CANONICAL ANALYSIS: SOME RELATIONS BETWEEN CANONICAL CORRELATION, FACTOR ANALYSIS, DISCRIMINANT FUNCTION ANALYSIS, AND SCALING THEORY

James J. McKeon
EDUCATIONAL TESTING SERVICE

ACKNOWLEDGEMENTS

The first version of this monograph was completed at the University of Illinois in 1962 United States Public Health Service Training Grant 2M-6961. I am indebted to Professor Leydyard R Tucker who encouraged this investigation of the relations between multivariate statistics and psychometric theory. During the course of the work I benefited from discussions with Dr. Tucker.

The revised manuscript was completed at Educational Testing Service. During the period of revision I was ably assisted by my wife in the editorial work and through her constructive suggestions on the text.

JAMES J. MCKEON

CONTENTS

1. CANONICAL ANALYSIS

The theory of canonical correlation (Hotelling, 1935, 1936) was originally developed to extract the most predictable criterion composite when several criteria and several predictors are available. The theory has since proved to have other applications. Included within the more general structure of canonical analysis are several important analytic techniques among which are canonical correlation in the ordinary sense, certain types of factor analysis, discriminant function analysis, methods for the scoring of categorical data and several quantification or scaling procedures used in the social sciences. A unified treatment is possible through the use of pseudo-variates and suitably defined conditional inverses. Canonical methods have two desirable properties. The results are invariant with respect to the choice of unit of measurement for any of the variables and they are often optimal by certain accepted measures of association and reliability.

Some Basic Theory

Given n observations on each of two sets of variates $\mathbf{x}$ and $\mathbf{y}$, where

$$\mathbf{x}_\alpha = x_{i\alpha}, \quad i = 1, 2, \cdots, p \text{ variates},$$

$$\mathbf{y}_\alpha = y_{i\alpha}, \quad j = 1, 2, \cdots, q \text{ variates},$$

$$\alpha = 1, 2, \cdots, n \text{ observations},$$

coefficients $\mathbf{a}$ and $\mathbf{c}$ are to be determined such that the correlation between $\mathbf{a}'\mathbf{x}$ and $\mathbf{c}'\mathbf{y}$ is maximum. In other words, letting

$$S_{11} = \sum_\alpha (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})',$$

$$S_{22} = \sum_\alpha (\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})',$$

$$S_{12} = \sum_\alpha (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})',$$

the quantity $\mathbf{a}'S_{12}\mathbf{c}$ is to be maximized subject to the conditions $\mathbf{a}'S_{11}\mathbf{a} = \mathbf{c}'S_{22}\mathbf{c} = 1$. Introducing Lagrange multipliers ρ and η, the function

$$\psi = \mathbf{a}'S_{12}\mathbf{c} - \tfrac{1}{2}\rho(\mathbf{a}'S_{11}\mathbf{a} - 1) - \tfrac{1}{2}\eta(\mathbf{c}'S_{22}\mathbf{c} - 1)$$

must be maximized with respect to $\mathbf{a}$ and $\mathbf{c}$. Differentiation with respect to $\mathbf{a}$ and $\mathbf{c}$ gives the two sets of equations,

$$(1.1) \qquad \frac{\partial \psi}{\partial \mathbf{a}} = S_{12}\mathbf{c} - \rho S_{11}\mathbf{a} = \mathbf{0},$$

$$\frac{\partial \psi}{\partial \mathbf{c}} = S_{21}\mathbf{a} - \eta S_{22}\mathbf{c} = \mathbf{0}. \tag{1.2}$$

Multiplying (1.1) on the left by $\mathbf{a}'$ and (1.2) by $\mathbf{c}'$ shows that

$$\rho = \eta = \mathbf{a}'S_{12}\mathbf{c},$$

$$S_{12}\mathbf{c} - \rho S_{11}\mathbf{a} = \mathbf{0}, \tag{1.3}$$

$$S_{21}\mathbf{a} - \rho S_{22}\mathbf{c} = \mathbf{0}. \tag{1.4}$$

Solving (1.4) for $\mathbf{c}$ and substituting in (1.3) yields

$$(S_{12}S_{22}^{-1}S_{21} - \lambda S_{11})\mathbf{a} = \mathbf{0}, \tag{1.5}$$

with $\lambda = \rho^2$. Premultiplying by S_{11}^{-1} this becomes

$$(R - \lambda I)\mathbf{a} = \mathbf{0}. \tag{1.6}$$

The canonical weights $\mathbf{a}$ are the elements of the principal vector of the canonical correlation matrix $R = S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$. From (1.4), the canonical weights for the $\mathbf{y}$ variates are, up to an arbitrary normalization,

$$\mathbf{c} = S_{22}^{-1}S_{21}\mathbf{a}. \tag{1.7}$$

The largest root of R is the squared canonical correlation between the two sets, $\lambda = \rho^2$. Premultiplying (1.5) by $\mathbf{a}'$,

$$\lambda = \frac{\mathbf{a}'S_{12}S_{22}^{-1}S_{21}\mathbf{a}}{\mathbf{a}'S_{11}\mathbf{a}}. \tag{1.8}$$

Equations (1.3) through (1.8) furnish the basic computational equations for all types of canonical analysis. Because of the symmetry between $\mathbf{x}$ and $\mathbf{y}$, there are a parallel set of equations in terms of $\mathbf{c}$ corresponding to (1.3) through (1.8). Any two canonical variates whether from the same or different sets are uncorrelated unless they correspond to the same root.

Pseudo-Variates and Conditional Inverses

One or both of the sets of variates $\mathbf{x}$ and $\mathbf{y}$ may consist of pseudo-variates, 0 or 1 scores indicating class membership. For example, in a one-way analysis of variance layout with p measurements on each of k groups, we would have

$$\mathbf{x}_\alpha = x_{i\alpha}, \quad i = 1, 2, \cdots, p \text{ variates},$$

$$\mathbf{y}_\alpha = y_{j\alpha}, \quad j = 1, 2, \cdots, k \text{ pseudo-variates (groups)},$$

where $y_{j\alpha} = 1$, if the αth observation belongs to the jth group, $y_{j\alpha} = 0$ otherwise. For categorical data both $\mathbf{x}$ and $\mathbf{y}$ are pseudo-variates. It should be noted that a set of pseudo-variates is not the same as a set of dichotomized variables. In the former all elements of the vector variate except one must be zero, while in the latter there are no restrictions on the number of ones and zeros. The sum of products (SP) matrix for a set of pseudo-variates is always singular while for a set of dichotomized variables the sum of products matrix will be nonsingular and may be handled as a set of continuous variables.

In the above example the SP matrix for the second (**y**) set of variates is

$$S_{22} = D_2 - \frac{1}{n}\mathbf{n}_2\mathbf{n}_2' , \tag{1.9}$$

where D_2 is a diagonal matrix and $\mathbf{n}_2$ is a vector, both with elements n_j , which is the number of observations in the jth group. Since $S_{22}\mathbf{1} = \mathbf{0}$, the matrix S_{22} is singular and cannot be inverted in the usual sense. This difficulty can be eliminated and symmetry preserved by defining a unique conditional inverse of S_{22} .

Let E be an idempotent matrix so that $E^2 = E$, where E is generally not of full rank. A vector **v** is in the space of E if $E\mathbf{v} = \mathbf{v}$. The matrix A will be considered in E if its row and column vectors are in E, that is, if $EA = AE = A$. If A and A^{-1*} are matrices in the space of E, then A^{-1*} is the conditional inverse of A if $AA^{-1*} = A^{-1*}A = E$. In general, for any square matrix A,

$$A^{-1*} = E[A + \tau(I - E)]^{-1}E \tag{1.10}$$

where $\tau \neq 0$ but is otherwise arbitrary. The matrix in brackets is nonsingular. The conditional inverse of A is made unique by inverting A within its own space. Equation (1.10) can be obtained by pre- and postmultiplying the identity

$$[A + \tau(I - E)][A + \tau(I - E)]^{-1} = I,$$

by E. If A is factored into its roots and vectors so that

$$A_{p\times p} = R_{p\times r}D_{r\times r}L'_{r\times p} , \qquad r \leq p,$$

then

$$E = RL' \quad \text{an} \quad A^{-1*} = RD^{-1}L'.$$

The idempotent E can be determined from the restrictions on A. If the columns of A satisfy the r restrictions $U'A = 0$ and the rows satisfy $AV = 0'$, and if $U'V = P$, then

$$E = I - VP^{-1}U'. \tag{1.11}$$

For the matrix S_{22} , the restrictions are $\mathbf{1}'S_{22} = \mathbf{0}'$ and $S_{22}\mathbf{1} = \mathbf{0}$, so that

$$E = J = I - \frac{1}{k}\,\mathbf{1}\mathbf{1}'. \tag{1.12}$$

This approach is convenient since the conditional inverse of S_{22} is simply

$$S_{22}^{-1*} = JD_2^{-1}J. \tag{1.13}$$

This can be verified directly, $JD_2^{-1}JS_{22} = JD_2^{-1}S_{22} = J$. Whenever $\mathbf{x}$ or $\mathbf{y}$ consist of pseudo-variates and conditional inverses are substituted for the ordinary inverses, the basic equations of canonical analysis, (1.3) through (1.8), remain valid. It should be noted that if $\mathbf{y}$ consists of pseudo-variates, then $\mathbf{1}'\mathbf{c} = 0$. It will often be convenient to adjust the category scores so that the score distribution has zero mean. The adjusted score vector $\tilde{\mathbf{c}}$ differs from $\mathbf{c}$ by a translation and satisfies the condition $\mathbf{n}_2'\tilde{\mathbf{c}} = 0$. Defining the idempotent matrix

$$Q_2 = \underline{I} - \frac{1}{n}\mathbf{1}\mathbf{n}_2' , \tag{1.14}$$

the following relations hold between J, Q_2, $\mathbf{c}$, and $\tilde{\mathbf{c}}$:

$$\begin{aligned} \tilde{\mathbf{c}} &= Q_2\mathbf{c}, & \mathbf{c} &= J\tilde{\mathbf{c}}, \\ Q_2J &= Q_2 , & JQ_2 &= J. \end{aligned} \tag{1.15}$$

Similar relations hold between the quantities J, Q_1, $\mathbf{a}$, and $\tilde{\mathbf{a}}$.

Multiple Measurements on Several Groups

For a set of p measurements on each of k groups the SP matrices are

$$S_{11} = \sum (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})',$$

$$S_{22} = D_2 - \frac{1}{n}\mathbf{n}_2\mathbf{n}_2' ,$$

$$S_{12} = GD_2,$$

where $\mathbf{1}'S_{21} = \mathbf{1}'S_{22} = \mathbf{0}'$ and $G = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$. Then the canonical weights for the $\mathbf{x}$ set satisfy

$$(S_{12}S_{22}^{-1*}S_{21} - \lambda S_{11})\mathbf{a} = \mathbf{0},$$

or

$$(GD_2G' - \lambda S_{11})\mathbf{a} = \mathbf{0}.$$

But $GD_2G' = B$, the between SP matrix, and $S_{11} = T$, the total SP matrix, so that the above is equivalent to

$$(B - \lambda T)\mathbf{a} = \mathbf{0}. \tag{1.16}$$

This is the equation for the discriminant function weights obtained by maximizing

$$\lambda = \frac{\mathbf{a}'B\mathbf{a}}{\mathbf{a}'T\mathbf{a}} ,$$

the ratio of the between to the total sum of squares or the correlation ratio.

Letting $W = T - B$, we obtain the alternative form

$$(B - \mu W)\mathbf{a} = \mathbf{0}, \tag{1.17}$$

where $\mu = \lambda/(1 - \lambda)$ is the largest root of $W^{-1}B$. The weights $\mathbf{c}$ are given by

$$\mathbf{c} = S_{22}^{-1*}S_{21}\mathbf{a} = JG\mathbf{a}.$$

Premultiplying by

$$Q_2 = I - \frac{1}{n}\mathbf{1}\mathbf{n}_2'$$

yields

$$\tilde{\mathbf{c}} = G\mathbf{a}, \tag{1.18}$$

a somewhat simpler expression. The adjusted weights $\tilde{\mathbf{c}}$ are the means of the canonical variates for the k groups. The weights for a set of pseudo-variates may be given the alternative interpretation of scores assigned to the categories in such a way as to maximize λ. This interpretation will generally be more meaningful and the vector $\tilde{\mathbf{c}}$ will generally be a more convenient form for these scores than $\mathbf{c}$. The tilde notation serves to indicate that the weights are being interpretated as scores as well as the change in origin.

