# SPECIAL TOPIC

*This article explains the French school of data analysis known as 'Analyse des Données'. It is written by two prominent French Region biometricians, Prof Yves Escoufier and Dr Susan Holmes, both of the Unité de Biométrie, ENSA M, INRA, Montpellier II, 9, Place Pierre Viala, 34060 Montpellier, France.*

## Data analysis in France

*by Yves Escoufier & Susan Holmes*

**Introduction**

The exact translation of the French expression 'Analyse des Données' should be taken as meaning the effective analysis of data (*statistical practice* versus *mathematical theory developed on a statistical problem*). In fact the expression more often refers to graphical techniques, their practice and basis, as opposed to the testing of hypotheses (*exploratory data analysis versus confirmatory data analysis*). Data Analysis (DA) makes intensive use of the computer, and, as a result, can be identified with the numerical approach to statistical problems, as opposed to the traditional probabilistic one (*computational statistics versus mathematical statistics*).

An explanation of DA's popularity must take the two following points into account:

1. DA provides users (biologists) with an easy-to-read graphical display of their multidimensional data together with a few numerical values quantifying its meaningfulness. One doesn't have to know too much mathematical vocabulary, and all that has to be learnt is how to recognize the sort of data one can analyse, what can be attained, and how to read the graphical displays.
2. On the other hand, mathematical justification of the methods has led to a formalisation that not only provides a homogeneous presentation of currently-used methods but has also led to original developments.

But perhaps the simplest explanation to the success of Data Analysis in France is that is enables the biologist to *talk* about his data.

**Factorial methods for analysing one triple**

Let's assume that the data to be studied is contained in a matrix $\boldsymbol{Y}_{nxp}$ of $p$ measurements made on $n$ observations. The idea is to find graphical displays where closeness between two points indicates similarities between the rows or the columns they represent. For a mathematician the only way to formalise these proximities is to associate the data $\boldsymbol{Y}$ with a positive semi-definite $\boldsymbol{Q}_{pxp}$ that defines distances between observations, and a diagonal matrix $\boldsymbol{D}_{nxn}$ of observations' weights $p_i$ used for computing the covariances. This presentation will stress the symmetry between the rows and columns of $\boldsymbol{Y}$.

The first thing the user has to learn is how to choose the method that leads to the relevant graphical displays: this will depend on the type of data and the study's objective. In order to learn how to read the graphical display one has to understand the meaning of the two sorts of numerical criteria provided with the solution:

- global criteria indicate the amount of data reconstructed and
- the individual criteria quantify the part played by each variable or observation in the definition of the representation spaces and symmetrically the amount of each row and column reconstructed by these spaces.

The transition formulae provide the link between the rows and columns of $\boldsymbol{Y}$ and the co-ordinates issued by the analysis thus making interpretation easier.

Specific names have been given to the various triples, as shown in the displayed box:

| | |
|---|---|
| 1. $\boldsymbol{Y}$ centred quantitative<br>$\boldsymbol{Q}$ identity matrix $\boldsymbol{I}_{pxp}$ | PCA (Principal Component Analysis) of variance-covariance matrix |
| 2. $\boldsymbol{Y}$ centred quantitative<br>$\boldsymbol{Q}$ diagonal matrix of the variances' inverses | PCA of correlation matrix |
| 3. $\boldsymbol{Y}$ centred quantitative<br>$\boldsymbol{D}$ a non-diagonal matrix | PCA on correlated observations |
| 4. $\boldsymbol{Y}$ Z $\boldsymbol{D}^{-1}\boldsymbol{P}\boldsymbol{Q}^{-1}$, $\boldsymbol{P}$ a contingency table<br>$\boldsymbol{Q}$ and $\boldsymbol{D}$ are the diagonal matrices made from $\boldsymbol{P}$'s marginals | Correspondence Analysis |
| 5. $\boldsymbol{Y}$ defined as above for an indicator matrix $\boldsymbol{P}$ | Multiple Correspondence Analysis |
| 6. $\boldsymbol{P}$ and $\boldsymbol{D}$ as in 4.<br>$\boldsymbol{Y}$ Z $\boldsymbol{D}^{-1}\boldsymbol{P}$ and $\boldsymbol{Q}$ Z $\boldsymbol{I}_{pxp}$ | Non-symmetric Correspondence Analysis |
| 7. $\boldsymbol{Y}$ is a row and column centred matrix of log ($\boldsymbol{P}_{ij}$), $\boldsymbol{Q}$ and $\boldsymbol{D}$ as in 4 | Mean-squares solution to the log-linear model |
| 8. $\boldsymbol{Y}$ a symmetrical $n \times n$ similarity matrix | Multidimensional scaling |

