
beds of argillaceous siltstone and silty deposit (Thompson, 1949)
shale

Salt Plain formation

Reddish-brown flakey siltstones, thin
sandy siltstones, and very fine-grained
sandstones

Harper formation

Brownish-red sandy siltstone and shale

Stone Corral dolomite

Dense to cellular, and grayish buff in color, containing many vugs
partly or completely filled with coarsely crystalline calcite or
gypsum, oolitic at the base

3.52 Stratigraphic factor map analysis: The 31 wells with six components, namely: total thickness, nonclastic thickness and the thickness of the four components mentioned above, were subjected to Q-mode factor analysis. Three stratigraphic factor maps were constructed (Imbrie, 1963; Krumbein and Imbrie, 1963), one for each of the first three factors, respectively. One reference well which has an identity factor loading in a particular factor is chosen for each map. The geological implications of the Q-mode analysis can be clarified by examination of the actual lithologic association of each reference well. Reference well R_1 of the first factor-vector map (fig. 4A) has maximum evaporite thickness and the minimum carbonate thickness of any well. Reference well R_2 of the second

factor-vector map (fig. 4B) has minimum thickness, shale content, and evaporite content of any well. The reference well of the third factor-vector map has minimum sandstone and carbonate content, and contains considerable shale, and some evaporite, Imbrie (1963) and Krumbein and Graybill (1965, p. 402-406) suggested that the first factor map represents an evaporite basin with rapid subsidence and clay influx, the second factor map indicates the distribution of shelf sedimentation, and the third factor map represents the distribution of an off-shore sedimentation of fine detritals adjacent to the evaporite basin.

The fundamental approach of the stratigraphic factor map analysis is to visualize the geologic implications of the distribution of a hypothetical variate through interpretations of the lithologic associations for a particular map.

3.53 Canonical trend surface analysis: The hypothesis we expect to establish is that the area shows the characteristics of a sedimentary basin. A sedimentary basin, in general, should show that some of sediments thicken toward the center, and the relative amount of shale and evaporite should be predominate over that of sand and carbonate, respectively, in the central part of the basin.

A sample of size 30 (excluding one well which did not contain any evaporite) was subjected to canonical trend analysis on two sets of different geological variates. The variates of the first set are thickness of sand, shale, carbonates, and evaporites, while the variates of the second set are total thickness, sand-shale ratio, and carbonate-evaporite ratio. The order of polynomial fitted to these two

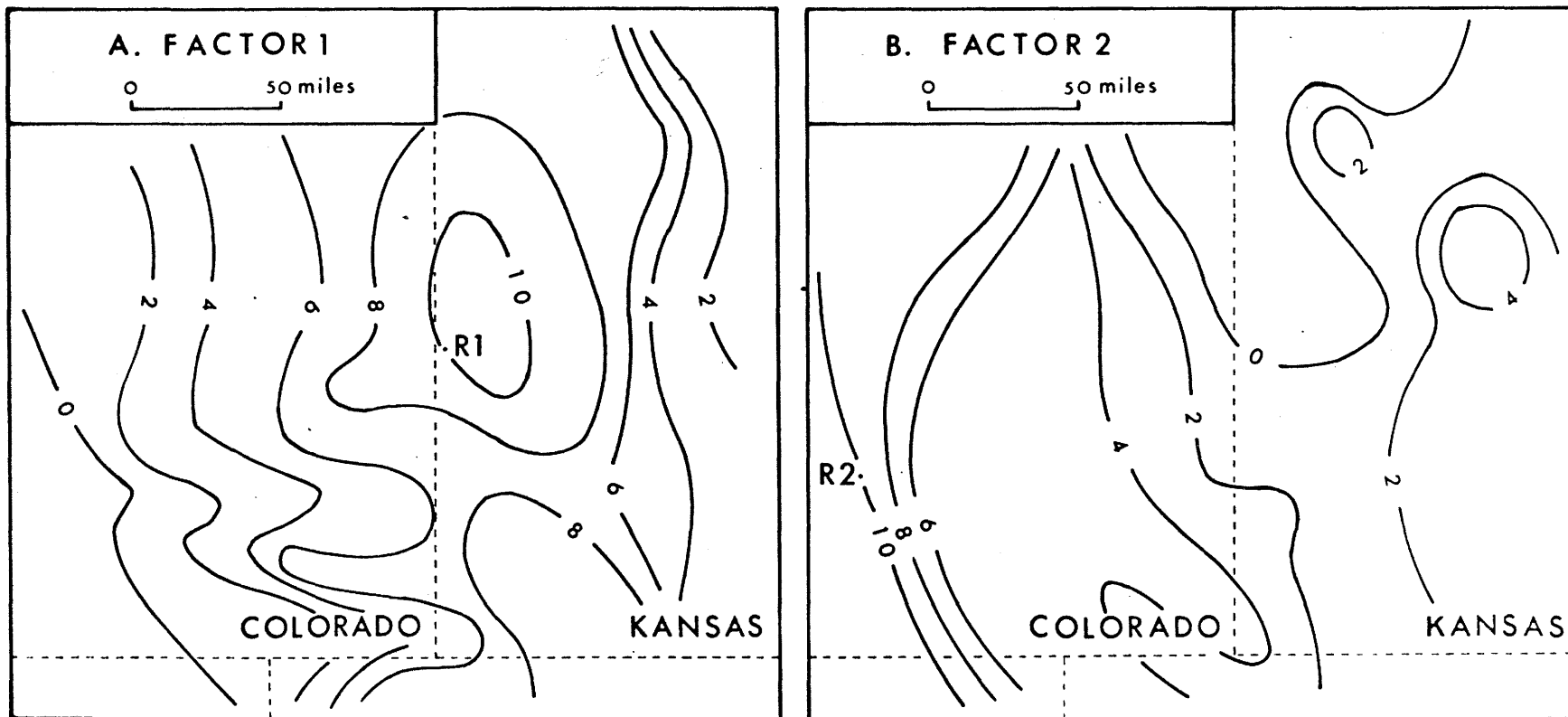


Figure 4 - Permian system of Western Kansas and Eastern Colorado stratigraphic factor maps.

A. Factor 1; B. Factor 2 (Data after Imbrie, 1963; redrawn by the present writer).

sets of variates cannot be greater than five. The raw data of this problem were published by Krumbein (1962, table VI).

Principal components were obtained from a covariance matrix of the thickness of sand, shale, carbonate, and evaporites. The first principal component contributes 73.8 percent of total variance, the second one contributes 20.6 percent, and the last one contributes 0.6 percent. The rotated, normalized eigenvectors are listed in table 1.

TABLE 1
ROTATED, NORMALIZED EIGENVECTORS OF SEDIMENT THICKNESS OF PERMIAN SYSTEM

Principal component Variate			
	1	2	3
Sand thickness	0.03	0.99	0.12
Carbonate thickness	0.06	-0.12	-0.99
Shale thickness	-0.99	-0.10	-0.10
Evaporite thickness	-0.97	0.06	0.23

The first principal component indicates a shale-evaporite association which is equivalent to the first factor obtained by Imbrie (1963) and also by Krumbein and Imbrie (1963). The second principal component explains the variance contributed by the thickness of sand, whereas the last principal component explains the variance contributed by the thickness of carbonates. These three principal components are mathematically orthogonal, but transitions between them are possible

from the geological point of view.

The result from the principal component analysis indicates that the carbonates behave quite differently from the evaporites, and the sand behaves in an opposite way from the shale and evaporites. Thus, the clastic ratio $(= (\text{sand} + \text{shale})/(\text{carbonates} + \text{evaporites}))$ should not be included in the analysis, because its geological meaning is ambiguous.

The first set of four variates, namely thickness of sand, shale, carbonates and evaporites were treated through canonical trend surface analysis. After three iterative processes, the highest canonical root that could be reached was 0.9464 which indicates that a cubic polynomial is the most predictable response surface (table 2).

TABLE 2
RECORD OF THE SUCCESSIVE EVALUATION OF THE ORDER OF POLYNOMIAL
FOR FOUR VARIATES

Canonical root	Order of polynomial
0.7804	Linear
0.8810	Quadratic
0.9464	Cubic

The canonical variate for the geological variates was

$$U = 0.516 (\text{sand}) + 0.408 (\text{shale}) - 0.104 (\text{carbonate}) + 0.746 (\text{evaporite}) \quad (45)$$

The canonical variate for the polynomial was

$$V = -0.407 X - 0.914 Y + 0.002 X^2 - 0.001 XY - 0.001 Y^2 \quad (46)$$

The canonical trend, equation (45), mainly indicates the variations in thickness of evaporite, sand and shale.

The second canonical root of the third order canonical trend surface is equal to 0.8676. The associated canonical variates are as follows:

$$U = 0.757 (\text{sand}) - 0.111 (\text{shale}) + 0.617 (\text{carbonate}) - 0.185 (\text{evaporite}) \quad (47)$$

and

$$V = 0.183 X - 0.983 Y - 0.001 X^2 + 0.001 XY - 0.002 Y^2 \quad (48)$$

The canonical trend, figure 5A, and equation (48), mainly summarizes the variations in the thickness of sand and carbonates. This figure shows that the sand and carbonate have a maximum thickness located around the southeast of Colorado.

A quadratic polynomial was fitted to these four variates. The canonical variates were

$$U = 0.575 (\text{sand}) + 0.425 (\text{shale}) + 0.154 (\text{carbonate}) + 0.682 (\text{evaporite}) \quad (49)$$

$$V = 0.357 X - 0.934 Y - 0.001 X^2 - 0.001 Y^2 \quad (50)$$

Judging from figure 5B and equation (50), it is concluded that sand, shale, carbonate, and evaporites thicken toward the southwest corner of Kansas. It is interesting to compare this canonical trend surface with net thickness maps of the same data. The general trend of the net thickness maps of sand (fig. 6A), shale (fig. 6B) carbonate (fig. 6C), and evaporite (fig. 6D) display patterns analogous to that

shown by the canonical trend surface. Figure 5A displays a trend pattern very similar to the thickness of the carbonate and sand, whereas Figure 5B displays a trend pattern almost identical with the thickness of the evaporites and shale. Figure 5B fails to show the maximum thickness of the sand in the northwest part of Colorado, but it indicates that the sand thickens toward the southwest of Kansas and thins toward the northeast of Kansas and the southwest of Colorado, respectively. The area having maximum thickness of carbonate and the thinning directions of the carbonates were also shown by Figure 5B.

Smaller canonical roots show different linear combinations of the four variates. These linear combinations are contradictory to the first two. It is generally true that the first or first two canonical roots have a higher probability of approximating real trends, while the interpretation of the smaller canonical roots may be obscured by local variation ("noise"). Thus, the other combinations are not discussed.

The second set of the three variates, namely, total thickness, sand/shale ratio, and carbonate/evaporite ratio were analyzed in the same way. The record of evaluation of the trend surface is listed in table 3.

TABLE 3

RECORD OF THE SUCCESSIVE EVALUATION OF THE ORDER OF POLYNOMIAL FOR
THREE VARIATES

Canonical root	Order of Polynomial
0.7829	Linear
0.9189	Quadratic
0.9612	Cubic

The maximum canonical root obtained is equal to 0.9612
indicating that a cubic polynomial is an adequate fitting surface.

The canonical variate for the geological variate is that

$$U = 0.803 (\text{total thickness}) - 0.254 (\text{sand/shale}) - 0.539 \\ (\text{carbonate/evaporite}) \quad (51)$$

The canonical variate for the polynomial is that

$$V = - 0.335 X + 0.843 Y + 0.233 X^2 - 0.088 XY + 0.335 Y^2 \\ - 0.026 X^3 + 0.018 X^2Y + 0.009 XY^2 + 0.041 Y^3 \quad (52)$$

The negative sign for the ratio indicates that the high values
in the map should be the low value for the ratios. Figure 7A shows
that the total thickness increases toward the most westsouthern part
of Kansas, whereas the sand/shale ratio and carbonate/evaporite ratio
decrease toward the same area. The isopach map and the maps of two
ratios (Krumbein, 1962) were also plotted for comparison (Figs. 7B,
7C, and 7D). It is also worthwhile to note how the canonical trend
fits closely to the total thickness and slightly to the carbonate/
evaporite and sand/shale ratios.

The smaller canonical roots also yield contradictory combinations to the first one. They are discarded here.

The question is whether is it possible to detect a gross trend representing variations of all or most of variates. The canonical trends (Figs. 5 and 7A) answer this question at least to the first approximation and also establish the hypothesis we expected before carrying out the analysis.

Judging from the nature of change of all the variates and their similarity in gross trend pattern to the two canonical trend surfaces, it was concluded that there was shelf sedimentation adjacent to an evaporite basin in southwest Kansas during Upper Permian time.

The essential approach to the interpretation of canonical trend surface is to visualize the geological implications of the variations of all or most of the variates being displayed realistically in a single map. The merit of a canonical trend surface is to evaluate a common trend to a set of variates and to condense a set of maps showing areal distributions of geological variates from a single population. Its maximum correlation property yields a high fidelity between the population response surface and the sample response surface.

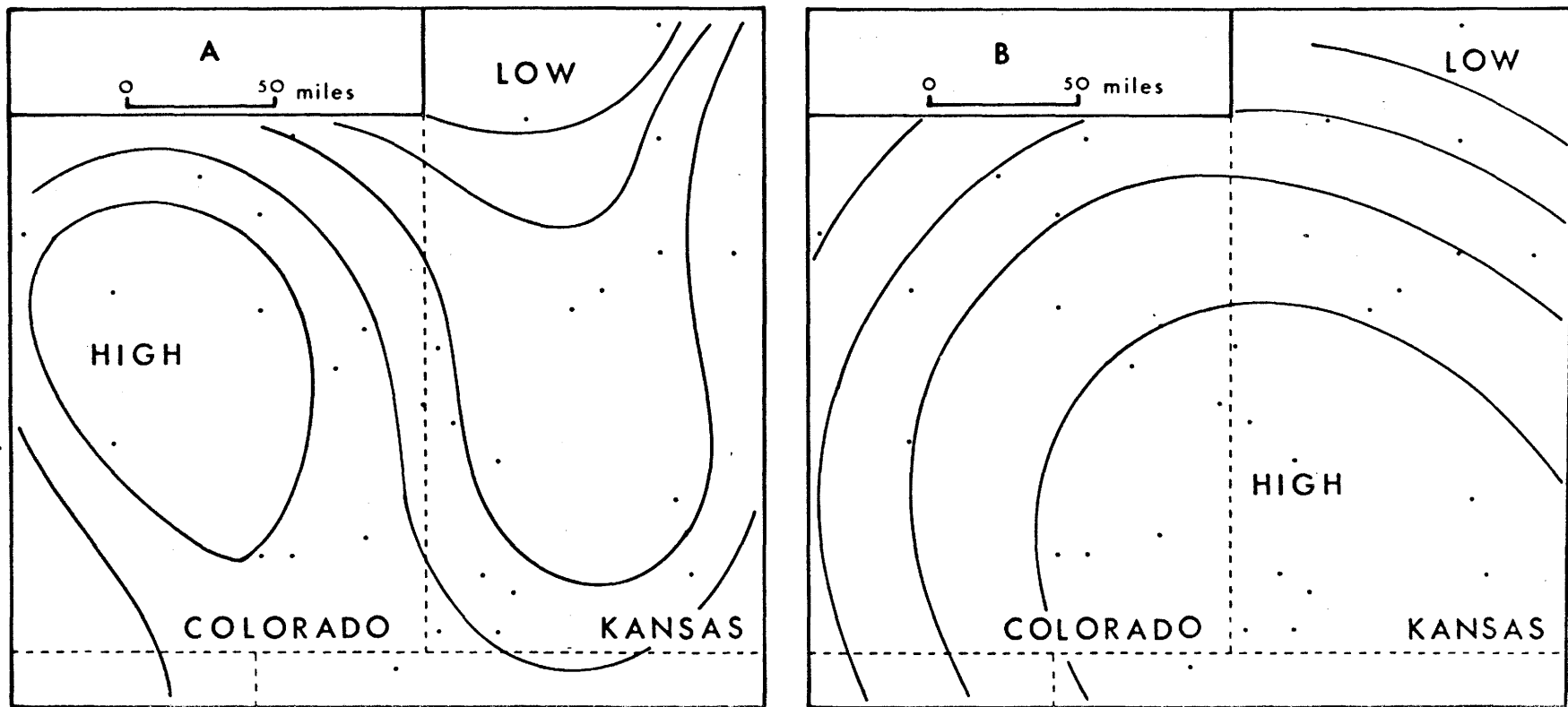


Figure 5 - Permian system of Western Kansas and Eastern Colorado. A. Third order canonical trend associated with second canonical root showing variations in thickness of the sand and the carbonate. B. Second order canonical trend associated with first canonical root showing variations in thickness of the evaporates, the sand, and the shale.