Categorical Data

For categorical data, both $\mathbf{x}$ and $\mathbf{y}$ are sets of pseudo-variates and the SP matrices are

$$S_{11} = D_1 - \frac{1}{n}\mathbf{n}_1\mathbf{n}_1',$$

$$S_{22} = D_2 - \frac{1}{n}\mathbf{n}_2\mathbf{n}_2',$$

$$S_{12} = N_{12} - \frac{1}{n}\mathbf{n}_1\mathbf{n}_2',$$

where N_{12} is the matrix of cell frequencies, n_{ij}, and $\mathbf{n}_1$, $\mathbf{n}_2$ are the marginal frequencies. Then

$$(S_{11}^{-1*}S_{12}S_{22}^{-1*}S_{21} - \lambda I)\mathbf{a} = \mathbf{0},$$

which reduces to

$$(JD_1^{-1}N_{12}D_2^{-1}N_{21} - \lambda I)\mathbf{a} = \mathbf{0}. \tag{1.19}$$

Premultiplying by $Q_1 = I - (1/n)\mathbf{1}\mathbf{n}_1'$,

$$\left[\left(D_1^{-1}N_{12}D_2^{-1}N_{21} - \frac{1}{n}\mathbf{1}\mathbf{n}_1'\right) - \lambda I\right]\tilde{\mathbf{a}} = 0. \tag{1.20}$$

The conjugate weights are

$$\mathbf{c} = S_{22}^{-1*}S_{21}\mathbf{a} = JD_2^{-1}N_{21}\mathbf{a}. \tag{1.21}$$

Premultiplying by Q_2 and putting $\mathbf{a} = J\tilde{\mathbf{a}}$ gives the score vector,

$$\tilde{\mathbf{c}} = D_2^{-1}N_{21}\tilde{\mathbf{a}}. \tag{1.22}$$

If the scores $\tilde{\mathbf{a}}$ are assigned to the first set of categories, then $\tilde{\mathbf{c}}$ is the vector of mean scores for the second set of categories.

2. CANONICAL FACTOR ANALYSIS

A principal distinguishing feature of different applications of canonical analysis is the normalization requirements imposed upon the canonical weights. Every such normalization implies a factorization of some matrix. The following theorem is of fundamental importance.

A Factoring Theorem

> If T and B are symmetric of order p, where all roots of T are positive and all roots of B are non-negative, then there is a square nonsingular matrix L and a diagonal matrix D such that $T = LL'$ and $B = LDL'$.

The matrices L and D may be identified by noting that

$$L'T^{-1}B = L'(LL')^{-1}LDL' = DL', \tag{2.1}$$

showing that the elements of D are the roots of $T^{-1}B$ (including zero roots) and the columns of L are the left-hand vectors of $T^{-1}B$ satisfying

$$l(T^{-1}B - \lambda I) = \mathbf{0}. \tag{2.2}$$

Note that for L to be of full rank there must be vectors corresponding to zero roots. Let L be the matrix of left-hand vectors. It may be readily verified that $A = T^{-1}L$ is the matrix of right-hand vectors. It follows that

$$L = TA. \tag{2.3}$$

The requirement that $T = LL'$ implies the normalization

$$L'T^{-1}L = I,$$

or equivalently,

$$A'TA = I. \tag{2.4}$$

Similarly, if A is normalized so that $A'BA = I$, then $L = BA$ provides a factorization of B. The following more general corollary may be established.

For any matrix $V = c_1T + c_2B$, where c_1 and c_2 are arbitrary constants subject only to the conditions that all roots of V are non-negative and if A is the matrix of right-hand vectors of $T^{-1}B$ normalized so that $A'VA = I$, then $L = VA$ provides a factorization of V.

Identification of Matching Factor Pairs

If A is the matrix of canonical weights for the first set of variates, normalized so that $A'S_{11}A = I$, then from the factoring theorem $L = S_{11}A$ is a factor matrix for S_{11}. The matrix C of weights for the second set of variates may be normalized, independently of A, so that $C'S_{22}C = I$. Then $K = S_{22}C$ is a factor matrix for S_{22}. This procedure has been suggested by Bartlett (1948) as a technique for matching the underlying factors for two sets of variates. The paired factor variables are the canonical variates. In effect, the factor matrices for both sets of variates are simultaneously rotated by orthogonal transformations until a factor from one set is maximally correlated with a factor from the other set, identifying the first factor pair. These factors are held fixed while the second pair is identified, and so forth. This method of identifying matching factor pairs has been applied in Appendix A, to data of K. J. Jones (1965). Factors underlying occupational classifications are related to factors of the Guilford-Zimmerman Temperament Survey.

A factorization of the S_{12} matrix is also implied by this procedure. From (1.3)

$$S_{12}C = S_{12}AD_\rho. \tag{2.5}$$

Assuming the rank of C is n_2 and postmultiplying by K' yields

$$S_{12} = LD_\rho K'. \tag{2.6}$$

This factorization remains valid when the rank of C is less than n_2.

Relation to Maximum-Likelihood Estimation in Common-Factor Analysis

The following connection between canonical analysis and maximum-likelihood estimation is due to Rao (1955). The common-factor analysis model assumes that the observed vector variate $\mathbf{x}$ is the sum of a specific part $\mathbf{s}$ and a common part $\mathbf{g}$ where the specific parts are uncorrelated with each other and with the common parts, that is,

$$\mathbf{x} = \mathbf{s} + \mathbf{g} \tag{2.7}$$

and

$$\mathcal{E}\begin{bmatrix}\mathbf{s}\\ \mathbf{g}\end{bmatrix}\begin{bmatrix}\mathbf{s}\\ \mathbf{g}\end{bmatrix}' = \begin{bmatrix}U & 0\\ 0 & \Theta\end{bmatrix}.$$

It follows that

$$\mathcal{E}S_x = U + \Theta. \tag{2.8}$$

A canonical analysis between the sets **x** and **g** will determine the linear composites **a**′**x** and **c**′**g** with maximum correlation. The SP matrices are

$$S_{11} = S, \qquad S_{22} = S_{12} = S - U.$$

Then the weights **a** satisfy

$$[S^{-1}(S - U) - \lambda I]\mathbf{a} = \mathbf{0},$$

or the equivalent form

$$(U^{-1}S - \eta I)\mathbf{a} = \mathbf{0},$$

with $\eta = 1/(1 - \lambda)$. For r such components the set of left-hand vectors L will satisfy

$$L'U^{-1}S = D_\eta L'. \tag{2.9}$$

For arbitrary U, these equations may be solved for L and by suitable normalization a factorization, $\Theta \cong LL'$ is possible. A unique U and L may be determined by requiring that the diagonal elements satisfy

$$\text{diag}\,(S) = U + \text{diag}\,(LL'). \tag{2.10}$$

These are the maximum-likelihood estimates of U and L (Lawley, 1940). A solution may be obtained by iterating between (2.9) and (2.10), provided the process converges. The factor scores **c**′**g** are unknown because the common parts **g** are undetermined. The best linear function of the observed variables **x** for estimating the factor scores are the variates **a**′**x** since they correlate most highly with the factor variables. The results are independent of the choice of units and therefore it does not matter whether the SP matrix, the covariance matrix, or the correlation matrix is used.

3. DISCRIMINANT FUNCTION ANALYSIS AND SCALING THEORY

Discriminant function analysis, introduced by Fisher (1936) as a classification procedure, supplies a theoretical basis for certain scaling procedures in the social and biological sciences (Rao, 1948). Most scaling methods amount to the location of a set of groups or objects within a Euclidean metric space given some direct or inferred measure of the intergroup distances. The fundamental assumption underlying this approach to scaling is that the distance between two groups is given by Mahalanobis' D^2. If a set of measurements $\mathbf{x}_j$, $j = 1, 2, \cdots, k$ groups, have means τ_j and a common covariance matrix Σ, the then squared distance between the groups j and h is

$$D_{jh}^2 = (\tau_j - \tau_h)'\Sigma^{-1}(\tau_j - \tau_h). \tag{3.1}$$

The sample estimate of D^2_{jh} is proportional to

$$d^2_{jh} = (\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_h)'W^{-1}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_h), \tag{3.2}$$

where W is the combined within-groups SP matrix. The coordinates of the groups for a set of orthogonal axes within a space of metric W, constitute a multidimensional representation of the intergroup configuration, or in psychometric terminology, a multidimensional scale for the groups. It follows that any factor matrix U of W provides a multidimensional scale with coordinates for the jth group $\mathbf{z}_j = U^{-1}\bar{\mathbf{x}}_j$ and intergroup distances

$$d^2_{jh} = (\mathbf{z}_j - \mathbf{z}_h)'(\mathbf{z}_j - \mathbf{z}_h) = \sum_r (z_{jr} - z_{hr})^2. \tag{3.3}$$

If the canonical weights A, which satisfy

$$(B - \mu W)\mathbf{a} = \mathbf{0},$$

are normalized so that $A'WA = I$, implying the normalization

$$C'S_{22}C = \tilde{C}'D_2\tilde{C} = D_\mu\,, \qquad \mu = \frac{\lambda}{1 - \lambda}\,, \tag{3.4}$$

then both C and $\tilde{C}$ are a set of coordinates for the groups on a multidimensional scale with

$$d^2_{jh} = \sum_r (c_{jr} - c_{hr})^2 = \sum_r (\tilde{c}_{jr} - \tilde{c}_{hr})^2. \tag{3.5}$$

These are the principal scale components since they successively account for the largest amount of discrimination between the groups. Like any orthogonal factor matrix, the coordinate matrix C may be rotated to an orthogonal simple structure.

Discriminant analysis was applied by Jones and Bock (1960) to ratings of six cultural groups on "ways to live" using data collected by Morris (Morris and Jones, 1955). Their results are given in Appendix B.

Rather than inferring the intergroup distances from measurements on the relevant variables, a multidimensional scale may be constructed from direct measurements of the intergroup distances. The latter approach has been taken in the development of multidimensional scaling methods in psychology (Richardson, 1938; Young and Householder, 1938; Torgerson, 1952). In distance observation methods the intergroup distances, d_{jh}, are determined by subjective estimation. The principal scale components are determined from a factorization of the matrix M with elements

$$m_{jh} = -\tfrac{1}{2}(d^2_{jh} - \bar{d}^2_{j.} - \bar{d}^2_{.h} + \bar{d}^2_{..}). \tag{3.6}$$

Under the assumption that there are a set of latent variables $\mathbf{x}$ such that

$$d^2_{jh} = (\mathbf{x}_j - \mathbf{x}_h)'W^{-1}(\mathbf{x}_j - \mathbf{x}_h),$$

the two types of scales are identical. To demonstrate this, let $W = UU'$ and $\mathbf{z}_j = U^{-1}\mathbf{x}_j$. The type of factorization is arbitrary. Then

$$d_{jh}^2 = (\mathbf{z}_j - \mathbf{z}_h)'(\mathbf{z}_j - \mathbf{z}_h),$$

and after some straightforward computation we find that

$$m_{jh} = (\mathbf{z}_j - \bar{\mathbf{z}})'(\mathbf{z}_h - \mathbf{z}) = (\mathbf{x}_j - \bar{\mathbf{x}})'W^{-1}(\mathbf{x}_h - \bar{\mathbf{x}})$$

or

$$M = G'W^{-1}G. \tag{3.7}$$

The group sizes are effectively equal so that $\mathbf{c} = \tilde{\mathbf{c}}$ and we can take $B = GG'$. The two scales are identical if the vectors $\mathbf{c}_r = \tilde{\mathbf{c}}_r = G'\mathbf{a}_r$ are the principal vectors of M. From a well-known theorem on characteristic roots

$$\text{ch}\,(G'W^{-1}G) = \text{ch}\,(W^{-1}GG') = \text{ch}\,(W^{-1}B)$$

except for zero roots, and

$$(G'W^{-1}G - \mu I)\mathbf{c} = G'(W^{-1}B - \mu I)\mathbf{a} = \mathbf{0},$$

so that the vectors $\mathbf{c}_r$ are, in fact, the principal components of M. If it is assumed that there is only one underlying variable or dimension then

$$D_{jh}^2 = \frac{(\tau_j - \tau_h)^2}{\sigma^2}$$

or

$$D_{jh} = \frac{\tau_j - \tau_h}{\sigma}. \tag{3.8}$$

Equation (3.8) together with the assumption that $D_{jh} = \text{normit}(P_{jh})$ constitutes Thurstone's (1927) Case 5 law of comparative judgment, where P_{jh} is the proportion of times object j is preferred to object h. When a direct estimate of the signed distances, D_{jh}, is available, rather than extracting the principal component of the M matrix, a least-squares solution for $\alpha_j = \tau_j/\sigma$ in the model

$$d_{jh} = \alpha_j - \alpha_h + \epsilon_{jh} \tag{3.9}$$

may be substituted.

