Recent Developments in Statistics
J.R.Barra et al., editors

Invited Paper

OPERATORS RELATED TO A DATA MATRIX

Y. Escoufier
Centre de Recherche en Informatique
et Gestion
Université des Sciences et Techniques
du Languedoc
Montpellier , France

I - INTRODUCTION

The results shown in this paper concern the analysis of data, and especially the comparison of several analyses. They have as a common starting point an accepted party which must be specified before anything else; a data analysis of n subjects is characterized by the couple (D,D_p) where D is the matrix n x n of distances between subjects and D_p a diagonal matrix of weights affected to each of the subjects.

The inevitable consequence of this point of view is that we are led to make a comparison of the data analyses given for the same subjects by comparing the couples (D,D_p) which characterize them ; paragraph 2 deals with the way in which this aim can be achieved. Paragraph 3 shows how this approach allows a new and unifying look at the different methods of multidimensional statistical analysis. Paragraph 4 touches the problem of variable choice while paragraph 5 deals with the joint treatment of several data matrices.

II - THE COMPARISON OF DATA ANALYSIS

II-1 Let us take, for this paragraph, a point of view which could be called "traditional" in which a data analysis of n subjects is defined by the triplet (X,Q,D_p) in which :

X is a matrix, p x n, containing the values taken by p numeric variables on each of the n subjects.

The column X^j of X contains the observations made on the individual j.

Q is a positive definite or semi-definite matrix, p x p, which allows the calculation of the distances between the individuals

$$D_{jk} = \left[(X^j - X^k)'Q(X^j - X^k)\right]^{1/2}$$

D_p is a diagonal positive matrix, n x n, of weights affected to the subjects.

Let us define the matrix W, the elements of which are : $W_{jk} = X^{j'} Q X^k$

We have : $$D_{jk} = \left[W_{jj} + W_{kk} - 2 W_{jk}\right]^{1/2}$$

It is obvious that the knowledge of W allows the calculation of D and that two analyses which lead to the proportional matrices W of proportionality $k > 0$ lead necessarily to the matrices D which have a coefficient of proportionality equal to $k^{1/2}$. This allows us, in this "traditional" point of view to substitute in the comparaison of the couples (D,D_p) those of the couples (W,D_p).

Remarks :

Let $\underset{\sim}{1}$ be the vector, n x 1, of components all equal to 1 and Y the centered matrix p x n associated with X. We can write : $Y = X(I - D_p \underset{\sim}{1}\,\underset{\sim}{1}')$.

We can easily verify that :

a) $Y'Q\,Y = (I - \underset{\sim}{1}\,\underset{\sim}{1}'\,D_p)\,W(I - D_p\,\underset{\sim}{1}\,\underset{\sim}{1}')$

b) The distances calculated from either W or Y'Q Y are the same.

c) Y' Q Y admits the eigenvector $D_p\,\underset{\sim}{1}$ associated with the zero eigenvalue so that for all other eigenvectors U, we have $U'D_p\,\underset{\sim}{1} = 0$.

Because of these remarks, we choose to calculate W and then to work with $(I - \underset{\sim}{1}\,\underset{\sim}{1}'\,D_p)\,W\,(I - D_p\,\underset{\sim}{1}\,\underset{\sim}{1}')$.

II-2 Let us now take a point of view less traditional in which the data are a couple (D^*, D_p) in which :

D_p has the same significance as in II-1.

D^* is a n x n matrix of dissimilarities between subjects obtained either as the result of manipulations of variables, eventually qualitatively, or in a purely subjective way.

Given $\mathcal{D}$ the matrix n x n the elements of which are the squares of those of D^* , the Torgerson method defines a matrix of scalar products W^* by the formula :

$$W^* = -(I - \underset{\sim}{1}\,\underset{\sim}{1}'D_p)\,\mathcal{D}\,(I - D_p\,\underset{\sim}{1}\,\underset{\sim}{1}')$$

which is such that :

$$D^*_{jk} = \left[W^*_{jj} + W^*_{kk} - 2\,W^*_{jk}\right]^{1/2}$$

but nothing assures that there exists a real configuration of n points, accepting as mutual distances the elements of D^*.

