Yugoslav Journal of Operations Research
7 (1997), Number 2, 267-284

ON THE VARIANCE, RELIABILITY AND IMPORTANCE OF CANONICAL VARIABLES

Konstantin MOMIROVIĆ
Institute of Criminological and Sociological Research and Faculty of Philosophy, University of Belgrade

Abstract: A new procedure is proposed to estimate the real variance of canonical variables and real redundancy measures derived from the formal definition of the standardized variance of a linear combination of standardized variables. On the basis of thus defined variance estimates, a measure of real reliability and two measures of the lower bound to reliability are derived. The application of reliability measures to the determination of the number of important canonical variables is also considered.

Keywords: Canonical analysis, reliability, importance

1. INTRODUCTION

The biorthogonal model of canonical correlation analysis (Hotelling, 1935; 1936) is, both from logical and mathematical point of view, the basic method for data analysis and statistical hypothesis testing. It is easy to prove that almost all standard statistical methods, including regression analysis, analysis of variance, discriminant analysis, factor analysis, and even some methods for cluster analysis and the analysis of stochastic processes are, in fact, special cases of the general model of canonical correlation analysis, and that most statistical tests can be reduced to tests of the significance of canonical correlations.

However, in spite of the central position of canonical correlation analysis in the field of modern statistics, many problems related to estimating the real importance of canonical variables, their reliability or generalizability, and to estimating the amount of information emitted by canonical variables have not been solved at all, or have been unsatisfactorily, or even wrongly solved.

The aim of this paper is to propose some acceptable solutions to the problems of estimating the variance and redundancy of canonical variables, the related problem of their reliability, and to propose some new criteria for the determination of the number of important canonical variables, independent of the outcome of the significance tests of canonical correlations.

2. DEFINITIONS

Let $E = (e_i;\ i = 1,....,n) \subset P$ be a random sample of objects from a homogeneous population P, and let $V_1 = (v_j;\ j = 1,....,m_1) \subset U_1$ and $V_2 = (v_k;\ k = 1,....,m_2) \subset U_2$ be samples of quantitative, normally distributed variables from homogeneous universes of variables U_1 and U_2 with logically different fields of meaning. Let e be the summation vector of order (n) and let $|$ be the symbol for "under condition of". Now, without loss of generalization, it is possible to define data matrices obtained by the description of set E over the sets V_1 and V_2 both in standard normal form, so that

$$Z_1 = E \otimes U_1 \,|\, Z_1^t e = 0,\ \ diag(Z_1^t Z_1) = I_1$$

and

$$Z_2 = E \otimes V_2 \,|\, Z_2^t e = 0,\ \ diag(Z_2^t Z_2) = I_2$$

where $\otimes$ is the symbol of Cartesian product, 0 zero vectors, and I_1 and I_2 identity matrices of order (m_1) and (m_2), respectively. In this metrics,

$$R_{11} = Z_1^t Z_1$$

and

$$R_{22} = Z_2^t Z_2$$

would be intercorrelation matrices of variables from V_1 and from V_2, estimated on the set E under the criterion of maximum likelihood, and

$$R_{12} = Z_1^t Z_2 = R_{21}^t$$

would be the cross correlation matrix of variables from V_1 and from V_2, estimated also under the maximum likelihood criterion.

3. CANONICAL CORRELATION ANALYSIS

Canonical correlation analysis can be defined as the solution of the problem

$$\left.\begin{array}{l} Z_1 x_{1p} = k_{1p} \\ Z_2 x_{2p} = k_{2p} \end{array}\right| \begin{array}{l} \rho_p = k_{1p}^t k_{2p} = \text{maximum} \\ p = 1,....,s;\ s = \min(m_1, m_2) \\ k_{1p}^t k_{1q} = \delta_{pq} \\ k_{2p}^t k_{2q} = \delta_{pq} \\ k_{1p}^t k_{2q} = 0 \,|\, p \neq q. \end{array}$$

where δ_{pq} is the Kronecker delta.

This problem can be solved in several ways (Anderson, 1984). Probably the simplest solution is the maximization of the function

$$f(x_{1p}, x_{2p}, \lambda_p, \eta_p) = \rho_p - \tfrac{1}{2}\lambda_p(k_{1p}^t k_{1p} - 1) - \tfrac{1}{2}\eta_p(k_{2p}^t k_{2p} - 1)$$
$$= x_{1p}^t R_{12} x_{2p} - \tfrac{1}{2}\lambda_p(x_{1p}^t R_{11} x_{1p} - 1) - \tfrac{1}{2}\eta_p(x_{2p}^t R_{22} x_{2p} - 1)$$

for $p = 1$ where x_{1p} is some unknown m_1- dimensional vector, x_{2p} some unknown m_2-dimensional vector, with λ_p and η_p some unknown Lagrangeian multipliers.

