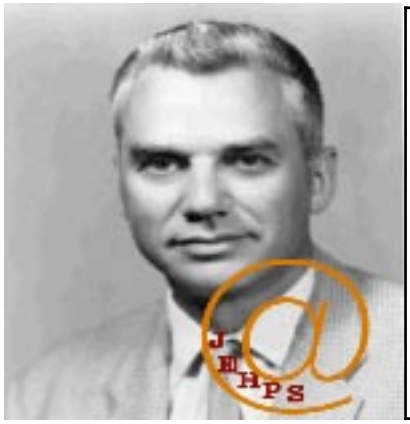

Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 4, n°2; Décembre/December 2008

www.jehps.net

Exploratory multivariate data analysis from its origins to 1980: Nine contributions.

LUDOVIC LEBART[1]

This issue is dedicated to the memory of Henry Rouanet who passed away during the preparation of this collection of contributions to which he was invited to participate. Henry Rouanet, an engineer, mathematician and psychologist Directeur de Recherches at the CNRS was a pioneer of multidimensional statistics and of "geometrical data analysis". Historical texts of Henry Rouanet concerning the theme of this special issue can be found in the "Document" section of this collection.

This thematic issue of the JEHPS has for object the origin and the development of exploratory methods of multidimensional statistics in the 20th century, without going beyond the 1980s. Since two (out of nine) contributions will be in French, let us note right away that the English equivalent of the French phrase "analyse des données" would be something like *exploratory multivariate data analysis*. Indeed, the English expression *data analysis* is used in a more general sense to designate applied statistics (with a pragmatic and computational connotation).

After the 1980s there was an explosion of methods together with the apparition of new paradigms which have not yet stabilized. The contributions presented here provide then material that could be reworked or whose foundations could be modified later on. Indeed, over these years we have

[1] ludovic.lebart@telecom-paristech.fr

seen the development of *neural networks*, s*elf organizing maps*, *data mining, learning theory*, *independent components analysis, and resampling methods*, which are all methods, schools or currents that are bound to contribute to the theme that we are interested in but which are still subject to debate or even controversy, and considerable terminological dispersion. Many of the authors of this thematic issue have been or are still major actors in *exploratory multivariate data analysis,* which gives this collection of testimonies an undeniable documentary interest.

Several countries or schools are represented by articles or documents but important gaps remain. Despite the internationalization and globalization of scientific activity, the geographical location of the scientific production but also of the observations remain important factors of explanation and interpretation in this field. As we will see, linguistic and even political barriers have had repercussions on the scientific exchange.

This issue of the Journal contains nine original contributions and three series of archive documents. All the authors are statisticians. However some of them have chosen to intervene from the point of view of the sociologist or historian.

1 Description, exploration, confirmation

Descriptive statistics allow us to represent statistical information in a graphical form by simplifying and schematizing it. Multidimensional descriptive statistics generalizes this idea in a natural way when the information concerns several variables or dimensions.

But multidimensionality induces an important qualitative change. To repeat a widespread analogy, microscopes and radiographic machines are not mere instruments of description they are also elements of observation and exploration and tools of research. In a similar way, the multidimensional reality is not just simplified because it is complex, it is also explored because it is hidden. The data must be prepared and coded, strict rules of interpretation must be used and the representations provided by the techniques used in the multidimensional case must be validated. These operations do not have the simplicity of elementary descriptive statistics. It is no more a matter of presentation but of analysis, of discovery and sometimes of verification and proof.

The new computational tools

The science of Statistics, born with the 20th century after works of precursors such as the astronomist Quetelet, the demographers, biometricians and statisticians Galton, Pearson, and then Fisher, have manipulated numbers for half a century without having at their disposal any real computational tools. The machines that can today be found in the pockets of school children and in most homes would have fulfilled the most ambitious dreams of statisticians up until 1960.

Faced with these new possibilities, John W. Tukey, founder of the field referred to as *Exploratory Data Analysis* (*EDA*), had a rather novel attitude (see Tukey, 1977, and the earlier publications of this author that are mentioned by the different authors of this thematic issue). With a more specifically multidimensional approach, J.-P. Benzécri, (1973) affirms that the computer requires that the very foundations of statistics be reconsidered.
These two pioneers did not have the immediate influence that one could have expected outside their direct spheres of influence. Without being entirely re-written, statistics has nevertheless grown progressively richer. Recent periods have brought about very major changes due to the diffusion of computational aids: existing tools have been improved and new tools have appeared,

new areas of application have been explored.

At present it is possible to process tables that correspond to tens of millions of observations and to hundreds or even thousands of variables. The change of scale in the data has rapidly lead to a change in the tools themselves and to the conception of new tools and approaches. However, statisticians know that there is sometimes no point in wanting to handle millions of observations when sampling is possible.
The removal of computational obstacles has had for result to spread the use of algorithmic techniques, among which major ones are the methods of automatic classification and methods involving costly algorithms. Other techniques, such as those of stepwise selection, the method of maximal likelihood estimates, of dynamic programming and the automatic search for rules in data bases are used on a large scale.