Because the obtaining of real configurations of points is the center of all methods of data analysis, we have chosen in this case to substitute for W^* the positive definite matrix W which is its approximation in the sense of least squares. If $\{U_i\ ;\ i = 1, \ldots, n\}$ is the set of eigenvectors (of unit norm) of W^* associated with the eigenvalues $\{\lambda_i\ ;\ i = 1, \ldots, n\}$ we know that :

$$W = \sum_{i \in I} \lambda_i\, U_i\, U_i' \quad \text{where} \quad I = \left\{ i / \lambda_i > 0 \right\}$$

We remark that for all numbers k, the matrix $W^* + kI$ has the same eigenvectors as W^* , the eigenvalues being $\lambda_i + k$, thus by taking $k^* = \max_{\lambda_i < 0} |\lambda_i|$, $W^* + k^*I$ is positive semi-definite and the distances corresponding to it are equal to :

$$\left[D^{*2}_{jk} + 2\,k^*\right]^{1/2} = D^*_{jk}\left[1 + \frac{2\,k^*}{D^{*2}_{jk}}\right]^{1/2}$$

Consequently, working with the matrix $W^* + k^* I$ (a method suggested to me by J.P. Pagès) does not modify the order of distances and thus is a happy alternative to the "additive constant method". Secondly, the eigenvectors associated with the largest eigenvalues of W and $W^* + k^* I$ are the same (those also of W^*). The configurations given by W and $W^* + k^* I$ are thus neighbours and are only differenciated by the fact that the eigenvectors of W have as norm $\lambda^{1/2}$ where those of $W^* + k^* I$ have as norm $(\lambda + k^*)^{1/2}$.

Thus we consider that in this less traditional point of view we can substitute for the couple (D^*,D_p) the couples (W,D_p) or $(W^* + k^* I, D_p)$. No matter which solution is chosen, the study is characterized by a couple which we call (W,D_p).

II-3 Let us now look at the operator on R^n defined by the matrix WD_p. Our previous choices put WD_p in the rank $k \leqslant n$.

If U_i is aneigenvector of WD_p, $D_p^{1/2} U_i$ is the eigenvector of $D_p^{1/2} W D_p^{1/2}$ for the same eigenvalue. It follows:

a) that the eigenvectors of $W D_p$ are D_p-orthogonal.

b) that if the U_i are chosen as D_p-orthogonal, then we have :

$$\sum_{i=1}^{n} \lambda_i U_i U_i' = W$$

At this point, the operator WD_p appears as characteristic of the couple (W,D_p), and comparing two studies E_1 and E_2 is the same as comparing the operators $W_1 D_1$ and $W_2 D_2$ which are associated with them.

It is known that for the square matrices A and B, Tr(AB) is a scalar product, the corresponding norm being $\left[\mathrm{Tr}(A^2)\right]^{1/2}$. If thus two studies E_1 and E_2 lead to the matrices $W_1 D_1$ and $W_2 D_2$, we can define a distance between E_1 and E_2 by :

$$\begin{aligned} d_1(E_1,E_2) &= \left[\mathrm{Tr}(W_1 D_1 - W_2 D_2)^2\right]^{1/2} \\ &= \left[\mathrm{Tr}(W_1 D_1)^2 + \mathrm{Tr}(W_2 D_2)^2 - 2\,\mathrm{Tr}(W_1 D_1 W_2 D_2)\right]^{1/2} \end{aligned}$$

In order to make the proportional matrices equivalent, we use as well

$$\begin{aligned} d_2(E_1,E_2) &= \left[\mathrm{Tr}\left[W_1 D_1 / \left[\mathrm{Tr}(W_1 D_1)^2\right]^{1/2} - W_2 D_2 / \left[\mathrm{tr}(W_2 D_2)^2\right]^{1/2}\right]^{1/2}\right]^{1/2} \\ &= 2\left[1 - 1\mathrm{Tr}(W_1 D_1 W_2 D_2) / \left[\mathrm{Tr}(W_1 D_1)^2\ \mathrm{Tr}(W_2 D_2)^2\right]^{1/2}\right]^{1/2} \end{aligned}$$

To clarify the significances of these distances, we must make the following remarks :

a) Let us represent a study by a point P of $R^{n \times n}$ of coordinates $p_{(i-1)n+j} = (W D_p)_{ij}$. Using in $R^{n \times n}$ the identity metric, if p^1 and p^2 are the points representative of the two studies E_1 and E_2, the distance between p^1 and p^2 in $R^{n \times n}$ is $d_1(E_1,E_2)$.

b) Let us now return to the point of view which we called "traditionnal" and suppose that we have to compare two studies (X,I,D_p) and (Y,I,D_p) given for the same individuals with the same weights and the identity metrics in both cases. If we remark that :

$X D_p X' = S_{11}$ the sample covariance matrix as estimated from the variables defining the rows of X,

$Y D_p Y' = S_{22}$ the sample covariance matrix as estimated from the variables defining the rows of Y,

$Y\ D_p\ X' = S_{21}$ and $X\ D_p\ Y' = S_{12}$,

then $Tr(W_1 D_p\ W_2 D_p) = Tr(S_{12}\ S_{21})$

and $Tr(W_1 D_p\ W_2 D_p) / \left[Tr(W_1 D_p)^2\ Tr(W_2 D_p)^2\right]^{1/2} = Tr(S_{12} S_{21}) / \left[Tr(S_{11} 2) Tr(S_{22})^2\right]^{1/2}$

