Recent Developments in Statistical Inference and Data Analysis; K. Matusita, editor

EXPLORATORY DATA ANALYSIS WHEN DATA ARE MATRICES

Yves Escoufier

C.R.I.G. - Av. d'Occitanie
34075 Montpellier Cedex - France

Le but du travail est de montrer comment K matrices peuvent être comparées visuellement; comment un compromis peut être défini entre les K matrices; et comment chacune de ces K matrices peut être comparée au compromis.

We consider a set of K matrices. We find a graphical representation which allows us to do an overall comparison of the matrices. We define a new matrix which can be considered as a good compromise between all the initial matrices. We compare the view of the individuals given by the compromise with the views given by the initial matrices.

1. Introduction

We consider the four matrices which are given below. They give four different measures of similarity between four individuals.

$$S_1 = \begin{vmatrix} 2 & 0 & -2 & 0 \\ 0 & 2 & 0 & -2 \\ -2 & 0 & 2 & 0 \\ 0 & -2 & 0 & 2 \end{vmatrix} \qquad S_2 = \begin{vmatrix} 5 & 3 & -5 & -3 \\ 3 & 5 & -3 & -5 \\ -5 & -3 & 5 & 3 \\ -3 & -5 & 3 & 5 \end{vmatrix}$$

$$S_3 = \begin{vmatrix} 5 & -3 & -5 & 3 \\ -3 & 5 & 3 & -5 \\ -5 & 3 & 5 & -3 \\ 3 & -5 & -3 & 5 \end{vmatrix} \qquad S_4 = \begin{vmatrix} 8 & 0 & -8 & 0 \\ 0 & 8 & 0 & -8 \\ -8 & 0 & 8 & 0 \\ 0 & -8 & 0 & 8 \end{vmatrix}$$

We are concerned with the three following problems.

a) We want to find a graphical representation which allows us to do an overall comparison of the four matrices. We must be able to recognize that two matrices are more similar than two others.

b) We want to define a new matrix which could be considered as a good compromise between all the initial matrices. The compromise is a weighted mean of the initial matrices. Its study will lead us to a global knowledge of the similarity between the individuals.

c) At the end, we want to be able to compare the view of the individuals given by the compromise with the views given by the initial matrices.

In this paper, we use this simple example of four matrices to present all the details of the calculus. In the general case, we will consider K similarity matrices relating to the same n individuals. We will see that the solution of the first problem is given by the eigenvectors of a particular K x K matrix. The solution of the second problem is given by the eigenvectors of a n x n matrix. And the solution of the third problem is obtained through the regression of the eigenvectors of each of the initial matrices on the eigenvectors of the compromise matrix. These results allow us to appreciate the complexity of the problems and to affirm that it is easy for a statistician to realize this type of study.

Later, we will give some indications on the extensions of the methodology when the initial matrices are not similarity matrices. The interested reader can find detailed applications of the method in the recent french papers refered to at the end of this article.

2. K Similarity matrices: an example

a) We consider $E = \{S_k / k = 1, \cdots, K\}$ a set of K similarity matrices relating to the same n individuals. We suppose that, for each k, S_k is a positive semi-definite or definite matrix. We know that an observed similarity matrix is seldom positive semi-definite; but we can find in the statistical literature [5] many ways to approximate a non definite matrix by a semi-definite one. So the positive definite condition is not a strong constraint. It will be important in the next paragraph.

E is a subset of F, the set of the K x K symmetric matrices and it is easy to see that the bilinear form b defined on F by $(U, V) = \mathrm{Tr}\,(UV)$ is an inner product on F.

Guided by our knowledge of the Multidimensional Scaling we define a K x K matrix C with elements: $C_{ij} = \mathrm{Tr}\,(S_i S_j)$.

C will be used as the basis for a graphical representation of the K matrices. We compute the eigenvectors $L_1, \cdots, L_K$ of C and the associated eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K$. It is known that

$C = \sum_{i=1}^{K} \lambda_i L_i L_i'$; we know also that $C^{(r)} = \sum_{i=1}^{r} \lambda_i L_i L_i'$ is the best approximation of rank r for C.

We must remark that the information on the S_k matrices contained in C can be studied through different technics: for example, cluster analysis and Indscal. We choose a method based on the eigenvectors because the solution of our first problem is, by this choice, consistent with the solutions of the following problems.

