THE DUALITY DIAGRAM :

A MEANS FOR BETTER PRACTICAL APPLICATIONS

Y. Escoufier
Unité de Biométrie
ENSA-INRA-USTL
9, place Pierre Viala
F-34060 Montpellier Cedex, France

Abstract - Producing the Principal Components Analysis of a data table requires choices which need to be explained in order to acquire complete understanding of the results. This explicitness opens the road to other possible choices, leading to theoretical research and many practical applications. Changes of scale, changes of variables, weighting of statistical units, decentering of the representations, and elimination of dependence between individuals are dealt with. After reviewing the usual methods from this perspective, it can be seen that it is possible to transform them in order to better adapt mathematical abstractions to concrete situations.

I - A REVIEW OF PRINCIPAL COMPONENTS ANALYSIS (PCA)

Let us put ourselves in the place of a scientist who runs a data table through a PCA Variance Matrix program. The program gives a representation of the variables and of the objects as well as the eigenvalues, which will allow him to estimate the overall validity of the graphs obtained. In addition, especially if the program is French, it will give quantities called absolute and relative contributions which will evaluate the role played by each of the variables and each of the objects, also called units or statistical units in this paper.

Let us try to specify what the program has done in order to go from the data table to the results. For this purpose, let X be an nxp matrix. The i-th row of X is denoted by X_i and contains the p measures made on the i-th statistical unit. The j-th column, denoted by X^j, contains the n values taken by the j-th variable.

NATO ASI Series, Vol. G14
Developments in Numerical Ecology
Edited by P. and L. Legendre

Representation of variables and statistical units. Looking closely at the program, we will see that the arithmetic mean values of each of the variables are calculated first :

$$\forall j = 1,\ldots,p \qquad \bar{X}^j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

Then X is replaced by the centered array $\hat{X}$ for which : $\hat{X}_i^j = X_i^j - \bar{X}^j$

Let I_{nxn} be the R^n identity matrix, and $\underset{\sim}{1}_n$ the column vector of R^n in which all the components are 1.

$$\text{We can write : } \hat{X} = (I_{nxn} - \frac{\underset{\sim}{1}_n \underset{\sim}{1}'_n}{n})\, X \qquad \text{I.1}$$

Secondly, the variance matrix associated with the array is calculated, i.e. : $V = \frac{\hat{X}'\hat{X}}{n}$ I.2

The third step consists in calculating the eigenvectors and eigenvalues of V, that is the p vectors ϕ_α of R^p and the p numbers λ_α satisfying

$$\forall \alpha = 1,\ldots,p \qquad V\phi_\alpha = \lambda_\alpha \phi_\alpha \qquad \text{with } \phi'_\alpha \phi_\alpha = 1 \qquad \text{I.3}$$

$$\lambda_1 \geqslant \lambda_2 \ldots \geqslant \lambda_p$$

Standard mathematical expansions show that

$$\forall \alpha = 1,\ldots,p \qquad \lambda_\alpha \geqslant 0$$

$$\forall \alpha = 1,\ldots,p \;;\; \alpha' = 1,\ldots,p \;;\; \lambda_\alpha \neq \lambda_{\alpha'} \;\Rightarrow\; \phi'_\alpha \phi_{\alpha'} = 0$$

If $\lambda_\alpha = \lambda_{\alpha'}$, ϕ_α and $\phi_{\alpha'}$ can be chosen in such a way that $\phi'_\alpha \phi_{\alpha'} = 0$.

$$\text{Then : } V = \sum_{\alpha=1}^{p} \lambda_\alpha \phi_\alpha \phi'_\alpha \qquad \text{I.4}$$

Finally consider for q < p, $V_q = \sum_{\alpha=1}^{q} \lambda_\alpha \phi_\alpha \phi'_\alpha$. It can be shown (see note after expression I.14) that for every pxp matrix A_q of rank q < p

$$\mathrm{Tr}((V - A_q)^2) \geqslant \mathrm{Tr}((V - V_q)^2) = \sum_{\alpha=q+1}^{p} \lambda_\alpha^2 \qquad \text{I.5}$$

where Tr represents the trace of the matrices, i.e. the sum of their diagonal elements.

This property justifies the interpretation that is made from the variables display. Suppose that for every variable k, we associate to it the point whose coordinates are

$$(\phi^*_{1k} = \sqrt{\lambda_1}\ \phi_{1k},\ \phi^*_{2k} = \sqrt{\lambda_2}\ \phi_{2k})$$

The scalar products that can be read from that representation are the elements of the matrix V_2, the best possible approximation of V by a matrix of rank 2.

If the sum $\sum_{\alpha=q+1}^{p} \lambda_\alpha^2$ is sufficiently small, the covariances and variances of the p variables can be visually appreciated.

