Appl. Statist. (1976),
25, *No.* 3, p. 257

A Unifying Tool for Linear Multivariate Statistical Methods: The *RV*-Coefficient

By P. ROBERT and Y. ESCOUFIER

Université de Montréal, Canada

Université des Sciences et Techniques du Languedoc, Montpellier, France

[Received October 1975. Revised May 1976]

SUMMARY

Consider two data matrices on the same sample of n individuals, $X(p \times n)$, $Y(q \times n)$. From these matrices, geometrical representations of the sample are obtained as two configurations of n points, in $\mathscr{R}^p$ and $\mathscr{R}^q$. It is shown that the RV-coefficient (Escoufier, 1970, 1973) can be used as a measure of similarity of the two configurations, taking into account the possibly distinct metrics to be used on them to measure the distances between points. The purpose of this paper is to show that most classical methods of linear multivariate statistical analysis can be interpreted as the search for optimal linear transformations or, equivalently, the search for optimal metrics to apply on two data matrices on the same sample; the optimality is defined in terms of the similarity of the corresponding configurations of points, which, in turn, calls for the maximization of the associated RV-coefficient. The methods studied are principal components, principal components of instrumental variables, multivariate regression, canonical variables, discriminant analysis; they are differentiated by the possible relationships existing between the two data matrices involved and by additional constraints under which the maximum of RV is to be obtained. It is also shown that the RV-coefficient can be used as a measure of goodness of a solution to the problem of discarding variables.

Keywords: LINEAR MULTIVARIATE ANALYSIS; PRINCIPAL COMPONENTS; MULTIVARIATE REGRESSION; CANONICAL VARIABLES; DISCRIMINANT ANALYSIS; DISCARDING VARIABLES; RV-COEFFICIENT; SPACIAL CONFIGURATIONS

1. INTRODUCTION

HAVING observed the values of p numerical variables on each individual of a given sample, it is customary to arrange the data into a $p \times n$ matrix $X = (x_{ij})$. The ith row of X, denoted by X_i, contains the n values of the ith variable while the jth column, denoted by X^j, contains the p observations recorded on the jth individual.

A common geometrical representation of the sample consists of a canonical mapping of the data matrix X into a "configuration" of n points in the p-dimensional space $\mathscr{R}^p$. Our interest will focus on the pattern of such a configuration or, equivalently, on the set of distances between its points. The distance between the jth and the kth points (individuals) is defined, by use of a positive semi-definite matrix Q, to be equal to $\{(X^j - X^k)' Q (X^j - X^k)\}^{\frac{1}{2}}$. Frequent choices for Q are the standardizing diagonal matrix having the inverses of the variances on the main diagonal and the inverse of the covariance matrix which makes distances independent of linear transformations on the data.

Given any positive semi-definite matrix Q one can find $p \times q$ matrices L(q not necessarily equal to p) such that $Q = LL'$. Therefore there is an equivalence between the choice of the metric defined by Q on points in $\mathscr{R}^p$ and the linear change of variables giving the new data matrix $Y = L'X$ followed by the use of the ordinary sum of squares metric on the configuration representing Y in $\mathscr{R}^q$. This equivalence is fundamental for the purpose of this paper.

We propose a unified view at some of the classical linear multivariate statistical methods (principal components, multivariate regression, canonical correlations, discriminant analysis) by showing that all can be interpreted as the search for linear transformations, L, of original variables, X, that will maximize, under constraints characteristic of each method, the "closeness" of the configurations of points associated with X and $Y = L'X$. The measure of closeness to be retained is based on the sample value of Escoufier's RV-coefficient (Escoufier, 1970, 1973) and is established in the next section. Our results are summarized in Section 3 with details given in Section 4. In Section 5 suggestions are made for the use of the RV-coefficient in the problem of discarding variables.

2. Comparison Between two Configurations of Points Representing the Same Individuals

Consider a given sample of n individuals on which two sets of observations have been made, giving a $p \times n$ data matrix X and a $q \times n$ data matrix Y. (The variables in the two sets may be partially or totally distinct.) Let $C(X)$ and $C(Y)$ be the two associated configurations, in $\mathscr{R}^p$ and $\mathscr{R}^q$, respectively. To see the extent to which the two sets of variables give similar images of the n individuals, we select a particular matrix to characterize each configuration and use a measure of closeness of these two matrices.