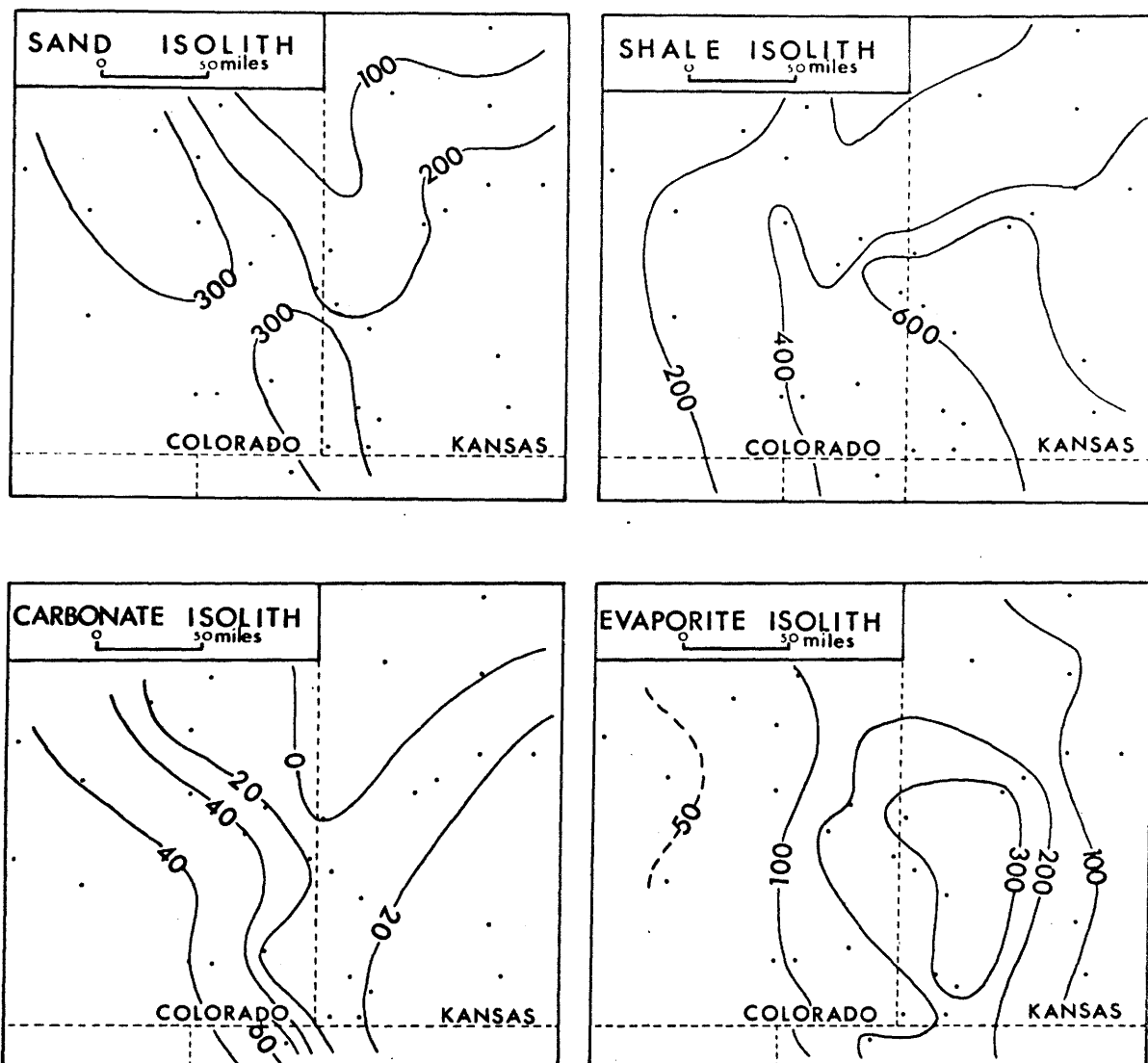


Figure 6 - Permian system of Western Kansas and Eastern Colorado.

Net thickness maps of the sand, the shale, the carbonate, and the evaporite. (Data after Krumbein, 1962; redrawn by the present writer).

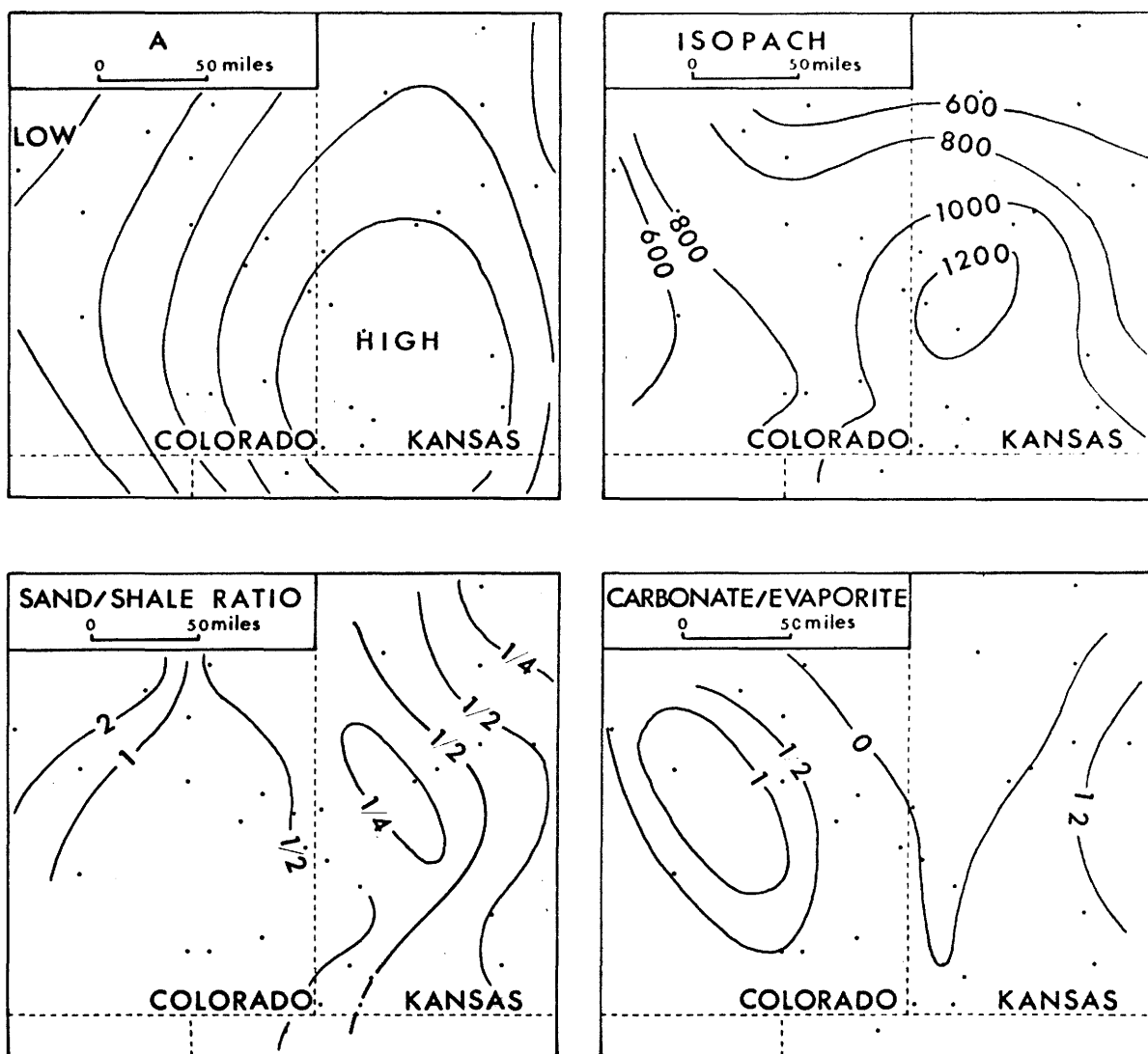


Figure 7 - Permian system of Western Kansas and Eastern Colorado.

A. Canonical trend surface for the total thickness, the sand/shale ratio, and the carbonate/evaporite ratio. Isopach map (right upper), sand/shale ratio map (left lower), and carbonate/evaporite ratio map (right lower). (Data after Krumbein, 1962; redrawn by the present writer).

CHAPTER 4

OTHER APPLICATIONS IN GEOLOGY

4.1 Relationships Between Two Sets of Variates

4.11 General Statement: Canonical correlation may be considered as a multivariate case of simple correlation. Suppose we have a set of p_1 predictor variates and a set of p_2 criterion variates for the same set of samples, then canonical correlation determines that linear function of the predictor variates and that of the criterion variates which gives the maximum correlation between these two sets.

Both canonical analysis and principal factor solution in factor analysis deal with linear functions. The most important difference between canonical and factor analysis is that in canonical analysis the criterion is external while in factor analysis the criterion is derived from the internal evidence of the variates. Thus, canonical analysis is called external factor analysis whereas conventional factor analysis is called internal factor analysis (Bartlett, 1948; Burt, 1948).

Canonical analysis or external factor analysis is a technique to reveal the maximum possible covariance between two sets of variates. We are not interested in the relations of the variates within either set itself, but in the relationship between the two sets. Internal factor analysis seeks underlying influence factors which explain maximum variance in any particular direction, but canonical analysis does not seek factors. Factor analysis expresses each variate in

terms of factors which are treated as independent variates, while canonical analysis expresses each canonical variate in terms of the geological variates which are treated as independent variates. Due to the difference between the two linear regression models similarity between the factor loadings and the coefficients of canonical variates is not to be expected. Comparisons of canonical correlation and factor analysis were made by Burt (1948) and Das (1965) by using psychological examples.

When we have two sets of variates relating to one set of samples, various statistical techniques may be adopted according to the nature of the problem discussed. We may, for example factorize the two sets together, or we may factorize each set independently, and try to correlate factors; or we may perform a canonical correlation analysis between the two sets of variate in order to yield a maximum correlation. The largest canonical root has the best chance of demonstrating the existence of a relation between the two sets if one actually exists, but each canonical root will help to elucidate the nature of a particular relation, which cannot be explained fully by any other canonical root.

In geology, a typical experimental design may be made to observe several variates on a set of rock specimens. For example, in a recent sediment study one set of variates may be the abundances of various kinds of animals or plants, while the other set of variates may be the physico-chemical parameters of the depositional environments, and the size parameters of the sediments which contain the animals.

The nature of the relations between the two sets of variates may be explained by some of the canonical roots between the two sets. There is a number of analogous problems such as correlations between modal and chemical analysis, between major and trace elements, between trace elements and clay minerals, between textural parameters (size, roundness, sphericity and so on) and composition parameters (mineral composition). These problems may be untangled by using Hotelling's idea to a certain extent. The following numerical example is given to demonstrate the usefulness of this mathematical model.

4.12 Recent Sediment from Buttonwood Sound: Recent sediment samples were collected from 19 stations located in Buttonwood Sound, Florida Bay by Lynts (1966). Three samples of size 19 were taken on August 14th, 17th, and 20th, 1962 respectively. A sample of size 16 was taken on February 9th, 1963. For each location, the following environmental parameters of the sediment-water interface were measured: depth of water, temperature, salinity, pH, and Eh. The sand, silt, and clay fractions were analyzed from each sample. The dominant element of microflora in the Sound was grasses, in particular, turtle grass. The distribution of turtle grass was mapped by using three grades, namely: very dense, dense-moderate, and patchy. The abundance of the turtle grass was coded by the present writer as follows: very dense--1; dense-moderate--0; patchy-- -1.

The problem here is to determine to what extent the environmental factors, namely: depth, temperature, salinity, pH and abundance of turtle grass are related to the size parameters of sediments. The

four samples were considered as a single sample of size 73, because the turtle grass map is not specified to any one of the four sampling dates.

Looking at the simple correlation coefficients of all possible pairs of the seven variates (table 4), one may conclude that the turtle grass is positively related to silt and also negatively related to sand. The simple correlation coefficients between the sediments and the depth, temperature, salinity, and pH are very small.

TABLE 4

*SIMPLE CORRELATION COEFFICIENTS OF SEDIMENTS FROM BUTTONWOOD SOUND

Turtle grass	1.000						
Depth	-0.061	1.000					
Temperature	0.091	-0.059	1.000				
Salinity	-0.085	0.086	-0.976	1.000			
pH	-0.194	0.242	0.122	-0.070	1.000		
Sand	-0.665	0.073	-0.192	0.194	0.173	1.000	
Silt	0.458	0.084	0.015	0.001	-0.139	-0.602	

A canonical analysis was performed on this set of data. The first canonical root is equal to 0.683 which is significant at the 1 percent level. The second canonical correlation (0.242) is not significant at the 5 percent level. This would suggest that in studying the relation between these two sets we can confine our attention to the first pair of canonical variates which is listed in

table 5.

It is clear that the distribution of sand is correlated closely with the abundance of turtle grass. Sands are more abundant in regions of less grass in Buttonwood Sound. Lynts (1966) obtained the same conclusion by comparison of turtle grass map with sediment-size distribution map. In addition to confirming this conclusion, canonical correlation analysis also indicates that there is no relation between the abundance of sand and measured physical parameters other than turtle grass.

TABLE 5
CANONICAL VARIATES FOR SEDIMENTS FROM BUTTONWOOD SOUND

Environmental factor		Sediment	
Turtle grass	-0.98	Sand	1.00
Depth	-0.00	Silt	-0.09
Temperature	-0.07		
Salinity	0.13		
pH	0.10		

The variate, sand, was discarded, then canonical correlation analysis was performed on the same data. The canonical root is 0.483, its associated canonical variate for the environmental factors is as follows:

$$V = 0.70 (\text{turtle grass}) + 0.19 (\text{depth}) + 0.47 (\text{temperature}) + 0.49 (\text{salinity}) - 0.15 (\text{pH}) \quad (1)$$

Equation (1) suggests that the silt is also moderately correlated with turtle grass and weakly to other parameters. The canonical root is significant at 1 percent level, but it gives a result contradictory to the principal component analysis which will be discussed later.

It is concluded from the canonical correlation analysis that

- (1) The abundance of sand is a better predictor of turtle grass than that of silt; but if the variate, sand is not used then the variate silt would give a less satisfactory result.
- (2) If the variate, clay, is inserted, the result may be a completely different linear function of environmental factors to predict clay or silt or sand or some combination of them.

Principal component analysis was also performed on the correlation matrix of the same set of data. The rotated and normalized eigenvectors of the last eigenvalue are practically zero. The eigenvectors of the sixth eigenvalues indicates the same relationship on sand, silt and turtle grass as that of the first eigenvalues. The analysis was repeated again and only the first five eigenvectors were orthogonally rotated. The result of the second run is listed in table 6.

The coefficients of the first component are highly loaded with turtle grass and sand, with coefficients of opposite sign. This shows that the first canonical variate is very similar to the first component obtained by an ordinary principal component analysis. This component is interpreted as a factor of sediment stabilizer, as Ginsburg and Lowenstam (1958) pointed out that a dense growth of grass may produce a layer of semi-motionless water over the bottom which traps fine

sediments. Actually these two variates are mutually dependent. The factor pattern (table 6) fails to show the relationship of equation (1). A more detailed discussion of the result of this factor analysis is given by Lee (1967).

TABLE 6
ROTATED AND NORMALIZED FIVE PRINCIPAL COMPONENTS FOR
SEDIMENT FROM BUTTWOOD SOUND

Component Variate	1	2	3	4	5
Turtle grass	0.98	0.03	-0.01	0.09	-0.16
Sand	-0.80	-0.18	0.08	-0.08	0.57
Temperature	0.06	1.00	-0.02	-0.07	-0.02
Salinity	-0.06	-1.00	0.04	0.02	0.00
Depth	-0.04	-0.05	0.99	-0.12	-0.04
Silt	0.30	-0.02	0.07	0.06	-0.95
pH	-0.11	0.07	0.13	-0.98	0.06
Eigenvalue	2.362	1.902	1.208	0.705	0.518
Cumulative Percentage	33.74	60.91	78.17	88.24	95.64

Canonical correlation gave the result that the only significant correlation between the sediment and environment variates was between sand and turtle grass. Inspection of components 2-5 confirms this

conclusion, since the only relationships revealed are those between environmental variates (components 2-4) or between sediment variates (component 5).

In this particular case, the loadings on first component (table 6) and the coefficients of the canonical variate (table 5) are very similar in magnitude and in sign. In this example, the principal factor solution in factor analysis has revealed more extensive information, and is likely to be a powerful tool for displaying the internal relations of a set of measures. On the other hand, when the purpose is to investigate the relations between two sets of variates, or for prediction in many dependent variates problems, the technique of canonical correlation is more appropriate.

4.2 Matching Two Factor Patterns

4.21 General Statement: Suppose a separate factor analysis has been conducted for a rock body at different geographic localities or different stratigraphic sections. It may be that the factor patterns obtained at the various localities appear quite different but that canonical transformations exist such that at least some of the factors in one set may be found to correspond closely to factors in the other set. In this case, the number of samples, N , is equal to number of variates, p_1 the number of factors for one set, and p_2 the number of factors for another set. Thus, the sum of p_1 and p_2 , in general, is greater than the value of N . In this application, the technique is used to determine to what extent the two factor patterns represent the same set of underlying factors. The technique has been generalized

for m sets of variates or m factor patterns by Horst (1961, 1961a) to find m canonical transformations which will yield new patterns having maximum similarity.

4.22 Factor patterns of sediment from Buttonwood Sound: Two samples of size 19 and 16 were taken from Buttonwood Sound, Floriday Bay (Lynts, 1966) in 14 August, 1962 and 9 February, 1963, respectively. Factor analysis was performed separately on these two sets of data. The factor patterns are listed in table 7.

The problem is to determine to what extent the two factor patterns are similar. Canonical analysis was carried out on the two patterns. In this case, $p_1 = 7$, $p_2 = 7$, $N = 8$ (number of variates). The result is that seven canonical roots are unity. All of them are significant at the 1 percent level. Their associated canonical variates are listed in table 8.

