Operator related to a data matrix: a survey

Yves Escoufier

Equipe de Probabilités et Statistique, Département des Sciences mathématiques, Université de Montpellier II, Case Courrier 051,Place Eugène bataillon, 34095 Montpellier cedex 5, France

Abstract: The reading of this article will allow the readers to appropriate the data analysis approach which is proposed. The first paragraph gives the basic tools: the triplet (X, Q, D), the operator related to a data matrix and the coefficient RV. The two following paragraphs show how these tools are used for reading out and solving the problems of joint analysis of several data matrices and of principal component analysis with respect to instrumental variables. The conclusion reminds of the construction of this approach along the past thirty five years.

1 The initial choices

1.1First choice: the triplet (X,Q,D)

When a researcher collects an nxp data array, X, of the values taken by n individuals on p variables, he generally has two objectives:

1. The comparison of the variables. If he chooses to conduct this comparison by the way of a linear correlation coefficient, he will use a positive diagonal matrix D which defines the weights attached to each individual.
2. The comparison of the individuals. If he chooses for that to compute a distance between the individuals, he will need a pxp symmetric positive definite matrix Q. In the simplest case, Q is a diagonal positive pxp matrix defining the scale of the different variables. In the general case, $Q=L^tL$ where L is a pxp matrix of rank p which can be

viewed as a linear transformation of X such that Y=XL is substituted to X.

From the preceding considerations it follows that when we speak of a data analysis, we must considered the triplet (X, Q, D) to describe the data and their use.

1.2 Second choice: the operator XQ^tXD

Let us consider now that we are mainly concerned by the dispersion of the individuals showed by the transformed data array Y. A usual way to study the dispersion is to realize a principal component analysis of Y. The mapping of the individuals in the space spanned by the principal components will give a way for studying the likeness of the individuals. For simplicity, we suppose that $Y= (I_{nxn}-1_nD^t1_n)\ Y$, which means that the columns of Y are centred for the weights given by D. Because Y= XL, it is the same for X. Let $S= {}^tYDY$, the covariance matrix of Y and $\{(z_{\alpha,}\lambda_\alpha),\ \alpha= 1,p\}$ the eigenvectors and eigenvalues of S such that $Sz_\alpha=\lambda_\alpha z_\alpha$ with ${}^tz_\alpha z_\beta= \delta_{\alpha\beta}$.. Then, $\{(\psi^\alpha= Yz_\alpha/\sqrt{\lambda_\alpha}),\ \alpha= 1, p\}$ are the principal components and ${}^t\psi^\alpha D\psi^\beta= \delta_{\alpha\beta}$.

Proposition 1.2.1:

For the principal components we have: $XQ^tXD\psi^\alpha= \lambda_\alpha\psi^\alpha$

Proof: from ${}^tYDYz_\alpha = \lambda_\alpha z_\alpha$ it comes: $Y^tYD(Yz_\alpha/\sqrt{\lambda_\alpha}) = \lambda_\alpha(Yz_\alpha/\sqrt{\lambda_\alpha})$ and thus $XQ^tXD\psi^\alpha = \lambda_\alpha\psi^\alpha$

So, as long as our interest for studying the dispersion of the individuals lies in the principal components of (X,Q,D), all the needed information is given by the eigenvectors and eigenvalues of the operator $WD= XQ^tXD$ which will be called the operator related to the study (X,Q,D).

Proposition 1.2.2

If $\varphi_\alpha= {}^tL^{-1}z_\alpha$ then:

1. ${}^tXDXQ\varphi_\alpha= \lambda_\alpha\varphi_\alpha$
2. ${}^t\varphi_\alpha Q\varphi_\beta= \delta_{\alpha\beta}$
3. $\psi^\alpha= XQ\varphi_\alpha/\sqrt{\lambda_\alpha}$
4. $\varphi_\alpha= {}^tXD\psi^\alpha/\sqrt{\lambda_\alpha}$

Proof:

1. ${}^tYDYz_\alpha = \lambda_\alpha z\alpha \Leftrightarrow {}^tL{}^tXDXLz_\alpha = \lambda_\alpha z\alpha$
   $\Leftrightarrow {}^tXDXL{}^tL({}^tL^{-1}z_\alpha) = \lambda\alpha\ ({}^tL^{-1}z_\alpha)$
2. ${}^t\varphi_\alpha Q\varphi_\beta = {}^tz_\alpha L^{-1}(L{}^tL){}^tL^{-1}z_\beta = {}^tz_\alpha z_\beta = \delta_{\alpha\beta}$
3. $\psi^\alpha = Yz_\alpha/\sqrt{\lambda_\alpha} = XL\ {}^tL\varphi_\alpha / \sqrt{\lambda_\alpha} = XQ\varphi_\alpha / \sqrt{\lambda_\alpha}$
4. ${}^tXD\psi^\alpha = {}^tXDXQ\varphi_\alpha / \sqrt{\lambda_\alpha} = \sqrt{\lambda_\alpha}\varphi_\alpha$

Proposition 1.2.3

Let Ψ (respectively Φ) be the matrix with ψ^α as column α (respectively φ_α) and Λ the diagonal matrix with $\Lambda_{\alpha\alpha}= \lambda_\alpha$. We will note $\Psi^{[k]}$ and $\Phi^{[k]}$ the k first columns of Ψ and Φ and $\Lambda^{[k]}$ the kxk diagonal matrix constructed from the first k rows and columns of Λ. Then:

1. $\Psi^{[k]}\Lambda^{[k]t}\Psi^{[k]}$ is the best approximation of $XQ{}^tXD$ and
   $Tr[(XQ{}^tXD - \Psi^{[k]}\Lambda^{[k]t}\Psi^{[k]}D)^2]= \Sigma_{\ i=k+1,\, I}\ \lambda_i^2$
2. $\Phi^{[k]}\Lambda^{[k]t}\Phi^{[k]}Q$ is the best approximation of ${}^tXDXQ$ and
   $Tr[{}^tXDXQ - \Phi^{[k]}\Lambda^{[k]t}\Phi^{[k]}Q)^2]= \Sigma_{i=\ k+1,I}\,\lambda_i^2$
3. $D^{½}\Psi^{[k]}\Lambda^{[k]t}\Phi^{[k]}Q^{½}$ is the best approximation of $D^{½}XQ^{½}$ and
   $Tr[D^{½}\Psi^{[k]}\Lambda^{[k]½t}\Phi^{[k]}Q - D^{½}XQ^{½})^2]= \Sigma_{\ i=k+1,\, I}\ \lambda_i$

The *proof* is a part of more general results given in (Sabatier et al. 1984).It is easy to see that the usual practices of principal components analysis on the covariance matrix and on the correlation matrix correspond respectively to the choices $Q= I_{pxp}$ and $Q= [diag({}^tXDX)]^{-1}$

1.

Consider now a contingency table $P= (P_{ij},\ i= 1,I;\ j= 1,J)$ with the usual notations for the margins $(P_{i.},i= 1,I)$ and $(P_{.j},\ j= 1,J)$.With the $P_{i.}$ (respectively the $P_{.j}$) we construct a diagonal matrix D_I (respectively D_J). Let $X= D_I^{-1}(P - D_I1_I{}^t1_JD_J)\ D_J^{-1}$. It is easy to see that $XD_J1_J = 0$ and ${}^t1_ID_IX = 0$.
The well-known correspondence analysis method can be viewed as the principal components analysis of the triplet (X, D_J, D_I).

1.3 Third choice: the RV coefficient

Consider now two studies $E_1 = (X_1, Q_1, D)$ and $E_2 = (X_2, Q_2, D)$ for the same individuals and the same D matrix. This is the usual situation when you want to study the links between two sets of variables.
The respective principal component analyses of E_1 and E_2 lead to two configurations of the individuals constructed with the two sets of principal components of W_1D and W_2D. It is natural to look for a comparison of these two operators.
Let S (D) be the set of the D – symmetric nxn matrices, i.e the set of matrices nxn A such that $DA = {}^tAD$. S (D) contains all the operators WD.
The symmetrical bilinear form Tr(AB) is positive on S (D). Hence, it defines a scalar product on S (D).By similarity with the usual statistical vocabulary, we define:

1. $COVV(W_1D, W_2D) = Tr(W_1DW_2D)$

2. $VAV(W_1D) = Tr[(W_1D)^2]$

3. $RV(W_1D,W_2D) = TR(W_1DW_2D)/[Tr[(W_1D)^2]Tr[(W_2D)^2]]^{½}$

The followings results help fort he understanding of the significance of RV. Their proofs are given in (Escoufier 1986)

1. For any (X_1, Q_1, D) and (X_2, Q_2, D): $0 \leq RV(W_1D,W_2D) \leq 1$
2. RV $(W_1D, W_2D) = 1$ if and only if $W_1 = kW_2$ for some non zero scalar k.
3. If Q_1 and Q_2 are positive definite, RV $(W_1D, W_2D) = 0$ if and only if ${}^tX_1DX_2 = 0$.
4. Let X_1 and X_2 be single variables and $Q_1 = Q_2 = 1$.Then:
   COVV $(W_1D, W2D) = [cov(X_1, X2)]^2$
   RV $(W_1D, W2D) = r^2(X_1, X2)$
5. Let X_1 be a single variable and $Q_1 = 1$. Let p_2 be the number of variables in X_2. We choose $Q_2 = ({}^tX_2DX_2)^{-1}$. Then:
   RV $(W_1D, W2D) = R^2_{X1/X2}/\sqrt{p_2}$
   where $R_{X1/X2}$ is the multiple correlation coefficient between X_1 and the variables in X_2.
6. Let $E_1 = (X_1, ({}^tX_1DX_1)^{-1}, D)$ and $E_2 = (X_2,({}^tX_2DX_2)^{-1}, D)$ then:
   RV $(W_1, W_2) = \Sigma_{i=1,p2}\, \rho_i^2/\sqrt{p_1p_2}$
   where ρ_i is the canonical correlation coefficient of rank I between X_1 and X_2.

7. Let X_1 be a single variable and $Q_1 = 1$. We suppose that X_2 is the nxp_2 array of the indicators variables for a qualitative variable x_2 with p_2 modalities. If the individual i takes the modality j, $X_{2i}^{j} = 1$ and $X_{2i}^{k} = 0$ for $k \neq j$. We choose $Q_2 = ({}^tX_2DX_2)^{-1} = D_2^{-1}$ the inverse of the diagonal matrix of the weights of the modalities. Then:
   $RV(W_1D, W_2D) = \eta_{X1/X2}{}^2/\sqrt{p_2}$
   where $\eta_{X1/X2}{}^2$ is the rate of correlation between the quantitative variable X_1 and the qualitative variable x_2.
8. X_1(respectively X_2) is the array nxp_1 (respectively nxp_2) of the indicator variables of the qualitative variable x_1(respectively x_2). We choose $Q_1 = D_1^{-1}$ and $Q_2 = D_2^{-1}$. Then:
   $RV(W_1D, W_2D) = (\chi^2/n+1)/\sqrt{P_1P_2}$
9. If moreover, the columns of X_1 and X_2 are centred for D, we have:
   $RV(W_1D, W_2D) = \chi^2/(n\sqrt{((p_1-1)(p_2-1))}) = T^2$
   where T^2 is the Tchupprov coefficient.

1.4 Bibliographical hints

The two concepts of operator related to a data matrix and RV coefficient have been first introduced in (Escoufier 1970) and (Escoufier 1973).The distribution of the RV coefficient has been studied in (Cléroux and Ducharme 1989). The study has been enlarged to rank data in (Cléroux et al. 1994).Going from matrices language to linear applications language allows to introduce the duality diagram for which the triplet (X, Q, D) can be seen as a summary. We do not follow this point of view here. Interested readers will find detailed presentation of this approach either in the book written by its pioneers (Caillez and Pagès 1976) either in (Escoufier 1987). A very complete recent **R** package and courses are available at http://pbil.univ-lyon1.fr/R/enseignement.html with comments in French and in English; a large collection of sets of data is proposed .See also (Chessel et al. 2004).

