CHOOSING VARIABLES AND METRICS BY OPTIMIZING THE RV-COEFFICIENT

Yves Escoufier

Centre de recherche en informatique et gestion
Université des Sciences et Techniques du Languedoc
34075 Montpellier, France

Pierre Robert

Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, Canada

In this paper, a statistical study is defined as a triplet (X,Q,D) where X is a data matrix, Q and D are matrices used to compute distances between individuals and variables respectively. It is shown that the problem of comparing statistical studies can be well posed and efficiently solved by use of the RV-coefficient.

Algorithms are given to choose metrics and perform variable selections by optimization of the RV-coefficient. The proposed methods are applicable to quantitative as well as qualitative variables.

In an Appendix, a conceptual framework is proposed in which the notion of statistical study, as defined in this paper, can be interpreted and further developed.

I. ANALYSIS OF DATA WITH CHARACTERISTIC OPERATORS

We consider a statistical study as characterized by a triplet (X,Q,D) where:

- X is a pxn matrix of measures of p variables on n individuals,

ISBN 0-12-604580-1

- Q is a pxp positive definite or semi-definite matrix which is used to compute the distances between the individuals,
- D is a nxn diagonal matrix, with positive diagonal elements, giving the weights attached to each of the individuals; the sum of the weights must be equal to 1.

In this work we assume that the rows of X are centered with respect to the weights in D = diag(d_j), i.e., $\sum_j d_j x_{ij}$ = 0 for each i. Then V = XDX' is the matrix of variances and covariances of the p variables.

The nxn matrix W = X'QX of cross products between individuals plays an important role in data analysis; its rank is p when Q and X are of rank p < n.

To define a statistical study by a triplet (X,Q,D) means that the study is completely specified when the statistician has an array of collected data (X), has decided on the weights to be attached to each one of the individuals in the sample (D) for the computation of the covariances from which variables are compared and, finally, has decided on how individuals are to be compared (Q).

To compare a statistical study (X,Q,D) with other studies, we shall make use of the associated operator WD.

Given (X,Q,D), WD = X'QXD is specified. Conversely, given WD, with a positive definite matrix W, one can construct a triplet (X,Q,D) such that X'QX = W.

To compare two statistical studies (X,Q_1,D) and (Y,Q_2,D) on the same individuals and with identical sets of weights D, we seek a mean of comparison between the matrices W_1D and W_2D. The following propositions set the mathematical background for this linking. No proofs are given; they are arrived at through simple algebraic calculus and can be found in [1], [2] or [7].

Let S(D) be the set of D-symmetric matrices, i.e., the set of matrices A such that DA = A'D. Note that S(D) contains all operators of the form WD; an arbitrary element A of S(D) cannot be considered as an operator associated to a statistical study unless (AD^{-1}) is positive-definite. Then, we have:

Proposition 1. The symmetrical bilinear from P defined on S(D) by

$$P(A,B) = tr(AB)$$

is positive. Hence, it defines a scalar product on S(D) and a distance d given by

$$d^2(A,B) = P(A-B, A-B) = \mathrm{tr}((A-B)^2)$$
$$= \mathrm{tr}(A^2) + \mathrm{tr}(B^2) - 2\,\mathrm{tr}(AB) .$$

By similarity with the usual statistical vocabulary, we define (on S(D)):

$$\mathrm{COVV}(A,B) = \mathrm{tr}(AB) ,$$
$$\mathrm{VAV}(A) = \mathrm{tr}(A^2) ,$$
$$\mathrm{RV}(A,B) = \mathrm{tr}(AB)/[\mathrm{tr}(A^2) \cdot \mathrm{tr}(B^2)]^{\frac{1}{2}} .$$

The RV-coefficient $\mathrm{RV}(W_1D,W_2D)$ can serve as a measure of comparison between the studies W_1D and W_2D. The following proposition helps to understand the significance of RV:

Proposition 2.

(i) For any (X,Q_1,D) and (Y,Q_2,D), $0 \leq \mathrm{RV}(W_1D,W_2D) \leq 1$;

(ii) $\mathrm{RV}(W_1D,W_2D) = 1$ if and only if $W_1 = kW_2$ for some non-zero scalar k;

(iii) If Q_1 and Q_2 are positive definite, $\mathrm{RV}(W_1D,W_2D) = 0$ if and only if $S_{12} = X'DY = 0$.