Judging from table 8, it is concluded as follows:

- (1) Factor F9-1 is negatively related to factor F14-1, because they have an opposite sign in their original factor patterns.
- (2) F9-2 is negatively related to F14-3 and partly to F14-1.
- (3) F9-3 is related to F14-2.
- (4) F9-4 is negatively related to F14-6.
- (5) F9-5 is negatively related to F14-5.
- (6) F9-6 is related to F14-4. The analysis shows that these two factor patterns represent the same set of underlying factors, although these two sets of samples were collected at different times.

TABLE 7
FACTOR PATTERNS FOR SEDIMENTS FROM BUTTONWOOD SOUND

Factor pattern for samples collected in 14 August, 1962

Factor Variate	F14-1	F14-2	F14-3	F14-4	F14-5	F14-6	F14-7
Sand	0.95	-0.07	0.19	0.11	-0.12	-0.01	-0.18
Turtle grass	-0.54	-0.14	-0.16	-0.22	0.25	0.00	0.74
Silt	-0.90	-0.02	0.39	-0.08	0.05	-0.09	0.11
Depth	-0.05	-0.23	0.18	-0.15	-0.09	-0.94	-0.00
pH	0.02	-0.95	0.14	-0.09	0.09	-0.22	0.07
Clay	0.05	0.14	-0.96	-0.03	0.09	0.17	0.10
Temperature	-0.14	-0.08	-0.02	-0.97	-0.01	-0.14	0.11
Salinity	0.13	-0.08	0.08	-0.01	-0.97	-0.08	-0.12
Eigenvalue	2.733	1.905	1.154	0.998	0.539	0.449	0.221
Cumulative percentage	34.16	57.98	72.41	84.89	91.63	97.24	100.00

Factor pattern for samples collected in 9 February, 1963

Factor Variate	F9-1	F9-2	F9-3	F9-4	F9-5	F9-6	F9-7
Silt	0.95	0.00	0.00	0.04	-0.17	-0.22	0.10
Sand	-0.75	-0.60	-0.10	0.08	0.19	0.07	-0.15
Clay	0.08	0.95	0.16	-0.17	-0.10	-0.14	0.11
Turtle grass	0.42	0.45	0.16	0.00	-0.35	-0.25	0.64

Table 7

Factor pattern for samples collected in 9 February, 1963 - (continued)

Factor Variate	F9-1	F9-2	F9-3	F9-4	F9-5	F9-6	F9-7
Depth	0.01	-0.15	-0.09	0.97	0.11	-0.10	-0.00
Salinity	-0.23	-0.14	-0.06	0.15	0.91	0.23	-0.13
pH	-0.03	-0.16	-0.97	0.09	0.06	0.12	-0.06
Temperature	0.24	-0.11	0.15	0.13	-0.24	-0.91	-0.10
Eigenvalue	3.545	1.631	0.996	0.825	0.556	0.276	0.170
Cumulative percentage	44.32	64.71	77.16	87.47	94.42	97.87	100.00

TABLE 8

CANONICAL VARIATES FOR TWO FACTOR PATTERNS

Factor Variate	1	2	3	4	5	6	7
F9-1	0.00	0.00	1.00	0.00	0.00	0.00	0.00
F9-2	0.00	0.00	0.00	0.00	0.00	0.00	1.00
F9-3	0.00	0.00	0.00	0.00	1.00	0.00	0.00
F9-4	0.00	0.00	0.00	1.00	0.00	0.00	0.00
F9-5	0.00	0.00	0.00	0.00	0.00	1.00	0.00
F9-6	1.00	0.00	0.00	0.00	0.00	0.00	0.00
F9-7	0.00	1.00	0.00	0.00	0.00	0.00	0.00

cont'd

Table 8

Canonical variates for two factor patterns - (continued)

Variate Factor	1	2	3	4	5	6	7
F14-1	-0.01	0.02	-0.95	0.13	0.09	0.09	-0.43
F14-2	-0.29	0.21	0.03	0.16	0.93	-0.12	0.01
F14-3	-0.32	0.01	0.02	0.04	0.04	-0.07	-0.26
F14-4	0.84	-0.01	-0.04	0.01	-0.16	0.21	0.16
F14-5	-0.33	0.09	0.20	-0.05	0.13	-0.95	-0.00
F14-6	-0.02	-0.04	0.09	-0.97	-0.13	-0.05	-0.03
F14-7	-0.02	0.97	-0.20	0.10	0.26	-0.04	0.09

4.3 Q-Technique Canonical Correlation

4.31 General statement: Q-technique canonical correlation is used for dealing with sample space. Suppose we have two sets of samples collected from different groups (such as localities, or different formations), the problem is to determine to what extent these two sets of samples are similar, based on observable variates. In this case, p_1 is the number of samples in one set, p_2 number of samples in another one, and N is the number of variates. The canonical root provides an index of the degree of similarity between the two groups of samples.

It must be noted, however, that canonical root indicates the maximum degree of similarity between the groups, and not the average degree of similarity. If there happen to be two closely similar individuals, one in each group, the canonical root will be very high even though the

majority of the samples in the two groups differ considerably. There is no statistical criterion for testing the coefficients of the canonical variates. Thus, the coefficients must be carefully examined. If the coefficients are "approximately" equal, this implies that the degree of similarity (the canonical root) is more or less reliable. On the other hand, if the coefficients are large in one or two samples, this indicates that these two groups may or may not be generally similar, even though the canonical root is high. On the other hand, if the canonical root is low for two groups, this strongly suggests that these two groups are quite different.

The Q-technique canonical root presents the same problem as in Q-technique factor analysis. The different variates may be measured in quite different units, and correlation of a pair of samples over N values of noncomparable units does not, in general, make sense. This difficulty cannot be overcome by standardization. Kendall (1966) suggested removing this difficulty by working with rank correlation, for example, presence of a variate is coded by 1, absence is coded by -1; or "rare", "common", and "abundant" are coded by using -1, 0, and 1, respectively.

4.32 Ordovician and Silurian Sediments: Two specimens of shale from each of four formations exposed in the Niagara escarpment, Ontario were analyzed for both major elements and trace elements (Cr, Mn, Ga, V, Ni, Ti, Zr, Sr, and Ba expressed in terms of ppm) (Candy, 1963). The formations were the Silurian Cabot Head (gray and green shale), Grimsby (red shale), and Thorold (gray and green shale), and the Ordovician Queenston (red

and green shale). The Q-technique canonical correlation matrix based on the trace elements is listed in table 9.

TABLE 9
CANONICAL CORRELATION MATRIX FOR THE FOUR FORMATIONS FROM ONTARIO

Thorold	1.000		
Grimsby	0.954	1.000	
Cabot Head	0.996	0.977	1.000
Queenston	0.940	0.886*	0.829*

NOTE: * not significant at 1 percent level.

Table 9 suggested that there is a clear geochemical distinction between the Ordovician Queenston shale and Silurian shales. This canonical correlation matrix also indicates that the Silurian shales are closely similar. Unfortunately, the coefficients of the canonical variates for the Grimsby formation are not approximately equal for all samples, the same is with the Thorold formation. Thus, the latter conclusion is still doubtful. The canonical variates of the Thorold and Grimsby formations, for example, are

$$U = 0.98 (\text{Grimsby sample 1}) - 0.18 (\text{Grimsby sample 2}) \quad (2)$$

$$V = -0.15 (\text{Thorold sample 1}) + 0.99 (\text{Thorold sample 2}) \quad (3)$$

The simple correlation coefficients between samples of the Grimsby and of the Thorold formation are positive. Thus, the opposite sign in the linear functions U and V would imply that these shale samples cannot be simply amalgamated in such a linear form.

The advantage of this technique is that it does not require a large sample size for each group. There is an elegant way to approach the same problem, and this will be discussed in the following section.

4.4 Discriminatory Analysis by Canonical Variates

4.41 General statement: Testing similarity of k groups can be carried out through canonical correlation analysis (Bartlett, 1938a, 1947, 1965). Suppose a discriminant function is defined as

$$Y = A' Z \quad (4)$$

where A is $p \times 1$ matrix to be determined, Z is $p \times 1$ variate matrix, then the analysis of variance of Y is

Between groups	$A' B A$
Within groups	$A' W A$
<hr/>	
Total	$A' T A$

where B , is the matrix of sums of squares and products 'between' the k groups, and w for the corresponding pooled 'within' matrix, and T for the corresponding 'total' matrix.

In order to make the variable Y as effective as possible in discriminating between these two groups, the ratio

$$R^2 = \frac{A' B A}{A' T A} \quad (5)$$

should be as large as possible. The solution to this problem may be obtained through canonical correlation analysis, that by solving the set of linear equations as

$$(B - R^2 T)A = 0 \quad (6)$$

where R^2 's are latent roots of the equation

$$|B - R^2 T| = 0 \quad (7)$$

Seal (1964, p. 127-144) used the following matrix equation instead of (7),

$$|B - R^2 (B + W)| = 0 \quad (8)$$

$$\text{or} \quad |B - \lambda^2 W| = 0 \quad (9)$$

where $\lambda^2 = R^2 / (1 - R^2)$.

The λ^2 's are the eigenvalues of the nonsymmetrical matrix $W^{-1}B$, and their normalized eigenvectors will be the columns of matrix A .

This means that the first transformed axis passes through the direction of the greatest variability 'between' the k means; then the second axis, at right angles to the first; passes in the direction of the next greatest variability and so on. This is essentially the case of several populations in discriminant analysis (Wilks, 1963, p. 576-581).

Equation (4) may be generalized for more than two groups, say k groups. In this case, Y is $m \times 1$ matrix, A is $m \times p$ matrix, Z is $p \times 1$ matrix, where m is the number of non-zero eigenvalues of equation (9), $m = p$, if $p \leq k - 1$, $m = k - 1$, if $p > k - 1$. Thus, if $k = 2$, equation (4) is known as the discriminant function. A new observation vector Z may be inserted into this function, and the unknowns will be allocated to one or other of the two groups depending on whether the value of Y is positive or negative, respectively. The set of linear equation (4) may be used for classification. After substituting each observation Z into the equation (4), we will have m

coordinates on the m -dimensional canonical axes for each observation. Finally, the canonical coordinates will display the m -dimensional relationship of all observations. If we emphasize the differences between the means of the k groups, the vector Z is replaced by a mean vector \bar{Z} . The canonical coordinates of the mean of each group will be obtained by placing the mean values of each variate in \bar{Z} .

Bartlett's test (1947) of significance of canonical roots is that

$$\chi^2 = \{(N - 1) - 0.5(p + k)\} \ln \Lambda \quad (10)$$

where $\Lambda = \prod_{j=i+1}^m (1 + \lambda_j^2)$, p = number of variates, k = number of groups, is approximately chi-square variate with $(p - i)(k - i - 1)$ degrees of freedom when $\lambda_{i+1} = \lambda_{i+2} = \dots = \lambda_m = 0$.

The assumptions are that (1) the variance-covariance matrix of p -variate groups are supposed equal (the classification of two multi-variate distributions with different covariance matrices was discussed by Anderson and Bahadur (1962)), (2) the unknown samples which are to be discriminated belong to one, and only one, of the groups and that the probabilities that the unknown sample belongs to any group are equal.

This technique has the great advantage that the p variates may be different types of measurements if all the variates are standardized. On the other hand, the disadvantage is that if p is large, and $k = 2$, the comparison between the two groups is contained in a straight line, because we only have one canonical root.

Suppose we have measured p variates on N_j samples of a given rock type (or fossil types) ($j = 1, 2, \dots, k$); $\sum N_j = N$, where these

variates may be measured in any one of nominal, ordinal, interval, or ratio scales, then canonical correlation analysis is the most reasonable procedure to discriminate between these k rock types (or fossil types) and to discover their underlying affinities. Canonical correlation analysis may be replaced by Q-technique principal component analysis if the following conditions are satisfied (Seal, 1964, p. 171-172): (1) the measured variates are uncorrelated, this would be done through a principal component analysis on each rock type, and (2) the corresponding diagonal elements of covariance matrices of "between rock type" and "within rock type" are equal. These are severe restrictions. The proper uses of factor analysis, canonical analysis, and principal component analysis have been discussed thoroughly in the recent literatures (Bartlett, 1948, Eysenck, 1952; Lindley, 1962, 1964; Rao, 1964; Cattell, 1965, 1965a; Gower, 1966, 1967). Seal (1964, p. 171) summarized the different purposes of these techniques as follows:

"(1) Principal component analysis is intended to achieve a parsimonious summarization of a random sample from a single universe of multivariate Normal measurements;

(2) Canonical analysis is a procedure of discriminating as clearly as possible between two or more multivariate Normal universes with the same variance-covariance matrix; and

(3) Factor analysis is an attempt to elicit the underlying Normal multivariate structure of a universe that can be sampled with respect to many correlated variates."

4.42 Ordovician limestones from Colborne quarry: The upper part of the Middle Ordovician Cobourg limestone was sampled at Colborne Quarry, Ontario (Lee, 1965). Three samples of size 78, 22, and 29 were collected from three stratigraphic sections, A, B, and C, respectively. The percentages of limestone variates, namely; matrix, sparry calcite, brachiopod shells, echinoderm fragments, trilobite fragments, bryozoa fragments, and ostracod shells were estimated from peel prints. Two peel prints were counted in order to make replicates for each sample.

One problem is whether or not these three stratigraphic sections are different, based on the seven limestone variates. The raw data were analyzed through canonical correlation analysis. The largest two canonical roots are equal to 0.426 and 0.320, respectively, which are significant at 5 percent level; whereas the last three roots are not significant at the same level. According to the canonical coordinates, sections A and C are slightly different from section B when based on the seven limestone variates (table 10). The individuals of the three sections are displayed on canonical coordinates (Fig. 8).

Another problem is whether or not these three stratigraphic sections are different, based only on brachiopod shells. The canonical coordinates indicate that sections A and C are very similar, while section B is different from the other two sections even though the canonical root (0.140) is not significant at the 5 percent level. The last conclusion is confirmed by the result obtained by using one-way analysis of variance and t-test (Lee, 1965; Lee and Winder, 1967).

TABLE 10
CANONICAL COORDINATES OF LIMESTONES FROM COLBORNE QUARRY, ONTARIO

Section no.	Based on 7 variates		Based on brachiopod shells
	Canonical axis 1 2		Canonical axis
Section A 1	39.238	7.516	8.590
Section A 2	38.488	7.396	7.397
Section B 1	38.400	7.786	10.136
Section B 2	37.975	7.565	10.364
Section C 1	38.406	7.531	8.414
Section C 2	38.212	7.450	8.310
Canonical root	0.426	0.320	0.140

Buzas (1966) applied the same principle to test the differences among the groups of Elphidium from Long Island Sound based on morphological characters.

The Nigerian Upper Cretaceous and Lower Tertiary ostracoda were analyzed through canonical correlation analysis by Reyment (1966).

This technique elegantly quantifies means of k groups, or all individuals, on the canonical coordinates according to their underlying affinities, whereas the multivariate analysis of variance and generalized T^2 -statistic show only the significance levels for difference between the groups.

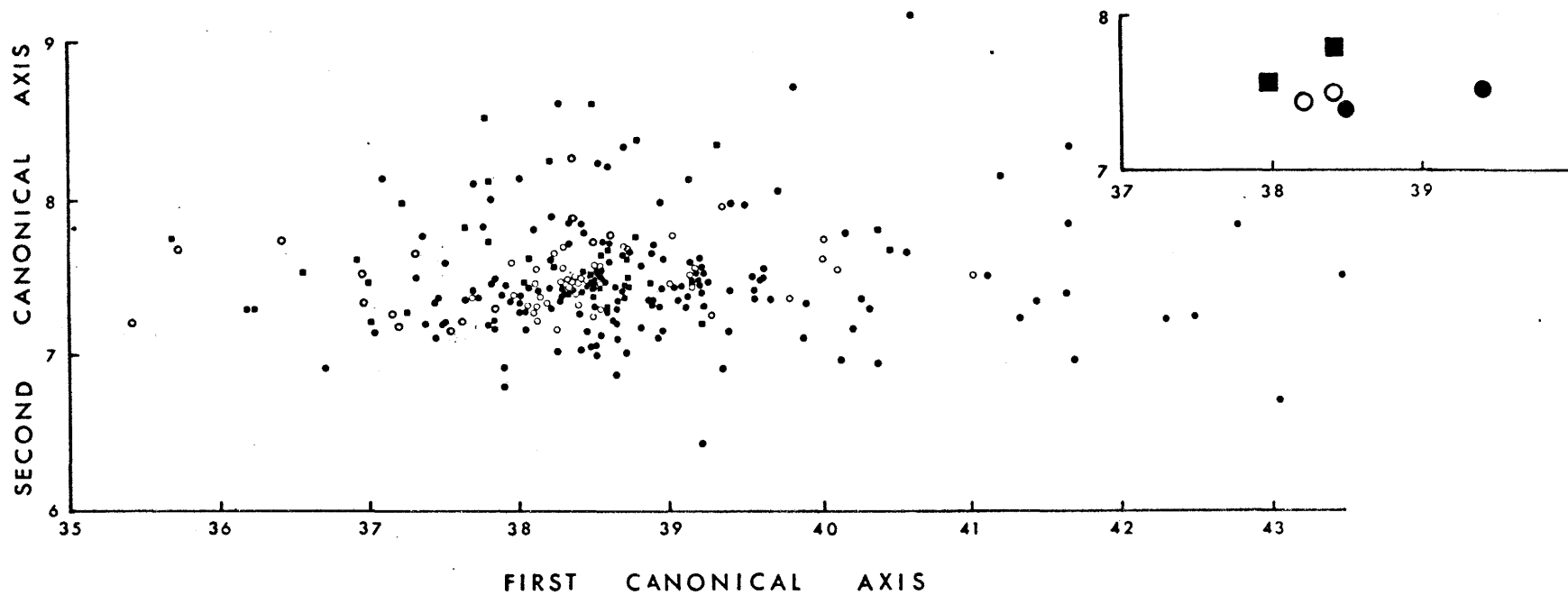


Figure 8 - Ordovician limestone specimens from Colborne Quarry (dot: section A; square: section B; and circle: section C) are shown on a canonical coordinates based on first two canonical variates. Mean values of all sections are shown in the upper right corner.