First, as a measure of the *relative* positions of points in a configuration, say $C(X)$, we could use the $n \times n$ distance matrix $D(X)$, with (j,k)th element equal to $\{(X^j - X^k)'(X^j - X^k)\}^{\frac{1}{2}}$, which is translation and rotation independent. Assuming that all variables have been centred to have means equal to 0, we prefer to use the matrix $S(X)/\{\mathrm{tr}\, S(X)^2\}^{\frac{1}{2}}$, where $S(X) = X'X$. This matrix is translation and rotation independent; the scalar denominator $\{\mathrm{tr}\, S(X)^2\}^{\frac{1}{2}}$ ensures that it is also independent of global changes of scale.

It is known that for square matrices A and B, $\mathrm{tr}\, A'B$ is a scalar product and that the corresponding norm for A is $\|A\| = (\mathrm{tr}\, A'A)^{\frac{1}{2}}$. (With this definition $\|S(X)/\{\mathrm{tr}\, S(X)^2\}^{\frac{1}{2}}\| = 1$.) We shall therefore measure the distance between the configurations $C(X)$ and $C(Y)$ by

$$\begin{aligned}\mathrm{dist}\{C(X), C(Y)\} &= \|S(X)/\{\mathrm{tr}\, S(X)^2\}^{\frac{1}{2}} - S(Y)/\{\mathrm{tr}\, S(Y)^2\}^{\frac{1}{2}}\| \\ &= \sqrt{(2)}\,[1 - \mathrm{tr}\{S(X)\,S(Y)\}/\{\mathrm{tr}\, S(X)^2 . \mathrm{tr}\, S(Y)^2\}]^{\frac{1}{2}} = \sqrt{(2)}\,[1 - RV(X, Y)]^{\frac{1}{2}},\end{aligned}$$

with

$$RV(X, Y) = \{\mathrm{tr}(X'X . Y'Y)\}/\{\mathrm{tr}(X'X)^2 . \mathrm{tr}(Y'Y)^2\} = \{\mathrm{tr}(XY' . YX')\}/\{\mathrm{tr}(XX')^2 . \mathrm{tr}(YY')^2\}.$$

Within a multiplicative factor of $1/n$, $S_{11} = XX'$, $S_{22} = YY'$, $S_{12} = XY'$ and $S_{21} = YX'$ are the sample covariance and cross covariance matrices of the variables defining X and Y. With these notations,

$$RV(X, Y) = \mathrm{tr}(S_{12} . S_{21})/(\mathrm{tr}\, S_{11}^2 . \mathrm{tr}\, S_{22}^2)^{\frac{1}{2}}.$$

This expression is analogous to the RV-coefficient originally defined by Escoufier (1970, 1973) for two random vectors on the same probability space. The link between the mathematical definition of RV using expectations and the intuitive, data-oriented approach taken in this paper is the fact that for sampled values the equality $\mathrm{tr}(S(X) . S(Y)) = \mathrm{tr}(S_{12} . S_{21})$ holds.

The coefficient $RV(X, Y)$ will be used as the actual measure of closeness of $C(X)$ and $C(Y)$. The value of $RV(X, Y)$ is in the closed interval $[0, 1]$ and the closer to 1 it is, the closer are the patterns and the better is Y (is X) as a substitute for X (for Y) to characterize the n individuals of the sample.

The reader will have noticed that our approach is the classical one, which assumes X known and D and S computed from X. The reversed approach is that of multidimensional scaling which attempts to determine X knowing D. A method due to Torgerson (1958) allows the computation of S knowing D. Provided S is positive semi-definite, X can be obtained by factorization. Without pursuing this point any further, let it be clear that the forthcoming

theory remains meaningful when the original data is a distance matrix D leading to a positive semi-definite matrix S.

The problem of comparing different multivariate analyses on the same individuals has been treated by a number of authors. In particular, Gower (1971) has used as a measure of closeness of two configurations a statistic equal to $\mathrm{tr}\,(S_{12})\,\mathrm{tr}\,(S_{12})$ under the assumption that the variables have been scaled in such a way that $\mathrm{tr}\,S_{11} = \mathrm{tr}\,S_{22} = 1$.

3. Summary of Results

Typically, the problems to be solved in Sections 4 and 5 can be stated as follows. Having observed two sets of variables on n individuals, giving a $p \times n$ data matrix X and a $q \times n$ data matrix Y, find linear combinations of the variables defining X to obtain a new data matrix $L'X$ and, also, linear combinations of the variables defining Y to obtain a new data matrix $M'Y$, in such a way that the "images" of the n individuals given by $L'X$ and $M'Y$ be as "similar" as possible. We shall interpret this as the search of the matrices L and M to maximize $RV(L'X, M'Y)$. Depending on the particular linear method of multivariate analysis under study, a choice of constraints will be imposed on the row-dimensions and the elements of L' and M'. (In certain cases the problem will be reduced by prespecifications, such as $Y = X$, $M = I$.)