The expressions on the right-hand side of the two preceding equalities are analogues of the coefficients COVV and RV introduced in (2) for two random vectors defined on the same probability space. By extension we thus use :

$$COVV(E_1, E_2) = Tr(W_1\ D_p\ W_2\ D_p)$$

$$RV(E_1, E_2) = Tr(W_1\ D_p\ W_2\ D_p) / \left[Tr(W_1 D_p)^2\ Tr(W_2\ D_p)^2\right]^{1/2}$$

III - MULTIDIMENSIONAL STATISTICAL ANALYSIS

Without going into the details of the demonstrations which the interested reader can find in a recent paper (5), we will now show how the different methods of multidimensional statistical analysis could be presented starting from the coefficient RV.

Note, first, that for every matrix Q, p x p, positive semi-definite, there exists a matrix L, p x t, $t \leq p$, such that Q = LL'. This allows us to describe a study on a matrix X, p x n in the form (X, LL', D_p). If we consider a second study (Y, MM', D_p) with Y of dimensions q x n, for the same individuals and the same weights, we can adopt the notation RV(L'X, M'Y). Note also that the matrix L is such that Q=LL' is determined to within a rotation in R^t. Inasmuch as we want to determine L, we are thus forced to introduce the conditions which allow its determination. In the following table, we can find the description of problems the solutions of which correspond with the methods of multidimensional analysis.

Problems	solutions
Find M, p x t, $t \leq p$ which maximises RV(X, M'X) under the constraint : $M'\ X\ D_p\ X\ M$ diagonal	First t principal components of X
Find M, q x t, $t \leq \inf(p,q)$ which maximises RV(X, M'Y) under the constraint : $M'Y\ D_p\ Y\ M$ diagonal	First t principal components of Y with respect to X
particular case : t = p	$M' = (X\ D_p\ Y')\ (Y\ D_p\ Y')^{-1}$ gives the maximum and $(X-M'Y)D_p Y' = 0$ Thus we have the regression of X with respect to Y.
Find L, p x t, and M, q x t which maximise RV(L'X, M'Y) under the constraints $L'\ X\ D_p\ X'L = M'\ Y\ D_p\ Y'M = I_t$	First t couples of canonical variables
Cas particulier : The individuals are divided into k groups. Y^j is the mean vector in the group containing the j^{th} indidual	First t discriminant functions (L = M)

N.B. - Our approach leads to the introduction of all the methods in terms of comparison of distance matrices. This gives sometimes an unusual point of view ; for example the regression of X with respect to Y is interpreted as the research for the metric MM' to be taken for Y, in such a way that the distances in the study (Y, MM', D_p) are as close as possible to the distances in the study (X,I,D_p).

Let us finish this paragraph by noting that for all the studies given under the form of a couple (D, D_p) we have associated a couple (W, D_p). The factorization of W in the form $W = X^{P}X$ allows the presentation of the study in the form (X,I,D_p) and thus the extension of all the classical methods of multidimensional analysis for data (D,D_p).

IV - CHOICE OF VARIABLES

The use of the RV coefficient and in particular the results of the previous paragraph allows us to study much more clearly the problem of the choice of variables by clearly showing that the choice is always accompanied by a choice of the metric to be used. Let us take two given studies (X,LL',D_p) and (Y,MM',D_p) with X, p x n, and Y, q x n, given for the same individuals (possibly Y = X and M = L).

The problem is to find Z , t x n, t $\leq$q, extracted from Y, which for a metric NN' where N is given or has to be found, realizes the maximum of RV(L'X, N'Z).

Problem 1 : $N = I_t$

We want to use the variables Z as they are. The distance between two individuals is the classical euclidean distance. The study (Z, I_t, D_p) is, from the point of view of the principal components method the closest possible to $(X. LL', D_p)$.

Problem 2 : N positive diagonal

The affecting of weight to the chosen variables is accepted (i.e. the changing of units) but we want to conserve the experimental significance of the variables. We look for both Z and N.

Problem 3 : Any N

Both Z and N have to be found, which means that the principal components analysis on (X, LL', D_p) and (Z, NN', D_p) must be as close as possible. Remark that for a given Z, N is given by the principal components of Z with respect to X.

Problem 4 : $N' = (L' X D_p Z') (Z D_p Z')^{-1}$

Following the results of the previous paragraph, we realize the regression of L'X with respect to Z. The problem is thus to find the best sub-set of Y from the point of view of the regression of L'X.

Problem 5 : $LL' = (X D_p X')^{-1}$, $MM' = (Z D_p Z')^{-1}$

One can see as well (by calculating RV for example), that the variables retained are those that are susceptible to allow the canonical analysis, with X the most satisfying.

Problem 6 : X matrix p x k of the averages of k groups $(LL')^{-1}$ inter-groups covariance matrix for X .