CHAPTER 5

CONCLUSIONS

The merit of canonical trend surface analysis is that it is a parsimonious summarization of a set of geographic variations of geological variates. The canonical trend surface has the elegant property of maximum correlation between geological variates and geographic coordinates. The succinct trend obtained does not show the absolute value of each variate, but it does indicate the nature of the variation for each variate. The linear function associated with the largest canonical root has highest probability of showing the true variations. The linear functions associated with smaller canonical roots explain other changes which cannot be summarized into the first linear function.

Examination of the canonical trend for a set of variates frequently leads to an understanding of the intrinsic trend, common to several variates, that would be difficult or impossible to obtain by other means.

The relation between two sets of variates may be untangled to a certain extent through canonical correlation analysis. The technique should have applications in prediction and exploratory studies.

The technique for matching two factor patterns is very interesting in some cases. It can help the investigator to visualize the variation of factors from place to place. Similar results may be obtained by examination of the factor loadings on particular variates, but canonical correlation provides more precise information and more rapidly.

The Q-technique canonical correlation analysis is applicable to samples of small size. In the case of samples of large size, the method -- where individuals are displayed on canonical coordinates through canonical correlation analysis -- has a strong mathematical rationale.

The use of canonical analysis as a procedure of discriminating between two populations is restricted by the assumption that the populations have the same covariance matrix. It has the advantage that variates may be different types of measurements because changing the scale of any one of them does not affect the results. Furthermore, experience from solving practical problems suggests that the number of dimensions required for a comparison of groups will generally be less than the number required by principal component analysis (Seal, 1964, p.123).

Principal factor analysis explains the maximum variance of a single sample whereas canonical factor analysis (Rao, 1955) yields factors which have maximum correlations between the observable variates. This result is the same as Lawley's (1940; Lawley and Maxwell, 1963) maximum-likelihood solution. Its application in geology has not yet been studied.

The use of canonical correlation analysis in multiple regression analysis is quite interesting. In this case, the number of variates in one set is equal to one (the dependent variable). The canonical root is identical with the multiple correlation coefficient, but the weights for each variate in the linear function may be different in these two approaches. To date, there has been no discussions on this topic from the geological point of view.

Another possible use of the canonical correlation analysis in geology is to differentiate sedimentary environments when based on grain size distribution. One set of variates would consist of the sedimentary environments such as beach, dune, lagoon, river and so on, the second set of variates would contain only one variate, the unknown individual to be determined. The grain size parameters would be treated as observations. The canonical correlation would give a maximum correlation between the unknown and the known environments. The largest coefficient attached to the known environment would indicate the nature of the unknown. This technique would be able to identify an individual which came from mixed environments, for instance dune sand and beach sand. The coefficients would be approximately equal on these two variates or environments.

The concept of two-dimensional canonical trend can easily be generalized into three-dimensional canonical trend, but it will need a large computer storage to handle the computation process.

Canonical and factor analysis are appropriate tools for geological data, but there are some disadvantages. Unfortunately, the existing tests of statistical significance do not in some cases help to make a reasonable decision. At the present time, the proper use of these methods is to generate hypotheses. Once a hypothesis has been created, the question should be whether or not it is interesting to conduct experiments to prove or disprove the hypothesis. One should not lean too much on the statistical tests of the number of factors or canonical roots, since these tests were derived from

specified assumptions which may not be reasonable in all cases. On the other hand, if there exists a hypothesis, obtained from experiments or experience, the use of this technique is to see whether or not the hypothesis can be confirmed by statistical analysis.

Canonical trend surface analysis represents a step toward the theoretical development of "geometrics". Its applications in geology open a promising field. More extensive studies on this field are desirable.

REFERENCES

- ANDERSON, T. W., 1958, "An Introduction to Multivariate Statistical Analysis", John Wiley & Sons, Inc., New York, 374 p.
- ANDERSON, T. W., 1963, "Asymptotic Theory for Principal Component Analysis", Ann. Math. Stat., V. 34, p. 122-148.
- ANDERSON, T. W. and BAHADUR, R. R., 1962, "Classification into Two Multivariate Normal Distributions with Different Covariance Matrix", Ann. Math. Stat., V. 33, p. 420-431.
- BARTLETT, M. S., 1938, "Further Aspects of the Theory of Multiple Regression", Proc. Cambridge Philosophy Soc., V. 34, p. 33-40.
- BARTLETT, M. S., 1941, "The Statistical Significance of Canonical Correlation", Biometrika, V. 32, p. 29-37.
- BARTLETT, M. S., 1947, "The General Canonical Correlation Distribution", Ann. Math. Stat., V. 18, p. 1-17.
- BARTLETT, M. S., 1947a, "Multivariate Analysis", Jour. Roy. Stat. Soc., Suppl. IX, p. 176-197.
- BARTLETT, M. S., 1947b, "The Use of Transformation", Biometrics, V. 3, p. 39-52.
- BARTLETT, M. S., 1948, "Internal and External Factor Analysis", British Jour. Psychology (Stat. Section), V. 1, p. 73-81.

- BARTLETT, M. S., 1965, "Multivariate Analysis" in Theoretical and Mathematical Biology, Edited by Waterman, T. H. and Morowitz, H. J., Blaisdell Publishing Company, London, 426 p.
- BEECH, H. R. and MAXWELL, A. E., 1958, "Differentiation of Clinical Groups Using Canonical Variates", Jour. Consulting Psychology, V. 22, p. 113-121.
- BURT, C., 1948, "Factor Analysis and Canonical Correlations", British Jour. Psychology (Stat. Section), V. 1, Pt. II, p. 95-106.
- BUZAS, MARTIN A., 1966, "The Discrimination of Morphological Groups of Elphidium (Foraminifer) in Long Island Sound through Canonical Analysis and Invariant Characters", Jour. Paleontology, V. 40, p. 585-594.
- CANDY, G. J., 1963, "The Geochemistry of some Ordovician and Silurian Shales from Southwestern Ontario", M. Sc. Thesis, Dept. Geology, McMaster University, unpublished, 175 p.
- CATTELL, R. B., 1965, "Factor Analysis: An Introduction to Essentials I: The Purpose and Underlying Models", Biometrics, V. 21, p. 190-215.
- CATTELL, R. B., 1965a, "Factor Analysis: An Introduction to Essentials II: The Role of Factor Analysis in Research", Biometrics, V. 21, p. 405-435.
- COCHRAN, W. G., 1963, "Sampling Techniques", 2nd Edition, John Wiley & Sons, Inc., New York, 413 p.
- CONSTANTINE, A. G. and JAMES, A. T., 1958, "On the General Canonical Correlation Distribution", Ann. Math. Stat., V.29, p. 1146-1166.

- COOLEY, W. W. and LOHNES, P. R., 1962, "Multivariate Procedures for the Behavioural Science", John Wiley & Sons, New York, 211 p.
- DAS, R. S., 1965, "An Application of Factor and Canonical Analysis to Multivariate Data", British Jour. Math. Stat. Psychology, V. 18, p. 57-67.
- DUNTEMAN, G. H., 1967, "A Canonical Correlation Analysis of the Strong Vocational Interest Blank and the Minnesota Multiphasic Personality Inventory for a Female College Population", Educational Psychol. Measurement, V. 27, p. 631-642.
- EBERLEIN, P. J., 1962, "A Jacobi-like Method for the Automatic Computation of Eigenvalues and Eigenvectors of an Arbitrary Matrix", Jour. Soc. Indust. App. Math., V. 10, p. 74-88.
- EYSENCK, H. J., 1952, "Uses and Abuses of Factor Analysis", App. Stat., V. 1, p. 45-49.
- FORGOTSON, J. M., Jr., 1960, "Review and Classification of Quantitative Mapping Techniques", Bull. Am. Assoc. Petroleum Geologists, V. 44, p. 83-100.
- GINSBURG, R. N. and LOWENSTAM, H. A., 1958, "The Influence of Marine Bottom Communities on the Depositional Environment of Sediments", Jour. Geology, V. 66, p. 310-318.
- GOWER, J. C., 1966, "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis", Biometrika, V. 53, p. 325-338.

- GOWER, J. C., 1967, "Multivariate Analysis and Multidimensional Geometry",
The Statistician, V. 17, p. 13-28.
- HILL, W. J. and HUNTER, W. G., 1966, "A Review of Response Surface
Methodology, A Literature Survey", Technometrics, V. 8, p.
571-590.
- HORST, P., 1961, "Generalized Canonical Correlation and Their Application
to Experimental Data", Jour. Clinical Psychology, Monograph
Supplement No. 14, p. 331-347.
- HORST, P., 1961a, "Relations Among m Sets of Measures", Psychometrika,
V. 26, p. 129-149.
- HOTELLING, H., 1936, "Relations Between Two Sets of Variates",
Biometrika, V. 28, p. 321-377.
- HOTELLING, H., 1957, "The Relations of the Newer Multivariate Statistical
Methods to Factor Analysis", British Jour. Stat. Psychology,
V. X, Pt. II, p. 69-79.
- HSU, P. L., 1941, "Canonical Reduction of the General Regression
Problem", Ann. Eugen. London, V. 11, p. 42-46.
- HSU, P. L., 1941a, "On the Limiting Distribution of the Canonical
Correlations", Biometrika, V. 32, p. 38-45.
- IMBRIE, J., 1963, "Factor and Vector Analysis Programs for Analyzing
Geological Data", Tech. Rept. No. 6, ONR Task no. 389-135,
Contract nonr 1228(26), Office of Naval Research, Geography
Branch, Northwestern Univ. Illinois, 83 p.
- KENDALL, M. G. 1961, "A Course in Multivariate Analysis", Griffin's Stat.
Mono. Cour., London, 185 p.

- KENDALL, M. G., 1966, "Discrimination and Classification", Proc. International Symp. held in Dayton, Ohio, June 14-19, 1965, Edited by Krishnaiah, Academic Press, New York, 592 p.
- KOONS, P. B., Jr., 1962, "Canonical Analysis" in "Computer Applications in the Behavioural Sciences", Edited by Borko, H., Englewood Cliffs, N. J., Prentice-Hall, New York, 633 p.
- KRUMBEIN, W. C., 1955, "Composite End Members in Facies Mapping", Jour. Sedimentary Petrology, V. 25, p. 115-122.
- KRUMBEIN, W. C., 1962, "Open and Closed Number Systems in Stratigraphic Mapping", Am. Assoc. Petroleum Geologists Bull., V. 46, p. 2229-2245.
- KRUMBEIN, W. C. and IMBRIE, J., 1963, "Stratigraphic Factor Maps", Bull. Am. Assoc. Petroleum Geologists, V. 47, p. 698-701.
- KRUMBEIN, W. C. and GRAYBILL, F. A., 1965, "An Introduction to Statistical Models in Geology", McGraw-Hill Book Company, 475 p.
- KSHIRSAGAR, A. M., 1962, "A Note on Direction and Collinearity Factors in Canonical Analysis", Biometrika, V. 49, p. 255-259.
- LAWLEY, D. N., 1940, "The Estimation of Factor Loadings by the Method of Maximum Likelihood", Proc. Roy. Soc. Edinburg, Series A, V. 60, p. 64-82.
- LAWLEY, D. N., 1959, "Tests of Significance in Canonical Analysis", Biometrika, V. 46, p. 59-66.

- LAWLEY, D. N. and MAXWELL, A. E., 1963, "Factor Analysis as a Statistical Method", London, Butterworths, 117 p.
- LEE, P. J., 1965, "Sedimentology of the Middle Ordovician Cobourg Limestone at Colborne, Ontario, Canada", M. Sc. Thesis, Univ. Western Ontario, unpublished, 70 p.
- LEE, P. J., 1967, "FORTRAN IV Program for Principal Component Analysis", Tech. Mem. 67-5, Dept. Geology, McMaster Univ., 24 p.
- LEE, P. J. and MIDDLETON, G. V., 1967, "Application of Canonical Correlation to Trend Analysis", Colloquium on Trend Analysis, Edited by Merriam, D. F. and Cocke, N. C., Computer Contribution 12, State Geological Survey, Kansas, p. 19-21.
- LEE, P. J. and WINDER, C. G., 1967, "Fabric of a Middle Ordovician Limestone at Colborne, Ontario", Can. Jour. Earth Sci., V. 4, p. 529-540.
- LINDLEY, D. V., 1962, "Factor Analysis: A Symposium", The Statistician, V. 12, no. 3.
- LINDLEY, D. V., 1964, "Factor Analysis: A Summary of Discussion", The Statistician, V. 14, p. 47-61.
- LYNTS, G. W., 1966, "Relationship of Sediment-size Distribution to Ecologic Factors in Buttonwood Sound, Florida Bay", Jour. Sedimentary Petrology, V. 36, p. 66-74
- MAHER, J. C., 1953, "Permian and Pennsylvanian Rocks of Southern Colorado", Am. Assoc. Petroleum Geologists, Bull., V. 37, p. 913-939.

- MAXWELL, A. E., 1961, "Canonical Variate Analysis when the Variables are Dichotomous", *Educational Psychol. Measurement*, V. 21, p. 259-271.
- MEREDITH, W., 1964, "Canonical Correlations with Fallible Data", *Psychometrika*, V. 29, p. 55-65.
- MIDDLETON, G. V. and LEE, P. J., 1967, "Applications of Canonical Correlation to Sedimentology", 7th Inter. Sedimentological Congress, England.
- MORRISON, D. F., 1967, "Multivariate Statistical Methods", McGraw-Hill Book Company, New York, 388 p.
- MUDGE, M. R., 1967, "Paleotectonic Investigations of the Permian System in the United States", Chapter F, Central Midcontinent Region, Geological Survey Professional Paper, 515-F, p. 93-123.
- O'CONNOR, H. G., 1963, "Changes in Kansas Stratigraphic Nomenclature", *Am. Assoc. Petroleum Geologists, Bull.*, V. 47, p. 1873-1877.
- PEARCE, S. C. and HOLLAND, D. A., 1960, "Some Applications of Multivariate Methods in Botany", *App. Stat.*, V. 9, p. 1-7
- PELTO, C. R., 1954, "Mapping of Multicomponent Systems", *Jour. Geology*, V. 62, p. 501-511.
- RAO, C. R., 1955, "Estimation and Tests of Significance in Factor Analysis", *Psychometrika*, V. 20, p. 93-111.
- RAO, R., 1964, "The Use and Interpretation of Principal Component Analysis in Applied Research", *Sankhyā, Series A*, V. 26, p. 329-358.

- REYMENT, R. A., 1966, "Studies on Nigerian Upper Cretaceous and Lower Tertiary Ostracoda: Part 3: Stratigraphic, Paleoecological and Biometrical Conclusions, Stockholm Contributions in Geology, Acta Universitatis Stockholmiensis, V. XIV, 151 p.
- ROY, S. N., 1957, "Some Aspects of Multivariate Analysis", John Wiley & Sons Inc., New York, 214 p.
- SEAL, H. L., 1964, "Multivariate Statistical Analysis for Biologists", Methuen and Co. Ltd., New York, 207 p.
- SWINEFORD, A., 1955, "Petrography of Upper Permian Rocks in South-central Kansas", State Geological Survey of Kansas, Bull. 111, 179 p.
- THOMSON, G., 1947, "The Maximum Correlation of Two Weighted Batteries", British Jour. Psychology (Stat. Sec.), Pt. I, V. 2, p. 27-34.
- THOMPSON, W. O., 1949, "Lyons Sandstone of Colorado Front Range", Am. Assoc. Petroleum Geologists, Bull., V. 33, p. 52-72.
- TINTNER, G., 1946, "Some Applications of Multivariate Analysis to Economic Data", Jour. Am. Stat. Assoc., V. 41, p. 472-500.
- WAUGH, F. V., 1942, "Regressions Between Sets of Variables", Econometrika, V. 10, p. 290-310.
- WILKS, S. S., 1963, "Mathematical Statistics", John Wiley & Sons, Inc., New York, 644 p.

APPENDIX I

FORTTRAN IV PROGRAM FOR TWO-DIMENSIONAL CANONICAL TREND

6.1 Instructions for Using the Program

6.11 Input to Program

I Instruction card

- (1) First 3 columns are reserved for punching the number of samples. The maximum sample size is 300.
- (2) Columns 4 and 5 are for punching the number of variates to be analyzed. The maximum number of variates is 15.
- (3) Columns 6 and 7 are for punching the number of variates to be read by the computer, it may or may not be the same as stated in (2) of I.
- (4) Columns 8 and 9 are for punching the number of description cards (see IV).
- (5) Column 10
 - Punch 1 Arc Sine transformation on the geological variate
 - Punch 2 Log transformation (base 10) on the geological variate
 - Punch 3 No transformation on the geological variate
- (6) Column 11
 - Punch the order of the polynomial. The maximum order is 6.
 - If the number 6 is punched on the column 11, the computer will stop computing the higher order polynomial (≤ 6) if the difference of two successive canonical correlations is less than 0.05.