Consider the statistic

$$RV(L'X, M'Y) = \{\mathrm{tr}\,(L'S_{12}MM'S_{21}L)\}/\{\mathrm{tr}\,(L'S_{11}L)^2 . \mathrm{tr}\,(M'S_{22}M)^2\}^{\frac{1}{2}}$$

and suppose that L is a $p \times t$ matrix. The choice of L to maximize RV is determined to within a rotation in $\mathscr{R}^t$ (i.e. L can be postmultiplied by any $t \times t$ matrix R such that $RR' = I_t$). Since a rotation in $\mathscr{R}^t$ is determined by $t(t-1)/2$ conditions, that number of *degrees of indeterminacy* can be eliminated by specifying the same number of constraints on L, or equivalently, on $L'X$. For example, we could require $L'S_{11}L$ to be a diagonal matrix. Similar sets of degrees of indeterminacy and constraints pertain to M or $M'Y$.

Table 1 gives a schematic list of the problems to be solved in Section 4, together with an indication of the nature of the solutions.

TABLE 1

Summary of results

X: $(p \times n)$; Y: $(q \times n)$; L and M: matrices to be determined

Initial data matrices	*Transformed data matrices (maximize RV of:)*	*Dimensions* L	*Dimensions* M	*Additional constraints to remove indeterminacy*	*Solution relates to:*
X	X $L'X$	$p \times t$		$L'S_{11}L$ is diagonal	First t principal components of X
X Y	X $M'Y$		$q \times t$	$M'S_{22}M$ is diagonal	First t principal components of Y with respect to X
X Y	X $M'Y$		$q \times p$	$S_{12}M - M'S_{22}M = 0$	Multivariate regression of X on Y
X Y	$L'X$ $M'Y$	$p \times t$	$q \times t$	$L'S_{11}L = M'S_{22}M = I_t$	First t pairs of canonical variables
X $Y=(Y_i^j)$	$L'X$ $M'Y$	$p \times t$	$p \times t$	$L'S_{11}L = I_t$; $M'S_{22}M$ is diagonal	First t discriminant hyperplanes ($L = M$)

where individuals are divided into g groups and Y_i^j = mean value of the ith variable in the group containing the jth individual.

In carrying the analyses of Section 4, we always reformulate any problem as that of the search for linear combinations of variables, leading to data matrices of the form $L'X$ and $M'Y$. We urge the reader to realize that the initial question asked by the statistician may be different but equivalent to this approach. It may be that of finding a proper metric, i.e. a positive semi-definite matrix Q, to apply to the initial data matrices; in accordance with our Introduction, we have implicitly selected to define Q by the product LL'. In other situations, it may be that of finding a proper set of t projection hyperplanes; each hyperplane in $\mathscr{R}^p$ being defined by a p-vector, such a set is represented by a $p \times t$ matrix L.

Our aim is only to show the unifying role that can be played by the RV-coefficient. Therefore, we do not develop any part of the classical theory of the methods referred to in Table 1. The reader should be familiar enough with this theory to see the possible extensions and benefits of our approach. (For references, see, for example, Anderson, 1958 and Cooley and Lohnes, 1971.)

4. Linear Multivariate Methods

4.1. *Principal Components of X*

Having observed p variables on each of n individuals, giving the $p \times n$ data matrix X, we seek a number of new variables, say $t \leqslant p$, which are linear combinations of the initial variables, with $Y = L'X$ (where L is a $p \times t$ matrix) as the new data matrix, such that the configurations $C(X)$ and $C(Y)$ be as similar as possible. To this aim, we shall maximize for L the statistic

$$RV(X, L'X) = \{\operatorname{tr}(S_{11}LL'S_{11})\}/\{\operatorname{tr} S_{11}^2 . \operatorname{tr}(L'S_{11}L)^2\}^{\frac{1}{2}}.$$

To remove the degrees of indeterminacy in this optimization problem we require that the new variables be uncorrelated, i.e. that

$$(L'X)(L'X)' = L'S_{11}L = \Delta = \operatorname{diag}(\delta_i), \tag{1}$$

where the δ_i's are to be determined.