Making use of conditional inverses where necessary, the computational procedures of this section may be applied to categorical data but the model of Section 4, which assumes an underlying bivariate normal distribution, will often be more appropriate.

Other Uses of Canonical Analysis Related to Discrimination

The notion of a linear discriminant function between k groups may be extended to more general formulations and methods. Given a measurement x and a set of pseudo-variates y_i corresponding to k groups, then for the two sets, $(x, x^2) : y_k$, the first canonical variate gives the best quadratic function of x for discriminating between the groups. More elaborate discriminant functions are easily determined in this manner.

The discrimination situation may be extended in another direction. Consider a two-way analysis of variance layout with a set of measurements $\mathbf{x}$ and pseudo-variates y_i and z_k corresponding to the two ways of classification. Then the two sets of variates $\mathbf{x} : (y_i \ , \ z_k)$ form the extension to the two-way classification. The SP matrix of the right-hand set lies within a space whose idempotent is

$$E = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}. \tag{3.10}$$

The canonical weights for the pseudo-variates will be the corresponding main effects.

The two preceding generalizations may be combined to obtain the best quadratic transformation for the elimination of interaction in a two-way design. The first canonical variate for the two sets $(x, x^2) : (y_i \ , \ z_k)$ is the required transformation on x.

4. SCORING CATEGORICAL DATA

The use of canonical scoring for categorical data was introduced by Fisher (1938) and important contributions to the theory and applications have been made by Guttman (1941, 1950) and Lancaster (1957, 1958). Without assumptions as to the nature of the underlying distribution, canonical scoring can be justified on the grounds that it maximizes discrimination and association between the ways of classification. A stronger and more directly interpretable model is based on the assumption of an underlying bivariate normal distribution. Under this assumption, the canonical scores provides estimates of the category locations on the two underlying variables and the canonical correlation estimates the correlation coefficient of the distribution. To demonstrate this it will be necessary to obtain an expansion of the joint frequency matrix in terms of the score vectors $\tilde{\mathbf{a}}_g$ and $\tilde{\mathbf{c}}_g$.

From (2.6), S_{12} can be factored into the form,

$$S_{12} = LD_\rho K'.$$

Making the substitutions

$$S_{12} = N - \frac{1}{n}\mathbf{n}_1\mathbf{n}_2' ,$$

$$L = S_{11}A = S_{11}(J\tilde{A}) = D_1A,$$

$$K = S_{22}C = S_{22}(J\tilde{C}) = D_2\tilde{C},$$

gives the identity

$$N_{12} - \frac{1}{n}\mathbf{n}_1\mathbf{n}_2' = D_1\tilde{A}D_\rho\tilde{C}'D_2 . \tag{4.1}$$

This may be written in the form

$$\frac{n_{ij}}{n} = \left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right)\left(1 + n \sum_{g=1}^{r} \rho_g \tilde{\mathbf{a}}_g \tilde{\mathbf{c}}_g'\right). \tag{4.2}$$

The bivariate normal distribution may be expanded in the tetrachoric series or Mehler identity,

$$f(x, y) = f(x)f(y)\left[1 + \sum_{g=1}^{\infty} \frac{\rho_0^g}{g!} H_g(x)H_g(y)\right], \tag{4.3}$$

where ρ_0 is the correlation in the distribution and $H_g(x)$ and $H_g(y)$ are Hermite polynomials, $H_0 = 1$, $H_1 = x$, $H_2 = x^2 - 1$, etc. Assuming that the frequencies n_{ij} result from an underlying bivariate normal distribution, as the number of categories becomes larger (4.2) approaches (4.3) (Lancaster, 1957). Therefore, the following approximate identifications may be made between the discrete and continuous distributions:

$$\mathbf{a}_g \rightarrow H_g(x)\frac{1}{\sqrt{ng!}}, \qquad \mathbf{c}_g \rightarrow H_g(y)\frac{1}{\sqrt{ng!}}, \tag{4.4}$$

$$\rho_g \rightarrow \rho_0^g , \qquad \lambda_g \rightarrow \lambda_0^g , \tag{4.5}$$

$$\operatorname{tr} R \rightarrow \sum_{g=1}^{r} \lambda_0^g = \frac{\lambda_0(1 - \lambda_0^r)}{1 - \lambda_0}, \tag{4.6}$$

where $r = \operatorname{rank} R = \min(k, l) - 1$. From the first component which identifies x and y, the higher components may be computed from the Hermite polynomials of x and y. Correspondence (4.6) allows a prediction of the largest root of R and so of all the roots, from the trace of R. The largest root will generally amount to only a small fraction of the total trace and the adequacy or goodness of fit of the assigned scores cannot be evaluated by the proportion of the trace removed. An example of canonical scoring for ratings of social class (Hollingshead 1949) is given in Appendix C.

In the previous discussion, the order among the categories was disregarded or assumed unknown but often prior knowledge of the category order exists. The categories may be ordered in one direction or in both directions. If the canonical scores conflict with this prior order, three alternatives are available.

1. The prior order can be abandoned in favor of the order of the canonical scores.
2. The correlation between the categories may be maximized subject to the order constraints on the scores. This amounts to a problem in nonlinear programming on the score differences. This type of solution requires access to an electronic computer.
3. The categories can be scored from the marginal distributions. If the underlying distribution is bivariate normal, then the marginal distributions are normal and the category mean estimated from Pearson's centroid formula may be used as the category score. The marginal scores may be regarded as approximations to the canonical scores since they converge to the same values as the number of categories becomes indefinitely large. The marginal scoring method was introduced by K. Pearson (1913) and has been in use for many years.

When the order of the categories in only one direction is known, as with "successive categories" data, marginal scores can be assigned only to those ordered categories. Scores for the other set may be obtained from what is in a sense one-half a canonical analysis, the mean category score. This method for successive categories data is computationally much simpler than Thurstone's "successive intervals" solution and appears to give results about as good provided the distribution of object scale values is approximately normal. Scoring the categories rather than the boundaries between the categories makes available simply computed estimates of the efficiency and reliability of the discrimination. Thurstone's method is compatible with the bivariate normal assumption. It amounts to assuming the normality of the conditional distribution $f(x \mid y_j)$, $j = 1, 2, \cdots, l$ objects. This is less restrictive since it does not require the normality of $f(y)$. Thurstone's method will therefore be appropriate even when the distribution of object scale values is seriously non-normal.

5. ASSOCIATION AND RELIABILITY

To include both continuous and pseudo-variates in the same formulation and for simplicity of presentation let

$$p = \text{rank } S_{11}, \qquad q = \text{rank } S_{22},$$

$$r = \text{rank } R = \min(p, q), \quad t = n - 1.$$

Association

For the regression of a variate x on a set of variates $\mathbf{y}$, the squared multiple correlation, $\rho^2(x, \mathbf{y})$ is the proportion of total variation accounted for by the regression and is essentially (except for the bias in small samples) independent of the sample size. The multiple correlation is, therefore, a measure of the efficiency of discrimination or prediction. When $\mathbf{y}$ is a set of pseudo-variates, then $\rho(x, \mathbf{y})$, the square root of the correlation ratio, is a measure of the efficiency of the discrimination between the k groups.

For two sets of variates, the canonical correlation, $\rho(\mathbf{x}, \mathbf{y})$, is one type of generalized association measure. But this depends only on the largest root of R. An alternative measure depending on all the roots is the root mean square canonical correlation, $\breve{\rho}(\mathbf{x}, \mathbf{y}) = (\bar{\lambda})^{1/2}$, where $\bar{\lambda} = (1/r)\ \mathrm{tr}\ R$. Cramér (1946) has suggested using $\breve{\rho}$ as a measure of association in categorical data. When the bivariate normal assumption is valid a more appropriate measure of association can be obtained from (4.6). At least approximately, the trace of R and its largest root are related by

$$\mathrm{tr}\, R = \sum_{g=1}^{r} \lambda_0^g = \frac{\lambda_0(1 - \lambda_0^r)}{1 - \lambda_0}.$$

This equation may be solved for λ_0 by a simple iterative technique. Take $\lambda_{0,0} = 0$ and

$$\lambda_{0,i+1} = \frac{\mathrm{tr}\, R}{\mathrm{tr}\, R + 1 - \lambda_{0,i}^r},$$

continuing until the change in λ_0 is sufficiently small. Then $\rho_0 = \lambda_0^{\frac{1}{2}}$ is an estimate of the correlation in the underlying population and is a measure of the association in the two-way table. For categorical data tr $R = \chi^2/n$ (see Section 8), so that ρ_0 is a monotonic function of the chi square for independence. The upper limit of ρ_0 is always one. From an application of L'Hôpital's rule,

$$\lim_{\lambda_0 \to 1} \frac{\lambda_0(1 - \lambda_0^r)}{1 - \lambda_0} = r, \tag{5.1}$$

which is the maximum value attainable by tr R. Pearson's contingency coefficient is a lower bound to ρ_0 and its limiting value as the number of categories in both directions becomes indefinitely large. For $\lambda < 1$, as $r \to \infty$, $\lambda_0^r \to 0$, and

$$\lambda_0 \to \frac{\mathrm{tr}\, R}{1 + \mathrm{tr}\, R} = \frac{\chi^2}{n + \chi^2}. \tag{5.2}$$

The often-mentioned inability of the contingency coefficient to reach the proper maximum of one is due to its asymptotic character. For a $k \times 2$ table, ρ_0 reduces to the phi-coefficient.

Reliability

The concept of reliability as used in psychological test theory, is a general term encompassing several specific types of reliability measures (Cronbach, 1947). These distinctions will not be made here and we will define a single reliability measure whose specific interpretation depends upon the nature of the replications.

Given the multiple regression situation with a single criterion variable and s predictors, the reliability is defined as the proportion of the estimation variance that is not error. Let

$$v_B = \frac{1}{s} S'_{21} S_{22}^{-1} S_{21} = \frac{1}{s}\,\text{SSB}$$

$$v_W = \frac{1}{t - s}(\text{SS}x - \text{SSB}),$$

then the reliability is estimated by

$$\psi = \frac{v_B - v_W}{v_B} = 1 - \frac{1}{F}, \tag{5.3}$$

where F is the variance ratio. Expressed as a function of λ

$$\psi = 1 - \frac{s}{t - s}\left(\frac{1 - \lambda}{\lambda}\right). \tag{5.4}$$

The quantity ψ estimates the correlation between the predicted values for two replications of the x variable. Since ψ is a monotonic function of λ, any scale constructed from the canonical approach has maximum reliability.

As an example of the use of these measures of association and reliability consider the ratings of l objects by m judges where there are $k < m$ rating categories. The judges are regarded as replications or parallel measurements so that $s = l - 1$ and $n = ml$. For arbitrary (not necessarily canonical) scores $\mathbf{a}$,

$$\lambda = \frac{\mathbf{a}' S_{12} S_{22}^{-1*} S_{21} \mathbf{a}}{\mathbf{a}' S_{11} \mathbf{a}} = \frac{\tilde{\mathbf{a}}' N_{12} D_2^{-1} N_{21} \tilde{\mathbf{a}}}{\tilde{\mathbf{a}}' D_1 \tilde{\mathbf{a}}}.$$

This expression may be simplified somewhat by noting that $D_2 = mI$ and $\tilde{\mathbf{c}} = (1/m)N_{21}\tilde{\mathbf{a}}$ are the object scale values, so that

$$\lambda = m \frac{\tilde{\mathbf{c}}'\tilde{\mathbf{c}}}{\tilde{\mathbf{a}}' D_1 \tilde{\mathbf{a}}}.$$

Then $\rho = \lambda^{1/2}$ is a measure of the efficiency of the discrimination and

$$\psi = 1 - \frac{s}{t - s}\left(\frac{1 - \lambda}{\lambda}\right)$$

is an index of scale reliability.