(7) Columns 12-19

In general we use an arbitrary coordinate system for sample location. In order to have an actual scale for the map printed by the printer, the distance in feet which is equal to 1 unit in the arbitrary coordinate system is punched on columns 12-19. The input format is F8.0.

II Title Card

The project title and investigator's name are punched on one card (no more than 80 characters).

III Map Title card

The title for the trend surface is punched on one card.

IV Description card(s)

A brief description of the project is allowed. The maximum number of cards is 50, whereas the minimum number is 1.

V Variate name card(s)

Twelve characters are used for the names of each variate. They should be punched in the same order as in the data matrix. Six names are punched on one card, (i.e. eight columns are left at the end of each card).

VI Data Format card

This format is used for the input data cards. No more than 80 characters are allowed.

VII Data cards

The X-, Y-coordinates must be in the first two places and

are followed by the geological variate data.

The origin of the coordinate may be placed anywhere, but the convention for the reference coordinate is the same as the Cartesian coordinate system. The right hand side of the X- axis is positive whereas the left hand side is negative. The upper part of the Y-axis is positive, whereas the lower part is negative. The maximum length of the X-axis should be greater than or equal to that of the Y-axis.

EACH SAMPLE HAS ITS OWN SEPARATE DATA CARD(S).

TWO OR MORE SAMPLES SHARING ONE DATA CARD IS NOT ALLOWED.

6.12 Output from Program

- (1) Project title and investigator's name.
- (2) Description of the project.
- (3) Mean, and standard deviation of the geological variates.
- (4) Covariance matrix for the geological variates.
- (5) Record of successive evaluation of the degree of the trend surface.
- (6) Each canonical correlation has one set of the following outputs.
 - (i) The trend surface equation
 - (ii) The observed, calculated, and residual values with X-, Y-coordinates.
 - (iii) The contoured trend surface map with title and legend.
 - (iv) The residual map.

6.2 Listing of the Program

THIS IS CANONICAL TREND SURFACE PROGRAM JULY 1966 P J LEE

```

DOUBLE PRECISION Q(903),R11(27,27),R22(15,15),R12(27,15),
1PRO(27,15),PRO1(15,15),SUMSQ(42)
REAL IPOLY
DIMENSION E(27),D(99),SUM(42), DOT(10), N1(108),VEC(15,15),
1PRO2(15,15),P(27),POLY(50,50),B(300),C(300),FMT(16),DM(300,17),
2TITLE(16),TITLE2(16),VARIAT(30),SOURCE(800),IPOLY(50,50),
3RESIDU(300),WORK(54),TERM(27),A(9,9),MAPING(300,2),SIG(300)
EQUIVALENCE (Q,POLY,IPOLY),(SUM,R22),(SUMSQ,PRO1,P),
1(TITLE,TITLE2),(R12,A)
DATA TERM/6H X      ,6H Y      ,6H X2    ,6H XY    ,6H Y2    ,6H X3    ,6H
1 X2Y    ,6H XY2    ,6H Y3    ,6H X4    ,6H X3Y    ,6H X2Y2 ,6H XY3    ,6H Y4
2      ,6H X5    ,6H X4Y    ,6H X3Y2 ,6H X2Y3 ,6H XY4    ,6H Y5    ,6H X6
3,6H X5Y    ,6H X4Y2 ,6H X3Y3 ,6H X2Y4 ,6H XY5    ,6H Y6    /
DATA SAMPLE,BLANK/2H *,2H /
DATA DOT/2H 0,2H 1,2H 2,2H 3,2H 4,2H 5,2H 6,2H 7,2H 8,2H 9/
READ(5,1) N,M,MV,NN,I1,NORDER,FEET
1 FORMAT(I3,3I2,2I1,F8.0)
S=N
READ(5,2) TITLE
2 FORMAT(16A5)
WRITE(6,3) TITLE
3 FORMAT(1H1,20X,16A5///)
READ(5,2) TITLE2
NU=NN*16
READ(5,4) (SOURCE(I),I=1,NU)
4 FORMAT(16A5)
WRITE(6,5) (SOURCE(I),I=1,NU)
5 FORMAT(1H ,25X,16A5)
MU=M*2
READ(5,6) (VARIAT(I),I=1,MU)
6 FORMAT(12A6)
READ(5,7) FMT
7 FORMAT(16A5)
XMAX=-100000.0
XMIN=100000.0
YMAX=-100000.0
YMIN=100000.0
M90=M+27
IM90=(27+M+1)*(27+M)/2
DO 20 I=1,M90
SUM(I)=0.0
20 SUMSQ(I)=0.0
DO 21 I=1,IM90,1
21 Q(I)=0.0
DO 131 ISAMPL=1,N

```

```

      READ(5,FMT) X,Y,(D(J),J=1,MV)
      GOTO (9,10,11), I1
19  WRITE(6,12)
12  FORMAT(1H0,30H WRONG CODE FOR TRANSFORMATION)
      GOTO 100
   9  DO 13 I=1,M
13  D(I)=ARSIN(D(I))
      GOTO 11
  10  DO 15 I=1,M
15  D(I)=ALOG10(D(I))
  11  DO 16 J=1,M
16  DM(ISAMPL,J)=D(J)
      DM(ISAMPL,M+1)=X
      DM(ISAMPL,M+2)=Y
      XMAX=AMAX1(X,XMAX)
      YMAX=AMAX1(Y,YMAX)
      XMIN=AMIN1(X,XMIN)
      YMIN=AMIN1(Y,YMIN)
      DO 130 I=1,M
      SUM(I)=SUM(I)+D(I)
130 SUMSQ(I)=SUMSQ(I)+D(I)*D(I)
131 CONTINUE
      DO 126 I=1,M
      SUMSQ(I)=SQRT((S*SUMSQ(I)-(SUM(I))**2)/(S*(S-1.0)))
126 SUM(I)=SUM(I)/S
      WRITE(6,96)
   96 FORMAT(1H1,30X,5H MEAN,5X,19H STANDARD DEVIATION/)
      DO 93 I=1,M
      IV=2*I-1
      IV1=IV+1
   93 WRITE(6,95) ((VARIAT(J),J=IV,IV1),SUM(I),SUMSQ(I))
   95 FORMAT(1H0,10X,2A6,5X,F10.3,7X,F12.3)
      WRITE(6,106)
106 FORMAT(///)
      DO 127 J=1,N
      DO 127 I=1,M
127 DM(J,I)=(DM(J,I)-SUM(I))/SUMSQ(I)
      DO 129 I=1,M
      SUM(I)=0.0
129 SUMSQ(I)=0.0
      DO 8 IS=1,N
      DO 128 J=1,M
128 D(J)=DM(IS,J)
      X=DM(IS,M+1)
      Y=DM(IS,M+2)
      L=M
      A(1,1)=1.0
      DO 17 I=1,NORDER
      DO 18 J=1,I
      A(J,I+1)=A(J,I)*X

```

```

      L=L+1
18  D(L)=A(J,I+1)
      A(I+1,I+1)=A(I,I)*Y
      L=L+1
17  D(L)=A(I+1,I+1)
      DO 22 I=1,M90
22  SUM(I)=SUM(I)+D(I)
      DO 8 I=1,M90
      DO 8 J=I,M90
      L=(J-1)*J/2+I
      Q(L)=Q(L)+D(J)*D(I)
      8 CONTINUE
      DO 25 I=1,M90
      DO 25 J=I,M90
      K=(J-1)*J/2+I
25  Q(K)=(S*Q(K)-SUM(I)*SUM(J))/(S*(S-1.0))
      J2=1
      J3=J2+1
      WRITE(6,102) N
102 FORMAT(1H0,10X,19H COVARIANCE MATRIX,10X,21H THE SAMPLE SIZE IS
      1,I3//)
      DO 103 J=1,M
      K=0
      DO 104 I=1,J
      KK=J*(J-1)/2+I
      K=K+1
104 E(K)=Q(KK)
      WRITE(6,105) ((VARIAT(J1),J1=J2,J3),(E(K1),K1=1,K))
105 FORMAT(1H0,5X,2A6,15F7.3)
      J2=J2+2
      J3=J3+2
103 CONTINUE
      ROOT=0.000
      ITERM=0
      WRITE(6,113)
113 FORMAT(1H1,10X,53H RECORD OF SUCCESSIVE EVALUATION TREND SURFACE D
      1EGREE///)
125 DO 27 NPOWER=1,NORDER
      IPOWER=NPOWER
      ITERM=IPOWER+ITERM+1
      IF(M.GT.ITERM) GOTO 37
      NL=ITERM
      NR=M
      DO 32 I=1,NR
      DO 32 J=1,NL
      K=(J+M-1)*(J+M)/2+I
32  R12(J,I)=Q(K)
33  DO 35 I=1,NR
      DO 35 J=I,NR
      L=J*(J-1)/2+I

```

```

      R22(I,J)=Q(L)
35  R22(J,I)=R22(I,J)
      CALL DMINVS(R22,15,NR,1.E-15,IERR,N1,WORK)
      IF(IERR.EQ.0) GOTO 29
39  WRITE(6,34)
34  FORMAT(1H0,45H GEOLOGICAL VARIATE MATRIX CANNOT BE INVERTED)
      GOTO 100
29  DO 31 I=1,NL
      DO 31 J=I,NL
      K=(J+M-1)*(J+M)/2+I+M
      R11(I,J)=Q(K)
31  R11(J,I)=R11(I,J)
      CALL DMINVS(R11,27,NL,1.E-15,IERR,N1,WORK)
      IF(IERR.EQ.0) GOTO 36
43  WRITE(6,30)
30  FORMAT(1H0,42H X-Y COORDINATES MATRIX CANNOT BE INVERTED)
      GOTO 100
37  NL=M
      NR=ITERM
      NTOTAL=NL+NR
42  DO 44 I=1,NR
      DO 44 J=I,NR
      K=(J+M-1)*(J+M)/2+I+M
      R22(I,J)=Q(K)
44  R22(J,I)=R22(I,J)
      CALL DMINVS(R22,15,NR,1.E-15,IERR,N1,WORK)
      IF(IERR.EQ.0) GOTO 56
      GOTO 43
56  DO 45 I=1,NL
      DO 45 J=1,NR
      K=(J+M-1)*(J+M)/2+I
45  R12(I,J)=Q(K)
      DO 40 I=1,NL
      DO 40 J=I,NL
      K=J*(J-1)/2+I
      R11(I,J)=Q(K)
40  R11(J,I)=R11(I,J)
      CALL DMINVS(R11,27,NL,1.E-15,IERR,N1,WORK)
      IF(IERR.EQ.0) GOTO 36
      GOTO 39
36  DO 46 I=1,NL
      DO 46 J=1,NR
      PRO(I,J)=0.0
      DO 46 K=1,NL
46  PRO(I,J)=PRO(I,J)+R11(I,K)*R12(K,J)
      DO 47 I=1,NR
      DO 47 J=1,NR
      PRO1(I,J)=0.0
      DO 47 K=1,NL
47  PRO1(I,J)=PRO1(I,J)+R12(K,I)*PRO(K,J)

```

```

DO 48 I=1,NR
DO 48 J=1,NR
PRO2(I,J)=0.0
DO 48 K=1,NR
48 PRO2(I,J)=PRO2(I,J)+R22(I,K)*PRO1(K,J)
DO 49 I=1,NR
DO 49 J=1,NR
VEC(I,J)=0.0
IF(I.EQ.J) VEC(I,J)=1.000
49 CONTINUE
CALL EBERVC(PRO2,NR,1,200,0.01,0.001,1000.0,15,VEC,1)
TEMP=-100.0
DO 50 J=1,NR
IF(TEMP.GT.PRO2(J,J)) GOTO 50
TEMP=PRO2(J,J)
JJ=J
50 CONTINUE
CANON=SQRT(PRO2(JJ,JJ))
DIF=CANON-ROOT
ROOT=CANON
WRITE(6,52) IPOWER,ROOT
52 FORMAT(1H ,10X,24H THE DEGREE IS EQUAL TO 12,5X,23H THE CANONICAL
1ROOT IS F10.4)
IF(DIF.LT.0.05.OR.ROOT.GE.0.95000) GOTO 51
27 CONTINUE
51 WRITE(6,53) IPOWER,ROOT
53 FORMAT(1H0,10X,57H THE DEGREE OF THE MOST PREDICTABLE SURFACE IS
1EQUAL TO 12,3X,48H THE CORRESPONDING CANONICAL ROOT IS EQUAL TO
2F7.4//)
NR1=NR-1
DO 171 I=1,NR1
JP=I+1
DO 171 J=JP,NR
IF(PRO2(I,I).GE.PRO2(J,J))GOTO 171
TEMP=PRO2(I,I)
PRO2(I,I)=PRO2(J,J)
PRO2(J,J)=TEMP
DO 174 K=1,NR
TEMP=VEC(K,J)
VEC(K,J)=VEC(K,I)
174 VEC(K,I)=TEMP
171 CONTINUE
TOTALX=XMAX-XMIN
TOTALY=YMAX-YMIN
INDEX=TOTALY/TOTALX*50.0
DELTA=TOTALX/50.0
TX=50.0/TOTALX
TY=FLOAT(INDEX)/TOTALY
DO 172 JJ=1,NR
INDEX1=1

```

```

PRO2(JJ,JJ)=SQRT(PRO2(JJ,JJ))
WRITE(6,173) PRO2(JJ,JJ)
173 FORMAT(1H1,10X,37H TREND SURFACE FOR THE CANONICAL ROOT,2X,F8.4//)
WRITE(6,114)
114 FORMAT(1H0,10X,34H THE EQUATION OF THE TREND SURFACE///)
DO 54 I=1,NL
E(I)=0.0
DO 54 J=1,NR
54 E(I)=E(I)+SNGL(PRO(I,J))*VEC(J,JJ)
RR=0.0
DO 55 I=1,NL
E(I)=E(I)/PRO2(JJ,JJ)
55 RR=RR+E(I)*E(I)
DO 57 I=1,NL
57 E(I)=E(I)/SQRT(RR)
J2=1
J3=J2+1
IF(M.GT.ITERM) GOTO 59
DO 60 J=1,M
WRITE(6,58) (VARIAT(J1),J1=J2,J3),VEC(J,JJ)
58 FORMAT(1H,5X,2A6,3X,F6.3)
J2=J2+2
60 J3=J3+2
GOTO 112
59 DO 132 J=1,M
WRITE(6,58) (VARIAT(J1),J1=J2,J3),E(J)
J2=J2+2
132 J3=J3+2
112 DO 64 I=1,ITERM,10
IPRINT=I+9
LIMIT=ITERM
IF(ITERM.GT.IPRINT) LIMIT=IPRINT
WRITE(6,115)
115 FORMAT(//)
WRITE(6,62) (TERM(J),J=I,LIMIT)
62 FORMAT(1H0,10(4X,A6))
IF(M.GT.ITERM) GOTO 65
61 WRITE(6,63) (E(IK),IK=I,LIMIT)
63 FORMAT(1H0,10F10.4///)
GOTO 64
65 WRITE(6,63) (VEC(J,JJ),J=I,LIMIT)
64 CONTINUE
WRITE(6,97)
97 FORMAT(1H0,40H NOTE X4Y2=X**4*Y**2,X3=X**3,AND SO ON)
BSUM=0.0
CSUM=0.0
BSS=0.0
CSS=0.0
BCR=0.0
DO 72 NSAMPL=1,N

```

```

L=0
A(1,1)=1.0
DO 73 I=1,IPOWER
DO 74 J=1,I
A(J,I+1)=A(J,I)*DM(NSAMPL,M+1)
L=L+1
74 P(L)=A(J,I+1)
A(I+1,I+1)=A(I,I)*DM(NSAMPL,M+2)
L=L+1
73 P(L)=A(I+1,I+1)
108 B(NSAMPL)=0.0
C(NSAMPL)=0.0
IF(M.GT.ITERM) GOTO 75
DO 76 J=1,ITERM
76 B(NSAMPL)=B(NSAMPL)+E(J)*P(J)
DO 77 I=1,M
77 C(NSAMPL)=C(NSAMPL)+VEC(I,JJ)*DM(NSAMPL,I)
GOTO 81
75 DO 78 J=1,ITERM
78 B(NSAMPL)=B(NSAMPL)+VEC(J,JJ)*P(J)
DO 79 I=1,M
79 C(NSAMPL)=C(NSAMPL)+E(I)*DM(NSAMPL,I)
81 BSUM=BSUM+B(NSAMPL)
BSS=BSS+B(NSAMPL)**2
CSUM=CSUM+C(NSAMPL)
CSS=CSS+C(NSAMPL)**2
72 BCR=BCR+B(NSAMPL)*C(NSAMPL)
DENON=S*BSS-BSUM**2
ALPHA=(BSS*CSUM-BSUM*BCR)/DENON
BETA=(S*BCR-BSUM*CSUM)/DENON
DO 133 I=1,N
133 B(I)=ALPHA+BETA*B(I)
RESIDU(I)=C(I)-B(I)
S2=0.0
DO 116 I=1,N
116 S2=S2+RESIDU(I)
S2=-1.0*S2/S
DO 117 I=1,N
117 RESIDU(I)=RESIDU(I)+S2
WRITE(6,118) S2
118 FORMAT(1H0,10X,55H THE CONSTANT OF THE TREND SURFACE EQUATION IS E
1QUAL TO,F10.4)
WRITE(6,80)
80 FORMAT(1H1,13H X-COORDINATE,5X,13H Y-COORDINATE,5X,15H OBSERVED VA
1LUE,5X,17H CALCULATED VALUE,5X,9H RESIDUAL///)
WRITE(6,82) ((DM(I,M+1),DM(I,M+2),C(I),B(I),RESIDU(I)),I=1,N)
82 FORMAT(1H ,F10.4,8X,F10.4,8X,F10.4,12X,F10.4,8X,F10.4)
DO 66 LENGTH=1,INDEX
Y=YMAX-DELTA*(FLOAT(LENGTH)-1.0)
DO 66 LWIDTH=1,50