Let λ_i $(i = 1, 2, \ldots, t)$ be arbitrary Lagrange multipliers. The problem is that of maximizing, under constraints (1), the function $\Phi(L) = \operatorname{tr}(S_{11}LL'S_{11}) - \sum \lambda_i [L'S_{11}L]_{ii}$. The derivative of $\Phi(L)$ with respect to L is equal to $2(S_{11}^2 L - S_{11}L\Lambda)$ where $\Lambda = \operatorname{diag}(\lambda_i)$. This derivative will vanish if $S_{11}L = L\Lambda$. Thus we can take, for the columns of L, t orthogonal eigenvectors of S_{11}, normalized to satisfy (1). Such a choice is possible so long as t does not exceed the number of non-zero δ_i's. The λ_i's will have to be the corresponding eigenvalues of S_{11} and we will have $\operatorname{tr}(S_{11}LL'S_{11}) = \operatorname{tr}(L'S_{11}^2 L) = \operatorname{tr}(L'S_{11}L\Lambda) = \operatorname{tr}(\Delta\Lambda)$,

$$RV(X, L'X) = \{(\operatorname{tr} S_{11}^2)^{\frac{1}{2}}\}^{-1} \left\{ \left(\sum_{i=1}^{t} \lambda_i \delta_i \right) \Big/ \left(\sum_{i=1}^{t} \delta_i^2 \right)^{\frac{1}{2}} \right\} = \left\{ \left(\sum_{i=1}^{t} \lambda_i^2 \right)^{\frac{1}{2}} \Big/ (\operatorname{tr} S_{11}^2)^{\frac{1}{2}} \right\} \left\{ \left(\sum_{i=1}^{t} \lambda_i \delta_i \right) \Big/ \left(\sum_{i=1}^{t} \lambda_i^2 \right)^{\frac{1}{2}} \right.$$

$$\left. \times \left(\sum_{i=1}^{t} \delta_i^2 \right)^{\frac{1}{2}} \right\}.$$

From the last ratio, it can be seen that given the λ_i's, an optimal choice for each δ_i is $\delta_i = \lambda_i$. Then, with the eigenvalues of S_{11} ordered so that $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$, the absolute maximum of RV is found to be

$$\max RV(X, L'X) = \left(\sum_{i=1}^{t} \lambda_i^2 \Big/ \sum_{i=1}^{p} \lambda_i^2 \right)^{\frac{1}{2}}.$$

If l^i is the ith column of L, we have $S_{11} l^i = \lambda_i l^i$ and the ith new variable is the ith principal component of X having variance equal to λ_i.

If t is equal to the rank of S_{11}, then the RV-coefficient is equal to 1. Further known results of principal component analysis could now be derived.

4.2. *Principal Components of Y with Respect to X*

The problem treated in this subsection is the following: given two data matrices X, $(p \times n)$ and Y, $(q \times n)$, on the same n-individual sample, find the optimal t linear combinations of the variables defining Y, to give a new matrix $M'Y$, in the sense that the geometrical representations of the sample $C(X)$ and $C(M'Y)$ will be as similar as possible.

This question will arise in practical situations when the population is essentially defined by the variables giving X but where, for some reasons, it appears more convenient to observe the variables in Y. We shall call the new variables the *principal components of Y with respect to X*. The problem has been treated under the title "principal components of instrumental variables" by Rao (1965).

With the RV-criterion of optimality, we must maximize

$$RV(X, M'Y) = \operatorname{tr}(S_{12} MM' S_{21}) / \{\operatorname{tr} S_{11}^2 . \operatorname{tr}(M' S_{22} M)^2\}^{\frac{1}{2}}.$$

Again, to remove the degrees of indeterminacy, we shall require the new variables to be uncorrelated, i.e. we impose on the $q \times t$ matrix M the constraint

$$M' S_{22} M = \Delta = \operatorname{diag}(\delta_i). \tag{2}$$

The analysis proceeds as in the previous subsection. Using Lagrange multipliers λ_i's and $\Lambda = \operatorname{diag}(\lambda_i)$, we must maximize the function $\Phi(M) = \operatorname{tr}(S_{12} MM' S_{21}) - \sum \lambda_i [M' S_{22} M]_{ii}$. The maximum is attained when

$$\tfrac{1}{2}(\partial \Phi / \partial M) = S_{21} S_{12} M - S_{22} M \Lambda = 0. \tag{3}$$