The basic elements introduced above can take place in a theoretical framework constructed on the usual vector spaces R^p and R^n, in which individuals and variables are commonly represented, as well as on their dual spaces of linear functionals. This duality scheme, developed by Cailliez and Pages [1], has already proved itself a useful tool for the analysis of statistical studies and could be instrumental in further research. The interested reader will find a summary of this duality scheme in the Appendix to this paper; it has been given this marginal place because its knowledge and understanding are not a prerequisite for the remaining sections of the paper.

II. CHOICE OF A METRIC

Consider a given pxn data array X and a statistical study on it: (X,I,D). D is nxn, non-singular. We assume the distances between individuals to be obtained from the metric

represented by the identity matrix I; it is easily seen in the sequel that this assumption bears no loss of generality.

One can think of many practical situations in which measurements will be taken on the same individuals but for a second set of variables (totally or partially different from those giving the array X). This second set of observations may be made for a number of purposes. For example, it may simply be to compare the apparent structure of the population (the n individuals as seen from two points of view); it may be with the idea of discarding variables (those or some from those giving X) which are difficult or costly to measure.

Suppose the new set of q variables gives a qxn data array Y on which the statistical study (Y,Q,D) is to be carried. The positive definite, qxq, matrix Q defines the new way of measuring distances between individuals.

The results of the two studies can be compared by use of the RV-coefficient. With $W_1 = X'X$, $W_2 = Y'QY$, the closer to the value 1 will $RV(W_1D,W_2D)$ be the closer will the studies (X,I,D) and (Y,Q,D) be. Note that, from Proposition 2, as RV approaches 1 the set of cross products between individuals, and therefore the set of distances between individuals, as computed in the two studies tend to be equal within a common multiplicative factor ($W_1 \simeq k\, W_2$ for some scalar k).

We shall show, using two particular cases, that it is always possible to find Q such that the studies (X,I,D) and (Y,Q,D) will be as close as possible. In the context of this paper, it means that $RV(W_1D,W_2D)$ must be maximized.

A. Selecting Q to be a Diagonal Matrix

In practice the restriction that Q be a diagonal matrix means that the statistician is ready to change the measuring scales of the variables defining Y. With $Q = \Delta = \mathrm{diag}(\delta_j)$, we have

$$RV(W_1D,W_2D) = t(X'XDY'\Delta YD)/[tr((X'XD)^2) \cdot tr((Y'\Delta YD)^2)]^{\frac{1}{2}} .$$

Let u_{ij} (i = 1,...,q; j = 1,...,p) be the elements of YDX'; v_{ij} (i = 1,...,q ; j = 1,...,q) be the elements of YDY'; $z_i = \sum_{j=1}^{p} u_{ij}^2$. It is easily shown [4] that

$$\mathrm{tr}(X'XDY'\Delta YD) = \sum_{i=1}^{q}\Big(\sum_{j=1}^{p} u_{ij}^2\Big)\ \delta_i = \sum_{i=1}^{q} z_i\,\delta_i\ ,$$

$$\mathrm{tr}((Y'\Delta YD)^2) = \sum_{i=1}^{q}\sum_{j=1}^{q} v_{ij}^2\ \delta_i\,\delta_j\ ,$$

so that the maximization of $RV(W_1D,W_2D)$ reduces to the simple quadratic problem:

- minimize $\sum_{i=1}^{q}\sum_{j=1}^{q} v_{ij}^2\ \delta_i\,\delta_i$

- subject to $\sum_{i=1}^{q} z_i\,\delta_i = 1$ and $\delta_i \geq 0$ for all i .

This problem can be written in a canonical form by setting $N_i = z_i\,\delta_i$, $C_{ij} = v_{ij}^2/z_iz_j$:

- minimize $\sum_{i=1}^{q}\sum_{j=1}^{q} C_{ij}\,N_i\,N_j$,

- subject to $\sum_{i=1}^{q} N_i = 1$ and $N_i \geq o$ for all i .

Since the variances and covariances matrix YDY' is positive definite, so are the matrices with elements v_{ij}^2 and C_{ij} [6]. Thus the above problem is totally convex and admits a unique solution. There exists a number of algorithms to compute the solution; one which was found particularly efficient is an iterative procedure given in [5]. Starting with an arbitrary vector $N^{(o)}$, the (m+1) approximation $N^{(m+1)}$ is obtained from $N^{(m)}$ as follows:

$$N_1^{(m+1)} = \max\Big[o,\frac{1}{C_{11}}\Big(1 - \sum_{j=2}^{q} C_{1j}\,N_j^{(m)}\Big)\Big]\ ,$$

$$N_i^{(m+1)} = \max[o, \frac{1}{C_{ii}}(1 - \sum_{j<i} C_{ij} N_j^{(m+1)} - \sum_{j>i} C_{ij} N_j^{(m)})]$$

for $1 < i < q$,

$$N_q^{(m)} = \max[o, \frac{1}{C_{qq}}(1 - \sum_{j=1}^{q-1} C_{qj} N_j^{(m+1)})] .$$