The fourth step is to calculate the coordinates of the units by the formula :

$$\forall \alpha = 1,\ldots,p \qquad \psi^*_\alpha = \hat{X}\ \phi_\alpha \qquad \text{I.6}$$

It can easily be checked that :

$$\forall \alpha = 1,\ldots,p \quad \frac{\hat{X}\hat{X}'}{n}\ \psi^*_\alpha = \lambda_\alpha\ \psi^*_\alpha \quad \text{with} \quad \frac{\psi^{*\prime}_\alpha\ \psi^*_\alpha}{n} = \lambda_\alpha \qquad \text{I.7}$$

This leads us to investigate the matrix $\frac{\hat{X}\hat{X}'}{n} = \frac{W}{n}$ that plays the same role for the units as V does for the variables.

We then set : $\psi_\alpha = \dfrac{\psi^*_\alpha}{\sqrt{\lambda_\alpha}}$

and we have : $$\frac{W}{n} = \sum_{\alpha=1}^{p} \lambda_\alpha\ \frac{\psi_\alpha\ \psi'_\alpha}{n} \qquad \text{I.8}$$

Let $$\frac{W_q}{n} = \sum_{\alpha=1}^{q} \lambda_\alpha\ \frac{\psi^*_\alpha\ \psi^{*\prime}_\alpha}{n} \qquad \text{I.9}$$

It has been shown (note after expression I.14) that for every nxn matrix A_q, of rank q < p

$$\mathrm{Tr}((\frac{W}{n} - A_q)^2) \geqslant \mathrm{Tr}((\frac{W}{n} - \frac{W_q}{n})^2) = \sum_{\alpha=q+1}^{p} \lambda_\alpha^2 \qquad \text{I.10}$$

This property justifies the interpretation that will be made from the representation of the objects. W_2 is the best rank 2 approximation of W, which means that the graphical representation obtained by giving the coordinates ($\psi^*_{1i} = \sqrt{\lambda_1}\ \psi_{1i}$; $\psi^*_{2i} = \sqrt{\lambda_2}\ \psi_{2i}$) to the i-th statistical unit allows to visualise the scalar products among the units, and thus their distances.

Absolute and relative contributions. We have seen in passing that $\phi^{*'}_\alpha\ \phi^*_\alpha = \frac{\psi^{*'}_\alpha\ \psi^*_\alpha}{n} = \lambda_\alpha$. The idea naturally arose to consider the quantities

$\frac{(\phi^*_{\alpha k})^2}{\lambda_\alpha}$ as the participation of the variable k in the definition of λ_α I.11

$\frac{(\psi^*_{\alpha i})^2}{n\ \lambda_\alpha}$ as the participation of the statistical unit i to the definition of λ_α . I.11'

These quantities are given the name of absolute contributions. They allow to estimate the part played by a variable or a statistical unit in the construction of the representations.

From $V = \sum_{\alpha=1}^{p} \lambda_\alpha\ \phi_\alpha\ \phi'_\alpha$ and $\frac{W}{n} = \sum_{\alpha=1}^{p} \lambda_\alpha \frac{\psi_\alpha\ \psi'_\alpha}{n}$ we can also

conclude that $V_{kk} = \sum_{\alpha=1}^{p} (\phi^*_{\alpha k})^2$ and $W_{ii} = \sum_{\alpha=1}^{p} (\psi^*_{\alpha i})^2$

Hence $\frac{(\phi^*_{\alpha k})^2}{V_{kk}}$ is the participation of the ϕ^*_α axis in the reconstruction of the variable K, in actual fact the reconstruction of the variance V_{kk}, I.12

while $\frac{(\psi^*_{\alpha i})^2}{W_{ii}}$ is the participation of axis ψ^*_α in the reconstruction of $\frac{W_{ii}}{n} = \frac{\hat{X}_i\ \hat{X}'_i}{n}$, that is, the inertia of the i-th statistical unit with respect to the mean point. I.12'

These quantities are given the name of relative contributions. They are criteria for the quality of the representation specifically concerning each variable and each statistical unit.

Reconstitution of data. Noting that $I_{pxp} = \sum_{\alpha=1}^{p} \phi_\alpha \phi'_\alpha$

then $\hat{X} = \hat{X}\, I_{pxp} = \sum_{\alpha=1}^{p} \hat{X}\, \phi_\alpha \phi'_\alpha = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\, \psi_\alpha \phi'_\alpha$ I.13

Then let $\hat{X}_q = \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha}\, \psi_\alpha\, \phi'_\alpha$. It has been shown (following note) that for each nxp matrix A_q of rank $q < p$:

$$\mathrm{Tr}\left(\frac{(\hat{X}-A_q)(\hat{X}-A_q)'}{n}\right) \geqslant \mathrm{Tr}\left(\frac{(\hat{X}-\hat{X}_q)(\hat{X}-\hat{X}_q)'}{n}\right) = \sum_{\alpha=q+1}^{p} \lambda_\alpha \qquad \text{I.14}$$

Note : Expressions I.5, I.10 and I.14 come from the well-known result by Eckart and Young (1936). It is of importance to remark that V_q, W_q and $\hat{X}_q$ are not only optimal for the least squares criterion given here by Tr(.) but also for an infinity of other criteria (Rao 1980 ; Sabatier et al. 1984).

II - CHANGES IN THE INITIAL CHOICES

II.1. Weighting of the statistical units. The preceding section weighted the units by $\frac{1}{n}$, first in the calculation of the mean values and second, for the calculation of the variance matrix.