Let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_q$ be the eigenvalues of the generalized eigenproblem (3). The columns of M should be the eigenvectors associated with $\lambda_1, \lambda_2, \ldots, \lambda_t$ $(t \leqslant q)$, normed in such a way that (2) is satisfied. (There is no gain in taking t larger than the number of non-zero eigenvalues.) The value of RV will then be

$$RV(X, M'Y) = \left(\sum_{i=1}^{t} \lambda_i \delta_i\right) \Big/ \left\{\operatorname{tr} S_{11}^2 . \left(\sum_{i=1}^{t} \delta_i^2\right)\right\}^{\frac{1}{2}}.$$

If the values of the variances of the new variables have not been preassigned, then an optimal choice for the δ_i's is given by $\Delta = \Lambda$ and the global maximum for RV will be attained:

$$\max RV(X, M'Y) = \left\{\left(\sum_{i=1}^{t} \lambda_i^2\right) \Big/ \operatorname{tr} S_{11}^2\right\}^{\frac{1}{2}}.$$

4.3. *Multivariate Linear Regression*

We consider again the search for linear combinations of the variables defining a $q \times n$ data matrix Y such that the pattern of the resulting configuration of points $C(M'Y)$ be as close as possible to the configuration $C(X)$ mapping the $p \times n$ data matrix X representing the same n individuals. We assume that $S_{22} = YY'$ is non-singular and we seek p linear combinations; thus we can speak about the "residual" variables and the corresponding residual data matrix $X - M'Y$.

Let H be a $p \times p$ orthogonal matrix which diagonalizes $S_{12} S_{22}^{-1} S_{21}$ and let

$$H S_{12} S_{22}^{-1} S_{21} H' = \Lambda = \operatorname{diag}(\lambda_i).$$

The solution to the problem of finding the principal components of Y with respect to X is given by $M' = H S_{12} S_{22}^{-1}$. Indeed, we have $M' S_{22} M = \Lambda$ and

$$S_{21} S_{12} (S_{22}^{-1} S_{21} H') - S_{22} (S_{22}^{-1} S_{21} H') \Lambda = 0,$$

which shows that (2) and (3) are satisfied.

The linear operator H is a rotation in $\mathscr{R}^p$. Since the RV-coefficient is independent of rotations, we conclude that $RV(X, M'Y)$ attains its maximum if we select $M' = S_{12}S_{22}^{-1}$. With this choice we have also

$$(X - M'Y)Y' = S_{12} - M'S_{22} = 0. \tag{4}$$

Thus it can be said that, the variables defining X being taken as regressands and those defining Y as regressors, the linear multivariate regression operator $S_{12}S_{22}^{-1}$ is one of the optimal linear transformations in the sense of the RV-coefficient.

It is interesting to note that the regression operator is arrived at in a totally non-parametric context. The same optimal value of RV is attained whether constraints (2) or (4) are imposed.

4.4. *Canonical Variables*

We now generalize the problem of Subsection 4.2 to the simultaneous search for t linear combinations of the variables defining X and t linear combinations of the variables defining Y. Let $L'X$ and $M'Y$ be the resulting data matrices. The configurations of points $C(L'X)$ and $C(M'Y)$ will be as similar as possible if we choose L and M so as to maximize

$$RV(L'X, M'Y) = \operatorname{tr}(L'S_{12}MM'S_{21}L)/\{\operatorname{tr}(L'S_{11}L)^2 . \operatorname{tr}(M'S_{22}M)^2\}^{\frac{1}{2}}.$$

The degree of indeterminacy will be removed by requesting that each set of new variables be uncorrelated, i.e. that

$$L'S_{11}L = \Delta_x = \operatorname{diag}(\delta_{xi}) \quad \text{and} \quad M'S_{22}M = \Delta_y = \operatorname{diag}(\delta_{yi}). \tag{5}$$

Let $\Lambda = \operatorname{diag}(\lambda_i)$ and $\Psi = \operatorname{diag}(\psi_i)$, where the λ_i's and ψ_i's are arbitrary scalars. The problem reverts to the maximization, under constraints (5), of the function of L and M:

$$\Phi(L, M) = \operatorname{tr}(L'S_{12}MM'S_{21}L) - \sum_{i=1}^{t} \lambda_i [L'S_{11}L]_{ii} - \sum_{i=1}^{t} \psi_i [M'S_{22}M]_{ii}.$$

The matrices L and M must satisfy

$$\tfrac{1}{2}(\partial\Phi/\partial L) = S_{12}MM'S_{21}L - S_{11}L\Lambda = 0,$$
$$\tfrac{1}{2}(\partial\Phi/\partial M) = S_{21}LL'S_{12}M - S_{22}M\Psi = 0.$$

If we premultiply the first equality by L' and the second by M' and let $A = L'S_{12}M$, we see that we must have $AA' = \Delta_x\Lambda$, $A'A = \Delta_y\Psi$ and, hence, $\Delta_x\Lambda A = A\Delta_y\Psi$. Thus $\Delta_x\Lambda = \Delta_y\Psi$ must hold.