The statistical study of categorical data is by its very nature more complex than that of continuous numerical variables which generally reposes on the normal distribution and on the simple formalizations that ensue (maximal likelihood, least squares, for example). Thus it is not surprising that the computational possibilities have allowed for much progress to be made in this area: simple and multiple correspondence analysis in the descriptive case, log-linear models, discriminant analysis and logistic regression in the case of inferential statistics.

One of the innovations in statistics after 1960 was the appearance of techniques in the form of "products": software developed with financial and commercial constraints on their conception, production, and distribution. Like any finalized product, the advantage of the software was its ability to diffuse and its inconvenience that it entailed a certain rigidity. Like any product intended to be used by specialists it induced new divisions of labour that were sometimes not very desirable in the knowledge process. Software that is accessible and easy to use allow methods to be widely spread but lead to a careless use in areas where much caution would be called for.

Two types of approaches may be distinguished in multidimensional statistics: *the descriptive and exploratory approach* (which are often the non supervised approaches of learning theory) and the *inferential and confirmatory approach* (to which the supervised approaches belong) that constitutes the most comprehensive and established section of statistical science.

Let us briefly recall the characteristics of these two families of methods which are complementary approaches.

-

Descriptive and exploratory statistics allows sets of statistical data to be described by more or less elaborate summaries and graphics, and to establish relationships between variables without giving more importance to any particular variable. It is to this family of methods that this issue is dedicated. Traditionally, in this phase of the study, the conclusions only concern the data that is analyzed and they are not inferred to a larger population. The exploratory analysis essentially reposes on graphical representations and on multidimensional description techniques (e.g.: principal components analysis, correspondence analysis, clustering). Today's re-sampling

methods allow structures to be validated and thus to articulate (with caution) the exploration and inference.
- *Inferential and confirmatory statistics* allow hypotheses that are formulated *a priori* (or after a separate explanatory phase) to be validated or invalidated on the basis of statistical tests or probabilistic models, and also allow for extrapolation, that is the extension of certain properties of a sample to a larger population : the conclusions obtained from data extend beyond this data. Confirmatory statistics mainly involve methods referred to as explanatory or predictive which are, as their names suggest, intended to explain and then to predict, following decision rules, one variable that is given a central role using one or several other explanatory variables (multiple and logistic regressions, variance analysis, discriminant analysis, segmentation, etc.)

These approaches are complementary, and generally the exploration and description must precede the explanatory and predictive phase.

However, the approaches themselves are not always easy to discern and identify. Pure exploration is rare because a priori information and knowledge about the table of data (meta-data) always exist to some degree. Similarly, general hypotheses or even expectations that are not explicitly formulated by the user often exist. In a noteworthy work of synthesis on significance tests Cox (1977) drew attention to the problems related to the articulation of exploration and inference. Obviously, one cannot statistically test on data new hypotheses suggested by the data in question. But one cannot deny that, in certain circumstances, collections of data (especially large collections of multidimensional data) could suggest hypotheses. Thus it remains to follow a specific discipline during the processing (fragmentation, replication, reproduction and renewal of the collected data) that is productive despite the constraints it imposes.

Four sections will structure what follow this introductory chapter: the sections 2 and 3 which are concerned with principal axes analysis and classification methods respectively do in fact account for the quasi totality of the nine contributions that constitute this issue. The sections 4 and 5 devoted to validation methods and to related themes deal with areas of research not explored or only briefly brought up by these contributors.

2 Principal axes methods

- The *principal axes methods*, and also *principal components analysis*, act by reducing certain "multidimensional" representations, thereby producing essentially planar or sometimes three dimensional *graphical visualizations* of the elements that are to be described. In the French statistics literature, the “analyse factorielle” includes all the representation techniques that use "principal axes ": principal components analysis, simple and multiple correspondences, the analysis in common and specific factors of Spearman and Thurstone, used mainly by psychologists and by psychometricians *(factor analysis)*.
At the foundation of these principal axes analyses is a theorem, *Singular Values Decomposition*, that was presented for the first time by Eckart and Young (1936) for rectangular matrices and which generalized the works of Sylvester (1889) concerning square matrices. Gifi (1981/1990) also mentions the earlier and independent works of Beltrami (1873) and Jordan (1874). The

problem that is at hand is a problem of pure numerical reduction, i.e. a problem of data compression: how to fit, in the least squares sense, a matrix by another matrix of inferior rank. Among the first articles that were published on the algebraic and geometric methods of principal axes methods, we should note: Gower (1966), Gabriel (1971).

The contribution of **John C. Gower** to this special number of JEHPS **[The biological stimulus to multidimensional data analysis]** thus covers the history of principal axes methods as well as that of clustering techniques which are discussed in section 3. A creator within this family of methods, J. Gower gives a first hand testimony. His contribution is presented in relation to principal axes methods since it seemed a priori that a large part of his works concerns this area. However, his recounting of the genesis of classification methods and their first hesitant steps, the incomprehension of how far reaching they were and how they were greeted in different communities will probably be considered as one of the highlights of this collection. As the title indicates, the emphasis is on the role of biometrics, anthropometrics, agronomics and natural sciences in the development of multidimensional analyses during and after the founding works of K Pearson and R Fisher. We will evoke again Gowers's contribution further on.