The quantity

$$\psi_0 = 1 - \frac{s}{t - s}\left(\frac{1 - \lambda_0}{\lambda_0}\right),$$

with λ_0 estimated from tr R, estimates what the reliability would be if canonical scores were assigned to the categories.

6. GENERALIZED CANONICAL ANALYSIS

Intra-Class Correlation and Generalized Association

As a starting point for the extension of canonical correlation to m sets of variates consider first the intra-class correlation r_I' between two measurements x and y based on a paired sample of size n,

$$r_I' = \frac{\sum_\alpha (x_\alpha - M)(y_\alpha - M)}{\frac{1}{2}[\sum_\alpha (x_\alpha - M)^2 + \sum_\alpha (y_\alpha - M)^2]}, \tag{6.1}$$

$$M = \frac{1}{2n}(\sum_\alpha x_\alpha + \sum_\alpha y_\alpha).$$

A generalized measure of association can be developed from the intra-class correlation of the deviation scores of x and y,

$$r_I(x, y) = \frac{\sum_\alpha (x_\alpha - \bar{x})(y_\alpha - \bar{y})}{\frac{1}{2}[\sum_\alpha (x_\alpha - \bar{x})^2 + \sum_\alpha (y_\alpha - \bar{y})^2]}. \tag{6.2}$$

The expression (6.2) will be referred to as the intra-class correlation between x and y even though it differs somewhat from (6.1).

The modified intra-class correlation between m variates is

$$r_I(x_1, x_2, \cdots, x_m) = \frac{\frac{2}{m(m-1)} \sum_{i<j} \text{SP}(x_i, x_j)}{\frac{1}{m} \sum_i \text{SS}x_i} \tag{6.3}$$

or the average sum of crossproducts over the average sum of squares. While the intra-class correlation is a measure of the association between m variates it is not a generalization of the product moment correlation coefficient since it does not reduce to the latter when $m = 2$. However, the maximum value of

$$r_I(ax, cy) = \frac{ac\,\text{SP}(x, y)}{\frac{1}{2}(a^2\,\text{SS}x + c^2\,\text{SS}y)}$$

with respect to variation of a and c is the product moment correlation between x and y, $\rho(x, y)$. For an arbitrary set of weights a_i,

$$r_I(a_1x_1\ , a_2x_2\ , \cdots\ , a_mx_m) = \frac{2}{m-1}\left[\frac{\sum_{i<j} a_ia_j \text{ SP } (x_i\ , x_j)}{\sum_i a_i^2 \text{ SS}x_i}\right]$$

$$= \frac{1}{m-1}\left[\frac{\sum_{ij} a_ia_j \text{ SP } (x_i\ , x_j)}{\sum_i a_i^2 \text{ SS}x_i} - 1\right].$$

A generalized product moment correlation among m variates may be defined as

$$\rho(x_1\ , x_2\ , \cdots\ , x_m) = \max r_I(a_1x_1\ , a_2x_2\ , \cdots\ , a_mx_m). \tag{6.4}$$

Generalized Canonical Correlation

It is now possible to define the generalized canonical correlation for m sets of variates as the maximum value of the generalized product moment correlation for the m linear composites $\mathbf{a}_i'\mathbf{x}_i$, $i = 1, 2, \cdots, m$, with respect to variation of the $\mathbf{a}_i$. Then

$$\begin{aligned} \rho(\mathbf{x}_1\ , \mathbf{x}_2\ , \cdots\ , \mathbf{x}_m) &= \max r_I(\mathbf{a}_1'\mathbf{x}_1\ , \mathbf{a}_2'\mathbf{x}_2\ , \cdots\ , \mathbf{a}_m'\mathbf{x}_m) \\ &= \max \frac{1}{m-1}\left[\frac{\sum_{ij} \mathbf{a}_i'S_{ij}\mathbf{a}_j}{\sum_i \mathbf{a}_i'S_{ii}\mathbf{a}_i} - 1\right]. \end{aligned} \tag{6.5}$$

This is equivalent to maximizing the quantity

$$\gamma = \frac{\sum_{ij} \mathbf{a}_i'S_{ij}\mathbf{a}_j}{\sum_i \mathbf{a}_i'S_{ii}\mathbf{a}_i} = (m-1)r_I + 1. \tag{6.6}$$

Let S be the sum of products matrix for the m sets combined and let S_d be a diagonal super-matrix with elements S_{ii} . Horst's term "super-matrix" will be used to refer to a matrix whose elements are matrices. Let $\mathbf{a}' = [\mathbf{a}_1'\ \mathbf{a}_2'\ \cdots\ \mathbf{a}_m']$ be the vector of combined weights, then

$$\gamma = \frac{\mathbf{a}'S\mathbf{a}}{\mathbf{a}'S_d\mathbf{a}}.$$

It follows that at the maximum γ and $\mathbf{a}$ satisfy

$$(S_d^{-1}S - \gamma I)\mathbf{a} = \mathbf{0}. \tag{6.7}$$

There is a single set of orthogonal canonical variates $\mathbf{a}_g'\mathbf{x}$, one for each root of $S_d^{-1}S$.

In dealing with the problem of how best to weight m similar measurements in the absence of a criterion, several authors (Horst, 1936; Edgerton and Kolbe, 1936; Wilks, 1938a; Lord, 1958) from different starting points have arrived at the same solution, namely, weighting each standardized variable by its loading on the principal axis of the correlation matrix. These are the generalized canonical weights for a single variable per set.

Alternative Generalizations

The extension of canonical correlation to more than two sets of variates has been based upon the generalized association measure,

$$r_I = \frac{2}{m-1} \frac{\sum_{i<j} \sigma_{ij}}{\sum_i \sigma_i^2} = \frac{\sigma_t^2 - \sum_i \sigma_i^2}{(m-1) \sum_i \sigma_i^2}. \tag{6.8}$$

In the above expression the σ_{ij} are the covariances of the variates and σ_t^2 the variance of their sum. This approach has the advantage of solutions which are the roots and vectors of a matrix but it is not the only possible generalization. For example, Horst (1961) suggests four generalized canonical correlations, one of which is equivalent to maximizing r_I. The other three definitions require more complex computational procedures and will not be given here. Still another generalization is possible from a maximization of

$$h = \frac{\sum_{i<j} \sigma_{ij}}{\sum_{i<j} \sigma_i \sigma_j} = \frac{\sigma_t^2 - \sum_i \sigma_i^2}{(\sum_i \sigma_i)^2 - \sum_i \sigma_i^2}. \tag{6.9}$$

The maximization problem for h has a solution in terms of roots and vectors for the case of a single variate per set but not otherwise. This association measure is closely related to Loevinger's coefficient of homogeneity discussed in Section 7. From the inequality

$$\sum_i \sigma_i^2 - \frac{1}{m} (\sum_i \sigma_i)^2 \geq 0$$

it follows that $h \geq r_I$.

Identifying Matching Factors for Several Sets of Variates

The technique described and illustrated in Section 2 for matching factors underlying two sets of variates can be extended to several sets of variates (Horst, 1961). The matrices of canonical weights A_i, $i = 1, 2, \cdots, m$, are normalized to fulfill the conditions

$$\operatorname{diag}(A_i' S_{ii} A_i) = I. \tag{6.10}$$

Then the matrices

$$V_{ij} = A_i' S_{ij} A_j$$

are the correlations between the matching composite variates, $\mathbf{a}_{ig}'\mathbf{x}$, where $i = 1, 2, \cdots, m$ sets of variates and $g = 1, 2, \cdots, r$ matching groups of composites. For two sets of variates

$$V_{11} = I, \qquad V_{22} = I, \qquad V_{12} = D_\rho ,$$

so that the composites for a given set of variates are uncorrelated and the composites are uncorrelated between sets except for matching pairs. For more than two sets of variates the composites will, in general, be correlated both within and between sets. The factors for more than two sets of variates will then be oblique. The oblique factors for the ith set can be interpreted from an examination of the oblique factor matrix (primary pattern),

$$F_i = S_{ii}A_i(A_i'S_{ii}A_i)^{-1}. \tag{6.11}$$

When matching factors across several sets have been identified, it may by hypothesized that there are common factors underlying all m sets which account for this congruence. The sum of the matching composite variables across sets gives an estimate of these underlying factors. If the original set of combined weights are normalized so that $A'SA = I$, then $L = SA$ is an orthogonal factor matrix of S, the SP matrix for all variables combined. The factor variables are generalized canonical variates $\mathbf{a}_g'\mathbf{x}$, $g = 1, 2, \cdots, r$.

Discrimination and Scaling

Since in generalized canonical analysis there are no within-group replications, the approach to scaling is not a generalization of scaling in canonical analysis but certain analogous quantities may be defined by considering the two-way table of m similar measurements on n individuals or experimental units. Let $x_{i\alpha}$ be the score of the αth individual on the ith variate. Since the mean on each variate is always zero, $\bar{x}_i = \bar{x} = 0$, and

$$\begin{aligned}
\text{SSB} &= \text{the between-individuals sum of squares} \\
&= m \sum_\alpha (\bar{x}_\alpha - \bar{x})^2, \\
\text{SSI} &= \text{the interaction sum of squares} \\
&= \sum_{i,\alpha} (x_{i\alpha} - \bar{x}_i - \bar{x}_\alpha + \bar{x})^2 \\
&= \sum_i \text{SS}_i - \text{SSB}, \\
\text{SST} &= \text{SSB} + \text{SSI} = \sum_{i,\alpha} (x_{i\alpha} - \bar{x})^2,
\end{aligned}$$

where

$$\text{SS}_i = \sum_\alpha (x_{i\alpha} - \bar{x}_i)^2.$$

Then the correlation ratio is

$$\theta^2 = \frac{\text{SSB}}{\text{SST}} = \frac{\text{SSB}}{\sum_i \text{SS}_i} = \frac{1}{m}\gamma. \tag{6.12}$$

It follows that generalized canonical analysis maximizes the correlation ratio and also the variance ratio

$$F = \frac{(m-1)(n-1)\ \mathrm{SSB}}{(n-1)\ \mathrm{SSI}} = \frac{(m-1)\gamma}{m-\gamma}. \tag{6.13}$$

Thus, we have quantities that are analogous to the correlation ratio and variance ratio in canonical analysis but are not generalizations of them. For an arbitrary set of weights $\mathbf{a}$, the individual scale values are $u_\alpha = \mathbf{a}'\mathbf{x}_\alpha$. The canonical weights optimally discriminate in the sense of maximizing the correlation ratio. If more than one scale component is extracted, the most appropriate normalization is

$$\mathbf{u}_g'\mathbf{u}_g = \left(\frac{\mathrm{SSB}}{\mathrm{SSI}}\right)_g = \frac{\gamma_g}{m-\gamma_g}, \tag{6.14}$$

where $\mathbf{u}_g$ is the vector of individual scores on the gth canonical variate and γ_g is the gth root of $S_d^{-1}S$. All the preceding results hold if each variate is a linear composite of a set of variates and so apply to the general case of m similar sets of variates.

Since $\bar{x}_i = \bar{x} = 0$, the interaction sum of squares reduces to

$$\mathrm{SSI} = \sum_{i,\alpha} (x_{i\alpha} - \bar{x}_\alpha)^2$$

which is the within-individuals sum of squares. Therefore, the canonical weights maximize the ratio of the between-individuals to the within-individuals sum of squares or the internal consistency of the measurements. In this regard it will be helpful to distinguish between within-set weights and between-set weights. Let the weights $\mathbf{a}$ for m sets of standardized variates be represented by the following model

$$\mathbf{a}_i = \beta_i \mathbf{w}_i, \qquad \mathbf{w}_i' S_{ii} \mathbf{w}_i = 1,$$

where β_i are the between-set weights and w_i are the within-set weights and $i = 1, 2, \cdots, m$. For the β's fixed and equal, the canonical weights $\mathbf{w}_i$ will maximize the average correlation between the m variates $\mathbf{w}_i'\mathbf{x}_i$. This will tend to produce variates that are most nearly equivalent measurements in the sense of uniform intercorrelation. For fixed values of the $\mathbf{w}_i$ the canonical β's tend to produce homogeneity by assigning small weights to the deviant sets and effectively eliminating them. The canonical weights $\mathbf{a}$ represent an equilibrium between these two tendencies.