Y matrix q x k of the averages of the same k groups $(NN')^{-1}$ inter-groups covariance matrix for Z extracted from Y. The maximisation of RV(L'X, N'Z) is the same thing as looking for the best sub-set of the variables Y, in the sense where it allows a discriminant analysis, the closest possible to that which allows X.

V - JOINT ANALYSES OF SEVERAL DATA MATRICES

V-1 Let us define $\{E_i \; ; \; i \in I\}$ a family of data analyses on the same individuals with the same weights. . With $\{E_i \; ; \; i \in I\}$ is associated the family of operators $\{W_i \, D_p \; ; \; i \in I\}$. We propose to study the proximities and the differences of E_i.

Consider the matrix C the elements of which are $C_{ij} = \mathrm{COVV}(E_i, E_j)$ C is the matrix of scalar products between the operators and, following the remark (a) in II_3, there exists a real configuration of points compatible with C. The canonical factorisation of C gives a visualisation of this configuration of points in which each study is represented by a point. The distances between the points are $(C_{ii} + C_{jj} - 2\, C_{ij})^{1/2}$. Particularly if $W_j \, D_p = k \, W_i \, D_p$, the origin and the points p^i and p^j associated with E_i and E_j are colinear and $\overrightarrow{Op^j} = k\overrightarrow{Op^i}$. The converse is true . Of course practical thought leads to limit the representation to two or three eigenvectors of C associated with the largest eigenvalues. The quality of this approximation is appreciated by the usual tools : rate between the extracted eigenvalues and Tr(C) for example.

Rather than work with C, we can choose to work with R the elements of which are $R_{ij} = \mathrm{RV}(E_i, E_j)$. In this case, studies leading to proportional operators are represented by confused points. Moreover, for all $i \in I$, $\|\overrightarrow{Op^i}\| = 1$ and $\langle \overrightarrow{Op^i}, \overrightarrow{Op^j} \rangle = \mathrm{RV}(E_i, E_j)$. If the representation by the first two eigenvectors of R keeps, for the projections of the vectors Op^i, the norms neighbour to the unity, then the value of $\mathrm{RV}(E_i, E_j)$ is rather equal to the cosine of the angle made by the projections of $\overrightarrow{Op^i}$ and $\overrightarrow{Op^j}$.

V-2 The coefficients $\mathrm{COVV}(E_i, E_j)$ and thus $\mathrm{RV}(E_i, E_j)$ are always non-negative. So, the matrices C and R have a first eigenvector , the elements of which are always non-negative. Let us take $\{\alpha_i \; ; \; i \in I\}$ as the components of this vector. We have the following theorem :

For all $\{\beta_i \; ; \; i \in I\}$ such that for all i , $\beta_i \geqslant 0$ and $\sum_{i \in I} \beta_i^2 = \sum_{i \in I} \alpha_i^2$, we have :

$$\sum_{i \in I} \left[\mathrm{RV}\Big(\sum_{j \in I} \beta_j \, W_j \, D_p\Big) \right]^2 \leqslant \sum_{i \in I} \left[\mathrm{RV}\Big(\sum_{j \in I} \alpha_j \, W_j \, D_p, \; W_i \, D_p\Big) \right]^2$$

Thus, $W \, D_p = \sum_{i \in I} \alpha_i \, W_i \, D_p$ is an operator which constitutes the best compromise between all the studies in the sense of the criteria used in the theorem. Because it is a positive linear combination of positive semi-definite operators, $W \, D_p$ is also positive semi-definite. We can thus obtain by a canonical factorisation a visualisation of the objects which constitutes a compromise between all the studies.

<u>Remark</u> : We have as well an analogous theorem for COVV.

V-3 The space of the representations of objects as they are seen by the compromise can be used as reference space into which each one of the initial studies is projected. This possibility allows us to see the way in which each one of the studies differs from the compromise. In the case of a chronological study, it allows us to visualise the evolution of the different objects during the time. Examples of the application of the theory developed in the fifth section are treated in (3). A program exists and the way in which it can be obtained will be given by the author.

REFERENCES

Caillez F. and Pages J.P. (1976), Introduction à l'Analyse des données - SMASH 9, rue Duban, 75016 PARIS.

Escoufier Y. (1973), Le traitement des variables vectorielles, Biometrics 29, p. 751-760.

L'Hermier des Plantes H. (1976), Structuration des tableaux à trois indices de la statistique. Thèse de 3ème cycle. Centre de Recherche en Informatique et Gestion, avenue d'Occitanie, Montpellier.

Pages J.P., A propos des opérateurs d'Y. Escoufier, séminaires IRIA, classification automatique et perception par ordinateur, 1974, p. 261-271.

Robert P. and Escoufier Y., A unifying tool for linear multivariate statistical methods : the RV. coefficients. Applied Statistics (à paraître).