Replacement of $\boldsymbol{R}^n$ and $\boldsymbol{D}$ by a space $L^2(\Omega, \boldsymbol{A}, \boldsymbol{P})$ brings us back to the PCA of $p$ random variables, whereas that of $\boldsymbol{P}^f$ and $\boldsymbol{Q}$ by $L^2(\boldsymbol{T}, \boldsymbol{B}, \mu)$ provides a way of studying the observations when the data are given as curves.

If we insert additional constraints on the co-ordinates (positivity, order, equality, etc.) we widen the scope of the possible approaches within the same framework.

**Factorial methods for analysing several triples**

If we have a triple $(\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$ that we want to use as a reference for a data matrix $\boldsymbol{X}$ of measurements on the same observations, it seems reasonable to look for the semi-definite positive matrix $\boldsymbol{M}$ such that $(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{D})$ gives row representations that are as close as possible to those provided by $(\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$. This general problem has a simple mathematical solution which provides co-ordinates and interpretation criteria in just the same way that PCA does. Different types of X and Y matrices lead to various classical methods as outlined in the displayed box shown overleaf.

Other couples $(\boldsymbol{X}, \boldsymbol{Y})$ can be considered within the same mathematical scheme leading to alternatives to the classical methods. For instance in (4) one could replace $\boldsymbol{Q}$ Z $\boldsymbol{S}_{YY}^{-1}$ thus obtaining new types of discriminant analysis. Once the part of $(\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$'s dispersion rebuilt by $(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{D})$ has been identified it seems natural to look at the residual dispersion, in fact this has led to recent work on the decomposition of variability. If $\boldsymbol{Z}$ is a data matrix of the same measurements on other observations, we can look for $\boldsymbol{R}$ such that $(\boldsymbol{Z}, \boldsymbol{Q}, \boldsymbol{R})$

| | |
|---|---|
| 1. $Y$ centred quantitative<br>$X$ centred quantitative | PCA with respect to instrumental variables |
| 2. $Y$ and $X$ as in 1.<br>but $Y$ with only one column | Multiple Regression |
| 3. $Y$ and $X$ as in 1.<br>$Q$ Z $S_{YY}^{-1}$ | Canonical Analysis |
| 4. $Y$ indicator matrix<br>$X$ and $Q$ as in 3. | Discriminant Analysis |
| 5. $Y$ issued from<br>a contingency table, $X$ as in 1. | Correspondence Analysis with respect to instrumental variables |
| 6. $Y$ as in 1.<br>$X$ indicator matrix<br>(one or several categorical variables) | Geometrical approach to MANOVA |
| 7. $X$ and $Y$ as in 1.<br>$M$ a rotation matrix | Procrustes Analysis |

gives column representations that are as close as possible to those provided by $(Y, Q, D)$. This approach can be combined with the preceding one.

Now, let us suppose that we have several triples $\{(Y_k, Q_k, D)_{k=1..K}\}$, relative to the same observations, it is quite easy to form a $K \times K$ similarity matrix based on the comparison of how the $K$ analyses represent the observations. Then any method used for analysing dissimilarity matrices can provide the multiway analysis of the $K$ triples.

**Clustering**

Clustering has a special place in our view of multidimensional analysis; the pioneers in this area have always kept close to applications, providing software with efficient algorithms that enable many types of clustering, with different indices. In partitioning methodology the most popular technique is the 'Nuées Dynamiques' method that generalizes the $k$-means procedure to many types of kernels: linear regressions, factorial spaces, density functions. From the outset, work on hierarchies has always given an important role to the algebraic properties of dissimilarities, in particular the bijection between ultrametrics and hierarchies is well-known, recent work has been done on similar equivalences between distances and other types of representations:

| | |
|---|---|
| 1. Ultrametric distances | Hierarchical trees |
| 2. Distances with a centre | Single knotted additive trees |
| 3. Quadrangular distances | Additive Trees |
| 4. Robinson distances | Pyramids |

Thus there are all sorts of methods available for faithful representations of a user's data. For hierarchical trees, algorithms are available that include constraints on the solution such as a certain order on the tree's branches or simultaneous clustering of two sets.