For our example we have:

$$C = \begin{vmatrix} 32 & 80 & 80 & 128 \\ 80 & 272 & 128 & 320 \\ 80 & 128 & 272 & 320 \\ 128 & 320 & 320 & 512 \end{vmatrix}$$

	i = 1	i = 2
λ_i	944	144
$\sqrt{\lambda}\ L_i$	5.66 14.14 14.14 22.63	0.0 8.49 -8.49 0.0

We see that we have an exact representation in a two dimensional space ($Tr\ (C) = 944 + 144$): the inner products matrix between the points of R^2 with coordinates $\sqrt{\lambda_1}\ L_1$ and $\sqrt{\lambda_2}\ L_2$ is exactly C.

If we look at the graphical representation (figure n°1) we see that S_1 and S_4 are proportional. We see also that S_2 and S_3 have the same norms. The distances between S_1 and S_2, S_1 and S_3, S_4 and S_2, S_4 and S_3 are equal and the distances between S_1 and S_4, S_2 and S_3 are equal. Therefore the four points are the vertices of a square.

b) Now, we deal with the compromise matrix.

When all the eigenvalues of the matrices S_k ($k = 1, \cdots, K$) are non negative, it can be shown that all the elements of C are positive. From that, it follows that all the elements of the first eigenvector of C can be chosen positive. So, if we define $L_1' = (L_{11}, \cdots, L_{1K})$, the matrix

$S = \sum_{i=1}^{K} L_{1k}\ S_k$ is also a positive semi-definite matrix and we have the following theorem:

Theorem: *Let* L_1 *be the first eigenvector of the matrix C. We suppose* $L_1' = (L_{11}, \cdots, L_{1K})$ *and* $L_1'\ L_1 = 1$.
Let α *be any vector of* R^K *such that* $\alpha' = (\alpha_1, \cdots, \alpha_K)$ *and* $\alpha'\alpha = 1$. *Then:*

a) $Tr\ (\sum_{k=1}^{K} \alpha_k\ S_k)^2 \leq Tr\ (\sum_{k=1}^{K} L_{1k}\ S_k)^2 = \lambda_1$,

b) $$\sum_{\ell=1}^{K} \left[Tr\ ((\sum_{k=1}^{K} \alpha_k\ S_k)\ S_\ell) \right]^2 \leq \sum_{\ell=1}^{K} \left[Tr\ ((\sum_{k=1}^{K} L_{1k}\ S_k)\ S_\ell) \right]^2 = \lambda_1^2 .$$

The demonstration is obvious if we remark that

$$Tr\ (\sum_{k=1}^{K} \alpha_k\ S_k)^2 = \alpha'\ C\ \alpha$$

and $$\sum_{\ell=1}^{K} \left[Tr\ ((\sum_{k=1}^{K} \alpha_k\ S_k)\ S_\ell) \right]^2 = \alpha'\ C^2\ \alpha .$$

These results look like the results relating to the first principal component in a principal components analysis. As a matter of fact, they are the same: The matrix C takes the place of the correlation matrix. In our example, we obtain the following S matrix:

$$S = \begin{vmatrix} 10.86 & 0.0 & -10.86 & 0.0 \\ 0.0 & 10.86 & 0.0 & -10.86 \\ -10.86 & 0.0 & 10.86 & 0.0 \\ 0.0 & -10.86 & 0.0 & 10.86 \end{vmatrix} .$$

It has two equal eigenvalues associated with the two following eigenvectors (we choose $Y_i'Y_i = \mu_i$)

	i = 1	i = 2
μ_i	21.72	21.72
Y_i	2.33	-2.33
	2.33	2.33
	-2.33	2.33
	-2.33	-2.33

The eigenvectors give the representation of the individuals as they are seen by the compromise matrix (figure n° 2).

It is important to understand the difference between the two figures. The figure n° 1 gives a representation of the matrices (K in the general case). The figure n° 2 gives a representation of the individuals (n in the general case). C is the basis for the figure n° 1. S is the basis for the figure n° 2.

We see that for the compromise matrix, the four points are the vertices of a square. But the question arises: What is the importance of the discrepancy between the representation of the individuals given by the compromise matrix and the representations which can be obtained from the initial matrices S_k?

c) To answer this last question, we define:

Y the matrix of the eigenvectors of S associated with the non-zero eigenvalues ($YY' = S$);

Y_k the matrix of the eigenvectors of S_k ($Y_k\ Y_k' = S_k$)

$\hat{Y}_k = Y\ (Y'Y)^{-1}\ Y'Y_k$ the orthogonal projection of Y_k on the space spanned by the columns of Y.