Scoring Categorical Data-Scalogram Analysis

The canonical method of scoring two-way contingency tables may be extended to m-way tables. There are m sets of pseudo-variates and the sum of product matrices are

$$S_{ij} = N_{ij} - \frac{1}{n}\mathbf{n}_i\mathbf{n}_j'$$

where $N_{ii} = D_i$, $i, j = 1, 2, \cdots, m$. Both S and S_d are of rank $\sum_i k_i - m$ and lie within the subspaces whose idempotent J_d is a diagonal super-matrix with elements J_i . The conditional inverse of S_d is

$$S_d^{-1*} = J_d D^{-1} J_d ,$$

where D is a $(\sum_i k_i) \times (\sum_i k_i)$ diagonal matrix whose elements are the marginal frequencies. The vector of canonical weights or scores $\mathbf{a}$ is the principal vector of

$$S_d^{-1*}S = J_d D^{-1} N \tag{6.15}$$

and satisfies

$$(J_d D^{-1} N - \gamma I)\mathbf{a} = \mathbf{0}. \tag{6.16}$$

Let Q_d be a diagonal super-matrix with elements $Q_i = I - \mathbf{1}\mathbf{n}_i'/n$, then

$$Q_d\mathbf{a} = \begin{bmatrix} Q_1\mathbf{a}_1 \\ Q_2\mathbf{a}_2 \\ \vdots \\ Q_m\mathbf{a}_m \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{a}}_1 \\ \tilde{\mathbf{a}}_2 \\ \vdots \\ \tilde{\mathbf{a}}_m \end{bmatrix} = \tilde{\mathbf{a}}. \tag{6.17}$$

Premultiplying (6.16) by Q_d gives

$$\left[\left(D^{-1}N - \frac{1}{n}\mathbf{1}\mathbf{n}'\right) - \gamma I\right]\tilde{\mathbf{a}} = \mathbf{0}, \tag{6.18}$$

where $\mathbf{n}' = [\mathbf{n}_1' \ \mathbf{n}_2' \ \cdots \ \mathbf{n}_m']$ and $\mathbf{n}'\mathbf{1} = mn$. The term $-\mathbf{1}\mathbf{n}'/n$ has the effect of subtracting out the largest root of $D^{-1}N$, which is always unity.

An attitude item for which each individual must indicate his opinion by making one of k mutually exclusive responses ranging from disagree strongly to agree strongly, has been used by Likert (1932) for the construction of attitude scales. The responses of n individuals on m items of this type form an m-way contingency table. The canonical scoring of such an m-way table was introduced by Guttman (1941) and is known in the psychological and sociological literature as scalogram analysis.

The bivariate normal model for two-way tables can be generalized by considering the joint distribution $f(u, x_i)$ of the scale values u and the scores on the ith item, x_i , for a population of individuals. Assume that, given the proper score assignments, the items are parallel measurements and that $f(u, x_i)$ is bivariate normal and identical for each of the m items. Then

$$f(u, x_1) = f(u, x_2) = \cdots = f(u, x_m) \tag{6.19}$$

and

$$f(u, x) = \frac{1}{m} \sum_i f(u, x_i)$$

is the same bivariate normal distribution where the set of values $\{x\}$ is the union of the sets $\{x_i\}$. It follows that $f(u, x)$ may be expanded in the form

$$f(u, x) = f(u)f(x)\left[1 + \sum_{g=1}^{\infty} \frac{\theta^g}{g!} H_g(u)H_g(x)\right], \tag{6.20}$$

where $H_g(u)$ and $H_g(x)$ are Hermite polynomials and θ is the correlation in the common distribution. The discrete frequency counterpart of $f(u, x)$ is the matrix Z_{21} with elements,

$$\begin{aligned} z_{\alpha\tau(i)} &= 1, \quad \text{if the } \alpha\text{th individual makes the } \tau\text{th response to the } i\text{th item,} \\ &= 0, \quad \text{otherwise,} \end{aligned}$$

where

$$\begin{aligned} \alpha &= 1, 2, \cdots, n \text{ individuals,} \\ \tau &= 1, 2, \cdots, k_i \text{ categories in the } i\text{th item,} \\ i &= 1, 2, \cdots, m \text{ items.} \end{aligned}$$

The matrix Z has dimensionality $(\sum_i k_i) \times n$ and marginal frequencies $z_\alpha = m$ and $z_j = n_j$ where j is an index running over all values of $\tau(i)$. Estimates of the category locations on the latent variables u and x may be obtained from a canonical scoring of the frequency matrix Z_{12}. From (1.20), the scores for the item categories satisfy

$$\left[\left(D_{z_1}^{-1}Z_{12}D_{z_2}^{-1}Z_{21} - \frac{1}{z}\,\mathbf{1z}'\right) - \theta^2 I\right]\tilde{\mathbf{a}} = \mathbf{0}.$$

But $D_{z_2} = mI$, $Z_{12}Z_{21} = N$, $\mathbf{z}_1 = \mathbf{n}$, and $z = mn$, where N and $\mathbf{n}$ are as previously defined. On substitution,

$$\left[\frac{1}{m}\left(D_1^{-1}N - \frac{1}{n}\,\mathbf{1n}'\right) - \theta^2 I\right]\tilde{\mathbf{a}} = \mathbf{0}, \tag{6.21}$$

which is identical to (6.18) with $\theta^2 = \gamma/m$. This result shows that any problem in the canonical scoring of an m-way contingency table is equivalent to the scoring of a related two-way table.

The estimate of the item intercorrelations is the generalized product moment correlation

$$\rho(x_1, x_2, \cdots, x_m) = \max r_I = \frac{\gamma - 1}{m - 1} = \frac{m\theta^2 - 1}{m - 1}. \tag{6.22}$$

It seems worthwhile to repeat some remarks made in connection with two-way tables. The largest root θ^2 will be only a small fraction of the trace even when the model is completely valid. No intensity effect is implied by the U-shaped character of the second component since this would be expected from the quadratic nature of $H_2(x)$. In the discussion of two-way tables it was pointed out that when the categories have a know order, marginal scores supply a convenient approximation to canonical scores. The response categories of attitude items are ordered and marginal scores could be substituted for the canonical within item scores. The between-item weights may be canonically determined from the principal component of the item intercorrelations. This "mixed" method of combining marginal within-item scores and canonical between-item weights provides a good approximation to a complete scalogram analysis.

Association and Reliability

Since the correlation ratio is $\theta^2 = \gamma/m$, a suitable measure of the efficiency of the discrimination is

$$\theta = \left(\frac{\gamma}{m}\right)^{1/2}. \tag{6.23}$$

Similarly, in a manner analogous to (5.3), we can take as the measure of reliability

$$\psi_m = 1 - \frac{1}{F} \tag{6.24}$$

with

$$F = \frac{(m-1)\gamma}{m-\gamma}.$$

Then

$$\psi_m = \frac{m}{m-1}\left(1 - \frac{1}{\gamma}\right). \tag{6.25}$$

Let

$$s_t^2 = \frac{1}{n-1}\,\mathbf{a}'S\mathbf{a} = \text{variance of the linear composite } \mathbf{a}'\mathbf{x},$$

$$s_i^2 = \frac{1}{n-1}\,\mathbf{a}_i'S_{ii}\mathbf{a}_i = \text{variance of the variate } \mathbf{a}_i'\mathbf{x}_i\,,$$

then

$$\gamma = \frac{\mathbf{a}'S\mathbf{a}}{\mathbf{a}'S_d\mathbf{a}} = \frac{s_t^2}{\sum_i s_i^2}$$

and

$$\psi_m = \frac{m}{m-1}\left[1 - \frac{\sum_i s_i^2}{s_t^2}\right] = \alpha. \tag{6.26}$$

The quantity ψ_m is equivalent to coefficient alpha (Kuder and Richardson, 1937; Guttman, 1945; Cronbach, 1951) which has been used extensively as a reliability measure in the area of psychological testing. It follows that maximizing γ is equivalent to maximizing alpha.

7. A CLASSIFICATION OF MEASUREMENT BASED ON CATEGORICAL RESPONSES

Much of the data in the social and biological sciences is categorical in nature. In this section the canonical scoring approach is related to other methods of quantifying categorical data which have been developed in the psychometric area. These methods are classified according to the context or purpose, the type of response or form of the data and the method of quantification.

A. Context

The context is determined by the nature of the variates and experimental units. The indefinitely large number of measuring instruments and things that may be measured has been dichotomized according to their human or non-human character. For variates we have either items or judges and for the experimental units either objects or subjects. The context determines only the interpretation of the results. The two dichotomies determine four types of context:

1. Items × objects—physical measurement,
2. Items × subjects—attitude measurement, mental testing, diagnostic testing, etc.,
3. Judges × objects—ratings of the attributes and performance of things, preference measurement,
4. Judges × subjects—ratings of the attributes and performance of individuals.

B. Response

The nature of the response required determines the form of the data. In each case there is a set of k categories ordered with respect to an underlying continuum. The labeling of the categories may differ, for example, a single Likert item with k response categories, a set of k Thurstone attitude statements and k problems graded in difficulty are analogous although the labeling of the response categories are considerably different. Although canonical scoring will be possible in principle, the simpler marginal scoring is usually more feasible. Position responses as opposed to order responses correspond to Thurstone's (1929) maximum probability and increasing probability items. Responses are grouped into four types, accompanied by typical instructions.

1. Single position response (Likert technique).
   Instruction: Check the statement that best represents your opinion.
2. Fixed number of position responses (Modified Thurstone technique).
   Instruction: Check the five statements that best represent your opinion.
3. Any number of position responses (Thurstone technique).
   Instruction: Check the statements that you agree with.
4. Order response (Method of constant stimuli, graded dichotomies).
   Instruction: Check all stimuli that are less than x_i (x is rated higher than or passes these categories).

While only response type B1 allows a direct application of canonical scoring the other three may be handled by transforming or interpreting the data as type B1. The data of response types B2 and B3 may be interpreted as replicate observations. Canonical scores would then maximize the internal consistency of the individual's responses. There is some evidence that fixing the number of statements to be endorsed produces better results than leaving the individual free to endorse any number (Guilford, 1954) and we would also expect better results from a canonical scoring since each individual receives equal weight. In response type B3, the individuals are weighted according to the number of statements they endorse.

Response type B4 is essentially the cumulative form of B1. The transformation from B4 to B1 amounts to the determination of the limen or threshold category. When the categories are numbered $1, 2, \cdots, k$, the number of positive responses is an estimate of the limen category and placing a one in this category and zeros elsewhere gives the corresponding B1 response. Without replications canonical scoring is not possible but marginal scoring is still appropriate.

Methods of Quantification

The methods of quantification are classified according to the way in which scores are assigned to the response categories within a given set or multiple choice item and the relative weight given to each item. Sometimes prior weights and scores are used, chosen from considerations external to the data. Prior weights are almost always equal and prior scores are often the natural numbers $1, 2, \cdots, k$. The alternatives are as follows.

C. Within-set scores

1. Canonical
2. Marginal
3. Prior

D. Between-set weights

1. Canonical
2. Prior

Several common measurement techniques can be located within this framework. For m-way tables there are six possibilities. For example:

1. Scalogram analysis; A2, B1, C1, D1. Both within-item scores and item weights are optimally assigned but the order among the categories is disregarded.
2. Likert scale; A2, B1, C2 or C3, D2. At first, Likert (1932) used marginal scoring but found that the natural numbers 1, 2, $\cdots$, k gave satisfactory results.
3. Most reliable Likert scale, approximate scalogram; A2, B1, C2 or C3, D1. Marginal item scoring is used but the items are weighted for maximum reliability as measured by KR 21. (The actual procedure in constructing a Likert scale is between (2) and (3) since the accompanying item analysis amounts to the optimal assignment of 0, 1 weights.)
4. Scoring response categories so as to maximize the average inter-item correlation; A, B, C1, D2. Equal prior weights and canonical scores is equivalent to maximizing the average inter-item correlation. This requires extensive computation and apparently has not been used. It is included here for completeness.