One shows that $\lim_{m \to \infty} N^{(m)} = N^*$ and

$$N = N^* / \sum_{i=1}^{q} N_i^* ; \text{ hence } \delta_i = N_i / z_i .$$

B. Unrestricted Q

We now seek a (symmetric positive semi-definite) matrix Q that will make the statistical studies (X,I,D) and (Y,Q,D) as similar as possible. For any Q, there exists a matrix M, qxr ($r \leq q$) such that $Q = MM'$. The matrix M is not unique. We must maximize

$$RV(W_1D, W_2D) = tr(X'XDY'MM'YD) / [tr((X'XD)^2) \cdot tr((Y'MM'YD)^2)]^{1/2} .$$

It is interesting to note that the same problem arises if the statistician substitutes for Y the linear combinations of variables which give the data array M'Y and then compares the studies (X,I,D) and (M'Y,I,D).

We define $S_{12} = XDY'$, $S_{21} = YDX'$, $S_{11} = XDX'$, $S_{22} = YDY'$ and write

$$RV(W_1D, W_2D) = tr(M'S_{21}S_{12}M) / [tr\, S_{11}^2 \cdot tr((M'S_{22}M)^2)]^{1/2} .$$

Note that the value of RV is not changed if we replace M by MR' and M' by RM' when R is an rxr orthogonal matrix ($RR' = I$) so that we can seek the maximum of RV with some degree of indeterminacy. We choose to impose that $M'S_{22}M$ be a diagonal matrix $\Delta = diag(\delta_i)$; then any rotation R applied to the solution found M will define the same solution matrix Q $Q = MR'RM' = MM'$.

Introducing the Lagrange multipliers $\lambda_i (i=1,\ldots,r)$, we are lead to the maximization of

$$\Phi(M) = \mathrm{tr}(M'S_{21}S_{12}M) - \sum_{i=1}^{r} \lambda_i [M'S_{22}M]_{ii} .$$

The maximum is attained when

$$\tfrac{1}{2} \frac{\partial \Phi}{\partial M} = S_{21}S_{12}M - S_{22}M\Lambda = o \qquad (1)$$

where $\Lambda = \mathrm{diag}(\lambda_i)$. The solution is to give to the λ_i's the values of the r largest eigenvalues of this last generalized eigenvalue problem. The columns of M will be the eigenvectors associated to $\lambda_1,\lambda_2,\ldots,\lambda_r$, normalized to satisfy $M'S_{22}M = \Delta$. It follows that

$$RV(W_1D,W_2D) = \sum_{i=1}^{r} \lambda_i \delta_i / [\mathrm{tr}\, S_{11}^2 \cdot (\sum_{i=1}^{r} \delta_i^2)]^{\frac{1}{2}}$$

and that the optimal choice consists in taking $\delta_i = \lambda_i$. Then

$$RV(W_1D,W_2D) = [\sum_{i=1}^{r} \lambda_i^2 / \mathrm{tr}\, S_{11}^2]^{\frac{1}{2}} .$$

Note that the problem solved here has been studied in [8] under the title of "principal component analysis of Y with respect to the instrumental variables X". Our approach and interpretation are different.

Another interpretation can be given, when r = p, in a regression context. Indeed, it can be verified by substitution that $HS_{12}S_{22}^{-1}$ is a solution of (1) if H is the pxp orthogonal matrix which diagonalizes $S_{12}S_{22}^{-1}S_{21}$: $HS_{12}S_{22}^{-1}S_{21}H' = \Lambda$. This shows that the regression operator $M^{*\prime} = S_{12}S_{22}^{-1}$ gives its maximum value to the RV-coefficient.

Relationships between this type of analysis of statistical studies and other classical linear multivariate techniques can be found in [9].

III. SELECTION OF VARIABLES

Again we consider two data arrays on the same individuals, X (pxn) and Y (qxn). The variables defining Y may be identical or they may be partially or totally distinct from those defining X. The array X and the statistical study (X,I,D) attached to it will serve as a reference study.