Assuming that S_{11} and S_{22} are non-singular, it can be verified that a solution for L, M, Λ and Ψ is as follows. Choose for the columns of L a set of t independent eigenvectors of the generalized eigenvalue problem

$$S_{12}S_{22}^{-1}S_{21}L = S_{11}L\Gamma, \tag{6}$$

where $\Gamma = \operatorname{diag}(\gamma_i)$, the γ_i's being the corresponding eigenvalues. The columns of L should be normalized so as to satisfy $L'S_{11}L = \Delta_x$. Then (assuming that t does not exceed the number of non-zero eigenvalues of problem (6)), $M = S_{22}^{-1}S_{21}L\Gamma^{-\frac{1}{2}}\Delta_x^{-\frac{1}{2}}\Delta_y^{-\frac{1}{2}}$; $\Lambda = \Delta_y\Gamma$ and $\Psi = \Delta_x\Gamma$. Note that in this solution the role of L and M are symmetrical, for we also have $S_{21}S_{11}^{-1}S_{12}M = S_{22}M\Gamma$ and $L = S_{11}^{-1}S_{12}M\Gamma^{-\frac{1}{2}}\Delta_y^{-\frac{1}{2}}\Delta_x^{-\frac{1}{2}}$.

The new variables defining the solution data matrices have the property that the jth variable for $L'X$ and the kth variable for $M'Y$ are uncorrelated when $j \neq k$; we find $(L'X)(M'Y)' = \operatorname{diag}\{(\gamma_i\,\delta_{xi}\,\delta_{yi})^{\frac{1}{2}}\}$.

For the selected values of Δ_x and Δ_y, the optimal value of the RV-coefficient is

$$RV(L'X, M'Y) = \left(\sum_{i=1}^{t} \gamma_i\,\delta_{xi}\,\delta_{yi}\right) \Big/ \left\{\left(\sum_{i=1}^{t} \delta_{xi}^2\right)\left(\sum_{i=1}^{t} \delta_{yi}^2\right)\right\}^{\frac{1}{2}}, \tag{7}$$

where the selected γ_i's must be the largest eigenvalues of problem (6).

The expression on the right in (7) is homogeneous in the δ_{xi}'s and also in the δ_{yi}'s. We may impose the additional constraints that $\sum \delta_{xi}^2 = \sum \delta_{yi}^2 = t$ without changing the value of RV, which is then equal to $RV(L'X, M'Y) = (\sum \gamma_i \delta_{xi} \delta_{yi})/t$. In particular, if we choose $\delta_{xi} = \delta_{yi} = 1$ for all i, we obtain $RV(L'X, M'Y) = \sum \gamma_i/t$ for the solution in L and M given above with $L'S_{11}L = M'S_{12}M = \Delta_x = \Delta_y = I_t$. This solution is precisely that formed by the first t pairs of canonical variables for the variables defining X and Y.

We conclude this subsection by a remark which will be useful in the next one. If we preassign Δ_x or Δ_y to be the identity matrix I_t then the optimal choice for the other is to make it equal to Γ. Indeed if, say, $\Delta_x = I_t$, then we see that the maximum of the right-hand side of (7) is attained for $\delta_{yi} = \gamma_i$ for all i. This maximal value is equal to $\{(\sum \gamma_i^2)/t\}^{\frac{1}{2}}$, and the solutions for L and Γ are given by (6) and $M = S_{22}^{-1} S_{21} L$.

4.5. *Discriminant Analysis*

Let us consider a $p \times n$ data matrix X from a sample of n individuals belonging to g distinct groups. The first n_1 individuals belong to the first group; the next n_2 belong to the second group, and so on; $n_1 + n_2 + \ldots + n_g = n$. As previously, we assume that each variable has been centred over the n sampled individuals. We shall denote by Y the data matrix obtained by substituting for each element X_{ij} of X the mean value of the ith variable within the group containing the jth individual.

With the terminology of Cooley and Lohnes (1971) and our previous notation, $S_{11} = XX'$ is the grand total sum of squares and cross-products and $S_{22} = YY'$ is the among-groups sum of squares and cross-products; $W = S_{11} - S_{22}$ is this sum within groups. The definition of Y is such that we have the particular simplification: $S_{22} = S_{12} = S_{21}$.