After equating the obtained results to zero in order to obtain maximum, the derivation of this function with respect to x_{1p} and x_{2p} gives,

$$\partial f / \partial x_{1p} = R_{12}x_{2p} - \lambda_p R_{11}x_{1p} = 0$$
$$\partial f / \partial x_{2p} = R_{21}x_{1p} - \eta_p R_{22}x_{2p} = 0.$$

Multiplying from the left first result with x_{1p}^t and second result with x_{2p}^t,

from the conditions $x_{1p}^t R_{11} x_{1p} = 1$

and $x_{2p}^t R_{22} x_{2p} = 1$

we obtain

$$x_{1p}^t R_{12} x_{2p} = x_{2p}^t R_{21} x_{1p} = \lambda_p = \eta_p.$$

The first result can now be written in the form

$$R_{12}x_{2p} = \lambda_p R_{11}x_{1p}$$

and the second in the form $R_{21}x_{1p} = \lambda_p R_{22}x_{2p}$.

Multiplying the first result with R_{11}^{-1}, after rearrangement

$$x_{1p} = R_{11}^{-1}R_{12}\lambda_p x_{2p}\lambda_p^{-1}$$

and multiplying the second result with λ_p

$$\lambda_p R_{21}x_{1p} = \lambda_p^2 R_{22}x_{2p}.$$

From the solution for x_{1p}

$$R_{21}R_{11}^{-1}R_{12}x_{2p} = \lambda_p^2 R_{22}x_{2p}$$

and the problem can be reduced to solving the characteristic equation

$$(R_{21}R_{11}^{-1}R_{12} - \lambda_p^2 R_{22})x_{2p} = 0, \quad p = 1, \ldots, s$$

Obviously, $\rho_p = \lambda_p$ and canonical correlation analysis is essentially the transformation of two oblique coordinate systems to two biorthogonal coordinate systems such that cosines of the angles between equally indexed coordinates are maximized.

Up to this point canonical correlation analysis is actually a data analysis, and not a statistical method; obviously, a data analysis method can be considered as a true statistical method if some possibilities exists to test hypotheses concerning a set of parameters. Of course, the most important is the set of canonical correlations $R = (r_p;\ p = 1,....,s)$.

The usual way to test the hypothesis that in population P canonical correlations $r_p;\ p = 1,....,s$ are equal to zero is sequential application of the Bartlett test (Bartlett, 1941; Lawley, 1959)

$$\chi_p^2 = -(n-(m_1+m_2+3)/2)\sum_p^s \log_e(1-\rho_p^2)$$

because under the hypothesis $H_{0p} : r_p = 0$, the variable χ_p^2 has χ^2 distribution with

$$\nu_p = (m_1 - p + 1)(m_2 - p + 1)$$

degrees of freedom.

However, thus obtained probabilities are not independent (Anderson, 1984) and, in addition, the statistical significance of a canonical correlation is not necessarily related to the importance of corresponding canonical variates. It is well known that it is sufficient for only two variables v_j from V_1 and v_k from V_2 to be related with a large coefficient of correlation for generation of two canonical variates related with a large and therefore significant canonical correlation, although the importance of so generated canonical variates may be very small or negligible.

The asymptotic variance of canonical correlation (Kendall & Stuart, 1976)

$$\zeta_p^2 = (1-\rho_p^2)^2 n^{-1},$$

is equal to the asymptotic variance of any Pearson-Bravais coefficient of correlation. This fact can produce confusion in decision making procedures, because the decision concerning the hypothesis that sets U_1 and U_2 are not related depends on the sampling of variables in the sets V_1 and V_2; one or more correlations r_{jk} from R_{12} can be "significant", although canonical correlation ρ_1 can be statistically equal to zero.

When canonical correlation analysis was first used, the identification of canonical variates was based on inspection of the pattern of vectors x_{1p} and x_{2p}, $p = 1,....t$ where t is the number of canonical variates related with significant canonical correlations on the predetermined level of type I error. However, the elements of vectors x_{1p} and x_{2p} are proportional to the coordinates of vectors k_{1p} and

k_{2p} in the space spanned by vectors from Z_1 and Z_2; and because the cosines of the angles among coordinates are equal to the correlation coefficients in R_1 and R_2 in most real cases identification based on x_{1p} and x_{2p} is a formidable task with very uncertain outcome. This, and obvious relations of canonical correlation analysis with factor analysis were the main reason for the practice, probably first proposed by Cooley and Lohnes, which consists of identifying canonical variates based on the pattern of structural vectors

$$f_{1p} = Z_1 k_{1p} = R_{11} x_{1p}$$

and

$$f_{2p} = Z_2^t k_{2p} = R_{22} x_{2p},$$

i.e. on the pattern of simple correlation coefficients of variables from Z_1 and canonical variates k_{1p}, and variables from Z_2 and canonical variates k_{2p}, and even on the pattern of cross structural vectors

$$c_{1p} = Z_1^t k_{2p} = R_{12} x_{2p}$$

and

$$c_{2p} = Z_2^t k_{1p} = R_{21} x_{1p}$$

i.e. on the pattern of simple correlation coefficients of variables from Z_1 and canonical variates k_{2p}, and variables from Z_2 and canonical variates k_{1p} (Cooley & Lohnes, 1971).