2 .Joint analysis of several data matrices (the STATIS method)

Let us consider a set of data analyses $\{(X_i, Q_i, D); i= 1,I\}$ on the same individuals provided with the same weights and the family of related operators $\{W_iD,\ i=1,I\}$. Our aim is to study the proximities and the differences between these I analyses.

2.1 Global comparison of the data analyses (Intrastucture)

Let C be the IxI matrix with elements C_{ij}=COVV (W_iD, W_jD). Let r be the rank of C ($r \leq I$). We note Γ the Ixr matrix of the eigenvectors of C and Θ the rxr diagonal matrix of the eigenvalues. By the spectral decomposition theorem, we have: $C = \Gamma\Theta{}^t\Gamma$ with ${}^t\Gamma\Gamma= I_{rxr}$.
So there exists a configuration of points $(P_i;i=1,I))$ in R^r such that each data analysis is represented by a point. The coordinates of P_i are the elements of the i^{th} row of $\Gamma\Theta^{½.}$ In this configuration the distance between P_i and P_j is $d(P_i,P_j) = (C_{ii} + C_{jj} - 2C_{ij})^{½}$. Of course practical thought leads to limit the representation to two or three eigenvectors of C associated with the largest eigenvalues. The quality of the approximation is appreciated by the usual tools: rate between the extracted eigenvalues and Tr(C) for example.
If the norms of the operators are very different, it could be better to conduct the same analysis with the matrix R with elements RV (W_id,W_jD). In this case the distance between P_i and P_j is $(2(1 - \mathrm{RV}\ (W_id,W_jD))^{½}$.

2.2 Looking for a summary (the compromise)

We have seen that the quantities COVV are always non negative. So, the matrix C has a first eigenvector, γ_1, the elements of which can be chosen non negative. Let $(\gamma_{1i}, i=1, I)$ these elements.

Proposition 2.2.1

For all $(\beta_i, i=1,I)$ such that $\beta_i \geq 0$ and $\Sigma_{i=1,I}\beta_i^2 = \Sigma_{i=1,I}\gamma_{1i}^2 = 1$, we have:

1. $\mathrm{VAV}(\Sigma_{i=1,I}\beta_iW_iD) \leq \mathrm{VAV}(\Sigma_{i=1,I}\gamma_{1i}W_iD) = \theta_1$
2. $\Sigma_{i=1,I}\ [\mathrm{COVV}\ (\Sigma_{j=1,I}\beta_jW_jD,\ W_iD)]^2$
$\leq \Sigma_{i=1,I}[\mathrm{COVV}(\Sigma_{j=1,I}\gamma_{1j}W_jD,W_iD)]^2 = \theta_1^2$

The *proof* comes from the two following equalities:

1. $VAV(\Sigma_{i=1,I}\beta_i W_i D) = Tr\,[(\Sigma_{i=1,I}\beta_i W_i D)^2] = {}^t\beta C\beta$
2. $\Sigma_{i=1,I}[COVV(\Sigma_{j=1,I}\beta_j W_j D, W_i D)]^2 = {}^t\beta C^2\beta$

These results look like the results obtained in principal component analysis for the first component. As a matter of fact, they are analogous. In principal component analysis, the objects are the variables and the inner product is the usual covariance. Here, the objects are the operators related to the statistical studies and the inner product is COVV.

2.

$WD = \Sigma_{i=1,I}\,\gamma_{1i}W_iD$ which has the largest norm and which maximizes the sum of squares of the inner products with the initial operators is named the compromise of the I studies. As a non negative linear combination of semi definite positive operators, WD is semi definite positive. So let ν the number of non zero eigenvalues of WD and let Ψ the $n x\nu$ matrix of its eigenvectors such that ${}^t\Psi D\Psi = I_{\nu x\nu}$ and Λ the $\nu x\nu$ diagonal matrix of its eigenvalues. The $\Psi\Lambda^{½}$ coordinates give a representation of n points. One point represents one initial individual. The proximity of two points is interpreted as an average likeness of the associated individuals.

2.3 Comparison of the initial studies with the compromise (Interstructure)

2.3.1 Representation of the individuals

Let $\lambda_\alpha^{½}\Psi^\alpha = WD\Psi^\alpha/\sqrt{\lambda_\alpha}$ the coordinates of the individuals on the axis α in the representation associated to the compromise.