To relate in some way the characterization of the sample by the individual values in X and its characterization by the group means forming Y, we shall search for projections on t orthogonal hyperplanes of the configuration of points $C(X)$ and for projections on t orthogonal hyperplanes of the configurations $C(Y)$ so as to optimize the RV-coefficient of the two projected sets of points. Each hyperplane being defined by a vector in $\mathscr{R}^p$, the problem amounts to that of finding the two $p \times t$ matrices L and M that will make the two configurations of points $C(L'X)$ and $C(M'Y)$ as similar as possible. To ensure that the hyperplanes are orthogonal, we shall request that the new variables defining $L'X$ be orthonormal, i.e. $L'XX'L = L'S_{11}L = I_t$, and those defining $M'Y$ be orthogonal: $M'YY'M = M'S_{22}M = \Delta_y$, where Δ_y is a diagonal matrix to be determined.

From the analysis done in Subsection 4.4, it can be seen that the problem is essentially that of finding the canonical correlations of X and Y under the above constraints. The last paragraph of this previous subsection provides the answer when we substitute S_{22} for S_{12} and S_{21} in (6).

The solution is $M = L$, where L is the eigenmatrix associated with the t largest eigenvalues of the problem $S_{22}L = S_{11}L\Gamma$. The matrix $\Gamma = \text{diag}(\gamma_i)$ is formed by those t largest eigenvalues and $\Delta_y = \Gamma$. The columns of L should be normalized to satisfy $L'S_{11}L = I_t$. Then, $RV(L'X, L'Y) = (\sum \gamma_i^2/t)^{\frac{1}{2}}$.

Note that since $W = S_{11} - S_{22}$, L also satisfies $S_{22}L = WL\tilde{\Gamma}$, where $\tilde{\Gamma} = \text{diag}\{\gamma_i/(1-\gamma_i)\}$.

One will have recognized that the solution matrix L defines the first t discriminant hyperplanes of discriminant analysis. It is interesting to note the relationship between canonical variables of X and Y and discriminant analysis when Y is defined by the within-group means. As in the case of multivariate regression, the results are arrived at in a totally non-parametric context.

5. Discarding Variables

A problem of great practical importance is that of discarding variables from a given set. The problem is two-fold. Firstly, one should choose the variables to be retained and, secondly,

one should find the proper metric to be applied to those variables so as to obtain an image of the population as close to the one he would have had if all variables had been kept.

A number of methods have been proposed to solve this problem. A limited survey of methods can be found in Jolliffe (1972). Most methods have the property, which we consider erroneously restrictive, that they use only the Euclidean metric on the selected variables.

In this section we shall describe a method using the RV-coefficient to solve this two-fold problem of discarding variables. Again, let X be a $p \times n$ data matrix giving the values of p variables for each of n individuals. Suppose $(p-t)$ variables are discarded and let Y be the $t \times n$ submatrix of X containing the values of the retained variables. Using the reduced information, Y, and a metric defined by the product MM', where M is a $t \times t$ matrix, the similarity of the configurations of the n-sample given by the total and the reduced information, is given, in accordance with the theory of Section 2, by $RV(X, M'Y)$.

Without loss of generality, assume that the variables defining Y are the first t. If we write $X' = (Y' \vdots Z')$, we have $S_{11} = XX'$, $S_{22} = YY'$, $S_{12} = XY'$, $S_{21} = YX'$ and

$$S_{11} = \left(\begin{array}{c|c} S_{22} & YZ' \\ \hline ZY' & ZZ' \end{array} \right) = \left(\begin{array}{c|c} S_{12} & YZ' \\ \hline & ZZ' \end{array} \right).$$

The solution for M is given by the principal components of Y with respect to X. It is the matrix of eigenvectors of the generalized eigenvalue problem (3): $S_{21} S_{12} M = S_{22} M \Lambda$. We then have

$$RV(X, M'Y) = \{(\sum \lambda_i^2)/\mathrm{tr}\, S_{11}^2\}^{\frac{1}{2}}.$$

A complete solution of the problem, in the scope of this theory, would then be to choose that set of t variables and the corresponding data submatrix Y in such a way that $RV(X, M'Y)$ will be maximized, i.e. the sum $\sum \lambda_i^2$ will be maximized. The optimal metric to use on Y is the resulting matrix MM' and a measure of quality of the particular choice of variables and metric is the resulting value of $RV(X, M'Y)$.