Identifying the first columns of $\hat{Y}_k$ with the corresponding columns of Y, we can represent in the space spanned by the column of Y (the compromise) the n individuals as they are seen by S_k.

Remark 1:

It is possible to compute the correlation between a column of Y and a column of Y_k. This is a tool to detect some permutation in the order of the eigenvectors of Y_k with respect to those of Y.

Remark 2:

If $Y_k = Y M_k$ and if M_k^{-1} exists, then $\hat{Y}_k = Y_k$.

In our example, it can be shown that we had used the following matrices M_k to construct the matrices S_k:

$$M_1 = \begin{vmatrix} 1/2.33 & 0 \\ 0 & 1/2.33 \end{vmatrix} \qquad M_2 = \begin{vmatrix} 2/2.33 & 0 \\ 0 & 1/2.33 \end{vmatrix}$$

$$M_3 = \begin{vmatrix} 1/2.33 & 0 \\ 0 & 2/2.33 \end{vmatrix} \qquad M_4 = \begin{vmatrix} 2/2.33 & 0 \\ 0 & 2/2.33 \end{vmatrix} .$$

The figure n° 3 gives the representation of the matrices $\hat{Y}_k$.
We see that S_1 and S_4 are squares and that S_2 and S_3 are rectangles stretched along the different axes. This is easily explained by the choice of the matrices M_i.

Remark 3:

In our example, the rows and the columns of the matrices S_k ($k = 1, \cdots, K$) and S are centered. It follows that the graphical representations of the individuals are also centered. This distinctive feature is not necessary for the progress of the method.

3. Some other results

We will give in this paragraph a general survey of the extensions of the method.

a) We consider first a set of K matrices $\{X_k / k = 1, \cdots, K\}$.

X_k is a n x P_k matrix of measures of P_k quantitative variables on n given individuals. The variables (and the P_k) are possibly different in the K studies. $S_k = X_k X_k'$ is the inner products matrix between the individuals deducted from X_k. All the results of the preceding paragraph still hold for the set $E = \{S_k / k = 1, \cdots, K\}$. If we decide to work with centered variables, then $X_k' X_k$ is proportional to the covariances matrix of the variables in X_k. So the columns of Y_k, which are the eigenvectors of $X_k X_k'$, are proportional to the wellknown principal components of X_k. We note Y the matrix, the columns of which are the eigenvectors of S. Then following the usual practice of principal components analysis, the matrices $(nY'Y)^{-1/2} Y'X_k$ and $(nY'Y)^{-1/2} Y'Y_k$ give the coordinates of the initial variables and of their principal components in the space spanned by the columns of Y.

b) We consider now K qualitative variables. U_k is the n x P_k matrix of indicator functions associated with the variable k defined as follows:

$(U_k)_{ij}$ is 1 if the individual i takes the modality j of the variable k.

$(U_k)_{ij}$ is 0 if the individual i does not take the modality j of the variable k.

Let D_k be the diagonal matrix of the weights of the modalities for the variable k : $D_k = U_k' U_k / n$.

We define $U_k^* = (I - \frac{\underset{\sim}{1}\,\underset{\sim}{1}'}{n})\, U_k$

$$S_k = U_k^*\, D_k\, U_k^{*\prime} / n$$

and C a K x K matrix, the elements of which are $C_{k\ell} = \mathrm{Tr}\,(S_k\, S_\ell)$.

It can be shown [2] that $C_{k\ell}$ is nothing other than the ϕ^2 coefficient between the variables k and ℓ.

The overall comparison of the K qualitative variables is made through the graphical representation given by the eigenvectors of C. The representation of the individuals given by the eigenvectors of S has the distinctive feature to be equal to the representation of the individuals in the Correspondence Analysis of the matrix:

$$M = (\sqrt{L_{11}}\, U_1 \,\vdots\, \sqrt{L_{12}}\, U_2 \,\vdots\, \cdots \,\vdots\, \sqrt{L_{1k}}\, U_k)\,.$$

It follows, that the detailed study of the variables is obtained by the representation of the modalities in this Correspondences Analysis.