However, because

$$x_{1p} = R_{11}^{-1} R_{12} x_{2p} \lambda_p^{-1}$$

and

$$x_{2p} = R_{22}^{-1} R_{21} x_{1p} \lambda_p^{-1}$$

cross structural vectors can be written in the form

$$c_{1p} = R_{12} R_{22}^{-1} R_{21} x_{1p} \rho_p^{-1}$$
$$c_{2p} = R_{21} R_{11}^{-1} R_{12} x_{2p} \rho_p^{-1}$$

so it is clear that c_{1p} and c_{2p} are actually structural vectors of the variables of one of the sets projected in the space spanned by variables of the other set. The characteristics of data analysis procedures which generate this type of latent dimensions were considered in the framework of generalized image transformations (Momirović, Štalec & Zakrajšek, 1973; Dobrić, Karaman & Momirović, 1983).

For the same reasons, structural vectors of canonical variates can be written in the form

$$f_{1p} = R_{12}x_{2p}\rho_p^{-1} = c_{1p}\rho_p^{-1}$$

$$f_{2p} = R_{21}x_{1p}\rho_p^{-1} = c_{2p}\rho_p^{-1};$$

so that

$$c_{1p} = f_{1p}\rho_p$$

$$c_{2p} = f_{2p}\rho_p$$

and structural and cross structural vectors are collinear and have only different norms. These simple findings are of great importance for an understanding of the real meaning of procedures strongly related to canonical correlation analysis, usually referred to as redundancy analysis.

4. REDUNDANCY ANALYSIS

Redundancy analysis is actually a generic term for several procedures for the determination of the amount of information emitted by latent dimensions of two sets of related variables and transferred from one set to the other.

Two of them are not directly related to canonical correlation analysis. The method of redundancy analysis proposed by Van den Wollenberg (1977) as an alternative to canonical correlation analysis generate sets of variables by maximizing of amount of information transferred from one to another and generating, of course, asymmetric relations between the two sets. The method, proposed in the framework of canonical covariance analysis (Momirović, Dobrić & Karaman, 1983; 1984; Prot, Bosnar & Momirović, 1983) is essentially an asymmetric approach to the basic method which consists in maximizing of covariances of not necessarily orthogonal latent dimensions from two sets of variables.

However, in the community of nonstatisticians and nonprofessional users of commercial statistical program packages, the term redundancy analysis is usually restricted to two additional operations associated to canonical correlation analysis, proposed by Stewart and Love (1968) and Miller (1969) together with a general measure of association between two sets of variables, known under the name canonical correlation index which is actually the simple mean of the canonical coefficients of determination. Both have been adopted and implemented in most program packages under the influence of the well-known book by Cooley and Lohnes (1971), probably the most influential text among people with a very modest mathematical and statistical education.

Thus, along with redundancy analysis, two additional measures, associated to canonical variates are considered, almost without exception. The first is usually referred to as variance and the second as the redundancy of canonical variates generated by the standard algorithm for canonical correlation analysis with the addition of structural and cross structural vectors of canonical factors.

Variance is defined by operations .

$$\xi_{1p}^2 = f_{1p}^t f_{1p} = x_{1p}^t R_{11}^2 x_{1p}$$

$$\xi_{2p}^2 = f_{2p}^t f_{2p} = x_{2p}^t R_{22}^2 x_{2p}$$

with the division of these values by m_1 and m_2, respectively. Redundancy is the result of operations

$$\phi_{1p}^2 = c_{1p}^t c_{1p} = x_{2p}^t R_{21} R_{12} x_{2p}$$

$$\phi_{2p}^2 = c_{2p}^t c_{2p} = x_{1p}^t R_{12} R_{21} x_{1p}$$

also with division of the obtained results by m_1 and m_2.

So-defined variance is considered to be a measure of discrimination of entities by canonical variables, and so-defined redundancy a measure of the amount of information transferred from one set of variables to another.