In this section we seek a selection of k variables from the q variables defining Y. Not only can we select variables but, using the RV-coefficient as a measure of quality, we can attach to the selected variables a metric that will make the corresponding statistical study as close as possible to the reference study (X,I,D). If K is a set of k distinct indices from $\{1,2,\ldots,q\}$ and Y_k is the subarray of Y which retains those rows with indices in K, we shall search for a metric Q_k to maximize $RV(X'XD, Y'_k Q_k Y_k D)$.

Theoretically one could consider all combinations of k variables from the q variables defining Y and, for each combination, use the algorithm described in Section II; the best pair (Y_k, Q_k) will be detected. Unless q is very small, the method is not practical.

The authors have designed sequential algorithms which are suboptimal but have proved to be numerically efficient and, at least from the test samples used, very close to optimality. We shall consider three cases depending on the type of metric Q_k which is looked for. Computer packages have been coded in FORTRAN which will perform the selected type of computation; they are available from the authors.

A. Ordinary Euclidian Metric on Y_k (Q_k=I)

The problem is simply that of selecting the k variables. We must maximize $RV(X'XD, Y'_k Y_k D)$ for the selected array Y_k.

Suppose a subset of (k-1) variables, corresponding to the array Y_{k-1}, have been selected. Let y be a not yet selected row of Y and let

$$Y_k = \begin{bmatrix} Y_{k-1} \\ \hline y \end{bmatrix} .$$

Y_{k-1} being known, it is easy to construct an algorithm to select the optimal y; one makes use of the simple expressions:

$$\operatorname{tr}(X'XDY_k'Y_kD) = \operatorname{tr}(X'XDY_{k-1}'Y_{k-1}D) + \operatorname{tr}(X'XDy'yD) ,$$

$$\operatorname{tr}(Y_k'Y_kD)^2 = \operatorname{tr}(Y_{k-1}'Y_{k-1}D)^2 + \operatorname{tr}(y'yD)^2$$

$$+ \operatorname{tr}(Y_{k-1}'Y_{k-1}Dy'yD) .$$

(Note that tr(A'ADy'yD) is equal to the sum of the squares of the elements of the vector ADy').

Variables are added sequentially. The first variable to be introduced must maximize

$$RV(X'XD, y'yD) = \operatorname{tr}(X'XDy'yD)/[\operatorname{tr}(X'XD)^2 \cdot \operatorname{tr}(y'yD)^2]^{\frac{1}{2}}$$

$$= \sum_{i=1}^{p} [COV(X_i,y)]^2/[\operatorname{tr}(X'XD)^2 \cdot \sigma_y^4]^{\frac{1}{2}} .$$

Thus, the process is initialized with the variable which maximizes

$$\sum_{i=1}^{p} [COV(X_i,y)]^2/\sigma_y^2 .$$

B. Selecting Q_k to be a Diagonal Matrix

To the problem of the selection of the variables is now added that of scaling the selected variables: we require Q_k to be a diagonal matrix, say Δ_k.

Again we assume that k-1 variables have been selected and their optimal weights computed; let Y_{k-1} and Δ_{k-1} be corresponding matrices. For each variable not yet selected the algorithm of Section II.A must be applied with Y replaced by

$$Y_k = \begin{bmatrix} Y_{k-1} \\ \hline y \end{bmatrix} .$$

The result will be a new set of weights Δ_k (the weights attached to the k-1 initial variables will generally be changed) and the resulting value of RV. The best choice for y is made by comparing the RV values.

Note that the first variable to be retained is the same as in III.A (with weight equal to 1).

C. *Unrestricted Q*

Lastly, we consider the general problem of selecting k variables, with data array Y_k, and an unrestricted (positive semi-definite) matrix Q_k to make the statistical studies (X,I,D) and (Y_k,Q_k,D) as close as possible.

As in Section II.B we write $Q_k = M_k M_k'$ and seek to maximize $RV(X'XD, Y_k'M_k M_k'Y_k D)$. When Y_k is given, we know from Section II.B that $M_k' = M_k^{*'} = XDY_k' (Y_k DY_k')^{-1}$ is optimal and then,

$$RV(X'XD, Y_k'M_k^*M_k^{*'}Y_k D) = \{tr[XDY_k' (Y_k DY_k')^{-1} Y_k DX']^2 / tr(X'DX)^2\}^{\frac{1}{2}} .$$

It can be shown that from k-1 known variables, it is relatively simple to select the k-th variable which will maximize the numerator of the above expression. The algorithm uses successive Choleski decompositions of the matrices $Y_{k-1} DY_{k-1}'$, $Y_k DY_k'$. If $Y_k DY_k' = S_k S_k'$, where S_k is lower triangular,

$$tr\,[XDY_k'(Y_k DY_k')^{-1} Y_k DX']^2 = tr(S_k^{-1} Y_k DX'XDY_k' S_k'^{-1})^2$$

and the matrices $S_k^{-1} Y_k$ are easily computed from $S_{k-1}^{-1} Y_{k-1}$.