For two-way tables, only the case where one way of classification is ordered will be considered. There are then three ways of scoring these ordered categories.

5. Canonical scoring of two-way contingency tables; A, B, C1. This is the procedure discussed at length in Section 4.
6. Preference measurement from ratings in successive categories; A3, B1, C2. The judges are regarded as replicate measurements and marginal scores are assigned to the rating categories. (This is not Thurstone's solution.)
7. Method of constant stimuli; A2, B4, C2. An order response must be made at all k levels. Often, as in bio-assay, a similar procedure is used but each individual or experimental unit responds to only one level. In this case, although the scale of measurement

can be determined in the sense of assigning marginal scores to the categories, the individuals or units cannot be located on this scale.

8. Thurstone scale; A2, B2 or B3, C3. In Thurstone's (1929) method of attitude scale construction, the scores for the categories (statements) are determined previous to the experiment from the ratings of judges.
9. A set of problems graded in difficulty; A2, B4, C3. The individual's score is the number of correct answers. This gives the location of his limen category. The categories for the corresponding B1 response would have scores 1, 2, $\cdots$, k.

A set of k dichotomized variables, as problems in an achievement test, may be classified in two different ways depending on whether they more closely approximate the requirements of a set of parallel dichotomies or a set of graded dichotomies. Parallel dichotomies are approximately equal in difficulty with uniform intercorrelations. They are equivalent to k two-category Likert items. The most appropriate measure of intercorrelation between the k items is r_I. A set of graded dichotomies has a definite order with the proportion of positive responses or passes decreasing from left to right, that is, $p_i > p_j$ when $i < j$. A set of k graded dichotomies is the same as a type B4 or "order" response to a set of k categories. In this case, an appropriate measure of association between the k items is Loevinger's coefficient of homogeneity, H_t (Loevinger, 1947). Perfect association occurs when each individual passes all items up to a point and then fails the remaining items. Under these circumstances H_t will be one but r_I will not, since it can attain the value of one only when $p_1 = p_2 = \cdots = p_k$.

The homogeneity coefficient may be derived by considering the previously introduced association measure

$$h = \frac{\sigma_t^2 - \sum_i \sigma_i^2}{(\sum_i \sigma_i)^2 - \sum_i \sigma_i^2} = \frac{1'(C - C_d)1}{1'(\boldsymbol{\sigma}\boldsymbol{\sigma}' - C_d)1},$$

where $C = [\sigma_{ij}]$. The matrix C_d is the covariance matrix of a perfectly uncorrelated (heterogeneous) set of variables and the matrix $\boldsymbol{\sigma}\boldsymbol{\sigma}'$ is the covariance matrix of a perfectly correlated (homogeneous) set of continuous variables with the same variances. When the variables are dichotomized the maximum correlation between two items is dependent upon the p's and is one only when they are equal. But since we are dealing with graded dichotomies the p's will differ considerably. The covariance matrix for a set of perfectly or maximally associated graded dichotomies with proportions p_i is

$$V_{ij} = V_{ji} = p_j q_i, \tag{7.1}$$

where $i < j$ and $p_i > p_j$ (see Guilford, 1954). Substituting V for $\sigma\sigma'$ in the expression for h, we have

$$(7.2) \qquad H_t = \frac{1'(C - C_d)1}{1'(V - C_d)1} = \frac{\sigma_t^2 - \sum_i p_i q_i}{2 \sum_{i<j} p_j q_i}.$$

From the fact that $p_j q_i \leq \sigma_i \sigma_j$ it follows that $H_t \geq h$ which gives the double inequality

$$(7.3) \qquad H_t \geq h \geq r_I .$$

The equality relations hold only when $p_1 = p_2 = \cdots = p_m$.

8. TESTS OF HYPOTHESES

To include both continuous and pseudo-variates in the same formulation let $p = \text{rank } S_{11}$, $q = \text{rank } S_{22}$, and $t = n - 1$. Hypotheses in canonical analysis may be classified as tests of independence, partial independence, or rank.

Independence

These tests evaluate the hypothesis that the population canonical correlations are all zero. Some hypotheses of independence in the general sense of canonical analysis are:

1. Independence of two sets of continuous variates,
2. Independence in contingency tables,
3. Non-significance of a discriminant function for k groups, or equivalently,
4. Equality of mean vectors for k groups (one-way analysis of variance).

If at least one set has a multivariate normal distribution, then the likelihood-ratio statistic

$$(8.1) \qquad \Lambda(p, q) = |I - R| = \prod_i (1 - \lambda_i)$$

is distributed as $L(p, q, t)$, a multivariate analogue of the F distribution. Even when there is considerable deviation from normality as in the case of categorical data, the test will still be asymptotically valid. The moments of this distribution were first obtained by Wilks (1932) and a series expansion for the distribution by Box (1949). The first term of the series provides a good approximation to the exact distribution (Bartlett, 1938). The quantity

$$-M \log_e L(p, q, t)$$

is approximately chi square with pq degrees of freedom, where

$$M = t - \tfrac{1}{2}(p + q + 1).$$

The distribution theory for $L(p, q, t)$ as well as its generalization is discussed in some detail by Anderson (1958), using slightly different notation. His $U(p, q, n_e) = L(p, q, n_e + q)$ is not symmetric in p and q. The following identities will be useful for obtaining some exact F tests.

$$L(p, q, t) = L(q, p, t),$$
$$L^{1/2}(2, q, t) = L(1, 2q, 2t - 2),$$
$$\frac{1 - L(1, q, t)}{L(1, q, t)} = \frac{q}{t - q} F(q, t - q).$$

If either p or $q = 1$ or 2 an exact F test is possible.

As an example, take the case of a set of q measurements on each of two groups and the hypothesis of no discrimination between the groups, or equivalently, equality of mean vectors. Then $p = \text{rank } S_{11} = k - 1 = 1$ and the R matrix has only one root. Then $\Lambda = 1 - \lambda$ is distributed as $L(1, q, t)$ and

$$\frac{1 - L}{L} = \frac{\lambda}{1 - \lambda} = \frac{q}{t - q} F(q, t - q),$$

allowing an exact F test. This can be expressed in terms of the distance between groups. Let

$$\mu = \text{ch } (W^{-1}B) = \frac{n_1 n_2}{n} d_{12}^2$$

where

$$d_{12}^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' W^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$$

is a measure of the distance between groups, then

$$\frac{\lambda}{1 - \lambda} = \mu = \frac{n_1 n_2}{n} d_{12}^2 = \frac{q}{t - q} F(q, t - q) \tag{8.2}$$

(Hotelling, 1931).

Partial Independence

Given three sets of variates $\mathbf{x}$ $(p \times 1)$, $\mathbf{y}$ $(q \times 1)$ and $\mathbf{z}$ $(r \times 1)$, then the combined SP matrix for $\mathbf{x}$ and $\mathbf{y}$ about their regression on $\mathbf{z}$ is

$$S_{(\mathbf{x},\mathbf{y})|\mathbf{z}} = S_{(\mathbf{x},\mathbf{y})}(I - R_{(\mathbf{x},\mathbf{y})\mathbf{z}}) = U.$$

The partial canonical correlation matrix of $\mathbf{x}$ and $\mathbf{y}$ conditional on $\mathbf{z}$ is

$$R_{\mathbf{xy}|\mathbf{z}} = U_{11}^{-1} U_{12} U_{22}^{-1} U_{21} . \tag{8.3}$$

The largest root of $R_{\mathbf{xy}|\mathbf{z}}$ is the squared partial canonical correlation of $\mathbf{x}$ and $\mathbf{y}$ conditional on $\mathbf{z}$, $\rho^2(\mathbf{x}, \mathbf{y}) \mid \mathbf{z}$. The following identity may be established,

$$I - R_{x(y,z)} = (I - R_{xz})(I - R_{xy|z}). \tag{8.4}$$

Tests of partial independence evaluate the hypothesis that the population partial canonical correlations are all zero. Some examples of hypotheses of partial independence are:

1. A subset z of the variates (y, z) is sufficient to account for the association with x or the discrimination between groups.
2. One or more hypothetical discriminant functions (coefficients determined by considerations external to the data) account for the discrimination between groups.
3. A subset of the variates (or groups) have equal canonical weights (or mean vectors).

The likelihood-ratio test of partial independence is

$$\Lambda(p, q \mid r) = |I - R_{xy|z}| = \frac{|I - R_{x(y,z)}|}{|I - R_{xz}|} \tag{8.5}$$

distributed as $L(p, q, t - r)$. The test will be exact when either **x** or **y** has a multivariate normal distribution about its regression on **z**.

For example,

$$|I - R_{xy|z}| = L(p, q, t - r)$$

gives a test of the hypothesis that the subset **z** $(r \times 1)$ of (**y**, **z**) $[(q + r) \times 1]$ is sufficient to account for the association with **x** $(p \times 1)$. Hypotheses of the sufficiency of one or more *hypothetical* discriminant functions are equivalent to the above hypothesis. Consider the two sets of variates **x** $(p \times 1)$ and **v** $[(q + r) \times 1]$ and the hypothesis that the hypothetical variates **z** $(r \times 1) = M_z\mathbf{v}$ account for the association or discrimination when **x** is a vector of pseudo-variates. We can define a set of variates **y** $(q \times 1) = M_y\mathbf{v}$ such that

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} M_y \\ M_z \end{bmatrix} \mathbf{v} = M\mathbf{v}$$

is a nonsingular transformation, otherwise **y** is arbitrary. Then

$$R_{(y,z)x} = M^{-1}R_{vx}M$$

and

$$\text{ch}\,[R_{(y,z)x}] = \text{ch}\,[R_{vx}].$$

The test for sufficiency of the hypothetical variates **z** is then

$$|I - R_{xy|z}| = \frac{|I - R_{x(y,z)}|}{|I - R_{xz}|} = \frac{|I - R_{xv}|}{|I - R_{xz}|}$$

distributed as $L(p, q, t - r)$. For a single hypothetical discriminant function

$z = \mathbf{m}'\mathbf{v}$ and two groups, $R_{\mathbf{xv}}$ and $R_{\mathbf{xz}}$ are of unit rank with nonzero roots $\lambda(\mathbf{v})$ and $\phi(z)$, respectively. Then

$$\frac{1 - \lambda(\mathbf{v})}{1 - \phi(z)}$$

is distributed as $L(1, q, t - 1)$, or

$$\frac{\lambda - \phi}{1 - \lambda} = \frac{q}{t - 1 - q} F(q, t - 1 - q).$$

When expressed in terms of the distance functions $d_{12}^2(\mathbf{v})$ and $d_{12}^2(z)$ this becomes

$$\frac{d_{12}^2(\mathbf{v}) - d_{12}^2(z)}{\dfrac{n}{n_1 n_2} + d_{12}^2(z)} = \frac{q}{t - 1 - q} F(q, t - 1 - q), \tag{8.6}$$

where

$$d_{12}^2(z) = \frac{[\mathbf{m}'(\bar{\mathbf{v}}_1 - \bar{\mathbf{v}}_2)]^2}{\mathbf{m}'W\mathbf{m}}$$

(Fisher, 1940). By taking

$$\mathbf{z} = \begin{bmatrix} \mathbf{v}_{\mathrm{I}} \\ \mathbf{1}'\mathbf{v}_{\mathrm{II}} \end{bmatrix},$$

we obtain a test of the hypothesis that the variates in the set $\mathbf{v}_{\mathrm{II}}$ have equal canonical weights.