If we denote $D = \frac{1}{n} I_{nxn}$ as the diagonal matrix of elements $\frac{1}{n}$, then

$$\hat{X} = (I_{nxn} - \underset{\sim}{1}_n \underset{\sim}{1}'_n D)\, X \qquad \text{II.1}$$

$$V = \hat{X}'\, D\, \hat{X} \qquad \text{II.2}$$

Re-reading section I with formulas II.1 and II.2 then shows that the fact that all the diagonal elements of D are equal to $\frac{1}{n}$ is never explicitly used in proofs.

Re-working the formulas in which n appears explicitly, we get :

$$\forall\alpha = 1,\dots,p \qquad \hat{X}\,\hat{X}'D\ \psi^*_\alpha = \lambda_\alpha\ \psi^*_\alpha \quad \text{with } \psi^{*\prime}_\alpha\ D\ \psi^*_\alpha = \lambda_\alpha \qquad \text{II.7}$$

$$W\ D = \sum_{\alpha=1}^{p} \lambda_\alpha\ \psi_\alpha\ \psi'_\alpha\ D \qquad \text{II.8}$$

$$W_qD = \sum_{\alpha=1}^{q} \lambda_\alpha\ \psi_\alpha\ \psi'_\alpha\ D \qquad \text{II.9}$$

$$\mathrm{Tr}((WD-A_q)^2) \geqslant \mathrm{Tr}((WD-W_qD)^2) = \sum_{\alpha=q+1}^{p} \lambda_\alpha^2 \qquad \text{II.10}$$

$$\frac{(\psi^*_{\alpha i})^2\ D_{ii}}{\lambda_\alpha} \qquad \text{II.11'}$$

$$\mathrm{Tr}((\hat{X}-A_q)(\hat{X}-A_q)'D) \geqslant \mathrm{Tr}((\hat{X}-\hat{X}_q)(\hat{X}-\hat{X}_q)'D = \sum_{\alpha=q+1}^{p}\lambda_\alpha \qquad \text{II.14}$$

It is possible to consider situations in which D is more general. Actually the mathematical results are stronger than those used by the usual computer programs. This lead to new applications which will be discussed in section IV.

II.2. Invertible linear transformations of variables. Consider now the case of an invertible, linear transformation M, applied to the data matrix X and write the elements of the PCA of XM.

Centering $\widehat{XM} = (I_{nxn} - \underset{\sim}{1}_n\ \underset{\sim}{1}'_n\ D)\ XM = \hat{X}\ M$ III.1

Variance matrix $V_{[M]} = M'\hat{X}'D\ \hat{X}\ M = M'V\ M$ III.2

Eigenvectors and eigenvalues of $V_{[M]}$

$$\forall\alpha = 1,\dots,p \qquad V_{[M]}u_\alpha = \lambda_\alpha\ u_\alpha \qquad \text{with } u'_\alpha u_{\alpha'} = \delta_{\alpha\alpha'}$$

$$\text{i.e.} \quad M'\ V\ M\ u_\alpha = \lambda_\alpha\ u_\alpha \qquad \text{with } u'_\alpha u_{\alpha'} = \delta_{\alpha\alpha'}$$

For each $\alpha = 1,\dots,p$, consider ϕ_α defined by $u_\alpha = M'\ \phi_\alpha$ and set $M\ M' = Q$

We have $M'V\,Q\,\phi_\alpha = \lambda_\alpha\,M'\phi_\alpha$ with $\phi'_\alpha\,M\,M'\,\phi_{\alpha'} = \delta_{\alpha\alpha'}$

$$\text{i.e.}\quad V\,Q\,\phi_\alpha = \lambda_\alpha\,\phi_\alpha \quad \text{with} \quad \phi'_\alpha\,Q\,\phi_{\alpha'} = \delta_{\alpha\alpha'} \qquad \text{III.3}$$

From $V_{[M]} = \sum_{\alpha=1}^{p} \lambda_\alpha\,u_\alpha\,u'_\alpha$, we have $M'VM = \sum_{\alpha=1}^{p} \lambda_\alpha\,M'\,\phi_\alpha\,\phi'_\alpha\,M$ hence

$$VQ = \sum_{\alpha=1}^{p} \lambda_\alpha\,\phi_\alpha\,\phi'_\alpha\,Q \qquad \text{III.4}$$

Moreover we have

$$\sum_{\alpha=q+1}^{p} \lambda_\alpha^2 = \mathrm{Tr}((V_{[M]} - V_{[M],q})^2)$$

$$= \mathrm{Tr}((M'VM - \sum_{\alpha=1}^{q} \lambda_\alpha\,M'\,\phi_\alpha\,\phi'_\alpha\,M)^2)$$

$$= \mathrm{Tr}((VQ - \sum_{\alpha=1}^{q} \lambda_\alpha\,\phi_\alpha\,\phi'_\alpha\,Q)^2) \qquad \text{III.5}$$

<u>The coordinates of the statistical units</u>

$$\text{Let}\quad \psi^*_\alpha = \hat{X}\,M\,u_\alpha = \hat{X}\,M\,M'\phi_\alpha = \hat{X}\,Q\,\phi_\alpha \qquad \text{III.6}$$