Remark. We have presented here forward selection techniques. Computer programs have been coded for backward elimination and step-wise selection (combined forward-backward). The relative efficiencies of the methods are presently being studied.

IV. QUALITATIVE VARIABLES

In a paper [6] of much practical interest J. C. Gower introduced a notion of similarity which allows the simultaneous treatment of quantitative, qualitative and dichotomous variables. For each one of such variables, an nxn similarity matrix S_i ($i = 1,2,\ldots,q$) is defined. Giving weight $\delta_i \geq 0$ to the i-th variable, a global similarity matrix S between the n individuals is defined as $S = \sum_{i=1}^{q} \delta_i S_i$. Gower discusses conditions under which S is positive definite.

Suppose a reference similarity matrix R, nxn, is known. Using the RV-coefficient as a measure of closeness, one can compute the optimal values for the δ_i's if S is to reproduce R as best as possible. Indeed, assuming that each S_i is positive semi-definite, R and S are analogous to the matrices of cross-products between variables, W_1 and W_2, respectively. (For example, referring to the notation of Section II.A and denoting by y_i the i-th row of Y, it can be seen that

$$Y'\Delta Y = \sum_{i=1}^{q} \delta_i \, y_i' y_i).$$

The maximization of RV (RD,SD) is easily seen to be equivalent to the quadratic programming problem:

- minimize $\sum_{i=1}^{q} \sum_{j=1}^{q} \operatorname{tr}(S_i D S_j D) \, \delta_i \delta_j$
- subject to $\sum_{i=1}^{q} \operatorname{tr}(S_i D R D) \, \delta_i = 1$ and $\delta_i \geq 0$ for all i.

Hence the algorithm of Section II.A can be used (with $v_{ij}^2 = \operatorname{tr}(S_i D S_j D)$ and $z_i = \operatorname{tr}(S_i D R D)$).

The proposed method assumes previous acceptance of a similarity matrix R and, using it as a reference, allows the determination of the relative weights to be given to qualitative variables to explain as well as possible the similarities in R.

APPENDIX: DUALITY DIAGRAM

Consider a pxn data array X giving the values of p variables on n individuals. As stated in Section I of this paper, a statistical study is completely defined when the statistician has decided on a metric, represented by a positive semi-definite matrix Q, to compute the distances between the individuals and has selected a set of weights d_j's to be attributed to the individuals. The matrix $D = \text{diag}(d_j)$ plays the role of a metric between the variables: the cross product between two variables is indeed their covariance computed using the density function defined by the selected weights ($V = XDX'$, assuming all variables centered).

The variables-individuals system is usually represented in the vector space $E = R^p$. There, the i-th axis corresponds to the i-th variable and the j-th individual is characterized by the p values on the j-th column of X. Considered as a point in E, the j-th individual is the vector

$$\sum_{k=1}^{p} x_{kj}\, \underline{e}_k, \text{ where } (\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p) \text{ is the basis of } E .$$

We can associate to the i-th variable a function $\underline{e}_i^*$ (a linear functional) which, when applied to any individual, gives the value of the i-th variable for this individual; i.e.:

$$\underline{e}_i^* \left(\sum_{k=1}^{p} x_{kj}\, \underline{e}_k \right) = \sum_{k=1}^{p} x_{kj} \cdot \underline{e}_i^* (\underline{e}_k) = x_{ij} .$$

Thus, variables also have a representation in the dual space E^* of linear functional on $E = R^p$. From a mathematical point of view, the functional $\underline{e}_i^*$ associated to the i-th variable is the i-th element of the canonical basis of E^*.