We define $\lambda_\alpha^{½}\Psi_k^\alpha = W_kD\Psi^\alpha/\sqrt{\lambda_\alpha}$.

If $W_kD = WD$ then $\Psi_k^\alpha = \Psi^\alpha$ and the representation of the individuals given by W_kD is exactly similar to the representation obtained from the compromise WD .When W_kD goes away from WD, the similarity of the representations decreases.

Moreover, $WD = \Sigma_{k=1,I}\gamma_{1k}W_kD$ and thus $\lambda_\alpha^{½}\Psi^\alpha = \Sigma_{k=1,I}\gamma_{1k}\lambda_\alpha^{½}\Psi^\alpha$

The coordinate of one individual given by the compromise is the barycentre of the coordinates of this individual given by the different studies on the same axis. When the index of the initial studies is the time, it is usual to speak of the trajectories of the individuals in the representation obtained by the compromise.

2.3.2 Representation of the variables

All the variables of the initial studies and all the linear combinations of these variables (for instance the principal components of the initial variables) can be represented as supplementary variables in the compromise. As in principal component analysis, the proximity between one variable's projection and one axis is used form the interpretation of the axis. The coordinate of a variable X^j on the axis α is given by the covariance $c_{\alpha j} = {}^tX^jD\Psi^\alpha$ between X^j and Ψ^α.

From a practical point of view we must keep in mind that at this step the number of points in the representations is very large: nx(I+1) for the individuals only .So there is a hard work at the border of mathematic and computer to construct efficient tools.

2.4 Bibliographical hints

The first publication on this topic is (Escoufier 1977). The two following papers consider the situation of a set of similarity or covariance matrices (Escoufier and L'Hermier 1978), (Escoufier 1980). A very detailed approach of the practical problems can be found in (Lavit et al. 1994); in an application , we must choose between the COVV approach and the RV approach; but choices are also necessary for the representations: they could be centred or not. All the situations are explained carefully .The book written by (Lavit 1988) gives many examples and suggest some software. An application in the field of sensometrics is the subject of the paper by (Schlich 1996). STATIS has been developed for a family of data array. This means that in STATIS we have three indices: one for each array, one for the individuals and one for the variables. So, we can use the term of three – way multiblock for the data. (Vivien and Sabatier 2003) and (Sabatier and Vivien 2OO4) explore extension of STATIS for the joint analysis of two three – way multiblocks or for a four – way multiblock.

3 Principal component analysis with respect to instrumental variables

3.1 The problem and its linear solution

We are concerned now with situations in which we have two sets of data observed on the same individuals provided with the same weights. We suppose that the two sets do not play the same role. One of them is a reference, a target. The objective is to know if the variables of the second set can reconstruct the principal component analysis of the target set. We will note (Y, Q, D) the target study where Y is nxp .The pxp Q matrix is known. Let $W_yD = YQ^tYD$ the operator related to this study. From the second set, we only know the data array X, nxq, and we use the same diagonal matrix of the weights D. We consider the following problem:

> Find M a qxq semi definite symmetric matrix such that
> $Tr[(YQ^tYD - XM^tXD)^2]$ is minimum.

Proposition 3.1.1

Let $R = ({}^tXDX)^{-1}\,{}^tXDYQ^tYDX({}^tXDX)^{-1}$ then :

3. $Tr[(YQ^tYD - XM^tXD)^2] =$
$Tr[(YQ^tYD - XR^tXD)^2] + Tr[(YR^tYD - XM^tXD)^2]$

The proof is given in (Bonifas et al. 1984).The result shows that the discrepancy between YQ^tYD and XM^tXD is the sum of two terms. The first one does not depend on M .It assesses the part of the representation of the individuals given by (Y, Q, D) which will never be reconstructed from a study based on X. The second term depends of the selected M. It shows obviously that the best choice for M is R.