```

```

X=XMIN+DELTA*(FLOAT(LWIDTH)-1.0)
L=0
A(1,1)=1.0
DO 67 I=1,IPOWER
DO 68 J=1,I
A(J,I+1)=A(J,I)*X
L=L+1
68 P(L)=A(J,I+1)
A(I+1,I+1)=A(I,I)*Y
L=L+1
67 P(L)=A(I+1,I+1)
107 POLY(LENGTH,LWIDTH)=S2
IF(M.GT.ITERM) GOTO 69
DO 70 J=1,ITERM
70 POLY(LENGTH,LWIDTH)=POLY(LENGTH,LWIDTH)+E(J)*P(J)
GOTO 66
69 DO 71 J=1,ITERM
71 POLY(LENGTH,LWIDTH)=POLY(LENGTH,LWIDTH)+VEC(J,JJ)*P(J)
66 POLY(LENGTH,LWIDTH)=ALPHA+BETA*POLY(LENGTH,LWIDTH)
PMAX=-100000.0
PMIN=100000.0
DO 83 I=1,N
PMAX=AMAX1(PMAX,B(I))
PMIN=AMIN1(PMIN,B(I))
83 CONTINUE
DIFF=PMAX-PMIN
86 GRAD=DIFF/10.0
DO 88 LENGTH=1,INDEX
DO 88 LWIDTH=1,50
DO 121 I=1,10
IGRAD=I
DISC=GRAD*(FLOAT(I)-1.0)+PMIN
IF(DISC.GE.POLY(LENGTH,LWIDTH)) GOTO 123
121 CONTINUE
IGRAD=10
123 IPOLY(LENGTH,LWIDTH)=DOT(IGRAD)
88 CONTINUE
DO 134 I=1,N
DTEMPX=(DM(I,M+1)-XMIN)*TX
DTEMPY=(YMAX-DM(I,M+2))*TY
MAPING(I,1)=DTEMPX
MAPING(I,2)=DTEMPY
IF(MAPING(I,1).EQ.0) MAPING(I,1)=1
IF(MAPING(I,2).EQ.0) MAPING(I,2)=1
134 CONTINUE
DO 137 I=1,INDEX
DO 137 J=1,50
DO 138 K=1,N
IF(MAPING(K,1).EQ.J.AND.MAPING(K,2).EQ.I) GOTO 139
138 CONTINUE

```



```

GOTO 137
139 IPOLY(I,J)=SAMPLE
137 CONTINUE
150 SCALE=TOTALX*4.0/100.0*FEET
WRITE(6,89) TITLE2
89 FORMAT(1H1,26X,16A5/)
SCALE=SCALE
WRITE(6,90) MSCALE
90 FORMAT(1H0,35X,5H ----,18,5H FEET,5X,18H * SAMPLE LOCALITY///)
155 WRITE(6,91)
91 FORMAT(1H ,13X,105H *****
,*****
)
WRITE(6,92) ((IPOLY(I,J),J=1,50),I=1,INDEX)
92 FORMAT(1H ,13X,2H *,101X,2H */14X,2H *,50A2,1X,2H *)
WRITE(6,124)
124 FORMAT(1H ,13X,2H *,101X,2H *)
WRITE(6,91)
WRITE(6,122)
122 FORMAT(1H0,10X,7H LEGEND/)
DO 109 I=1,10
DISC=GRAD*(FLOAT(I)-1.0)+PMIN
109 WRITE(6,110) (DOT(I),DISC)
110 FORMAT(1H0,11X,A2,2H =,F8.3)
IF(INDEX1.EQ.2) GOTO 172
PMAX=-100000.0
PMIN=100000.0
DO 151 I=1,N
PMAX=AMAX1(PMAX,RESIDU(I))
PMIN=AMIN1(PMIN,RESIDU(I))
151 CONTINUE
GRAD=(PMAX-PMIN)/10.0
DO 161 I=1,N
DO 159 J=1,10
JDOT=J
DISC=GRAD*(FLOAT(J)-1.0)+PMIN
IF(DISC.GE.RESIDU(I)) GOTO 160
160 CONTINUE
JDOT=9
160 SIG(I)=DOT(JDOT)
161 CONTINUE
DO 152 I=1,INDEX
DO 152 J=1,50
DO 157 K=1,N
IF(MAPING(K,1).EQ.J.AND.MAPING(K,2).EQ.I) GOTO 156
IPOLY(I,J)=BLANK
157 CONTINUE
GOTO 152
156 IPOLY(I,J)=SIG(K)
158 CONTINUE
EX1=2

```

```
      WRITE(6,158)
158  FORMAT(1H1,10X,17H THE RESIDUAL MAP//)
      GOTO 153
172  CONTINUE
100  STOP
      END
```

Note: If this program is to be used outside McMaster University, one has to insert two subroutines, (1) double precision matrix inversion, (2) eigenvalues and eigenvectors of an arbitrary matrix.

APPENDIX II
FORTRAN IV PROGRAM FOR CANONICAL CORRELATION

7.1 Instructions For Using the Program.

7.11 Input to Program

I Instruction card 1.

The function of this card is to tell the computer how many sets of different data are going to be subjected to canonical correlation analysis. Each data set should have the following input cards (from II to VII).

Columns 1 to 10 are reserved for punching the number, digits are right justified.

II Instruction card 2.

- (1) First 5 columns are reserved for punching the number of samples. There is no reasonable limit on the sample size ($N \leq 99999$ at present case). Digits are right justified.
- (2) Columns 6 to 7 are reserved for punching the number of variates on the left hand set.
- (3) Columns 8 to 9 are reserved for punching the number of variates on the right hand set.

The total number of variates is 100. The number of variates on the left hand set is greater than that of the right hand set. Neither of them can exceed 50. Otherwise, the Dimension statement should be changed.

- (4) Columns 10, 11 and 12 are to specify the number of variates read in by the computer. This number is not necessarily equal to the number of total variates to be subjected to canonical correlation analysis.
- (5) Columns 13, 14 and 15 are used to indicate the number of description cards (see IV).
- (6) Column 16, punch 1 if input is correlation matrix,
punch 2 if input is raw data matrix.
- (7) Column 17, punch 1 Arc sine square root transformation.
The user must make sure that all input data are less than 1.0, and positive when arc sine square root transformation is used.

punch 2 Log transformation (base 10)

The user must make sure that all input data are greater than zero.

punch 3 No transformation

III Title card.

The project title and the investigator's name will be punched on one card (no more than 80 characters).

IV Description card(s).

A brief description about the project is allowed. The maximum number of cards is 50, whereas the minimum number is 1.

V Variate Name card(s).

Twelve characters are used for the names of each variate.

They should be punched in a continuous sequence and also in the same order as in the data matrix or correlation matrix.

Six names are punched on one card.

VI Data Format card.

This is used for the input data card. For example, the original data cards contain 14 variates, such as Si_2O , TiO_2 , Al_2O_3 , Fe_2O_3 , FeO , MgO , CaO , Na_2O , K_2O , CO_2 , quartz, feldspar, micaceous materials, and carbonates. Five columns are used for each variate. If one wishes to study the set of five variates, namely MgO , CaO , CO_2 , Feldspar, and carbonates, the variable format card should be

(30X, 2F5.2, 10X, F5.2, 5X, F5.2, 5X, F5.2)

The sample numbers are punched at the first 5 columns in this case.

VII Data cards.

One may punch the data cards in one's own way. EACH SAMPLE HAS ITS OWN SEPARATE DATA CARD(S). TWO OR MORE SAMPLES SHARING ONE DATA CARD IS NOT ALLOWED. The correlation matrix must be stored row by row, each row going up to the diagonal elements.

For example:

1.000			
0.979	1.000		
0.807	0.818	1.000	
0.462	0.549	0.731	1.000

The input data cards should be as follows:

1000097910000807081810000462054907311000

The input data format is (10F4.3).

7.12 Output from program

- (1) Project title and the investigator's name,
- (2) Description about the project,
- (3) Correlation matrix,
- (4) Mean, variance, and standard deviation of each variate,
- (5) Canonical roots, their chi-square values, and significant levels,
- (6) Canonical vectors both of right hand set and left hand set.

Note: If this program is to be used outside McMaster University, one has to insert two subroutines, (1) double precision matrix inversion, (2) eigenvalues and eigenvectors of an arbitrary matrix.

7.2 Listing of the Program

THIS IS CANONICAL CORRELATION PROGRAM AUGUST 1966 P J LEE

```

DOUBLE PRECISION Q(1830),COR(30,30),A2(30,30),B(30,30),
1SUMSQ(60),B1(30,30),A(465)
  DIMENSION TITLE(16),VARIAT(120),SOURCE(800),FMT(16),SUM(60),D(60)
1,R(30),VEC(30,30),B2(30,30),CHI(30)
  DATA CHI/6.635,9.210,11.345,13.277,15.086,16.812,18.475,20.090,
121.666,23.209,24.725,26.217,27.688,29.141,30.578,32.000,33.409,
234.805,36.191,37.566,38.932,40.289,41.638,42.980,44.314,45.642,
346.963,48.278,49.588,50.892/
  READ(5,91) ISET
91 FORMAT(I10)
  DO 90 NSET=1,ISET
  READ(5,1) N,NL,NR,MM,NS,I1,I2
  1 FORMAT(I5,2I2,2I3,2I1)
  READ(5,2) TITLE
  2 FORMAT(16A5)
  WRITE(6,3) TITLE
  3 FORMAT(1H1,30X,16A5)
  WRITE(6,4)
  4 FORMAT(///)
  NN=NS*16
  READ(5,5) (SOURCE(I),I=1,NN)
  5 FORMAT(16A5)
  WRITE(6,6) (SOURCE(I),I=1,NN)
  6 FORMAT(1H0,20X,16A5)
  M=NL+NR
  NV=2*M
  S=N
  M1=(M+1)*M/2
  READ(5,7) (VARIAT(I),I=1,NV)
  7 FORMAT(12A6)
  READ(5,8) FMT
  8 FORMAT(16A5)
  IF(I1.EQ.1) GOTO 9
  DO 10 J=1,M
  SUM(J)=0.0
10 SUMSQ(J)=0.0
  DO 11 J=1,M1
11 Q(J)=0.0
  DO 20 NSAMPL=1,N
  READ(5,FMT) (D(J),J=1,MM)
  IF(I2.EQ.3) GOTO 13
  IF(I2.EQ.2) GOTO 14
  IF(I2.EQ.1) GOTO 15
  WRITE(6,16)
16 FORMAT(1H0,37H WRONG CODE NUMBER FOR TRANSFORMATION)

```

```

      GOTO 83
14 DO 17 J=1,M
17 D(J)=ALOG10(D(J))
      GOTO 13
15 DO 18 J=1,M
18 D(J)=ARSIN(SQRT(D(J)))
13 DO 19 J=1,M
      SUM(J)=SUM(J)+D(J)
19 SUMSQ(J)=SUMSQ(J)+D(J)*D(J)
      M2=M-1
      DO 20 K=1,M2
      DO 20 K1=K,M
      L=K1*(K1-1)/2+K
88 IF(K1-K) 21,22,21
22 Q(L)=1.000
      GOTO 20
21 Q(L)=Q(L)+D(K)*D(K1)
20 CONTINUE
      DO 23 I=1,M2
      DO 23 J=I,M
      K=J*(J-1)/2+I
      IF(I-J) 24,25,24
25 Q(K)=1.000
      GOTO 23
24 Q(K)=(S*Q(K)-SUM(I)*SUM(J))/SQRT((S*SUMSQ(I)-SUM(I)*SUM(I))*
      1(S*SUMSQ(J)-SUM(J)*SUM(J)))
23 CONTINUE
      Q(K+1)=1.000
      WRITE(6,99)
99 FORMAT(1H1,30X,5H MEAN,11X,9H VARIANCE,5X,19H STANDARD DEVIATION)
      DO 101 I=1,M
      AMEAN=SUM(I)/S
      VAR=(S*SUMSQ(I)-SUM(I)**2)/(S*(S-1.0))
      DEV=SQRT(VAR)
      IV=2*I-1
      IV1=IV+1
101 WRITE(6,100) ((VARIAT(J),J=IV,IV1),AMEAN,VAR,DEV)
100 FORMAT(1H ,10X,2A6,5X,F8.3,9X,F10.3,9X,F8.3)
      GOTO 26
      9 READ(5,FMT) (Q(I),I=1,M1)
26 DO 27 L=1,M,14
      WRITE(6,30) N
30 FORMAT(1H1,10X,41H CORRELATION MATRIX      THE SAMPLE SIZE IS,I5///)
      IF(I2-2) 92,93,94
94 WRITE(6,95)
95 FORMAT(1H0,30H NO TRANSFORMATION ON RAW DATA//)
      GOTO 96
92 WRITE(6,97)
97 FORMAT(1H0,36H ARC SINE TRANSFORMATION ON RAW DATA//)
      GOTO 96

```



```

93 WRITE(6,98)
98 FORMAT(1H0,31H LOG TRANSFORMATION ON RAW DATA//)
96 J2=(L-1)*2+1
   J3=J2+1
   DO 27 J=L,M
     K=0
     DO 29 I=L,J
       LIMIT=L+13
       IF(I.GT.LIMIT) GOTO 86
       KK=J*(J-1)/2+I
       K=K+1
29  A(K)=Q(KK)
86  WRITE(6,84) ((VARIAT(J1),J1=J2,J3),(A(K1),K1=1,K))
84  FORMAT(1H0,5X,2A6,14F8.3)
     J2=J2+2
     J3=J3+2
27  CONTINUE
28  L=0
     NLH=(NL+1)*NL/2
     DO 32 J=1,NLH
       L=L+1
32  A(L)=Q(L)
     CALL DSYMIN(A,NL,IERR)
     IF(IERR.EQ.0) GOTO 34
     WRITE(6,85)
85  FORMAT(1H0,85H CORRELATION MATRIX ON THE LEFT HAND SIDE CANNOT BE
1  INVERTED TRANSFORMATION IS NEEDED)
     GOTO 90
34  DO 49 I=1,NL
     DO 49 J=I,NL
       K=J*(J-1)/2+I
       A2(I,J)=A(K)
49  A2(J,I)=A2(I,J)
     NL1=NL+1
     L=0
     DO 53 J=NL1,M
     DO 53 I=NL1,J
       L=L+1
       K=(J-1)*J/2+I
53  A(L)=Q(K)
     DO 50 J=1,NR
     DO 50 I=1,NL
       K=(J+NL-1)*(J+NL)/2+I
50  COR(I,J)=Q(K)
     DO 51 K=1,NL
     DO 51 I=1,NR
       B(K,I)=0.0
     DO 51 J=1,NL
51  B(K,I)=B(K,I)+A2(K,J)*COR(J,I)
     DO 52 K=1,NR

```

```

DO 52 I=1,NR
  B1(K,I)=0.0
DO 52 J=1,NL
52 B1(K,I)=B1(K,I)+COR(J,K)*B(J,I)
  CALL DSYMIN(A,NR,IERR)
  IF(IERR.EQ.0) GOTO 55
  WRITE(6,56)
56 FORMAT(1H1,85H CORRELATION MATRIX ON THE RIGHT HAND SIDE CANNOT BE
  1INVERTED TRANSFORMATION IS NEEDED)
  GOTO 90
55 DO 58 I=1,NR
  DO 58 J=I,NR
    K=J*(J-1)/2+I
    A2(I,J)=A(K)
58 A2(J,I)=A2(I,J)
  DO 59 I=1,NR
  DO 59 J=1,NR
    B2(I,J)=0.0
    DO 59 K=1,NR
59 B2(I,J)=B2(I,J)+A2(I,K)*B1(K,J)
  DO 60 I=1,NR
  DO 60 J=1,NR
    IF(I.EQ.J) GOTO 61
    VEC(I,J)=0.0
    GOTO 60
61 VEC(I,J)=1.00
60 CONTINUE
  CALL EBERVC(B2,NR,1,200,0.01,0.001,1000.0,30,VEC,1)
  M3=NR-1
  DO 62 I=1,M3
    IJ=I+1
    DO 62 J=IJ,NR
      IF(B2(I,I).GE.B2(J,J)) GOTO 62
      TEMP=B2(I,I)
      B2(I,I)=B2(J,J)
      B2(J,J)=TEMP
    DO 63 K=1,NR
      R(K)=VEC(K,I)
      VEC(K,I)=VEC(K,J)
63 VEC(K,J)=R(K)
62 CONTINUE
  WRITE(6,65)
65 FORMAT(1H1,10X,47H CHI SQUARE TESTS OF SUCCESSIVE CANONICAL ROOTS)
  WRITE(6,64)
64 FORMAT(1H0,16H CANONICAL ROOTS,5X,7H LAMBDA,8X,18H CHI SQUARE VALU
  1ES,5X,18H DEGREE OF FREEDOM,5X,18H SIGNIFICANT LEVEL)
  WL=1.000
  NROOT=0
  DO 66 I=1,NR
    JJ=NR-I+1