Similarly, the system can be represented in $F = R^n$. In this representation the i-th variable corresponds to the vector

$$\sum_{k=1}^{q} x_{ik}\, \underline{f}_k, \text{ where } (\underline{f}_1, \underline{f}_2, \ldots, \underline{f}_n) \text{ is the basis of } F .$$

Also, it can be argued that individuals have a representation in the dual space F* of F: the j-th individual corresponds to the functional $\underline{f}_j^*$ which is the j-th element of the canonical basis of F*. As a function applied to the vector of F representing a variable, it selects the value of this variable for the j-th individual:

$$\underline{f}_j^* \left(\sum_{k=1}^{q} x_{ik} \underline{f}_k \right) = \sum_{k=1}^{q} x_{ik} \underline{f}_j^* (\underline{f}_k) = x_{ij} .$$

This conceptualization gives us two representations for the individuals: one in $E = R^p$ (as column of X) and the second in F* (as a basis vector). The matrix associated to the corresponding mapping from F* onto E is obviously X.

Similarly for the two representations of the variables in $F = R^n$ and E*, the mapping is represented by the matrix X':

- for the individuals $E = R^p \longleftarrow X \longleftarrow F^*$
- for the variables $E^* \longleftarrow X' \longrightarrow F = R^n$.

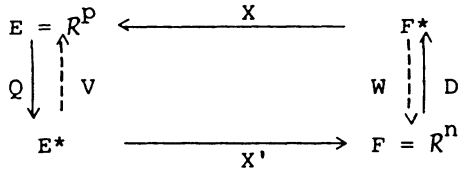
To compute distances between individuals, considered as points in $E = R^p$, a quadratic form Q has been selected. From a mathematical standpoint the choice of a quadratic form is known to be equivalent to the choice of a linear mapping from a space to its dual. Thus Q can be thought of as a mapping from E to E*. Similarly, D is considered to be a mapping from F to F*.

We have the following diagram:

$$\begin{array}{ccc} E = R^p & \xleftarrow{\quad X \quad} & F^* \\ Q \downarrow & & \uparrow D \\ E^* & \xrightarrow[\quad X' \quad]{} & F = R^n \end{array}$$

It can be seen that V = XDX' is a mapping from E* to E. V is the variance-covariance matrix for the p variables. Similarly, the matrix W = X'QX represents a mapping from F* to F; it is the matrix of cross products between individuals.

The above diagram is then augmented into a complete duality diagram in which all initial components (X,Q,D) of the statistical study, together with V and W, do appear:



Experience has shown that this simple diagram can be an excellent conceptual aid in the analysis of statistical methods. Here are two examples:

(i) If $Q = V^{-1}$ (Mahalanobis distances) then $WD = X'V^{-1}XD$ is the orthogonal projection operator on the subspace of R^n generated by the rows of X.

(ii) One can see that a principal components analysis on the triplet (X,Q,D) consists in the computation of each eigenvector h of QV followed by that of the corresponding principal component X'h; $h \in E^*$ and $X'h \in F$ are two representations of the same variable. If Q is selected to be the identity matrix I, one is lead to the diagonalization of the variance-covariance matrix V. If Q is the diagonal matrix of the inverses of the variances, to work on QV is equivalent to working directly on the correlation matrix.

To convince himself of the advantages of using the duality diagram, the reader is invited to refer to [1] or [3].

REFERENCES

1. Cailliez, F., and Pages, J. P., *Introduction à l'analyse des données,* Smash, Paris (1976).
2. Escoufier, Y., "Operators Related to a Data Matrix," in *Recent Developments in Statistics,* Ed., Barra, J. R., et al., North-Holland Publ. Co. (1977).
3. Escoufier, Y., Cailliez, F., and Pages, J. P., "Geometry and Special Procedures in Factor Analysis," European Meeting on Psychometrics and Mathematical Statistics; Uppsala, Sweden (1978).
4. Escoufier, Y., Robert, P., and Cambon, J., "Construction of a Vector Equivalent to a Given Vector from the Point of View of the Analysis of Principal Components," *Compstat 1974,* Physica-Verlag, Vienna (1974).
5. Ferland, J. A., Lemaire, B., and Robert, P., "Analytic Solutions for Non-Linear Programs with One or Two Equality Constraints," (to appear).

6. Gower, J. C., "A General Coefficient of Similarity and Some of Its Properties," *Biometrics, XXVII,* 857-874 (1971).
7. Pages, J. P., Escoufier, Y., and Cazes, P., "Opérateurs et Analyse des Tableaux à Plus de Deux Dimensions," Cahier B.U.R.O. No. 25, Institut de Statistique, Universite Pierre et Marie Curie, Paris, 61-89 (1976).
8. Rao, C. R., "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya, A, 26,* 329-358 (1966).
9. Robert, P., and Escoufier, Y., "A Unifying Tool for Linear Multivariate Statistical Methods: the RV-Coefficient," *Appl. Stat., C, 25* (3), 257-265 (1976).