However, from the relations between vectors x_{1p} and x_{2p}, variances can be written in the form

$$\xi_{1p}^2 = x_{2p}^t R_{21} R_{12} x_{2p} \rho_p^{-2}$$

$$\xi_{2p}^2 = x_{1p}^t R_{12} R_{21} x_{1p} \rho_p^{-2}$$

so that redundancy measures can be defined simply as

$$\phi_{1p}^2 = \xi_{1p}^2 \rho_p^2$$

$$\phi_{2p}^2 = \xi_{2p}^2 \rho_p^2.$$

Therefore, this approach to redundancy analysis depends on the definition of the real variance of canonical variables.

The question of the adequacy of ξ_{1p}^2 and ξ_{2p}^2 as the estimates of real variances of canonical variables was posed more than 20 years ago (Nicewander & Wood, 1974). Although the answer to this question has not been based on valid arguments (Miller, 1975), application of relative variance estimates ξ_{1p}^2/m_1, and ξ_{2p}^2/m_2, and redundancy measures ϕ_{1p}^2/m_1 and ϕ_{2p}^2/m_2 has been generally adopted, especially in the field of psychometrics and computational statistics. The reason for that was the incorrect assertion of Nicewander and Wood that the variance of canonical variables is by definition equal to 1, and the generally accepted misconcep-

tion that the sum of the squares of correlations of a set of variables with any linear combination of standardized variables from this set is the variance of this linear combination. However, the variances of canonical variables are only equal to 1 by convention, and not by definition; and because the squared norms of structural vectors are equal to the variances of latent dimensions only if the latent dimensions are defined as principal components, it is sensible to reconsider the solution of the problem of real variances of canonical variables.

5. REAL VARIANCES OF CANONICAL VARIABLES

Definition 1: The standardized variance of a linear combination of standardized variables induced by a vector y, $y^t y = 1$ is

$$\sigma^2 = y^t R y,$$

the quadratic form of the intercorrelation matrix of variables induced by vector y.

Now let vectors x_{1p} and x_{2p} be standardized by operations

$$y_{1p} = x_{1p}(x_{1p}^t x_{1p})^{-1/2}$$

$$y_{2p} = x_{2p}(x_{2p}^t x_{2p})^{-1/2}.$$

Define the unstandardized canonical variables as

$$h_{1p} = Z_1 y_{1p}$$

$$h_{2p} = Z_2 y_{2p}.$$

The variances of these variables are

$$\sigma_{1p}^2 = h_{1p}^t h_{1p} = y_{1p}^t R_{11} y_{1p} = (x_{1p}^t x_{1p})^{-1}$$

$$\sigma_{2p}^2 = h_{2p}^t h_{2p} = y_{2p}^t R_{22} y_{2p} = (x_{2p}^t x_{2p})^{-1}$$

since

$$x_{1p}^t R_{11} x_{1p} = x_{2p}^t R_{22} x_{2p} = 1.$$

But, because the canonical correlation can be defined as

$$\rho_p = v_{1p}^t R_{11}^{-1/2} R_{12} R_{22}^{-1/2} v_{2p}$$

where v_{1p}, $v_{1p}^t v_{1p} = 1$ are the left, and v_{2p}, $v_{2p}^t v_{2p} = 1$ the right eigenvectors of the cross correlation matrix of variables from V_1 and V_2 transformed to the Mahalanobis form

$$M_{12} = R_{11}^{-1/2} R_{12} R_{22}^{-1/2},$$

associated to singular values of this matrix (Hadžigalić, Bogdanović, Tenjović & Wolf, 1994), the standardized variances of canonical variables can be written in the form

$$\sigma_{1p}^2 = (v_{1p}^t R_{11}^{-1} v_{1p})^{-1}$$

$$\sigma_{2p}^2 = (v_{2p}^t R_{22}^{-1} v_{2p})^{-1}$$

because vectors x and x_{1p} and x_{2p} are equal to

$$x_{1p} = R_{11}^{-1/2} v_{1p}$$

and

$$x_{2p} = R_{22}^{-1/2} v_{2p}.$$

The quasi variances ξ_{1p}^2 and ξ_{2p}^2 can now be written in the form

$$\xi_{1p}^2 = v_{1p}^t R_{11} v_{1p}$$

$$\xi_{2p}^2 = v_{2p}^t R_{22} v_{2p}$$

and it is clear that ξ_{1p}^2 and ξ_{2p}^2 are not equal to the real variances of the canonical variables, because vectors v_{1p} are not eigenvectors of R_{11}, nor are v_{2p} eigenvectors of R_{22}.

This fact has some serious consequences on the procedures for the estimation of redundancies as well as on the procedures for the estimation of the generalizability of canonical variables, and therefore on the procedures for the estimation of the amount of emitted information of any canonical variable. Let us consider first some rational procedures for the estimation of redundancy.

6. MEASURES OF THE REAL REDUNDANCY OF CANONICAL VARIABLES

A rational measure of the redundancy of canonical variables can be derived from the formal definition of redundancy.