Rank or Dimensionality

When the canonical variates are estimated from the data, a test of sufficiency is equivalent to a test of the rank of the population counterpart of R. Tests of rank evaluate the hypothesis that the first r significant estimated canonical variates or discriminant functions are sufficient to account for the association or discrimination between $\mathbf{x}$ $(p \times 1)$ and $\mathbf{y}$ $(q \times 1)$. Bartlett (1941) has suggested that after the elimination of the association due to the first r canonical variates, from the symmetry between the two sets of variates, the residual determinant

$$\Lambda(p - r, q - r) = \frac{|I - R|}{\prod_{i \le r} (1 - \lambda_i)} = \prod_{i > r} (1 - \lambda_i) \tag{8.7}$$

is distributed approximately as $L(p - r, q - r, t - r)$ in large samples.

This test of rank does not apply to categorical data under the bivariate normal assumption since the residual population roots will not be zero unless the largest root is zero. A test of goodness of fit for the bivariate normal model based on the first component would be more appropriate.

Largest Root and Trace Criteria

There are two other useful test criteria. If $|I - R|$ is distributed as $L(p, q, t)$ then λ_1, the largest root of R has the largest-root distribution. The marginal distribution of the largest sample root has been determined by Roy (1945) from the joint distribution of all the roots. Upper percentage points of this distribution have been computed by Heck (1960). This criterion is most appropriate as a test of the significance of a single canonical variate or discriminant function since it will tend to have greatest power against the alternative of unit rank, for a fixed value of the sum of the population roots.

Since all the roots of R lie within the range from zero to one, the expansion

$$-\log_e (1 - \lambda_i) = \lambda_i + \tfrac{1}{2}\lambda_i^2 + \tfrac{1}{3}\lambda_i^3 + \cdots$$

will be valid with probability one for all roots, and

$$-\log_e |I - R| = \operatorname{tr} R + \tfrac{1}{2} \operatorname{tr} R^2 + \tfrac{1}{3} \operatorname{tr} R^3 + \cdots . \tag{8.8}$$

Under the null hypothesis, as $n \to \infty$,

$$-n \log_e |I - R| \to n \operatorname{tr} R$$

which will be asymptotically chi square with pq df. Using the trace criterion in the test of independence for a $k \times l$ contingency table,

$$\begin{aligned} n \operatorname{tr} R &= n \operatorname{tr} (S_{11}^{-1*} S_{12} S_{22}^{-1*} S_{21}) = n \operatorname{tr} (J D_1^{-1} N_{12} D_2^{-1} N_{21}) \\ &= n \operatorname{tr} (D_1^{-1} N_{12} D_2^{-1} N_{21} J) = n \sum_{i,j} \left(\frac{n_{ij}^2}{n_i n_j} - 1 \right) \\ &= \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i n_j}{n} \right)^2}{\frac{n_i n_j}{n}} \end{aligned} \tag{8.9}$$

is distributed as chi square with $pq = (k - 1)(l - 1)$ df. This is the familiar chi-square test of independence in contengency tables (Williams, 1952).

Independence for Several Sets of Variates

If there are m sets of variates $\mathbf{x}_i$, each set with a multivariate normal distribution, then the hypothesis of the simultaneous independence of all m sets of variates is equivalent to the hypothesis that all the generalized canonical correlations,

$$\rho_r(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m) = \frac{\gamma_r - 1}{m - 1},$$

are zero. The likelihood-ratio criterion

$$\Lambda(\mathbf{p}) = \frac{|S|}{\prod_i |S_{ii}|} = \frac{|S|}{|S_d|} = |S_d^{-1}S| \tag{8.10}$$

is distributed as $L(\mathbf{p}, t) = L(p_1, p_2, \cdots, p_m, t)$, a generalization of $L(p, q, t)$ symmetric in all the p_i. The moments of the distribution of $L(\mathbf{p}, t)$ were found by Wilks (1935) and a series expansion for the distribution by Box (1949). The first term of the series provides a good approximation to the exact distribution. Let $p = \sum_i p_i$,

$$g(\mathbf{p}) = \tfrac{1}{2}(p^2 - \sum_i p_i^2),$$

$$h(\mathbf{p}) = (p^3 - \sum_i p_i^3),$$

$$M = t - \frac{1}{2}\left(\frac{h(\mathbf{p})}{g(\mathbf{p})} + 1\right),$$

then $-M \log_e L(\mathbf{p}, t)$ is approximately chi square with $g(\mathbf{p})$ df.

Partial Independence for Several Sets of Variates

Let $S_{\mathbf{x}|\mathbf{z}}$ be the SP matrix of the vector variate

$$\mathbf{x}'\ (1 \times p) = [\mathbf{x}_1'\ \mathbf{x}_2' \cdots \mathbf{x}_m']$$

about its regression on $\mathbf{z}$ $(r \times 1)$. Then the generalized partial canonical correlation between the m sets $\mathbf{x}_i$ conditional on $\mathbf{z}$ is

$$\rho(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m \mid \mathbf{z}) = \frac{\gamma - 1}{m - 1}, \tag{8.11}$$

where γ is the largest root of $(S_{\mathbf{x}|\mathbf{z}})_d^{-1} S_{\mathbf{x}|\mathbf{z}}$. A test of partial independence evaluates the hypothesis that $\mathbf{z}$ accounts for all of the association between the m sets or that all the generalized partial canonical correlations are zero. When $\mathbf{x}$ has a multivariate normal distribution about its regression on $\mathbf{z}$, under the null hypothesis

$$\Lambda(\mathbf{p} \mid r) = |(S_{\mathbf{x}|\mathbf{z}})_d^{-1} S_{\mathbf{x}|\mathbf{z}}|$$

is distributed as $L(\mathbf{p}, t - r)$.

Hypothesis of Rank in Common-Factor Analysis

The common-factor analysis model assumes there exists a set of r latent variables $\mathbf{f}$ that account for the association between the observed variables $\mathbf{x}$. Let $\mathbf{f}$ be an arbitrary set of latent variables. With no loss of generality the latent variables $\mathbf{f}$ are required to be uncorrelated with unit variance. Then

$$S_{\mathbf{x}|\mathbf{f}} = S(I - R_{\mathbf{x}f}) = S - S_{\mathbf{x}\mathbf{z}}S_{\mathbf{z}\mathbf{x}}.$$

Letting $S_{\mathbf{x}\mathbf{z}} = L$, then

$$S_{x|f} = S - LL'. \tag{8.12}$$

Under the hypothesis of the independence of the observed variables **x** about their regression on **f**,

$$S_{x|f} = U$$

is a diagonal matrix. It follows that if r latent variables are sufficient, then

$$S_0 = U + LL'. \tag{8.13}$$

If U and L are maximum-likelihood estimates under the model (8.13), then

$$|S_0^{-1}S| = |U^{-1}(S - LL')| = |(S_{x|f})_d^{-1}S_{x|f}| \tag{8.14}$$

is the likelihood-ratio test statistic for the hypothesis that r factors are sufficient. For large samples $-n \log_e |U^{-1}(S - LL')|$ is distributed approximately as chi square with $\frac{1}{2}[(p - r)^2 - p - r]$ degrees of freedom (Lawley, 1940). The criterion is not distributed as $L(1, 1, \cdots, 1, t - r)$ because the regression variables **f** have been estimated from the data.

General Trace Criterion

All the tests we have considered may be regarded as hypotheses on the structure of a covariance matrix, H_0: $\Sigma = \Sigma_0$. The likelihood-ratio criterion is in each case of the form

$$\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} = |S_0^{-1}S| \tag{8.15}$$

and from the theory for such tests (see Wilks, 1938b), it is known that $-n \log_e |S_0^{-1}S|$ is asymptotically chi square with degrees of freedom, v, equal to the number of independent parameters in Σ minus the number of independent parameters in Σ_0. Under the null hypothesis as $n \rightarrow \infty$, $(1/n) S \rightarrow (1/n)S_0$, so that for sufficiently large n the conditions $0 < \gamma_i < 2$, will be valid for all roots of $S_d^{-1}S$ with probability approaching one. Under these conditions the expansions

$$-\log_e \gamma_i = (1 - \gamma_i) + \tfrac{1}{2}(1 - \gamma_i)^2 + \tfrac{1}{3}(1 - \gamma_i)^3 + \cdots$$

for all γ_i, and

$$-\log_e |S_0^{-1}S| = \operatorname{tr}(I - S_0^{-1}S) + \tfrac{1}{2}\operatorname{tr}(I - S_0^{-1}S)^2 + \tfrac{1}{3}\operatorname{tr}(I - S_0^{-1}S)^3 + \cdots \tag{8.16}$$

will be valid. The first term vanishes for maximum-likelihood estimates and as $n \rightarrow \infty$,

$$-n \log_e |S_0^{-1}S| \rightarrow \frac{n}{2}\operatorname{tr}(I - S_0^{-1}S)^2$$

which will be asymptotically chi square with v degrees of freedom.

For a test of the independence of m sets of variates $S_0 = S_d$ and

$$\frac{n}{2} \operatorname{tr} (I - S_d^{-1}S)^2 = n \sum_{i<j} \operatorname{tr} R_{ij} \tag{8.17}$$

is distributed as chi square with $g(\mathbf{p}) = \sum_{i<j} p_i p_j$ degrees of freedom, where R_{ij} is the canonical correlation matrix between the ith and jth sets. The separate chi-square tests of independence between two sets are pooled to obtain the over-all criterion.

The determinantal criterion (8.14) for a test of the sufficiency of r factors may be replaced by a trace criterion. Let $E = S - LL'$ be the matrix of residuals after the removal of r factors. The likelihood-ratio criterion

$$-n \log_e |U^{-1}(S - LL')|$$

is approximately equal to

$$\frac{n}{2} \operatorname{tr} [I - U^{-1}(S - LL')]^2 \underline{=} n \sum_{i<j} \frac{e_{ij}^2}{u_i u_j} \tag{8.18}$$

which is asymptotically chi square with $\frac{1}{2}[(p - r)^2 - p - r]$ degrees of freedom.

APPENDIX A

Canonical Factor Analysis of Intercorrelations Between Personality Characteristics and Occupational Preference

To investigate the importance of personality characteristics in the determination of occupational preference, K. J. Jones (1965) administered the Guilford-Zimmerman Temperament Survey and an occupational preference inventory based on Roe's occupational categories to a small group of high school students. Canonical factor analysis was applied to Jone's data to provide an example of this method. To reduce the amount of computation, only four of the eight Roe categories were used:

1) business contact,
2) technology,
3) outdoor,
4) general cultural.

For the same reason, only five of the ten G–Z scales were included:

5) social,
6) emotional stability,
7) objectivity,
8) thoughtfulness,
9) personal relations.

Four occupational factors were in this way paired with four personality factors.

TABLE A1
Matrix of Intercorrelations

	1	2	3	4	5	6	7	8	9
1	1.00	.00	−.14	.09	.08	.03	−.10	−.16	−.15
2	.00	1.00	.43	−.37	−.19	.09	.05	−.17	.01
3	−.14	.43	1.00	−.15	−.22	.18	.12	.08	.15
4	.09	−.37	−.15	1.00	−.08	−.09	−.22	.06	−.30
5	.08	−.19	−.22	−.08	1.00	.24	.22	−.36	.22
6	.03	.09	.18	−.09	.24	1.00	.71	.13	.57
7	−.10	.05	.12	−.22	.22	.71	1.00	.12	.53
8	−.16	−.17	.08	.06	−.36	.13	.12	1.00	−.20
9	−.15	.01	.15	−.30	.22	.57	.53	−.20	1.00

TABLE A2
Canonical Weights

	A					*C*			
	I	II	III	IV		I*	II*	III*	IV*
1	.46	.08	−.30	.83	5	−.41	.92	−.28	.43
2	.82	−.42	−.35	−.70	6	.92	−.37	.43	.96
3	−.32	−.65	.56	.70	7	−.30	−.29	−.63	−.72
4	.73	.30	.68	−.32	8	−.78	.42	.78	.31
					9	−1.02	−.30	−.05	.24
ρ =	.56	.38	.22	.16					

TABLE A3
Factor Matrices

	L					*K*			
	I	II	III	IV		I*	II*	III*	IV*
1	.54	.19	−.34	.71	5	−.20	.55	−.60	.45
2	.48	−.78	−.30	−.31	6	−.07	−.48	−.01	.74
3	.01	−.83	.48	.27	7	−.37	−.46	−.32	.23
4	.49	.55	.68	−.08	8	−.35	.06	.87	.14
					9	−.58	−.55	−.36	.44

APPENDIX B

Discriminant Analysis Applied to Ratings of Morris's "Ways to Live"

During a world tour Morris asked six groups of college students to indicate their preferences for thirteen "ways to live." The six student groups participating were: U. S. (white), U. S. (negro), India, Japan, China (pre-communist), and Norway. The list of "ways to live" is as follows.