```

```

      IF(B2(JJ,JJ).LT.0.000) B2(JJ,JJ)=0.0
      WL=WL*(1.00-B2(JJ,JJ))
      L=(NR-JJ+1)*(NL-JJ+1)
      CHISQ=-1.00*(S-0.5*(FLOAT(M)+1.00))*ALOG(WL)
      IF(B2(JJ,JJ).NE.0.000) GOTO 67
      GOTO 68
67  NROOT=NROOT+1
      B2(JJ,JJ)=SQRT(B2(JJ,JJ))
68  IF(L.GE.30) GOTO 69
      IF(CHISQ.GE.CHI(L)) GOTO 70
71  F=-0.01
      GOTO 66
70  F=0.01
      GOTO 66
69  IF(CHISQ.GE.CHI(30)) GOTO 70
      GOTO 71
66  WRITE(6,72) B2(JJ,JJ),WL,CHISQ,L,F
72  FORMAT(1H0,6X,F8.3,5X,F8.3,15X,F8.3,18X,I4,12X,F6.2)
      WRITE(6,12)
12  FORMAT(1H0,56H 0.01 SIGNIFICANT AT 1 PERCENT LEVEL -0.01 INSIGNIFI
      ICANT)
      WRITE(6,73)
73  FORMAT(1H1,10X,41H CANONICAL VECTORS ON THE RIGHT HAND SIDE///10X,
      131H LARGEST VECTOR AT FIRST COLUMN//)
      KKK=2*NL
      DO 74 I=1,NR
      KK=KKK+1
      KKK=KK+1
74  WRITE(6,75) ((VARIAT(K),K=KK,KKK),(VEC(I,J),J=1,NROOT))
75  FORMAT(1H0,2A6,4X,14F6.2)
      DO 76 I=1,NL
      DO 76 J=1,NROOT
      COR(I,J)=0.0
      DO 76 K=1,NR
76  COR(I,J)=COR(I,J)+B(I,K)*VEC(K,J)
      DO 77 I=1,NROOT
      DO 77 J=1,NL
77  COR(J,I)=COR(J,I)/B2(I,I)
      DO 78 I=1,NROOT
      R(I)=0.0
      DO 78 J=1,NL
78  R(I)=R(I)+COR(J,I)*COR(J,I)
      DO 79 I=1,NROOT
      DO 79 J=1,NL
79  COR(J,I)=COR(J,I)/SQRT(R(I))
      WRITE(6,89)
89  FORMAT(///)
      WRITE(6,80)
80  FORMAT(1H0,10X,40H CANONICAL VECTORS ON THE LEFT HAND SIDE///10X,
      131H LARGEST VECTOR AT FIRST COLUMN//)

```

```
    KKK=0  
    DO 81 I=1,NL  
      KK=KKK+1  
      KKK=KK+1  
81  WRITE(6,82) ((VARIAT(K),K=KK,KKK),(COR(I,J),J=1,NROOT))  
82  FORMAT(1H0,2A6,4X,14F6.2)  
90  CONTINUE  
83  STOP  
    END
```

APPENDIX III

FORTRAN IV PROGRAM FOR DISCRIMINATORY ANALYSIS BY CANONICAL VARIATES

8.1 Instructions For Using the Program

8.11 Input to Program A

I Instruction card

- (1) Columns 1 and 2 are for punching the number of groups (≤ 20).
- (2) Columns 3 and 4 are for punching the number of variates (≤ 50).
- (3) Columns 5 and 6 are for punching the number of description card(s) ($1 \leq$ and ≤ 20).

II Description card(s)

The title of project, name of investigator, and a brief description about the project are punched on the description cards.

III Variate Name card(s)

Twelve characters are used for one variate name. Six names are punched on one card. All groups should have the same variate(s).

IV Group Name card(s)

Twelve characters are used for one group name. Six names are punched on one card.

V Data Format card

This is used for the input data. EACH SAMPLE HAS ITS OWN SEPARATE DATA CARDS. TWO OR MORE SAMPLES SHARING ONE DATA CARD IS NOT ALLOWED. Data of all groups should be punched in the same format.

VI Data cards

Each group should have the following cards: First card is used for punching the number of samples of this group. First five columns are reserved for this number. Digits are right justified. Data cards follow the first card immediately.

8.12 Output from Program A

I Description about project

II Canonical root and result of statistical test

III Canonical transformation matrix (print as well as punch)

IV Means of all groups expressed in terms of canonical coordinates.

8.13 Input to Program B

If one wishes to plot each individual sample on canonical coordinates, the program B should be used.

I Instruction card

(1) Columns 1 and 2 are for punching the number of groups (≤ 20).

(2) Columns 3 and 4 are for punching the number of variates (≤ 50).

(3) Columns 5 and 6 are for punching the number of canonical axes (≤ 50).

II Data format card (same as used in program A)

III Canonical transformation matrix

These are the output cards from program A.

IV Data cards

These are exactly same as used in program A.

8.14 Output from Program B

Canonical coordinates for each sample are printed and group 1, 2, ..., are also labelled.

Note: If the program A is used outside McMaster University, two subroutines should be inserted into the main program (1) matrix inversion, (2) eigenvalues and eigenvector of a nonsymmetrical matrix.

8.2 Listing of the Program

```

C   THIS IS DISCRIMINATORY ANALYSIS BY CANONICAL VARIATES
C   JULY 1967   P. J. LEE
C   THIS IS PROGRAM A
C
  DIMENSION SUM(20,50),SUMB(50),R(50),VEC(50,50),FMT(16),Z(50),
1 SOURCE(320),TITLE(16),CHI(37),WORK(100),N1(350),VARIAT(100),
2B(50,50),T(50,50),SUMA(50,50),BPOOL(50,50),NUMBER(50),GROUP(40)
  DATA NUMBER/ 1, 2, 3, 4, 5, 6, 7, 8, 9,10,11,12,13,14,15,16,17,18,
119,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,
241,42,43,44,45,46,47,48,49,50/
  DATA CHI/ 6.6349, 9.2103, 11.3449, 13.2767, 15.0863, 16.8119, 18
1.4753, 20.0902, 21.6660, 23.2093, 24.7250, 26.2170, 27.6883, 29.14
213, 30.5779, 31.9999, 33.4087, 34.8053, 36.1908, 37.5662, 38.9321,
3 40.2894, 41.6384, 42.9798, 44.3141, 45.6417, 46.9630, 48.2782, 49
4.5879, 50.8922, 63.6907, 76.1539, 88.3794,100.4250,112.3290,124.11
560,135.8070/
  READ(5,1) NOG,NOV,NOD
1 FORMAT(3I2)
  NOC=NOV*2
  NOD=NOD*16
  READ(5,3) (SOURCE(I),I=1,NOD)
3 FORMAT(16A5)
  WRITE(6,42) (SOURCE(I),I=1,NOD)
42 FORMAT(1H0,20X,16A5)
  NOV1=2*NOV
  READ(5,4) (VARIAT(I),I=1,NOV1)
4 FORMAT(12A6)
  NOG1=NOG*2
  READ(5,4) (GROUP(I),I=1,NOG1)
  READ(5,3) FMT
  NTOTAL=0
  DO 6 I=1,NOV
  DO 5 J=I,NOV
    T(I,J)=0.0
    T(J,I)=0.0
    BPOOL(I,J)=0.0
5 BPOOL(J,I)=0.0
6 R(I)=0.0
  DO 7 I=1,NOG
  DO 7 J=I,NOV
    SUM(I,J)=0.0
7 SUM(J,I)=0.0
  DO 8 NGROUP=1,NOG
  READ(5,9) NN
9 FORMAT(I5)
  NTOTAL=NTOTAL+NN
  S1=NN
  DO 10 I=1,NOV

```



```

DO 11 J=I,NOV
  B(I,J)=0.0
11 B(J,I)=0.0
10 SUMB(I)=0.0
  DO 12 IREAD=1,NN
    READ(5,FMT) (Z(J),J=1,NOV)
    DO 12 I=1,NOV
      DO 13 J=I,NOV
        B(I,J)=B(I,J)+Z(I)*Z(J)
13 B(J,I)=B(I,J)
12 SUMB(I)=SUMB(I)+Z(I)
  DO 14 I=1,NOV
    DO 15 J=I,NOV
      T(I,J)=T(I,J)+B(I,J)
15 T(J,I)=T(I,J)
      R(I)=R(I)+SUMB(I)
      SUMB(I)=SUMB(I)/S1
14 SUM(NGROUP,I)=SUMB(I)
  DO 16 I=1,NOV
    DO 16 J=I,NOV
      SUMA(I,J)=SUMB(I)*SUMB(J)*S1
16 SUMA(J,I)=SUMA(I,J)
  DO 17 I=1,NOV
    DO 17 J=I,NOV
      BPOOL(I,J)=BPOOL(I,J)+B(I,J)-SUMA(I,J)
17 BPOOL(J,I)=BPOOL(I,J)
  8 CONTINUE
  SS=NTOTAL
  DO 18 I=1,NOV
18 R(I)=R(I)/SS
  DO 19 I=1,NOV
    DO 19 J=1,NOV
19 SUMA(I,J)=R(I)*R(J)*SS
  DO 20 I=1,NOV
    DO 20 J=I,NOV
      B(I,J)=T(I,J)-SUMA(I,J)-BPOOL(I,J)
20 B(J,I)=B(I,J)
  CALL MINVSE(BPOOL,50,NOV,0.0,IERR,N1,WORK)
  IF(IERR.NE.0) GOTO 99
  DO 21 I=1,NOV
    DO 21 J=1,NOV
      T(I,J)=0.0
    DO 21 K=1,NOV
21 T(I,J)=T(I,J)+BPOOL(I,K)*B(K,J)
  DO 22 I=1,NOV
    DO 22 J=I,NOV
      VEC(I,J)=0.0
      VEC(J,I)=0.0
      IF(I.EQ.J) VEC(I,J)=1.000
22 CONTINUE

```

```

CALL EBERVC(T,NOV,1,200,0.01,0.001,1000.0,50,VEC,1)
M3=NOV-1
DO 23 I=1,M3
  IJ=I+1
  DO 23 J=IJ,NOV
    IF(T(I,I).GE.T(J,J)) GOTO 23
    TEMP=T(I,I)
    T(I,I)=T(J,J)
    T(J,J)=TEMP
  DO 24 K=1,NOV
    R(K)=VEC(K,I)
    VEC(K,I)=VEC(K,J)
24 VEC(K,J)=R(K)
23 CONTINUE
  WRITE(6,25)
25 FORMAT(1H1,10X,47H CHI SQUARE TESTS OF SUCCESSIVE CANONICAL ROOTS)
  WRITE(6,26)
26 FORMAT(1H0,16H CANONICAL ROOTS,5X,7H LAMBDA, 8X,18H CHI SQUARE VAL
  UES,5X,18H DEGREE OF FREEDOM,5X,18H SIGNIFICANT LEVEL)
  WL=1.000
  NROOT=0
  NOV2=NOV
  NOG1=NOG-1
  IF(NOV.GT.NOG1) NOV2=NOG1
  DO 31 I=1,NOV2
    JJ=NOV2-I+1
    IF(T(JJ,JJ).LT.0.000) T(JJ,JJ)=0.0
    WL=WL*(1.000+T(JJ,JJ))
    L=(NOV-JJ+1)*(NOG-JJ)
    CHISQ=(SS-1.0-0.5*(FLOAT(NOV)+FLOAT(NOG)))*ALOG(WL)
    IF(T(JJ,JJ).EQ.0.000) GOTO 27
    NROOT=NROOT+1
    T(JJ,JJ)=SQRT(T(JJ,JJ))
27 IF(L.GT.30) GOTO 28
    IF(CHISQ.GE.CHI(L)) GOTO 29
30 F=-0.01
    GOTO 31
29 F=0.01
    GOTO 31
28 IF(L.GT.100) GOTO 32
    L=(L-25)/10+30
    IF(CHISQ.GE.CHI(L)) GOTO 29
    GOTO 30
32 CHI1=0.5*(2.3263+SQRT(2.0*FLOAT(L)-1.000))**2
    IF(CHISQ.GE.CHI1) GOTO 29
    GOTO 30
31 WRITE(6,34) T(JJ,JJ),WL,CHISQ,L,F
34 FORMAT(1H0,6X,F8.3,5X,F8.3,15X,F8.3,18X,I4,12X,F6.2)
  WRITE(6,35)
35 FORMAT(1H0,61H NOTE 0.01 SIGNIFICANT AT 1 PERCENT LEVEL -0.01 INSI

```

```

1GNIFICANT)
WRITE(6,36)
36 FORMAT(1H1,10X,29H THE CANONICAL TRANSFORMATION//)
CALL ROSIE(VEC,VARIAT,NUMBER,NROOT,NOV)
DO 41 J=1,NOV
41 WRITE(7,2) (VEC(J,I),I=1,NROOT)
2 FORMAT(10F8.4)
DO 37 I=1,NOG
DO 37 J=1,NROOT
T(I,J)=0.0
DO 37 K=1,NOV
37 T(I,J)=T(I,J)+VEC(K,J)*SUM(I,K)
WRITE(6,38)
38 FORMAT(1H1,10X,68H THE GROUP MEAN COORDINATES OF EACH VARIATE ALON
1G THE CANONICAL AXES//)
CALL ROSIE(T,GROUP,NUMBER,NROOT,NOG)
WRITE(6,39)
39 FORMAT(///)
WRITE(6,40)
40 FORMAT(1H0,40X,68H *****OUR PROGRAM IS ONLY TO PRODUCE WHAT IT CON
1TAINS THANK YOU*****)
WRITE(6,44)
44 FORMAT(1H1)
GOTO 100
99 WRITE(6,43)
43 FORMAT(1H0,51H WITHIN GROUP DISPERSION MATRIX CAN NOT BE INVERTED)
100 STOP
END

```

```

SUBROUTINE ROSIE(VEC,VARIAT,NUMBER,NROOT,NOV)
DIMENSION VEC(50,50),VARIAT(100),NUMBER(50),GROUP(40)
DO 1 L=1,NROOT,10
LIMIT=L+9
IPRINT=NROOT
IF(NROOT.GT.LIMIT) IPRINT=LIMIT
WRITE(6,2) (NUMBER(J),J=1,IPRINT)
2 FORMAT(1H0,7X,15H CANONICAL AXIS,5X,10I10//)
DO 1 K=1,NOV
IV=2*K-1
IV1=IV+1
1 WRITE(6,3) ((VARIAT(J),J=IV,IV1),(VEC(K,J),J=L,IPRINT))
3 FORMAT(1H ,16X,2A6,10F10.4)
RETURN
END

```

C
C
C
THIS IS PROGRAM B

```
DIMENSION T(50),VEC(50,50),D(50),FMT(12)
READ(5,1) NOG,NOV,NROOT,FMT
1 FORMAT(3I2/12A6)
DO 2 J=1,NOV
2 READ(5,3) (VEC(J,K),K=1,NROOT)
3 FORMAT(10F8.4)
WRITE(6,10)
10 FORMAT(1H1,42H THE CANONICAL COORDINATES FOR ALL SAMPLES///)
DO 7 NGROUP=1,NOG
READ(5,6) N
6 FORMAT(I5)
DO 8 I=1,N
READ(5,FMT) (D(J),J=1,NOV)
DO 5 K=1,NROOT
T(K)=0.0
DO 5 J=1,NOV
5 T(K)=T(K)+VEC(J,K)*D(J)
8 WRITE(6,9) NGROUP,(T(K),K=1,NROOT)
9 FORMAT(1H0,8H GROUP 12,10X,(15F8.2))
7 CONTINUE
END
```

APPENDIX IV

FORTRAN PROGRAM FOR PRINCIPAL COMPONENT ANALYSIS IV

9.1 Instructions For Using The Program

9.11 Input to Program

I Instruction card 1.

The function of this card is to tell the computer how many sets of different data are to be subjected to component analysis. Each data set should have the following input cards (from II to VII). Columns 1 to 10 are reserved for punching the number, digits are right justified.

II Instruction card 2.

- (1) First 5 columns are reserved for punching the number of samples. There is no reasonable limit on the sample size ($N \leq 99999$ at the present case). Digits are right justified.
- (2) Columns 6, 7 and 8 are reserved for punching the number of variates. The maximum number of variates is 100.
- (3) Columns 9, 10 and 11 are reserved for punching the total number of variates. It is not necessary that all variates are to be subjected to principal component analysis.
- (4) Columns 12, 13 and 14 are used to indicate the number of description cards.
- (5) Column 15

punch 1 if input is correlation matrix

punch 2 if input is raw data matrix

(6) Column 16

punch 1 Arc sine square root transformation

The user must make sure that all input data are less than 1.00, and positive when this transformation is used.

punch 2 Log transformation (base 10)

The user must make sure that all input data are greater than zero.

punch 3 no transformation

(7) Column 17

punch 1 number of components to be rotated is specified by the user (see no. 8).

punch 2 number of components to be rotated is equal to number of components (or eigenvalues) of the correlation matrix which are greater than zero.

punch 3 number of components to be rotated is equal to the number of components (or eigenvalues), which are greater than 1.000.