Definition 2: Redundancy is the part of the variance of an unstandardized canonical variable generated from a set of variables which can be attributed to the variables from another set of variables.

Let

$$h_{2p} = Z_2 x_{2p} \sigma_{2p}$$

be some canonical variable generated from the set V_2.

Define the regression problem

$$Z_1 \beta_{1p} = h_{2p} - e_{2p} \mid \varepsilon_{2p}^2 = e_{2p}^t e_{2p} = \text{minimum}$$

where β_{1p} is some unknown m_1 - dimensional vector.

Because

$$\varepsilon_{2p}^2 = (h_{2p} - Z_1 \beta_{1p})^t (h_{2p} - Z_1 \beta_{1p}) =$$
$$= \sigma_{2p}^2 - 2\sigma_{2p} \beta_{1p}^t R_{12} x_{2p} + \beta_{1p}^t R_{11} \beta_{1p},$$

by derivation of this function with respect to β_{1p}

$$\partial f(\beta_{1p}) / \partial \beta_{1p} = -2\sigma_{2p} R_{12} x_{2p} + 2R_{11} \beta_{1p};$$

dividing by 2 and equating to zero, we get

$$\sigma_{2p} R_{12} x_{2p} = R_{11} \beta_{1p}.$$

Multiplying by R_{11}^{-1}

$$\beta_{1p} = R_{11}^{-1} R_{12} x_{2p} \sigma_{2p};$$

but because

$$R_{11}^{-1} R_{12} x_{2p} = x_{1p} \rho_p,$$

$$\beta_{1p} = x_{1p} \rho_p \sigma_{2p}.$$

The residual variance of canonical variable h_{2p} is therefore

$$\varepsilon_{2p}^2 = (Z_2 x_{2p} \sigma_{2p} - Z_1 x_{1p} \rho_p \sigma_{2p})^t (Z_2 x_{2p} \sigma_{2p} - Z_1 x_{1p} \rho_p \sigma_{2p}) =$$
$$= \sigma_{2p}^2 - \sigma_{2p}^2 \rho_p^2,$$

so that

$$\omega_{2p}^2 = \sigma_{2p}^2 \rho_p^2$$

is the measure of redundancy of canonical variable h_{2p}. In the same way it can be demonstrated that

$$\omega_{1p}^2 = \sigma_{1p}^2 \rho_p^2$$

are redundancy measures of canonical variables h_{1p}.

Although formally similar to ϕ_{1p}^2 and ϕ_{2p}^2 the measures ω_{1p}^2 and ω_{2p}^2 are different not only because they are defined on the basis of real variances of canonical variables, but also because they are derived independently of the norms of structural and cross structural vectors; obviously, the norms of both structural and cross structural vectors have nothing in common with the variances and redundancy measures of canonical factors.

Of course, because the covariances of unstandardized canonical variables are

$$\vartheta_p = h_{1p}^t h_{2p} = \rho_p \sigma_{1p} \sigma_{2p},$$

the redundancy measures of canonical variables can also be derived in an another way, solving the regression problem

$$h_{1p}\psi_{1p} = h_{2p} - g_{2p} \mid \varepsilon_{2p}^2 = g_{2p}^t g_{2p} = \text{minimum}$$

where ψ_{1p} is some unknown scalar.

Because

$$\varepsilon_{2p}^2 = \sigma_{2p}^2 - 2\psi_{1p}\vartheta_p + \psi_{1p}^2 \sigma_{1p}^2$$

the derivation of this function with respect to ψ_{1p} gives

$$\partial f(\psi_{1p}) / \partial \psi_{1p} = -2\vartheta_p + 2\psi_{1p}\sigma_{1p}^2$$

and dividing by 2 and equating to zero, we get

$$\psi_{1p}\sigma_{1p}^2 = \vartheta_p.$$

Multiplying by σ_{1p}^{-2}

$$\psi_{1p} = \vartheta_p \sigma_{1p}^{-2} = \rho_p \sigma_{2p} \sigma_{1p}^{-1}$$

so that

$$\varepsilon_{2p}^2 = (h_{2p} - h_{1p}\psi_{1p})^t (h_{2p} - h_{1p}\psi_{1p}) = \sigma_{2p}^2 - \rho_p^2 \sigma_{2p}^2$$

we have

$$\omega_{2p}^2 = \rho_p^2 \sigma_{2p}^2$$

which is the redundancy measure of canonical variable h_{2p} with canonical variable h_{1p}.

In the same way,

$$\omega_{1p}^2 = h_{2p}^t h_{2p} \psi_{2p}^2 = \rho_p^2 \sigma_{1p}^2$$

is the redundancy measure of canonical variable h_{1p} with canonical variable h_{2p}.