1. Refinement, moderation, restraint; preservation of the best attainments of man.
2. Self-sufficiency, understanding of self; avoidance of outward activity.
3. Sympathy, concern for others; restraint of one's self-assertiveness.
4. Abandonment, sensuous enjoyment of life; solitude and sociality both are necessary.
5. Energetic, cooperative action for the purpose of group achievement and enjoyment.
6. Activity; constant striving for improved techniques to control nature and society.
7. Flexibility, diversity within self; accept something from all other paths of life.
8. Carefree, relaxed, secure enjoyment.
9. Quiet receptivity to nature yields a rich self.
10. Dignity, self-control; but no retreat from the world.
11. Give up the world and develop the inner self.
12. Outward, energetic activity; use of the body's energy.
13. Let oneself be used; remain close to persons and to nature.

Discriminant analysis was applied to Morris's data by L. V. Jones and R. D. Bock (1960) and their results are summarized in Tables B1 through B5. One "way" was omitted from the analysis because of an error in translation into Chinese.

TABLE B1
Between-Groups Crossproducts Matrix, B

Way	1	2	3	4	5	6	7	8	9	10	11	12
1	28.22											
2	−3.37	71.27										
3	1.79	56.22	67.73									
4	1.58	−1.53	−15.02	9.82								
5	.64	−13.91	6.15	−6.16	51.58							
6	16.86	−8.80	8.97	−4.51	23.59	25.70						
7	29.65	−37.12	−34.28	5.38	−14.46	7.89	59.14					
8	11.61	−25.37	−18.14	−1.22	−3.34	3.54	29.42	17.72				
9	−4.56	70.91	48.38	4.36	−22.83	−10.43	−37.40	−29.13	79.48			
10	42.34	27.88	32.72	2.58	8.95	32.87	18.73	−.78	29.96	89.92		
11	9.41	61.02	46.78	2.41	3.73	.95	−23.26	−18.89	55.73	44.31	69.60	
12	7.36	−5.03	6.42	−6.09	.38	7.82	8.86	6.23	−7.47	8.36	−6.44	7.55

TABLE B2

Within-Groups Crossproducts Matrix, W

Way	1	2	3	4	5	6	7	8	9	10	11	12
1	637.11											
2	113.30	743.69										
3	175.56	61.89	688.05									
4	−4.72	107.79	−49.93	824.85								
5	20.32	−178.91	109.77	42.08	830.32							
6	.44	−29.12	30.60	3.65	195.16	638.44						
7	.56	44.47	−18.34	142.24	−22.61	7.89	786.57					
8	50.71	32.93	7.43	214.72	11.31	−20.52	131.96	927.45				
9	55.40	185.50	94.19	62.86	−77.43	−77.79	67.27	110.58	608.76			
10	153.84	122.69	133.30	−115.66	17.19	91.17	1.90	−89.30	37.13	849.38		
11	38.54	236.74	79.96	17.55	−109.75	−29.65	67.92	36.56	182.66	125.13	690.77	
12	4.70	12.21	73.61	44.34	169.21	197.63	−48.04	9.98	10.40	63.90	−6.85	797.38

TABLE B3
Discriminant Function Weights, A

Way	a_1	a_2	a_3	a_4
1	.013139	.013618	−.002754	.003038
2	−.014712	−.004098	.000487	−.006875
3	−.015051	−.000774	.001093	−.028165
4	−.001095	.003030	−.003466	.016797
5	−.004946	.001029	.027915	.008204
6	−.000457	.011249	.005901	−.004736
7	.016187	.008244	−.008534	−.005995
8	.007683	.000912	.002130	−.009582
9	−.018562	.003452	−.017691	.011624
10	.001343	.023713	−.002583	.003899
11	−.008339	.010496	.012308	.010210
12	.005681	−.001607	−.007641	−.011609

TABLE B4
Group Mean Scores on Canonical Variates

		Canonical Variate			
	Group	1	2	3	4
1	U.S.w	.5669	3.8803	.3521	−1.3281
2	U.S.n	.6105	4.4896	.7844	−1.5961
3	India	−.3245	4.8085	.6591	−1.7434
4	Japan	−1.0336	4.1608	.6359	−1.2574
5	China	−.2183	3.6052	1.0256	−1.8301
6	Norway	−.3537	4.0315	.0750	−1.9074
	Variance	.3243	.1553	.0927	.0598

TABLE B5
Intergroup Distances

Group	U.S.w	U.S.n	India	Japan	China	Norway
U.S.w	———					
U.S.n	.6319	———				
India	1.9229	1.0133	———			
Japan	2.7258	2.9479	1.1591	———		
China	1.3978	1.5820	1.6010	1.4533	———	
Norway	1.2827	1.7397	.9727	1.2161	1.1097	———

APPENDIX C

Canonical Scoring of Ratings of Social Class Membership by Nationality

In his study of the social structure of "Elmtown," Hollingshead (1949) asked a group of raters to classify families into groups according to prestige in the community. Table C1 is a classification of the adolescents of the community according to their families' social class and nationality. A group labeled "American" has been omitted. The scores assigned to the class categories should conform to the assumed order of the class structure. The optimum scoring compatible with this constraint assigned the same value to the first two social classes. This indicates that in this case a distinction between the first two social classes does not contribute to the discrimination. The same results would be obtained by combining classes 1 and 2 and ignoring the order constraint.

TABLE C1
Frequency of Class Membership by Nationality

	Social Class					
Nationality	1	2	3	4	Total	$\bar{c}$
Norwegian	4	46	68	35	153	.43
Irish	2	30	50	20	102	.58
German	2	16	40	21	79	−.20
Polish	0	0	13	20	33	−3.25
Total	8	92	171	96	367	
$\bar{a}$	4.59	4.59	.54	−5.72		$\rho = .28$

REFERENCES

Anderson, T. W. (1958) *Introduction to multivariate analysis.* New York: Wiley.

Bartlett, M. S. (1938) Further aspects of the theory of multiple regression. *Proc. Camb. phil. Soc.*, **34**, 33–40.

Bartlett, M. S. (1941) The statistical significance of canonical correlation. *Biometrika*, **32**, 29–37.

Bartlett, M. S. (1947) Multivariate analysis. *J. roy. statist. Soc.*, **9**, 176–197.

Bartlett, M. S. (1948) Internal and external factor analysis. *Brit. J. Psychol.*, Statist. Sec., **1**, 73–81.

Bartlett, M. S. (1952) The goodness of fit of a single hypothetical discriminant function. *Ann. Eugenics*, **16**, 199–214.

Box, G. E. P. (1949) A general distribution theory for a class of likelihood ratio criteria. *Biometrika*, **36**, 317–346.

Burt, C. (1948) Factor analysis and canonical correlations. *Brit. J. Psychol.* Statist. Sec., **1**, 95–106.

Cramér, H. (1946) *Mathematical methods of statistics.* Princeton Univ. Press.

Cronbach, L. J. (1947) Test "reliability": Its meaning and interpretation. *Psychometrika*, **12**, 1–16.

Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Edgerton, H. and Kolbe, E. L. (1936) The method of minimal variation for the combination of criteria. *Psychometrika*, **1**, 183–187.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, **7**, 179–188.

Fisher, R. A. (1938) *Statistical methods for research workers.* (7th ed.) Edinburgh: Oliver and Boyd.

Fisher, R. A. (1940) The precision of discriminant functions. *Ann. Eugenics*, **10**, 422–429.

Guilford, J. P. (1954) *Psychometric methods.* New York: McGraw-Hill.

Guttman, L. (1941) The quantification of a class of attributes. In P. Horst, *et al. The predication of personal adjustment*, Soc. Sci. Res. Council, Bull. No. 48.

Guttman, L. (1945) A basis for analyzing test-retest reliability. *Psychometrika*, **10**, 255–282.

Guttman, L. (1950) Components of scale analysis. In Stouffer *et al.*, *Measurement and prediction.* Princeton Univ. Press.

Harman, H. H. (1960) *Modern factor analysis.* Chicago: Univ. Chicago Press.

Heck, D. L. (1960) Charts of some upper percentage points of the largest characteristic root. *Ann. math. Statist.*, **31**, 625–642.

Hollingshead, A. B. (1949) *Elmtown's youth.* New York: Wiley.

Horst, P. (1936) Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, **1**, 53–60.

Horst, P. (1961) Generalized canonical correlations. *J. clin. Psychol.*, Mon. Suppl. No. 14.

Hotelling, H. (1931) The generalization of Student's ratio. *Ann. math. Statist.*, **2**, 360–378.

Hotelling, H. (1935) The most predictable criterion. *J. educ. Psychol.*, **26**, 139–142.

Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Jones, K. J. (1965) Occupational preference and social orientation. *Personnel and guid. J.*, **43**, 574–579.

Jones, L. V. and Bock, R. D. (1960) Multiple discriminant analysis applied to "Ways to live" ratings from six cultural groups. *Sociometry*, **23**, 162–176.

Kendall, M. G. and Stuart, A. (1961) *The advanced theory of statistics.* Vol. 2: *Inference and relationship.* London: Griffin.

Kuder, G. F. and Richardson, M. W. (1937) The theory of the estimation of test reliability. *Psychometrika*, **2**, 151–160.

Lancaster, H. O. (1957) Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, **44**, 289–292.

Lancaster, H. O. (1958) The structure of bivariate distributions. *Ann. math. Statist.*, **29**, 719–736.

Lawley, D. N. (1940) The estimation of factor loadings by the method of maximum likelihood. *Proc. roy. Soc. Edinburgh*, **60**, 64–82.

Likert, R. (1932) A technique for the measurement of attitudes. *Arch. Psychol.*, No. 140, 1–55.

Loevinger, J. (1947) A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monog.*, **61**, No. 4.

Loevinger, J. (1948) The technic of homogeneous tests compared with some aspects of scale analysis. *Psychol. Bull.*, **45**, 507–530.

Lord, F. M. (1958) Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, **23**, 291–295.

Morris, C. and Jones, L. V. (1955) Value scales and dimensions. *J. abnorm. soc. Psychol.*, **51**, 523–535.

Pearson, K. (1913) On the measurement of the influence of broad categories on correlation. *Biometrika*, **9**, 116–139.

Rao, C. R. (1948) The utilization of multiple measurements in problems of biological classification. *J. roy. statist. Soc.*, **10**, 159–203.

Rao, C. R. (1955) Estimation and tests of significance in factor analysis. *Psychometrika*, **20**, 93–111.

Richardson, M. W. (1938) Multidimensional psychophysics, *Psychol. Bull.*, **35**, 659–660.

Roy, S. N. (1945) The individual sampling distribution of the maximum, minimum and any intermediate of the p-statistics on the null hypothesis. *Sankyā*, **7**, 133–158.

Roy, S. N. and Roy, J. (1958) Analysis of variance with univariate or multivariate, fixed or mixed classical models. Inst. Statistics, Mimeo. Ser. No. 208, Univ. North Carolina, Chapel Hill.

Thurstone, L. L. (1927) A law of comparative judgment. *Psychol. Rev.*, **34**, 273–286.

Thurstone, L. L. and Chave, E. J. (1929) *The measurement of attitude.* Chicago: Univ. Chicago Press.

Torgerson, W. S. (1952) Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–419.

Torgerson, W. S. (1958) *Theory and methods of scaling.* New York: Wiley.

Wilks, S. S. (1932) Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471–494.

Wilks, S. S. (1935) On the independence of k sets of normally distributed variables. *Econometrika*, **3**, 309–326.

Wilks, S. S. (1938a) Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, **3**, 23–40.

Wilks, S. S. (1938b) The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. math. Statist.*, **9**, 60–62.

Williams, E. J. (1952) Use of scores for the analysis of association in contigency tables. *Biometrika*, **39**, 274–289.

Young, G. and Householder, A. S. (1938) Discussion of a set of points in terms of their mutual distances. *Psychometrika*, **6**, 331–333.