It can be verified that

$$\forall\alpha = 1,\ldots,p \quad \hat{X}\,M\,M'\hat{X}'D\,\psi^*_\alpha = \hat{X}\,Q\,\hat{X}'D\,\psi^*_\alpha = \lambda_\alpha\,\psi^*_\alpha$$

$$\text{with}\quad \psi^{*\prime}_\alpha\,D\,\psi^*_\alpha = \phi'_\alpha\,Q\,\hat{X}'D\,\hat{X}\,Q\,\phi_\alpha$$

$$= \phi'_\alpha\,Q\,\phi_\alpha\,\lambda_\alpha = \lambda_\alpha \qquad \text{III.7}$$

and the properties II.8, II.9, II.10, which do not explicitly involve Q, remain valid. The matrix $\hat{X}\,Q\,\hat{X}'$ which could be noted $W_{[M]}$, is the matrix of the scalar products between statistical units when the space R^p is given the positive bilinear form defined by $Q = M\,M'$.

Thus, studying linear data transformations is similar to choosing a means of calculating distances between statistical units. Most of the current programs choose $Q = I_{pxp}$, avoiding the problem of choice by previous processing (the variables are standardized in order to use the correlation matrix).

We can see that this choice in mathematical equations does not present great difficulties, but it appears important not to hide it.

Absolute and relative contributions. For absolute contributions, the starting point is $\phi^{*\prime}_\alpha Q\ \phi^*_\alpha = \lambda_\alpha$ so that absolute contributions are only obtained when Q is diagonal like D. Thus we have

$$\frac{(\phi^*_{\alpha k})^2\ Q_{kk}}{\lambda_\alpha}$$

Formula III.4 implies $V = \sum_{\alpha=1}^{p} \lambda_\alpha\ \phi_\alpha\ \phi'_\alpha$ so that relative contributions can always be calculated.

Reconstitution of data.

$$\text{We have } \widehat{XM} = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\,\psi_\alpha\ u'_\alpha = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\ \psi_\alpha\ \phi'_\alpha\ M \qquad \text{III.13}$$

Supposing that $(\widehat{XM})_q = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\ \psi_\alpha\ \phi'_\alpha\ M$, we obtain

$$\sum_{\alpha=q+1}^{p} \lambda_\alpha = Tr((\widehat{XM} - (\widehat{XM})_q)(\widehat{XM} - \widehat{XM})'_q)D)$$

$$= Tr((\hat{X} - \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha}\,\psi_\alpha\ \phi'_\alpha)Q(\hat{X} - \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha}\,\psi_\alpha\ \phi'_\alpha)'D) \qquad \text{III.14}$$

III - THE DUALITY DIAGRAM

The above section presents the idea of a PCA which is a function of the triplet (X, Q, D) instead of the usual presentation of the PCA of the X array based on some implicit choices : $Q = I_{pxp}$, $D = \frac{1}{n} I_{nxn}$. Cailliez and Pagès (1976) popularized this point of view in France by giving it a rigourous mathematical formalization that we are going to review. We will keep in mind that our first objective is to bring out the choices to be made in order to carry out a study : the data (X), the weighting of statistical units necessary for the calculation of relationships between the variables (D), and the way of quantifying the resemblances between the statistical units (Q).

Our second objective is to define the mathematical nature of the objects dealt with in order to make the best use of their properties.

The first step consists of considering the i-th unit as a vector of $E = R^p$. It will be written as

$$\sum_{j=1}^{p} X_i^j \underline{e}_j ,$$

where ($\underline{e}_1$, ..., $\underline{e}_p$) is a system of n linearly independent vectors of E, i.e. a basis of E.

Symmetrically the j-th variable is considered as a vector of $F = R^n$. It will be written as

$$\sum_{i=1}^{n} X_i^j \underline{f}_i \quad \text{where } (\underline{f}_1, \ldots, \underline{f}_n) \text{ is the basis of F.}$$

The second step consists in associating a linear mapping $\underline{e}_j^*$ with the j-th variable, which makes the i-th statistical unit correspond to the value X_i^j, that it has taken for that variable :

$$\underline{e}_j^* \left(\sum_{k=1}^{p} X_i^k \underline{e}_k \right) = \sum_{k=1}^{p} X_i^k \underline{e}_j^* (\underline{e}_k) = X_i^j$$

Thus variables also have a representation in E*, the dual space of E. In fact ($\underline{e}_i^*$, ..., $\underline{e}_p^*$) is the basis of E*, the dual basis of ($\underline{e}_1$, ..., $\underline{e}_p$) which is the basis of E. In a similar way ($\underline{f}_1^*$, ..., $\underline{f}_n^*$), the basis of F*, the dual of ($\underline{f}_1$, ..., $\underline{f}_n$), can be defined. $\underline{f}_i^*$ is the representation of the i-th statistical unit.

This construction gives two representations for each unit : one in E, the second one in F*. Consider then the linear mapping defined by

$$\forall i = 1,\ldots,p \qquad \underline{f}_i^* \longrightarrow \sum_{j=1}^{p} X_i^j \underline{e}_j$$

Its associated matrix is X'.