Note, by the way, that

$$\psi_{2p} = \rho_p \sigma_{1p} \sigma_{2p}^{-1}$$

so that coefficients ψ_{1p} and ψ_{2p} are very sensitive measures of the asymmetry of relations between canonical variables.

7. ESTIMATION OF THE RELIABILITY OF CANONICAL VARIABLES

Let N_1 be some, perhaps unknown, matrix of measurement errors or estimation of variables from V_1, and let N_2 be some, also perhaps unknown, matrix of measurement errors or estimation of variables from V_2. In this case, the elements of matrices

$$T_1 = Z_1 - N_1$$

and

$$T_2 = Z_2 - N_2$$

would be true scores of objects from E on the variables from V_1 and V_2.

Suppose that the postulates of the classic theory of measurements are valid, so that

$$E_1^2 = N_1^t N_1$$

and

$$E_2^2 = N_2^t N_2$$

are diagonal matrices, and

$$T_1^t N_1 = 0$$

and

$$T_2^t N_2 = 0$$

are zero matrices. Then

$$C_{11} = T_1^t T_1 = R_{11} - E_1^2$$

and

$$C_{22} = T_2^t T_2 = R_{22} - E_2^2$$

would be covariance matrices of true scores of objects from E on the variables from V_1 and V_2.

True variances of canonical variables, induced by vectors y_{1p} and y_{2p}, would be, therefore,

$$\tau_{1p}^2 = y_{1p}^t C_{11} y_{1p} = \sigma_{1p}^2 - \gamma_{1p}^2$$

and

$$\tau_{2p}^2 = y_{2p}^t C_{22} y_{2p} = \sigma_{2p}^2 - \gamma_{2p}^2$$

where

$$\gamma_{1p}^2 = y_{1p}^t E_1^2 y_{1p}$$

and

$$\gamma_{2p}^2 = y_{2p}^t E_2^2 y_{2p}$$

are error variances of canonical variables.

On the basis of formal definitions of reliability measures

$$\alpha = \sigma_t^2 / \sigma^2 = 1 - \sigma_e^2 / \sigma^2$$

where σ_t^2 is the variance of true scores, σ_e^2 the variance of error scores, and σ^2 the total variance of some variable, the coefficients of reliability of canonical variables would be

$$\alpha_{1p} = 1 - \gamma_{1p}^2 / \sigma_{1p}^2 = 1 - y_{1p}^t E_1^2 y_{1p} / y_{1p}^t R_{11} y_{1p}$$

and

$$\alpha_{2p} = 1 - \gamma_{2p}^2 / \sigma_{2p}^2 = 1 - y_{2p}^t E_2^2 y_{2p} / y_{2p}^t R_{22} y_{2p}.$$

Of course, the matrices N_1 and N_2 are generally unknown so that matrices E_1^2 and E_2^2 are also unknown. However, if the data are obtained by adequate measuring instruments, some estimation of the reliability of these instruments is almost always known. Let the elements of diagonal matrices

$$A_1 = (\alpha_j) \quad j = 1,\ldots,m_1$$

and

$$A_2 = (\alpha_k) \quad k = 1,\ldots,m_2$$

be coefficients of reliability of measuring instruments which generate variables from V_1 and V_2. The matrices E_1^2 and E_2^2 can be estimated by

$$E_1^2 = I_1 - A_1$$

$$E_2^2 = I_2 - A_2$$

so that

$$\beta_{1p} = 1 - (y_{1p}^t y_{1p} - y_{1p}^t A_1 y_{1p}) / (y_{1p}^t R_{11} y_{1p})$$

$$= 1 - (1 - y_{1p}^t A_1 y_{1p}) / \sigma_{1p}^2$$

and

$$\beta_{2p} = 1 - (y_{2p}^t y_{2p} - y_{2p}^t A_2 y_{2p}) / (y_{2p}^t R_{22} y_{2p})$$

$$= 1 - (1 - y_{2p}^t A_2 y_{2p}) / \sigma_{2p}^2$$

are reliability estimates of canonical variables.

In the case of when the elements of matrices A_1 and A_2 are unknown but sets V_1 and V_2 are samples from the universes U_1 and U_2 defined over different but homogeneous fields, the upper bound of the variance of variable measurement error can be estimated by (Guttman, 1945)

$$E_1^2 = (diag\ R_{11}^{-1})^{-1} = U_1^2$$

and

$$E_2^2 = (diag\ R_{22}^{-1})^{-1} = U_2^2$$

so that the estimates of the lower bound to the reliability of canonical variables are