Let $P_X = X\,({}^tXDX)^{-1}\,{}^tXD$ the D-orthogonal projector on the subspace of R^n spanned by the columns of X .We have:

$$XR^tXD = P_XYQ^t(P_XY)D$$

Thus, the operators related to the studies (X, R, D) and (P_XY, Q, D) are identical .They give the same representation of the individuals. This means that the best reconstruction of the representation of the individuals given by (Y, Q, D) is obtained when applying the same Q matrix to the projections of the variables Y on the subspace of R^n spanned by the variables in X.

3.2 Quality of the linear solution

From the properties of the projectors and of the trace, we obtain:

Proposition 3.2.1

1. $Tr({}^tYDYQ) =$
$Tr({}^t(P_XY)D(P_XY)Q)+Tr({}^t((I_{nxn}-P_X)Y)D((I_{nxn}-P_X)Y)Q)$

2. $Tr(YQ^tYD) =$
$Tr((P_XY)Q^t(P_XY)D) + Tr(((I_{nxn} - P_X)Y)Q^t((I_{nxn} - P_X)Y)D)$

The proposition shows that the total inertia of the study (Y, Q, D) can be cut in two parts, one given by the projections of the variables Y on the subspace of R_n spanned by the variables X and one part given by the projections in the orthogonal sub-space.
The second result of the proposition says that the decomposition is true for each diagonal element of YQ^tYD which is the norm of this element multiplied by its weight. This quantity is the inertia of the individual with respect to the origin.
The first result shows that, when Q is diagonal, an analogous decomposition is available for the variances and this is well-known.

The ratio $Tr({}^t(P_XY)D(P_XY)Q) / Tr({}^tYDYQ)$ can be used for appreciating the quality of the reconstruction. When Q = Ipxp, this ratio is the coefficient of Stewart and Loeve. It can be used as a basis for a permutation test of significance for the reconstruction. The following proposition shows that RV $(YQ^tYD, (P_XY)\ Q^t\ (P_XY)\ D)$ can be also used for such an appreciation.

Proposition 3.2.2

$TR((YQ^tYD - (P_XY)Q^t(P_XY)D)^2) =$
$Tr((YQ^tYD)^2) \times (1 - (RV(YQ^tYD, (P_XY)Q^t(P_XY))^2)$

Let Z a third data array observed on the same individuals provided with the same weights .We can do for P_Z and($I_{nxn} - P_X$)Y what we did for P_X and Y. We will obtain a decomposition of the inertia contained in the subspace orthogonal to the space spanned by the variables in X. It is clear that all the results traditionally used in variance analysis for the decomposition of the variance with respect to factors, orthogonal or not, can be called here for the decomposition of the inertia of (Y, Q, D).

3.3 Non linear solution

We consider a finite set of functions (b_k;k=1,l) and F the set of the linear combinations of these functions. Let B^j the matrix nxl in which $B_i^{jk} = b_k(X_i^j)$ where X_i^j is the value taken by the individual i for the variable j. Let t_j a vector with l elements. We consider the non linear transformation of X^j defined by: $f(X^j) = \Sigma_{k=1,l}\, t_{jk}B^{jk} = B^j t_j$.

Let B the nx(qxl) matrix obtained by the juxtaposition of the matrices B^j and T, the (qxl)xq matrix constructed with the (t_j; j=1,q) in such a way that the column j of BT is $B^j t_j$.

We can look for the solution of the following problem: Find T and R such that $Tr((YQ^tYD - BTR^tT^tBD)^2)$ is minimum.

For a given T, we obtain from the preceding paragraphs an explicit solution for R. When R is known, T can be computed through a numerical algorithm. The solution will be obtained by an iterative algorithm based on these two steps.

3.4 Bibliographical hints

The method of principal components with respect to instrumental variables is a part of the basic paper by (Rao 1965).It has been presented through the use of the RV coefficient in (Robert and Escoufier 1976). Canonical analysis and discriminant analysis are also presented in that paper through a problem of RV optimisation. Here we have followed the presentation published in (Bonnifas et al. 1984).For the paragraph 3.2, we recommend (Fraile et al. 1993) which gives the details of an application in the correspondence analysis context and also the paper by (Kazi – Aoual et al. 1995) which gives the explicit first three moments for the distribution of the per-