In the same way, X is associated with the linear mapping

$$\underline{e}^*_j \longrightarrow \sum_{i=1}^{n} X_i^j \, \underline{f}_i$$

The calculation of distances between objects considered as points of E, entails the choice of a positive definite bilinear form Q which is considered to be a mapping from E into E*. Similarly, the calculation of covariances between the variables in F depends on a quadratic form D that maps F into F*.

This can be summarized by the following diagram which illustrates the choices to be made for a study.

$$\begin{tikzcd}
E=R^p \arrow[d, "Q"'] & & F^* \arrow[ll, "X'"'] \\
E^* \arrow[rr, "X"'] & & F=R^n \arrow[u, "D"']
\end{tikzcd}$$

The calculation of scalar products between two variables $\underline{e}^*_k$ and $\underline{e}^*_\ell$ in E* must give the same result as the calculation made between the two same variables X^k and X^ℓ in F for the positive definite bilinear form D. This leads to the fact that E* must be provided with the metric V = X' D X. For symmetrical reasons, F* has the metric W = X Q X'. The diagram can then be completed as follows :

$$\begin{tikzcd}
E=R^p \arrow[d, shift right, "Q"'] & & F^* \arrow[ll, "X'"'] \arrow[d, shift right, "W"'] \\
E^* \arrow[u, shift right, "V"'] \arrow[rr, "X"'] & & F=R^n \arrow[u, shift right, "D"']
\end{tikzcd}$$

Expressions III.4 and III.7 show that the solutions of the PCA are given by the eigenvalues and eigenvectors of VQ and WD, which appear on the diagram.

IV - ON APPLICATIONS CONCERNING D

IV.1. Special centering

Since the duality diagram just described coïncides exactly with sections I and II using the matrix $\hat{X}$, the weights D can be included as follows :

$$\begin{tikzcd}
E=R^p \arrow[d, shift right, "Q"'] & & F^* \arrow[ll, "X'(I_{nxn} - D \underset{\sim}{1}_n \underset{\sim}{1}'_n)"'] \arrow[d, shift right, "W"'] \\
E^* \arrow[u, shift right, "V"'] \arrow[rr, "(I_{nxn} - \underset{\sim}{1}_n \underset{\sim}{1}'_n D)X"'] & & F=R^n \arrow[u, shift right, "D"']
\end{tikzcd}$$

IV.1.1. It is possible that one of the units has a very unusual behaviour. The representation of the units will tend to show on the first axis that this individual is in opposition to the others. While this unit can be eliminated and the PCA repeated, the duality diagram allows for another possibility. Let Δ be a diagonal matrix whose diagonal elements are all zero except for that corresponding to the unusual object, which is set to 1. In the following diagram,

$$\begin{tikzcd}
E=R^p \arrow[d, shift right, "Q"'] & & F^* \arrow[ll, "X'(I_{nxn} - \Delta \underset{\sim}{1}_n \underset{\sim}{1}'_n)"'] \arrow[d, shift right, "W"'] \\
E^* \arrow[u, shift right, "V"'] \arrow[rr, "(I_{nxn} - \underset{\sim}{1}_n \underset{\sim}{1}'_n \Delta)\ X"'] & & F=R^n \arrow[u, shift right, "D"']
\end{tikzcd}$$

the principal components will be the eigenvectors of $WD = (I_{nxn} - \underset{\sim}{1}_n \underset{\sim}{1}'_n \Delta)X\ D\ X'(I_{nxn} - \Delta\ \underset{\sim}{1}_n \underset{\sim}{1}'_n)D$. They are clearly

a) centered for Δ (because $\underset{\sim}{1}'_n\ \Delta\ W\ D = 0$) ;

and b) orthogonal for D.

In practice, this means that the unusual object is located at the origin, and the representation of the other points is studied in relation to it. The matrix $(I_{nxn} - \underset{\sim}{1}_n \underset{\sim}{1}'_n \Delta)X$ expresses the deviations from that statistical object. Note that the weighting assigned to that object in D is unimportant.

IV.1.2. This procedure can be further modified by a matrix Δ having more than one diagonal element different from zero. Thereby, representations of the objects, centered for Δ and orthogonal for D, are obtained. This means giving more importance to the representations of some objects. Here, the relative weighting of these objects in D cannot be ignored.

IV.2. Analysis of partial covariances (Lebart et al. 1979, p.300)

Equation $\underset{\sim}{1}_n \, \underset{\sim}{1}'_n \, D = \underset{\sim}{1}_n \, (\underset{\sim}{1}'_n \, D \, \underset{\sim}{1}_n)^{-1} \, \underset{\sim}{1}'_n \, D$ is the basis for the interpretation of centering in terms of projecting on the line of constants (Cailliez and Pagès 1976, p. 146).