**Where are we going next?**

It must be admitted that the computer has played a crucial role in developing these techniques. Making multidimensional methods available on all the commonest configurations was the first step, but now the ever increasing computational power influences today's research:

- First of all, current advances in numerical analysis have enlarged the scope of the initial optimization problem. Solutions to the factorial methods have, up to now, been given as linear combinations of the original variables. Today it is possible to look for non-linear functions of such a combination or even a linear combination of smooth transformations of the original variables.
- A second contribution of the computer has been to produce procedures based on resampling, such as the bootstrap and cross-validation. Combination of geometry and these resampling techniques has shed a new light on the inferential approach because initial hypotheses can now be left out.
- Finally, the computer enables description of a phenomenon not only through a few numerical values but also through non-numerical operations such as those developed in the analysis of symbolic data, a domain close to artificial intelligence.

**If you want to go a little further**

A small bibliography is provided that should make a first contact easy. The French books have been chosen for their popularity, the English books for their proximity to our presentation, and the articles because they are in direct relation with the results stated here.

**a few useful books:**

**Benzécri, J. P.** (1973) *L'Analyse des Données*, Vols. 1 & 2, Ed: Dunod, Paris.
**Bouroche, J. M. & Saporta, G.** (1980) *L'Analyse des Données*, Collection Que-sais-je, Puf, Paris.
**Caillez & Pages** (1976) *Introduction à l'analyse des données*, SMASH, Paris.
**Lavit, C.** (1988) *Analyse Conjointe de Plusieurs Tableaux Quantitatifs*, Masson, Paris.
**Lermann, I. C.** (1970) *Les bases de la classification automatique*, Gauthiers Villars, Paris.
**Roux, M.** (1986) *Algorithmes de Classification*, Masson, Paris.

**for those who don't read French:**

**Greenacre, M. J.** (1984) *Correspondence Analysis*, Academic Press, NY.
**Jambu, M. & Lebeaux, M. O.** (1983) *Cluster Analysis and Data Analysis*, North Holland, New York.
**Lebart, L., Morineau, A., Warwick** (1984) *Multivariable descriptive statistical Analysis*, Wiley, New York.

**and a few typical articles:**

**Barthélémy, J. P., Leclerc, B., Monjardet, B.** (1986) On the use of ordered sets in problems of comparison and consensus of classifications, *Journ. of Class.*, **3**, 187–224.
**Besse, P. H. & Ramsay, J. O.** (1986) Principal components analysis of Sampled Functions, *Psychometrika*, **51**, 285–311.
**Diday, E.** (1986) Orders and Overlapping Clusters by Pyramids in *Multidimensional Data Analysis*, pp. 201–234, Ed: Jan de Leeuw & al., DSWO Press, Leiden.
**Escoufier, Y.** (1987) The Duality Diagram: A means for better practical applications, in *Developments in Numerical Ecology*, pp. 139–156, Ed: P. & L. Legendre, Springer-Verlag, Heidelberg.
**Escoufier, Y.** (1987) Beyond Correspondence Analysis, in *Classification and Related Methods of Data Analysis*, pp. 505–514, Ed: H. Bock, Elsevier, Amsterdam.
**Escoufier, Y. & Junca, S.** (1986) Least Squares Approximations of Frequencies and their logarithms, *Int. Stat. Rev.*, **54**, 279–283.
**Fichet, B.** (1986) Distances and Euclidian Distances for Presence-Absence Characters and their Application to Factor Analysis, in *Multidimensional Data Analysis*, pp. 23–46, Ed: Jan de Leeuw & al., DSWO Press, Leiden.
**Lebreton, J. D., Sabatier, R., Chessel, D.** (1988) Principal Component Analysis with Instrumental Variables as a Tool for modelling Composition Data, in *Multiway Data Analysis*, pp. 341–352, Ed: Coppi et Bolasco, North Holland.
**Le Calvé** (1988) Similarities Functions, in *Compstat 1988*, pp. 341–347, Ed: D. Edwards & N. E. Raun, Physica-Verlag, Heidelberg.
**Robert, P. and Escoufier, Y.** (1976) A Unifying Tool for Linear Multivariate Methods: the RV-coefficient, *Applied Statistics*, **25**, 257–265. ■