$$\beta_{61p} = 1 - (y_{1p}^t U_1^2 y_{1p}) / (y_{1p}^t R_{11} y_{1p})$$

and

$$\beta_{62p} = 1 - (y_{2p}^t U_2^2 y_{2p}) / (y_{2p}^t R_{22} y_{2p}).$$

Of course, it is now easy to derive the estimation of the absolute lower bound to the reliability if we define $A_1 = 0$ and $A_2 = 0$, so that

$$\beta_{11p} = 1 - (y_{1p}^t y_{1p}) / (y_{1p}^t R_{11} y_{1p})$$

$$= 1 - (y_{1p}^t R_{11} y_{1p})^{-1}$$

$$= 1 - \sigma_{1p}^{-2}$$

and

$$\beta_{22p} = 1-(y_{2p}^t y_{2p})/(y_{2p}^t R_{22} y_{2p})$$

$$= 1-(y_{2p}^t R_{22} y_{2p})^{-1}$$

$$= 1-\sigma_{2p}^{-2}$$

8. INFORMATION MEASURES

Definition 3: The measure of information emitted by some linear combination of standardized variables, induced by vector y, $y^t y = 1$ is

$$\iota^2 = (1-\alpha)^{-1} = y^t C y + 1$$

where α is the coefficient of reliability of the linear combination and C is the covariance matrix of true scores on the standardized variables.

Therefore, the information measure of some linear combination of standardized variables is defined as the ability of the so-defined composite variable to differentiate the objects at least as well as any element of the composite.

On the basis of reliability measures β, β_6 and β_1 the following measures of the informativeness of canonical variables can be defined:

(1) True measures of informativeness

$$\iota_{1p}^2 = (1-\beta_{1p})^{-1} = y_{1p}^t (R_{11}-E_1^2) y_{1p} + 1$$

and

$$\iota_{2p}^2 = (1-\beta_{2p})^{-1} = y_{2p}^t (R_{22}-E_2^2) y_{2p} + 1;$$

(2) Measures of the lower bound to informativeness

$$\iota_{61p}^2 = (1-\beta_{61p})^{-1} = y_{1p}^t (R_{11}-U_1^2) y_{1p} + 1$$

and

$$\iota_{62p}^2 = (1-\beta_{62p})^{-1} = y_{2p}^t (R_{22}-U_2^2) y_{2p} + 1;$$

and, finally

(3) Measures of the absolute lower bound to informativeness

$$\iota_{11p}^2 = (1-\beta_{11p})^{-1} = y_{1p}^t R_{11} y_{1p} = \sigma_{1p}^2$$

and

$$\iota_{12p}^2 = (1-\beta_{12p})^{-1} = y_{2p}^t R_{22} y_{2p} = \sigma_{2p}^2.$$

The information measures, as the exponential functions of reliability measures, clearly express the real values of canonical variables for the determination of the position of objects in the space spanned by canonical factors, and can be therefore a firm basis for decisions concerning the real value of results obtained by canonical correlation analysis.

9. AN INFORMATION CRITERION FOR DETERMINATION OF THE NUMBER OF IMPORTANT CANONICAL FACTORS

The rule to accept and retain only canonical variables related by significant canonical correlations is not always a wise strategy for several different, but mutually related reasons.

(1) Probabilities associated to the sequential Bartlett test of the hypotheses $H_{Op}: r_p = 0$: $p = 1,\ldots,s$ are not independent, so that decisions based on probabilities p_{Op} are also not independent;

(2) Some canonical correlations may be very high, and therefore significant only as the consequence of very high correlations between variables v_j from V_1 and v_k from V_2, although the reliabilities and informativeness of canonical variables related by such spurious canonical correlations may be very low;

(3) If $(n-(m_1+m_2)/2)$ is not a sufficiently large number in many instances the hypothesis H_{o1} cannot be rejected, but is quite possible that some hypothesis of the type $H_{Ojk}: r_{jk} = 0$:, where r_{jk} are correlations between v_j and v_k in P, can be rejected on the same type I error; in this case it is impossible to overcome the logical confusion and to decide if the universes U_1 and U_2 are related or are stochastically independent;

(4) From the scientific, as well as from the application point of view, the information that two well defined and reliably estimated variables are stochastically independent can be as important as the information that two, possibly not very well-defined, latent variables are related by canonical correlation different from zero.

Therefore, it is sensible to consider some possible criteria for determining the importance of canonical variables.

Of course, the distinction between the concepts of significance and importance must be clearly defined. Obviously, only canonical variables with proven existence can be considered important, that is with sufficiently high coefficients of reliability.