Let us consider an nxq matrix of data X_2, dealing with the same objects as X. We define X_3 as the matrix obtained by the juxtaposition of $\underset{\sim}{1}_n$ and X_2

$$X_3 = (\underset{\sim}{1}_n \; \vdots \; X_2)$$

Let $P_3 = X_3(X_3' \, D \, X_3)^{-1} \, X_3' \, D$. Based on the orthogonality of $\underset{\sim}{1}_n$ and the columns of $\hat{X}_2 = (I_{nxn} - \underset{\sim}{1}_n \, \underset{\sim}{1}'_n \, D) \, X_2$,

$$I_{nxn} - P_3 = (I_{nxn} - \hat{X}_2(\hat{X}_2' \, D \, \hat{X}_2)^{-1} \, \hat{X}_2' \, D)(I_{nxn} - \underset{\sim}{1}_n \, \underset{\sim}{1}'_n \, D)$$

$$= (I_{nxn} - \underset{\sim}{1}_n \, \underset{\sim}{1}'_n \, D)(I_{nxn} - \hat{X}_2(\hat{X}_2' \, D \, \hat{X}_2)^{-1} \, \hat{X}_2' \, D)$$

In the next duality diagram,

$$\begin{tikzcd}
E=R^P \arrow[d, shift right, "Q"'] & & F^* \arrow[ll, "X'(I_{nxn} - P_3')"'] \arrow[d, shift right, "W"'] \\
E^* \arrow[u, shift right, "V"'] \arrow[rr, "(I_{nxn} - P_3) \, X"'] & & F=R^n \arrow[u, shift right, "D"']
\end{tikzcd}$$

i) $(I_{nxn} - P_3)X = (I_{nxn} - \hat{X}_2(\hat{X}_2' \, D \, \hat{X}_2)^{-1} \, \hat{X}_2' \, D) \, \hat{X}$

We do the PCA of the residuals of $\hat{X}$ in the regression on $\hat{X}_2$, i.e., V is the residual variance matrix.

ii) $WD = (I_{nxn} - P_3) \, X \, Q \, X'(I_{nxn} - P_3') \, D$

Because P_3 is idempotent, $\underset{\sim}{1}'_n \, P_3 \, W \, D = 0$ and the principal components satisfy

$$\underset{\sim}{1}'_n \, P_3 \, \psi_\alpha = 0$$

i.e. $\underset{\sim}{1}'_n \, D \, \psi_\alpha = 0$, the principal components are centered for D, and

$\hat{X}_2' \, D \, \psi_\alpha = 0$, the principal components are orthogonal to the sub-space of F generated by the columns of $\hat{X}_2$.

iii) Finally, the principal components are orthogonal for D.

IV.3. Correlated objects

One of the consequences made apparent by the duality diagram is that any change in the weights D and in $\hat{X}$ modifies V. Thus to modify V, the changes in D and $\hat{X}$ which will produce the appropriate V are needed.

This problem arises, for example, if the observations $\hat{X}_i$ are linked to the observations $\hat{X}_{i-1}$ by

$$\forall i = 2,\dots,n \qquad \hat{X}_i = \rho\, \hat{X}_{i-1} + e_i \qquad \text{with } |\rho| < 1$$

It is clear that here, V mixes the correlations of objects with the correlations of the variables, and that it is desirable to eliminate the effect of the correlations between objects.

In order to do this, Aragon and Caussinus (1980) suggest studying the following diagram

$$\begin{tikzcd}
E=R^p \arrow[d, shift right, "Q"'] & & F^* \arrow[ll, "\hat{X}'"'] \arrow[d, shift right, "W"'] \\
E^* \arrow[u, shift right, "V"'] \arrow[rr, "\hat{X}"'] & & F=R^n \arrow[u, shift right, "C^{-1}"']
\end{tikzcd}$$

where C is the matrix of auto-correlations

$$C = \begin{bmatrix}
1 & \rho & \rho^2 & \cdots\cdots & \rho^{n-1} \\
\rho & 1 & \rho & & \vdots \\
\rho^2 & \rho & 1 & & \vdots \\
\vdots & & & & \vdots \\
\rho^{n-1} & & & \cdots\cdots\cdots & 1
\end{bmatrix}$$

with inverse

$$C^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+\rho^2 & -\rho \\ 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

If $A = \begin{pmatrix} 0 & . & 0 & . & 0 \\ +\rho & \ddots & \ddots & \ddots & 0 \\ & \ddots & \ddots & \ddots & 0 \\ 0 & . & . & +\rho & 0 \end{pmatrix}$ and $\Delta = \frac{1}{1-\rho^2}\begin{pmatrix} 1-\rho^2 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$

then $C^{-1} = (I - A')\ \Delta\ (I - A)$

Thus the analysis is equivalent to the following :

$$\begin{tikzcd} E=R^p \arrow[d, shift right, "Q"'] & & F^* \arrow[ll, "X'(I-A')"'] \arrow[d, shift right, "W"'] \\ E^* \arrow[u, shift right, "V"'] \arrow[rr, "(I - A)\hat{X}"'] & & F=R^n \arrow[u, shift right, "\Delta"'] \end{tikzcd}$$

The first object is associated with $\hat{X}_1$ given the weight 1 and those that follow are associated with $\hat{X}_i - \rho\ \hat{X}_{i-1}$ given the weight $1/(1-\rho^2) > 1$. The sum of the diagonal terms of Δ can be made equal to unity by multiplying Δ by the necessary constant. The principal components, i.e. the eigenvectors of $(I-A)\ \hat{X}\ Q\ \hat{X}'(I-A')\ \Delta$, are orthogonal for Δ. If there is a matrix D such that $\underset{\sim}{1}'_n\ D\ (I-A)\hat{X} = 0$, the principal components are also centered for D (this would be true if $\hat{X}$ was centered with respect to a matrix D, giving a weight of 0 to the first object).