Remark 4:

Let M be a p x q positive matrix. We define Δ_p as the diagonal matrix, the elements of which are the sums of the rows of M, and Δ_q as the diagonal matrix, the elements of which are the sums of the columns of M.
Looking only at the results we can say that in the Correspondences Analysis of M, the p items associated with the rows of M are represented by points, the coordinates of which are the components of the eigenvectors Ξ of $\Delta_p^{-1}\, M\, \Delta_q^{-1}\, M'$; the q items associated with the columns of M are represented from the eigenvectors T of $\Delta_q^{-1}\, M'\, \Delta_p^{-1}\, M$. If Λ is the diagonal matrix of the non-zero eigenvalues of the two preceding matrices, Ξ and T are choosen in such a way that: $\Xi'\, \Delta_p\, \Xi = T'\, \Delta_q\, T = \Lambda$. The two matrices have an eigenvalue equal to 1 which is irrelevant for the representations.

c) At the end of this general survey, we consider K contingency tables: For two given qualitative variables, X_k is a p x q contingency table and we have K such tables.

Let $D_p\,(k)$ and $D_q\,(k)$ be the diagonal matrices of marginal weights in X_k.
We define the so called Burt's tables:

$$S_k = \begin{vmatrix} D_p\,(k) & X_k \\ X_k' & D_q\,(k) \end{vmatrix}\,.$$

Let C be the K x K matrix with elements

$$C_{k\ell} = \mathrm{Tr}\,(S_k\, S_\ell)\,.$$

The S_k tables are again symmetrical matrices so, the eigenvectors of the matrix C lead to a representation of the K contingency tables.

We define S from the first eigenvector of C by

$$S = \sum_{k=1}^{K} L_{1k} S_k \Big/ \sum_{k=1}^{K} L_{1k} = \begin{vmatrix} D_p & X \\ X' & D_q \end{vmatrix}.$$

S is the compromise matrix and can be utilized as a Burt's table. The Correspondences Analysis of S is equivalent to the Correspondences Analysis of X and gives a representation of the modalities of the two variables. The tables X_k can be compared together by the way of the projections of their rows and columns as supplementary points in the Correspondences Analysis of X.

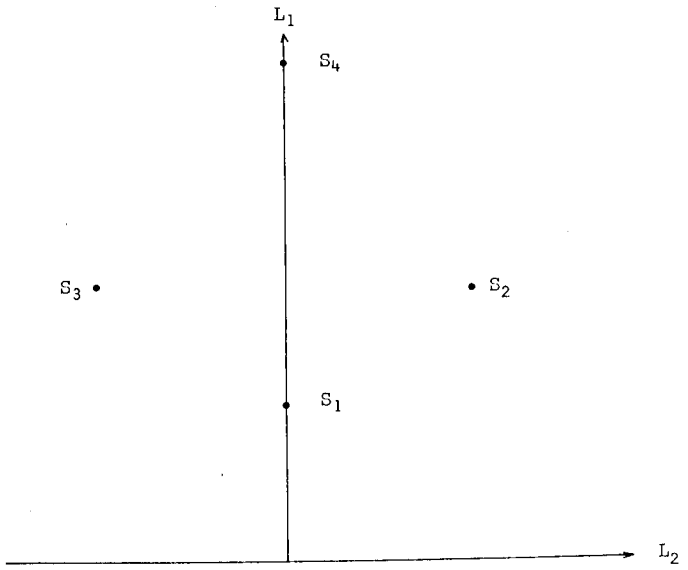

Figure n° 1 :

Overall comparison of the matrices

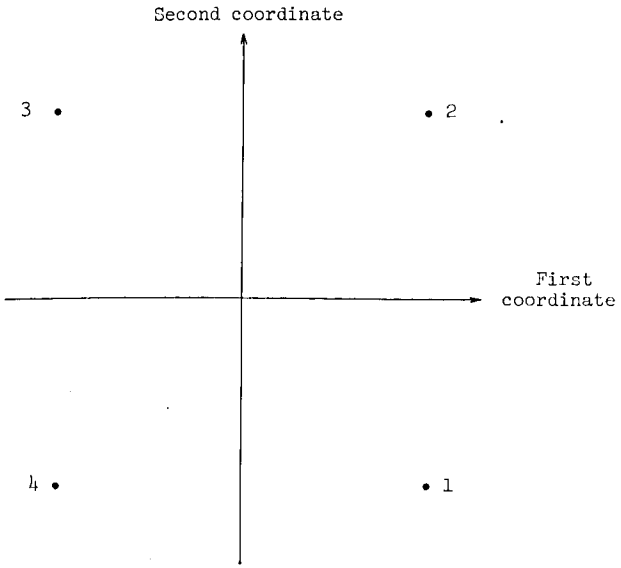

Figure n° 2 :