When the numbers t and p are small, one could conceivably try all combinations of t variables. For large values of t or p, the authors have not succeeded, up to now, in finding a numerically efficient algorithm to select the optimal subset of variables. However, a suboptimal but numerically efficient technique will now be summarized.

For the purpose of this description, we call a diagonal matrix with non-negative elements a weight matrix. The optimal weight matrix Δ to be used on the data submatrix Y of X is the one that maximizes $RV(X, \Delta Y)$. As for the case of an arbitrary metric, the authors have not yet found an algorithm that they would call numerically efficient for the two-fold problem of selecting the best set of t variables together with their optimal weight matrix. They have found and coded a suboptimal algorithm which makes a sequential selection of variables. This algorithm, described in Escoufier *et al.* (1974), will be called the Sequential-Weight-Selection algorithm. It selects the one best variable with its corresponding weight; then, having selected k variables at the kth step, it proceeds to determine which one of the remaining $p-k$ variables, when added to the k previously selected variables, will, with the optimal weight matrix, maximize RV. (Of course, at each step, the weights of the variables previously selected have to be recomputed.) The choice of the $(k+1)$th variable is dependent on the choice of the first k; as with sequential regression, it is this sequentiality that restricts the choice algorithm to suboptimality.

For the general problem of discarding $p-t$ variables, we may then proceed in two phases:

Phase 1. Use the Sequential-Weight-Selection algorithm to select t variables and the associated optimal weight matrix. Let Y be the selected data submatrix, Δ its weight matrix and $\rho_1 = RV(X, \Delta Y)$.

Phase 2. Use the results of Subsection 4.2 to compute the optimal metric MM' to be used on the variables defining the data matrix Y selected in Phase 1. Let $\rho_2 = RV(X, M'Y)$.

Since the choice of M is not restricted to that of a diagonal matrix, ρ_2 is at least equal to ρ_1 and will usually be greater. The value of ρ_2 serves as a measure of the goodness of the solution to the discarding problem (with proper choice of metric on the retained variables).

In practice, assuming the RV-coefficient is to be used as a criterion to estimate the quality of the result, it is most likely that the problem of discarding variables will be posed in the following terms: discard as many as possible among the variables defining the data matrix X, while ensuring that the remaining variables (Y) with a proper metric (MM') will represent the total information to a "level" at least equal to α ($RV(X, M'Y) \geqslant \alpha$, α relatively close to 1).

The Sequential-Weight-Selection algorithm coded by the authors is designed so that, given ρ_1, it will select variables until the resulting $RV(X, \Delta Y)$ attains the value ρ_1. For a solution of the problem stated in the previous paragraph, one would use this option in Phase 1 with ρ_1 equal to or slightly smaller than α. Then Phase 2 will give a value of ρ_2 which should reach the required level.

The authors are presently gathering numerical evidence, to be published in the near future, on the following properties of the proposed method: computational performance, degree of suboptimality, comparison with other methods. We conclude by pointing out that the Sequential-Weight-Selection algorithm, using the simpler transformation Δ instead of M, is of interest in its own right since it provides good solutions to the problem of retaining variables with only an appropriate change of scales.

Acknowledgements

The authors wish to thank their colleague Robert Cléroux for numerous stimulating and constructive discussions on the topics of this paper. Comments by the referees are also appreciated. The work was partly supported by National Research Council of Canada, Grant A-4093.

References

Anderson, T. W. (1958). *Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Cooley, W. W. and Lohnes, P. R. (1971). *Multivariate Data Analysis*. New York: Wiley.

Escoufier, Y. (1970). Echantillonnage dans une population de variables aléatoires réelles. *Publ. Inst. Stat. Univ. Paris*, **19**, (fasc. 4) 1–47.

—— (1973). Le traitement des variables vectorielles. *Biometrics*, **29**, 751–760.

Escoufier, Y., Robert, P. and Cambon, Y. (1974). Construction of a vector equivalent to a given vector from the point of view of the analysis of principal components. In *Computational Statistics* (G. Bruckmann *et al.*, eds), pp. 155–164. Wien: Physical Verlag.

Gower, J. C. (1971). Statistical methods of comparing different multivariate analyses of the same data. In *Mathematics in the Archaeological and Historical Sciences* (F. R. Hodson *et al.*, eds), pp. 138–149. Edinburgh: University Press.

Jolliffe, I. T. (1972). Discarding variables in principal component analysis. I: artificial data. *Appl. Statist.*, **21**, 160–173.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

Rao, C. R. (1965). The use and interpretation of principal component analysis in applied research. *Sankhyā*, A, **26**, 329–358.

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