V - PRACTICAL CONSEQUENCES OF THE USE OF Q

V.1. In the first place, the explicit use of the metric Q allows an explicit discussion of the choice of the scale of measurement, and, in particular, the replacement of initial data by standardized data. It will be noted, however, that there is a slight difference between the PCA on the correlation matrix, as in conventional software, and that here. The first is the PCA of the triplet $(\hat{X}\,[\mathrm{Diag}(V)]^{-1/2},\ I_{pxp},\ \frac{1}{n}\,I_{nxn})$. The second is the PCA of the triplet $(\hat{X},\ [\mathrm{Diag}(V)]^{-1},\ \frac{1}{n}\,I_{nxn})$. They both yield the same WD, and therefore the same representation of the units. However, the variables are represented differently. The first leads to the diagonalization of

$$[\mathrm{Diag}(V)]^{-1/2}\ \hat{X}'D\ \hat{X}\ [\mathrm{Diag}(V)]^{-1/2}$$

The second to $\hat{X}'D\ \hat{X}\ [\mathrm{Diag}(V)]^{-1}$

Obviously the two solutions are related.

Recent works on the choice of a metric in special cases includes that of Karmierczak (1985), which considers the choice of distances between profiles, and that of Besse and Ramsay (1986) on the distances between curves.

V.2. Correspondence Analysis of a nxp contingency table, P, has been shown (Escoufier 1982) as the PCA of the triplet

$$(D_I^{-1}(P - D_I\ \underset{\sim}{1}_n\ \underset{\sim}{1}'_n\ D_J)\ D_J^{-1},\ D_J,\ D_I)$$

It is easy to see that the product of the sum of the eigenvalues by the total number of statistical units under study is simply the χ^2 statistic describing the contingency between the qualitative variable defining the rows of P and the qualitative variable defining the columns. Correspondence analysis can be considered as a means of bringing out the modalities of the variables which differ the most from the model of independence. Lauro and D'Ambra (1983) have shown how χ^2 could be replaced by the asymmetric criterion of Goodman and Kruskal (1954). Here again the use of a special PCA is justified because of the natural asymmetry between the two qualitative variables being studied.

The problem is no longer that of the deviation from the independence model, but that of the difference between the conditional distributions of a variable and its marginal distribution.

These approaches suggest that the comparison of an experimental variance matrix $V = \hat{X}'D\,\hat{X}$ with a theoretical variance matrix Σ can be developed by the PCA of the triplet (X, Σ^{-1}, D). The eigenvalues of $V\,\Sigma^{-1}$ will be computed. They can be used for testing the hypothesis that the variance matrix is equal to Σ (Anderson 1958, p. 265). The PCA will indicate those objects that contribute most to the different eigenvalues i.e., those that are mainly responsible for the difference between V and Σ. Since Σ^{-1} in general is not diagonal, it is no longer possible to consider the absolute contribution of the variables. However, the variables having large relative contributions are considered to be responsible for any difference between V and Σ.

Similarly, Discriminant Analysis can also be considered as a PCA of the triplet (M, T^{-1}, D_q) in which M is the qxp matrix of the means of p variables in each of q classes, D_q is the qxq diagonal matrix of the weights of the classes, and T is the variance matrix calculated over the set of units. Let $B = M'D_q\,M$ be the between-class variance matrix. The sum of the eigenvalues is $\mathrm{Tr}(B\,T^{-1})$, the criterion which is referred to by Morrison (1967, p. 198), to test the equality of the means among the different groups. Evaluating the contributions of objects (mean points per class) will reveal which groups contribute most towards rejecting the hypothesis of equality.

V.3. Now let us look at a situation in which two sets of quantitive variables have been observed for the same objects.

The first set leads to a completely determined PCA, that of the triplet $(\hat{X}, Q, D)$. For the second PCA we use the data Y and we agree to give the same weight D to all statistical units. What metric R should be chosen so that the PCA of $(\hat{Y}, R, D)$ "resembles the closest" the PCA of $(\hat{X}, Q, D)$? In order to answer that question, it is necessary to give a precise meaning to "resembles the closest".