The compromise

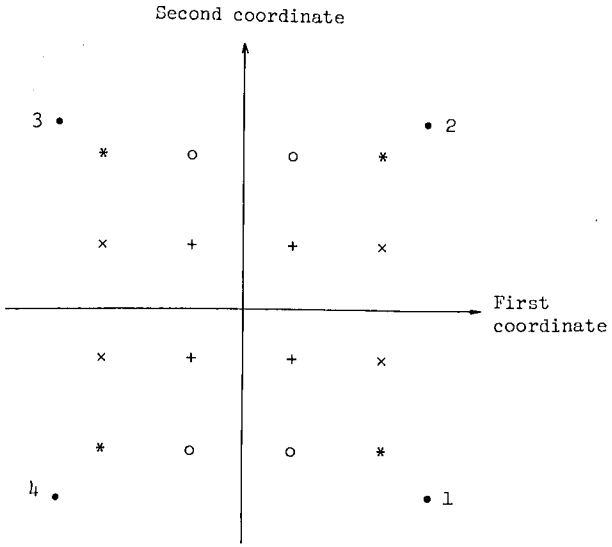

Figure n° 3 :

Graphical representations
of the matrices $\hat{Y}_k$
in the space of the compromise matrix
(S_1 = + ; S_2 = × ; S_3 = o ; S_4 = *)

REFERENCES

[1] BERNARD, M.C., DIAZ-LLANOS, F.J. et ESCOUFIER, Y., La méthode Statis: une application à l'évolution des campagnes langue-dociennes, C.R.I.G. Av. d'Occitanie, 34075 Montpellier Cedex, Rapport Technique n° 7905 (1979).

[2] CAILLEZ, F. et PAGES, J.P., Introduction à l'Analyse des données, SMASH, 9, rue Duban, 75016 Paris (1976).

[3] CAZES, P., BONNEFOUS, S., BAUMERDER, A. et PAGES J.P., Description cohérente des variables qualitatives prises globalement et de leurs modalités, S.A.D. 2 (1976) 48-62.

[4] ESCOUFIER, Y., Opérateur associé à un tableau de données, Annales de l'Insee n° 22-23 (1976) 165-179.

[5] ESCOUFIER, Y., Cours d'Analyse des Données, C.R.I.G. Av. d'Occitanie, 34075 Montpellier Cedex (1979).

[6] FOUCARD, T., Sur les suites de tableaux de contingence indexés par le temps, S.A.D. 2 (1978) 67-85.

[7] FOUCARD, T., Prévision d'une suite de tableaux de probabilités. Ajustement à des marges données, S.A.D. 2 (1979) 51-71.

[8] FOUCARD, T., Structure des tableaux de probabilités, Description et Prévision, Thèse, Université des Sciences et Techniques du Languedoc, Montpellier (1979).

[9] HILL, M.O., Correspondence Analysis: a neglected multivariate method Applied Statistics, 23 (3) (1974) 340-354.

[10] JAFFRENOU, P.A., Sur l'Analyse des familles finies de variables vectorielles, Thèse, Université Claude Bernard, Lyon I (1978).

[11] L'HERMIER DES PLANTES, H., Structuration des tableaux à trois indices de la statistique, Thèse, Université des Sciences et Techniques du Languedoc, Montpellier II (1976).

[12] L'HERMIER DES PLANTES, H., et ESCOUFIER, Y., A propos de la comparaison graphique des matrices de variance, Biom. J. vol. 20, n° 5 (1978) 477-483.

[13] L'HERMIER DES PLANTES, H., et THIEBAUT, B., Etude de la pluviosité au moyen de la méthode STATIS, R.S.A. vol. XXV n° 2 (1977) 57-58.

[14] MAILLES, J.P., Analyse factorielle des tableaux de dissimilarités, Thèse, Université Pierre et Marie Curie, Paris VI (1978).

[15] STEMMELEN, Prévision des tableaux d'échanges, les marges étant connues, Cahier du Buro n° 28 (1977).