mutation test. The non linear approach is mainly based on the work of (Durand 1992, 1993). (Imam et al. 1998; Schlich and Guichard 1989) are applications. In the beginning of the research on RV, the choice of variables in principal components analysis was a cause of concern. This topic has been presented as a contributed paper in Compstat 1974 in Vienna (Escoufier et al. 1974).See also (Escoufier and Robert 1979).
We have considered before the operator $(I_{nxn}-P_XY)\ Q\ {}^t(I_{nxn} - P_XY)D$ related to the study $((I_{nxn} - P_X)Y, Q, D)$.In such a study, all the information orthogonal to the X variables has been deleted. Consider Z a qxr matrix and $P_Z = Z\ ({}^tZQZ)^{-1}ZQ$. The array $(I_{nxn} - P_X)\ Y\ {}^t(I_{qxq} - P_Z)$ has its columns D – orthogonal to X and its rows Q – orthogonal to Z. Some authors use these results in correspondence analysis to avoid linear, quadratic or cubic components. They introduce suitable constraints matrices X and Z (Beth 1997), (D'Ambra et al. 2002).

4 Conclusions

It could be useful in this survey to remind of the construction of the results along the years.

1. First Steps: The initial work (Escoufier 1970) was concerned by the sampling of variables in a family of variables. The aim was to quantify the discrepancy between the principal component analysis of the family and the principal component analysis of the sample of variables. The not yet called RV coefficient was proposed. The immediate consequence was an interest for the choice of variables in principal component analysis (Escoufier et al. 1974). The applied orientation of these works must be underlined.
2. Then two theoretical orientations appeared. The first was the use of the RV coefficient as a unifying tool for the presentation of the different methods of multivariate analysis (Robert and Escoufier 1976).They were presented as solutions of optimization problems under various constraints. The second orientation sprang to mind

from collaboration with JP. Pagés and F. Caillez. It appeared that the operator related to a data matrix found a natural place in the duality diagram which was at the centre of their own work on data analysis. All the works made after for presenting the different multivariate analysis methods through a particular triplet (X, Q, D) have their beginning in this convergence.

3. STATIS came from on other convergence. Two topics were often discussed in the statistical publications: multidimensional scaling and joint analysis of several data matrices. The operators related to data matrices and their scalar product COVV gave a very straightforward solution for the global comparison of the studies. The property of the solution (the compromise is also an operator) has been exploited for the definition of the two other steps of the method.
4. In France, the use of supplementary individuals and supplementary variables was frequent in principal component analysis and correspondence analysis. This practice which uses at the end of a study information known at its beginning is rather questionable from a logical point of view even if it is useful. The principal component analysis with respect to instrumental variables method allows taking in account the instrumental variables from the beginning of the study and moreover gives a quantification of their effects. This is the reason of its development.

The reader will recognize three types of references in this article. The oldest, often in French, are given for an historical reason. They had opened a line. The second, in English, can be found more easily. They have been chosen for the interested readers who will want to go further in this approach of data analysis. The third, the most recent, concern colleagues younger than me, who have been actors of this story and who are always very active. When I name then, I know that I commit two injustices: One towards them because their works do not find in my article a sufficient place and one towards other researchers who made important contributions to the topic. I hope that they will forgive me.

Acknowledgements: I wish to thank the scientific committee of Compstat 2006 and its president, Professor Alfredo Rizzi, who invited me for this contribution. They have given to me the opportunity to go back over what has been written about operator related to a data matrix. This exercise has been a pleasure for me. Several colleagues help me for finding appropriate references: I thank them very much.

References

Beh E.J (1997) Simple correspondence analysis of ordinal cross - clasifications using orthogonal polynolials. Biometrical Journal 39, 589 - 613

Bonifas L, Escoufier Y, Gonzalez P.L, Sabatier R (1984) Choix de variables en analyse en composantes principales. Rev. Statistique Appliquée XXXII(2) 5 – 15

Caillez F, Pages J.P (1976) Introduction à l'analyse des données SMASH, Paris

Chessel D, Dufour A.B, Thioulouse J (2004) The ade4 package – I: One – table methods. R News 4: 5 - 10

Cléroux R, Ducharme G (1989) Vector correlation for elliptical distribution. Comm. Stat. A, 18, 1441 - 1454

Cléroux R, Lazraq A, Lepage Y (1995) Vector correlation based on ranks and a nonparametric test of no association between vectors.Communications in Stat.,24, 713 – 733.