Choosing the resemblance of representations of the objects, it is natural to quantify the distance between the two PCAs by :

$$\mathrm{Tr}((\hat{X}\ Q\ \hat{X}'D - \hat{Y}\ R\ \hat{Y}'D)^2)$$

Bonifas *et al.* (1984) show that the best choice is :

$$R = (\hat{Y}'D\ \hat{Y})^{-1}\ \hat{Y}'\ D\ \hat{X}\ Q\ \hat{X}'\ D\ \hat{Y}\ (\hat{Y}'\ D\ \hat{Y})^{-1}$$

which goes back, from the point of view of the statistical units under consideration, to the representation given by the PCA of

$$(\hat{Y}(\hat{Y}'\ D\ \hat{Y})^{-1}\ \hat{Y}'\ D\ \hat{X},\ Q,\ D)$$

Note that the sum of the eigenvalues equals $\mathrm{Tr}(VQ) = \mathrm{Tr}((\hat{Y}'D\ \hat{Y})^{-1}\ \hat{Y}'\ D\ \hat{X}\ Q\ \hat{X}'D\ \hat{Y})$ and that $\hat{Y}'D\ \hat{X} = \hat{Y}'D\ X$.

Consider the case where X is a nxq response pattern array associated with a qualitative variable with q categories. We know that $\hat{X}'D\ \hat{Y} = D_q M$ where M is the qxp matrix of q mean vectors calculated for each category and D_q is the weight matrix of each. The choice $Q = D_q^{-1}$ leads to :

$$\hat{Y}'D\ X\ Q\ X'D\ \hat{Y} = M'\ D_q\ M = B,$$

so that setting $(\hat{Y}'D\ \hat{Y}) = T$ we get $\mathrm{Tr}(VQ) = \mathrm{Tr}(T^{-1}\ B)$. In other words, discriminant analysis measures the distance between the i-th and i'-th units by the quantity $(X_i - X_{i'})\ D_q^{-1}(X_i - X_{i'})'$. It is possible to question this choice of D_q^{-1}, and to consider other possibilities.

VI - CONCLUSION

A deeper mathematical understanding of the steps taken in a normal PCA program based upon the variance matrix opens up numerous paths for theoretical and practical work.

This does not challenge the usual methods of data analysis, which are still a reasonable compromise between current knowledge and what the user is willing to do in terms of cost, whether it be the cost of the mathematical training necessary for understanding, or for computations.

This formalization allows anyone, who is willing to make the effort to acquire the necessary knowledge (and ultimately to pay for the expense of special programs), to be able to choose the mathematical abstractions best adapted to the concrete problem under study.

REFERENCES

ANDERSON, T.W. 1958. An introduction to multivariate statistical analysis. John Wiley & Sons, New York, NY.

ARAGON, Y., and H. CAUSSINUS. 1980. Une analyse en composantes principales pour des unités statistiques corrélées, p. 121-131. In E. Diday et al. [ed.] Data analysis and informatics. North Holland Publ. Co. New York, NY.

BESSE, Ph., and S.O. RAMSAY. 1986. Principal components analysis of sampled functions. Psychometrika (in press).

BONIFAS, L., Y. ESCOUFIER, P.L. GONZALEZ, and R. SABATIER. 1984. Choix de variables en analyses en composantes principales. Revue de Statistique Appliquée, Vol. XXXII n° 2 : 5-15.

CAILLIEZ, F., and J.P. PAGES. 1976. Introduction à l'analyse des données. SMASH, 9, rue Duban, Paris 75010.

ECKART, C., and G. YOUNG. 1936. The approximation of one matrix by another of lower rank. Psychometrika, Vol. 1 n° 3 : 211-218.

ESCOUFIER, Y. 1982. L'analyse des tableaux de contingence simples et multiples. Metron, Vol. XL n° 1-2 : 53-77.

ESCOUFIER, Y. 1985. L'analyse des correspondances : ses propriétés et ses extensions. Institut International de Statistique. Amsterdam : 28.2.1-28.2.16.

ESCOUFIER, Y., and P. ROBERT. 1979. Choosing variables and metrics by optimizing the RV-coefficient, p. 205-219. In J.S. Rustagi [ed.] Optimizing methods in statistics. Academic Press Inc.

GOODMAN, L.A., and W.H. KRUSKAL. 1954. Measures of association for cross-classifications. J. amer. stat. Ass., Vol. 49 : 732-764.

KARMIERCZAK, J.B. 1985. Une application du principe du Yule : l'analyse logarithmique. Quatrièmes Journées Internationales Analyse des données et informatique. Versailles. France. (Document provisoire : 393-403).

LAURO, N., and L. D'AMBRA. 1983. L'analyse non symétrique des correspondances, p. 433-446. In E. Diday et al. [ed.] Data analysis and informatics III. Elsevier Science Publ. BV. North Holland.

LEBART, L., A. MORINEAU, and J.P. FENELON. 1979. Traitement des données statistiques. Dunod.

MORRISON, D.F. 1967. Multivariate statistical methods. Mc Graw-Hill Bock Co.

PAGES, J.P., F. CAILLIEZ, and Y. ESCOUFIER. 1979. Analyse factorielle: un peu d'histoire et de géométrie. Revue de Statistique Appliquée, Vol. XXVII n° 1 : 6-28.

RAO, C.R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis, p. 3-22. In P.R. Krishnaiah [ed.] Multivariate analysis V. North-Holland Publ. Co.

SABATIER, R., Y. JAN, and Y. ESCOUFIER. 1984. Approximations d'applications linéaires et analyse en composantes principales, p. 569-580. In E. Diday et al. [ed.] Data analysis and informatics III. Elsevier Science Publ. BV. North Holland.