(8) Columns 18, 19 and 20

To specify the number of components to be rotated. It is used for the case 1 of (7).

(9) Column 21

punch 1 for using correlation matrix

punch 0 for using covariance matrix

If the variates are in widely different units, covariance of the original quantities would have little meaning and

the correlation matrix should be used. Conversely, if the variates are reasonably comparable, the covariance form has a greater statistical appeal (Anderson, 1963).

(10) Column 22

punch 1 for R-mode

punch 0 for Q-mode

III Title card

The project title and investigator's name will be punched on one card (no more than 80 characters).

IV Description card(s)

A brief description about your project is allowed. The maximum number of cards is 50, whereas the minimum number is 1.

V Variate Name Cards

Twelve characters are used for the name of each variate. They should be punched in a continuous sequence and also in the same order as in the data matrix or correlation matrix. Six names are on one card.

VI Data Format cards

This is used for your input data card. See "Instructions for using the Canonical Correlation Program).

VII Data cards

One may punch the data cards in one's own way. The

correlation matrix must be stored row by row, each row going up to the diagonal elements (see "Instructions for using the canonical correlation programme"). Each sample has its own separate data card(s). Two or more samples sharing one data card is not allowed.

Note: If the variates are punched in columns whereas the samples are punched in rows, this data matrix will only be used for R-mode principal component analysis. If we wish to carry out both R- and Q- mode on the same set of data, the raw data matrix should be transposed and repunched by matrix transpose program in order to arrange the sample in columns and variates in rows. Again, each variate has its own separate data card(s). Two or more variates sharing one data card is not allowed.

9.12 Output from Program

- I Project title and investigator's name.
- II Description of the project.
- III Mean, variance, and standard deviation of each variate for R-mode.
- IV Correlation matrix.
- V Components (eigenvalues) and their contributions to the total variance.
- VI The number of cycle for varimax rotation and varimax

criterion.

VII The rotated normalized component loadings arranged in column for each factor.

VIII Punch the normalized component loadings (6E12.4) and the rotated normalized component loadings (10F8.4) on the cards.

Note: If this program is to be used outside McMaster University, one has to insert a subroutine for evaluating eigenvalues and eigenvectors of a symmetric matrix.

9.2 Listing of the Program

THIS IS PRINCIPAL COMPONENT ANALYSIS PROGRAM JUNE 1966 P J LEE

```

C
C
C
  DIMENSION TITLE(16),VARIAT(200),SOURCE(800),FMT(16),SUMSQ(100),
1 SUM(100),Q(5050),D(100),R(100),VEC(100,100),WORK(702),A(5050),
2 H(100),R1(100),R2(100),NUMBER(100)
  EQUIVALENCE (TITLE,SOURCE,VARIAT),(Q,VEC),(A,VEC(5051)),(SUM,H),
1 (SUMSQ,R1)
  DO 200 I=1,100
200 NUMBER(I)=I
  READ(5,91) ISET
91 FORMAT(I10)
  DO 90 NSET=1,ISET
  READ(5,1) N,M,NOV,NS,I1,I2,I3,IFAC,I4,I5
1 FORMAT(I5,3I3,3I1,I3,2I1)
  READ(5,2) TITLE
2 FORMAT(16A5)
  WRITE(6,3) TITLE
  WRITE(7,2) TITLE
3 FORMAT(1H1,30X,16A5)
  WRITE(6,4)
4 FORMAT(///)
  NN=NS*16
  READ(5,5) (SOURCE(I),I=1,NN)
5 FORMAT(16A5)
  WRITE(6,6) (SOURCE(I),I=1,NN)
6 FORMAT(1H ,20X,16A5)
  S=N
  M1=(M+1)*M/2
  NV=M*2
  READ(5,7) (VARIAT(I),I=1,NV)
7 FORMAT(12A6)
  READ(5,8) FMT
8 FORMAT(16A5)
  IF(I1.EQ.1) GOTO 9
  DO 10 J=1,M
  SUM(J)=0.0
10 SUMSQ(J)=0.0
  DO 11 J=1,M1
11 Q(J)=0.0
  DO 20 NSAMPL=1,N
  READ(5,FMT) (D(J),J=1,NOV)
  IF(I2.EQ.3) GOTO 13
  IF(I2.EQ.2) GOTO 14
  IF(I2.EQ.1) GOTO 15
  WRITE(6,16)
16 FORMAT(1H0,37H WRONG CODE NUMBER FOR TRANSFORMATION)
  GOTO 90

```

```

14 DO 17 J=1,M
17 D(J)=ALOG10(D(J))
   GOTO 13
15 DO 18 J=1,M
18 D(J)=ARSIN(SQRT(D(J)))
13 DO 19 J=1,M
   SUM(J)=SUM(J)+D(J)
19 SUMSQ(J)=SUMSQ(J)+D(J)*D(J)
   M2=M-1
   DO 20 K=1,M2
   DO 20 K1=K,M
   L=K1*(K1-1)/2+K
21 Q(L)=Q(L)+D(K)*D(K1)
20 CONTINUE
   IF(I4.EQ.1) GOTO 138
   IF(I4.EQ.0) GOTO 139
   WRITE(6,140)
140 FORMAT(1H0,48H WRONG CODE FOR COVARIANCE OR CORRELATION MATRIX)
   GOTO 90
139 DO 142 I=1,M2
   DO 142 J=I,M
   K=(J-1)*J/2+I
142 Q(K)=(S*Q(K)-SUM(I)*SUM(J))/(S*(S-1.0))
   GOTO 141
138 DO 23 I=1,M2
   DO 23 J=I,M
   K=(J-1)*J/2+I
   IF(I-J) 24,25,24
25 Q(K)=1.000
   GOTO 23
24 Q(K)=(S*Q(K)-SUM(I)*SUM(J))/SQRT((S*SUMSQ(I)-SUM(I)*SUM(I))*
1(S*SUMSQ(J)-SUM(J)*SUM(J)))
23 CONTINUE
   Q(K+1)=1.000
141 IF(I5.EQ.1) GOTO 143
   IF(I5.EQ.0) GOTO 26
   WRITE(6,145)
145 FORMAT(1H0,29H WRONG CODE FOR R OR Q MATRIX)
   GOTO 90
143 WRITE(6,51)
51 FORMAT(1H1,30X,5H MEAN,11X,9H VARIANCE,5X,19H STANDARD DEVIATION)
   DO 28 I=1,M
   AMEAN=SUM(I)/S
   VAR=(S*SUMSQ(I)-SUM(I)**2)/(S*(S-1.0))
   DEV=SQRT(VAR)
   IV=2*I-1
   IV1=IV+1
28 WRITE(6,49)((VARIAT(J),J=IV,IV1), AMEAN,VAR,DEV)
49 FORMAT(1H ,10X,2A6,5X,F8.1,9X,E12.4,9X,E12.4)
   IF(I4.EQ.0) GOTO 150

```

```

      GOTO 26
    9 READ(5,FMT) (Q(I),I=1,M1)
  26 DO 27 L=1,M,14
      WRITE(6,30) N,M
  30 FORMAT(1H1,10X,55H CORRELATION OR COVARIANCE MATRIX      THE SAMPLE
    1SIZE IS,15,5X,26H THE NUMBER OF VARIATES IS,13/)
      IF(I2-2) 122,121,120
  120 WRITE(6,123)
  123 FORMAT(1H0,30H NO TRANSFORMATION ON RAW DATA///)
      GOTO 124
  121 WRITE(6,125)
  125 FORMAT(1H0,31H LOG TRANSFORMATION ON RAW DATA///)
      GOTO 124
  122 WRITE(6,127)
  127 FORMAT(1H0,36H ARC SINE TRANSFORMATION ON RAW DATA///)
  124 J2=(L-1)*2+1
      J3=J2+1
      DO 27 J=L,M
      K=0
      DO 29 I=L,J
      LIMIT=L+13
      IF(I.GT.LIMIT) GOTO 86
      KK=J*(J-1)/2+I
      K=K+1
  29 A(K)=Q(KK)
  86 WRITE(6,84) ((VARIAT(J1),J1=J2,J3),(A(K1),K1=1,K))
  84 FORMAT(1H ,5X,2A6,14F8.3)
      J2=J2+2
      J3=J3+2
  27 CONTINUE
  150 K=0
      DO 35 I=1,M
      DO 35 J=I,M
      K=K+1
      L=J*(J-1)/2+I
  35 A(K)=Q(L)
      REWIND 0
      CALL HOUSE2(M,M,0.000001,A,R,WORK,M)
      REWIND 0
      DO 37 I=1,M
  37 READ(0) (VEC(I,J),J=1,M)
      DO 38 I=1,M
      SUM(I)=0.0
      DO 38 J=1,M
  38 SUM(I)=SUM(I)+VEC(I,J)*VEC(I,J)
      DO 39 I=1,M
      DO 39 J=1,M
  39 VEC(I,J)=VEC(I,J)*SQRT(ABS(R(I)))/SQRT(SUM(I))
      NFAC=0
      DO 40 I=1,M

```

```

      IF(R(I).LE.0.001) GOTO 40
      NFAC=NFAC+1
40  CONTINUE
      PER=0.0
      DO 92 I=1,NFAC
92  PER=PER+R(I)
      DO 93 I=1,NFAC
93  R1(I)=(R(I)/PER)*100.0
      WRITE(6,43)
43  FORMAT(1H1,10X,66H TABLE FOR COMMON FACTORS AND THEIR CONTRIBUTION
      1 TO TOTAL VARIANCE//17X,8H FACTORS,2X,30H PER CENT OF TOTAL COMMUN
      2ALITY,2X,20H CUMULATIVE PER CENT//)
      R2(1)=0.0
      DO 116 I=1,NFAC
116 R2(I+1)=R2(I)+R1(I)
      DO 115 I=1,NFAC
115 WRITE(6,94) (R(I),R1(I),R2(I+1))
94  FORMAT(1H ,15X,E12.4,12X,F8.2,19X,F8.2)
      IF(I3.EQ.1) GOTO 109
      IF(I3.EQ.2) GOTO 110
      IF(I3.EQ.3) GOTO 111
      WRITE(6,112)
112 FORMAT(1H0,10X,24H WRONG CODE FOR ROTATION)
      GOTO 90
109 LFAC=IFAC
      GOTO 113
110 LFAC=NFAC
      GOTO 113
111 KFAC=0
      DO 114 I=1,M
      IF(R(I).GE.1.000) KFAC=KFAC+1
114 CONTINUE
      LFAC=KFAC
113 NFAC1=LFAC-1
      WRITE(7,146)
146 FORMAT(26H NORMALIZED FACTOR LOADING)
      DO 147 J=1,M
147 WRITE(7,135) (VEC(I,J),I=1,LFAC)
      IF(LFAC.EQ.1) GOTO 201
      WRITE(6,136)
136 FORMAT(1H1,15X,38H RECORD OF SUCCESSIVE VARIMAX ROTATION///)
      WRITE(6,128) NFAC
128 FORMAT(1H0,17X,34H THE NUMBER OF POSITIVE FACTORS IS,I3/)
      WRITE(6,129) LFAC
129 FORMAT(1H0,17X,33H THE NUMBER OF ROTATED FACTORS IS,I3///)
133 CRITEI=0.0
      DO 75 I=1,20
      DO 73 J=1,NFAC1
      J12=J+1
      DO 73 K=J12,LFAC

```

```

DO 71 J2=1,M
H(J2)=0.0
DO 71 JP=1,LFAC
71 H(J2)=H(J2)+VEC(JP,J2)*VEC(JP,J2)
U=0.0
V=0.0
AU=0.0
BV=0.0
CW=0.0
DW=0.0
DO 72 L=1,M
VEC(J,L)=VEC(J,L)/SQRT(H(L))
VEC(K,L)=VEC(K,L)/SQRT(H(L))
U=VEC(J,L)**2-VEC(K,L)**2
V=VEC(J,L)*VEC(K,L)*2.0
AU=AU+U
BV=BV+V
CW=CW+U**2-V**2
72 DW=DW+2.0*U*V
T1=DW-2.0*AU*BV/FLOAT(M)
T2=CW-(AU**2-BV**2)/FLOAT(M)
IF(T2.EQ.0.0) GOTO 90
PHI=ATAN(ABS(T1)/ABS(T2))
IF(T1.GT.0.0.AND.T2.GT.0.0) GOTO 101
IF(T1.GT.0.0.AND.T2.LT.0.0) GOTO 102
IF(T1.LT.0.0.AND.T2.LT.0.0) GOTO 103
IF(T1.LT.0.0.AND.T2.GT.0.0) GOTO 104
101 PHI=PHI/4.0
GOTO 105
102 PHI=0.78539813-PHI/4.0
105 T12=-1.0*SIN(PHI)
T21=SIN(PHI)
GOTO 106
103 PHI=0.78539813-PHI/4.0
GOTO 107
104 PHI=PHI/4.0
107 T12=SIN(PHI)
T21=-1.0*SIN(PHI)
106 T11=COS(PHI)
T22=COS(PHI)
DO 74 L=1,M
WEC1=VEC(J,L)*T11+VEC(K,L)*T21
WEC2=VEC(J,L)*T12+VEC(K,L)*T22
VEC(J,L)=WEC1
74 VEC(K,L)=WEC2
ANGLE=PHI*57.29577867
73 CONTINUE
V1=0.0
V2=0.0
DO 76 J=1,LFAC

```

```

SUM(J)=0.0
SUMSQ(J)=0.0
DO 76 J1=1,M
SUM(J)=SUM(J)+VEC(J,J1)**2
76 SUMSQ(J)=SUMSQ(J)+VEC(J,J1)**4
DO 78 J=1,LFAC
V1=V1+SUM(J)**2
78 V2=V2+SUMSQ(J)
VARIMA=FLOAT(M)*V2-V1
CRITIC=VARIMA-CRITIC
WRITE(6,79) (I,VARIMA)
79 FORMAT(1H0,10X,15H ROTATION CYCLE,2X,I2,4X,18H VARIMAX CRITERION,
1F12.4)
IF(CRITIC.LE.0.10) GOTO 117
CRITIC=VARIMA
75 CONTINUE
117 JEN=1
WRITE(7,148)
148 FORMAT(34H ROTATED NORMALIZED FACTOR LOADING)
DO 137 K=1,M
137 WRITE(7,151) (VEC(J,K),J=1,LFAC)
151 FORMAT(10F8.4)
201 NF=0
M4=M-1
DO 62 I=1,LFAC
NFAC3=0
DO 63 J=JEN,M
IF(ABS(VEC(I,J)).GT.0.500) NFAC3=NFAC3+1
63 CONTINUE
IF(NFAC3.EQ.0) GOTO 62
DO 61 J=JEN,M4
J1=J+1
DO 61 K=J1,M
IF(ABS(VEC(I,J)).GT.ABS(VEC(I,K))) GOTO 61
DO 95 L=1,LFAC
D(L)=VEC(L,J)
VEC(L,J)=VEC(L,K)
95 VEC(L,K)=D(L)
TEMP=VARIAT(2*J-1)
TEMP1=VARIAT(2*J)
VARIAT(2*J-1)=VARIAT(2*K-1)
VARIAT(2*J)=VARIAT(2*K)
VARIAT(2*K-1)=TEMP
VARIAT(2*K)=TEMP1
61 CONTINUE
NFAC2=NFAC3-1+NF
NFAC4=NFAC2+1
DO 64 J=JEN,NFAC2
J1=J+1
DO 64 K=J1,NFAC4

```

```

      IF (VEC(I,J).GT.VEC(I,K)) GOTO 64
      DO 96 L=1,LFAC
      D(L)=VEC(L,J)
      VEC(L,J)=VEC(L,K)
96   VEC(L,K)=D(L)
      TEMP=VARIAT(2*J-1)
      TEMP1=VARIAT(2*K)
      VARIAT(2*J-1)=VARIAT(2*K-1)
      VARIAT(2*K)=VARIAT(2*J)
      VARIAT(2*K-1)=TEMP
      VARIAT(2*J)=TEMP1
64   CONTINUE
      JEN=NFAC3+1+NF
      NF=NFAC3+NF
      IF (JEN.EQ.M) GOTO 134
62   CONTINUE
134  DO 34 L=1,LFAC,15
      WRITE(6,52)
52   FORMAT(1H1,10X,74H THE ROTATED NORMALIZED FACIOR LOADINGS ARRANGED
1    IN COLUMN FOR EACH FACTOR//)
      LIMIT=L+14
      IPRINT=LFAC
      IF (LFAC.GT.LIMIT) IPRINT=LIMIT
      WRITE(6,130) (NUMBER(I),I=L,IPRINT)
130  FORMAT(1H0,7X,7H FACTOR,2X,15I7)
      DO 34 K=1,M
      IV=2*K-1
      IV1=IV+1
34   WRITE(6,50)((VARIAT(J),J=IV,IV1),(VEC(J,K),J=L,IPRINT))
50   FORMAT(1H ,4X,2A6,15F7.2)
135  FORMAT(6E12.4)
90   CONTINUE
83   STOP
      END

```