D'Ambra L, Beh E.J, Amenta P (2005) Catanova for two – way contingency tables with ordinal variables using orthogonal polynomials. Communications in statistics, theory and methods, 34, 1755 – 1770

Durand J.F (1992) Additive spline discriminant analysis. In: Y. Dodge and J.C. Whittakers (eds) Computational Statistics, Heidelberg: Physica – Verlag,I,145-150

Durand J.F (1993) Generalized principal component analysis with respect to instrumental variables via univariate spline transformations. Computational Statistics and Data Analysis, 16, 423 – 440

Escoufier Y (1970) Echantillonnage dans une population de variables aléatoires réelles.Publl. Inst.Statist. Univ. Paris 19, 4, 1 – 47

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

Escoufier Y (1977) Operators related to a data matrix. In: J.R. Barra (ed) Recents developpements in Statistics: North – Holland Publishing Company, 125 – 131

Escoufier Y (1980) Exploratory data analysis when data are matrices. In: K. Matusita (ed) Recent developments in Statistical inference and data analysis: North – Holland Publishing Company

Escoufier Y (1986) A propos du choix de variables en analyse des données. Metron XLIV, n°1 - 4, 31 – 47

Escoufier Y (1987) The duality diagramm : a means of better practical applications. In: Legendre P. and Legendre L (eds) Development in numerical ecology,Nato ASI series, Vol.G14, Springer – Verlag, Berlin Heidelberg, 139 - 156

Escoufier Y, L'Hermier H (1978) A propos de la comparaison graphique des matrices de variance. Biom.J. vol.20,n°5, 477 – 483

Escoufier Y, Robert P (1979) Choosing variables and metrics by optimizing the RV – coefficient. In: Optimizing methods in Statistics: Academic Press,Inc.

Escoufier Y, Robert P, Cambon J (1974) Construction of a vector equivalent to a given vector from the point of view of the analysis of principal components: Compstat,Vienne

Fraile L, Escoufier Y, Raibaut A (1993) Analyse des correspondances de données planifiées : étude de la chémotaxie de la larve infestante d'un parasite. Biometrics 49, 1142 – 1153.

Iman W, Abdelkbir S, Escoufier Y (1998) Quantification des effets spatiaux linéaires et non linéaires dans l'explication d'un tableau de données concernant la qualité des eaux souterraines. Rev. Statistique Appliquée, XLVI (3) 37 – 52

Lavit Ch (1988) Analyse conjointe des tableaux quantitatifs : Masson, Paris

Lavit Ch, Escoufier Y, Sabatier R, Traissac P (1994) The ACT (STATIS method). Computational Statistics &data Analysis 18, 97 – 119

Kazi – Aoual F, Hitier S, sabatier R, Lebreton J.D (1995) Refined approximations to permutation tests for multivariate inference. Computational Statistics & Data Analysis 20, 643 – 656

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

Robert P, Escoufier Y (1976) A unifying tool for linear multivariate statistical methods: the RV – coefficient. Appl.Statist. 25 n°3, 257 -265

Sabatier R, Vivien M (2004) A new linear method for analyzing four – way multibloks tables: STATIS – 4 submitted to Computational Statistics & Data analysis

Sabatier R, Jan Y, Escoufier Y (1984) Approximations d'applications linéaires et analyse en composantes principales In: E.Diday et al.(eds) Data Analysis and informatics III. Elsevier Science Publishers B.V. (North – Holland)

Schlich P (1996) Defining and validating assessor compromises about product distances and attributes correlations. In: Naes T. and Risvik E (eds) Multivariate analysis of data in sensory science. Elesevier Science B.V.

Schlich P, Guichard E (1989) Selection and classification of volatile compounds of abricot using the RV coefficient. Journal of Agricultural and Food Chemistry, 37, 142 -150

Vivien M, Sabatier R (2004) A generalization of STATIS – ACT strategy: Do ACT for multiblocks tables. Computational Statistics & data Analysis 46, 155 - 171