Let β^* be some selected lower limit of reliability such that a pair of canonical variables can be considered important. Some pair (k_{1p}, k_{2p}) of canonical variables are important under the condition that

$$\beta_{1p} > \beta^* \vee \beta_{2p} > \beta^*$$

if coefficients β are known, or under the condition that

$$\beta_{61p} > \beta^* \vee \beta_{62p} > \beta^*$$

if coefficients β are unknown, but samples of variables V_1 and V_2 are representative samples from the universes U_1 and U_2 so that coefficients of type β_6 can be sensible measures of reliability. In the case of unknown β coefficients and non representative samples, a pair (k_{1p}, k_{2p}) of canonical variables can be considered important under the condition that

$$\beta_{11p} > 0 \vee \beta_{12p} > 0,$$

because the negative value of the absolute lower bound to reliability of any linear combination of standardized variables permits the conclusion that nothing similar to this linear combination exists with a probability close to 1.

The question of the significance of a canonical correlation is a sensible question only for pairs of canonical variables whose real existence has been proven with reasonably high probability. Of course, the significance of such a canonical correlation can be tested by the Bartlett procedure, but if sample E is reasonably large, decisions on the basis of confidence interval

$$\rho_p \mp \zeta_p * t_{\alpha/2}$$

where $t_{\alpha/2}$ is the value of the t - distribution with $(n-(m_1+m_2))$ degrees of freedom for a reliability of inference of 1 - α and ζ_p^2 is asymptotic variance of coefficient ρ_p, can also be taken into consideration.

REFERENCES

[1] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis* (2 edition), Wiley, New York, 1984.

[2] Bartlett, M. S., "The statistical significance of canonical correlation", *Biometrika*, 32 (1941) 29-38.

[3] Cooley, W.W., Lohnes, P.R., *Multivariate Data Analysis*, Wiley, New York, 1971.

[4] Dobrić, V., Karaman, Ž., Momirović, K., "LSD: A method, algorithm and program for latent structure decomposition, *Proceedings of the 7th Symposium on Informatics 'Jahorina '83'*, 1983, 1-7.

[5] Guttman, L., "A basis for analysis test-retest reliability", *Psychometrika*, 10 (1945) 255-282.

[6] Hadžigalić, S., Bogdanović, M., Tenjović, L., Wolf, B., "O nekim svojstvima Mahalanobisovih prostora", *Zbornik radova 8 Sekcije za klasifikacije SSDJ*, Savezni zavod za statistiku, 1994, 99-132.

[7] Horst, P., *Factor Analysis of Data Matrices*, Holt, Rinehart and Winston, New York, 1965.

[8] Hotelling, H., "The most predictable criterion". *Journal of Educational Psychology*, 26 (1935) 139-142.

[9] Hotelling, H., "Relations between two sets of variates", *Biometrika*, 28 (1936) 321-377.

[10] Lawley, D. N., "Tests of significance in canonical analysis", *Biometrika*, 46 (1959) 59-66.

[11] Kendall, M. G., *A Course in Multivariate Analysis* (4th edition), Griffin, London, 1968.

[12] Kendall, M. G., Stuart, A., *Mnogomernyj statističeskij analiz i vremennie rjady (perevod E. L. Presmana i V. I. Rotarja)*, Nauka, Moskva, 1976.

[13] Miller, J.K., "The development and application of bi-multivariate correlation: a measure of statistical association between multivariate measurement sets", Dissertation, State University of New York and Buffalo, 1969.

[14] Miller, J.K.," In defense of the general canonical correlation index: Reply to Nicewander and Wood", *Psychological Bulletin*, 82 (1975) 207-209.

[15] Momirović, K., Štalec, J., Zakrajšek, E., "Primjena generaliziranih image transformacija u analizi relacija skupova varijabli", *Kineziologija*, 3 (2) (1973) 57-61.

[16] Momirović, K., Dobrić, V., Karaman, Ž.," Canonical covariance analysis", *Proceedings of the 5th International Symposium "Computer at the University"*, 1983, 463-473.

[17] Momirović, K., Dobrić, V., Karaman, Z., "Algorithm and program for multicriterial selection with consistent linear constraints", in: N. Wolansky and A. Siniarska (eds.), *Genetics of Psychomotor Traits in Man*, Polish Academy of Science, Warsaw, 283-293.

[18] Nicewander, W.A., Wood, D.A., "Comments on "A general canonical correlations index"", *Psychological Bulletin*, 81 (19) (1974) 92-94.

[19] Prot, F., Bosnar, K., Momirović, K., "An algorithm for redundancy analysis of two sets of quantitative variates", *Proceedings of the 5th International Symposium "Computer at the University"*, 1983, 475-484.

[20] Rao, C. R., *Linear Statistical Inference and Its Application*, Wiley, New York, 1973.

[21] Stewart, D.K., Love, W.A., "A general canonical correlation index", *Psychological Bulletin*, 70 (1968) 160-163.

[22] Van den Wollenberg, A.L., "Redundancy analysis-an alternative model for canonical correlation analysis", *Psychometrika*, 42 (1977) 207-219.