2.1 Principal components analysis

Principal components analysis is the oldest and most established of the methods of principal axes visualization. Conceived for the first time in a limited setting by Karl Pearson in 1901, and integrated into mathematical statistics by Harold Hotelling in 1933, principal components analysis has not really been used before the arrival and diffusion of computational aids. For the traditional statistician, it is about searching for the principal axes of a multivariate normal distribution from a sample. This is the initial presentation of Hotelling (l933), and later that of classic manuals of multivariate analysis, such as the fundamental treatise by Anderson (1958). To classic factorialists, it is a special case of factor analysis (case of null or equal specificities; cf. Horst, 1965; Harman, 1967). Finally, from a more recent data analysis perspective, it is a technique for the presentation of data which has an optimal character according to certain algebraic and geometric criteria and that could be used without involving hypotheses of a statistical nature. This point of view which is currently widespread is perhaps the most ancient one. It is the one adopted by Pearson (1901). A presentation that is closer to current preoccupations can be found in the synthetic article by Rao (1964).

2.2 Factor analysis

Among the founders of the method that comes mainly from psychologists and psychometricians we find Spearman (1904) (one latent factor), then Garnett (1919) and Thurstone (1947) (several latent factors). Although it is a specific statistic model and not an exploratory method, it is closely linked to principal components analysis. The developments that it has given rise to are complex and diverse. On this topic, one may consult the works of Harman (1967), Mulaik (1972). To conclude this brief recall let us mention the historical works of Anderson and Rubin (1956) and of Lawley and Maxwell (1963) who immersed factor analysis in a classical inferential setting.

2.3 Correspondence analysis

The other fundamental technique is correspondence analysis. Most other techniques are derived from these two basic ones in order to adapt to specific areas of application. One of the most used ones is multiple correspondences analysis that is applicable to large sets of nominal variables.
Correspondence analysis was presented and developed under this name for the first time by Escofier-Cordier (1965) and Benzécri (1969). It has several precursors, among which we should mention Guttman (1941), Hayashi (1956). These two authors have independently proposed the technique as a way of analyzing data.

The contribution of **Fionn Murtagh**, to this issue of JEHPS**, [Origins of Modern Data Analysis Linked to the Beginnings and Early Development of Computer Science and Information Engineering]** analyzes in greater detail the specific approach of Chikio Hayashi, who was also, by the way, the one who coined the concept of " Data Science ". The document in appendix 13 contains a short biography of Professor Hayashi. We will return to the contribution of F. Murtagh further on. Correspondence analysis can indeed be presented from different points of view, it is difficult to give a precise history. The theoretical foundations probably go back to the works of Fisher (1940) on contingency tables, in a classical inferential statistics setting. Since the works of Benzécri (1973) and of Escofier-Cordier (1965), it is mainly the algebraic and geometric properties of the descriptive tool provided by the analysis that are used. More distant ancestors of correspondence analysis would be, in a completely independent manner, Richardson and Kuder (1933) and Hirschfeld (1935). Richardson and Kuder aimed to achieve a better selection of salesmen for the company *Procter and Gamble*, and empirically they discovered the method of " reciprocal averaging " while Hirschfeld discovered an important property of mathematical statistics (concerning this point, see the contribution of J. Gower). Jan de Leeuw (1983) showed that Karl Pearson had been close to discovering correspondence analysis in 1906, but that he lacked (as all his contemporaries) knowledge of the spectral properties of matrices. These varied contexts are typical for correspondence analysis, a method that is as useful in practice as it is simulating from a theoretical point of view. Cf. the reference Escofier (2003) that is a posthumous collection of works from Brigitte Escofier-Cordier. Cf. also the historical references of Hill (1974), and then the historical work of Benzécri (1982), scrutinized by Michel Armatte.

The article of **Michel Armatte** in this special issue **[Histoire et Préhistoire de l'Analyse des données par J.P. Benzécri: un cas de généalogie rétrospective]** analyzes from the point of view of a science historian (and when thirty years have gone by) the work completed by J.P. Benzécri after the publication of the aforementioned article of M.O. Hill (*Correspondence Analysis: a neglected multivariate method*) in 1974. That work (*Histoire et Préhistoire de l'analyse des données*) which came in the form of lecture notes was published in the *Cahiers de l'Analyse des données* en 1976/77, and then published as a book by Dunod (Paris) in 1982. It is perhaps the title of the article by Hill (" *a neglected method* ") that irritated a community that on the contrary had undergone over a period of several years a frenetic and often excessive use of correspondence analysis in every conceivable area (" *an overused method* " would have been better received). This article by a historian is also a lively testimony from a former student of Benzécri, which makes the reading as pleasant as it is instructive. Michel Armatte's witty remark that " The history of the sciences is a much too important area to be left to the scientists…whom we want to study! " is certainly very much in its place in this issue of the JEHPS in which most of

the contributors are scientists who are deeply and often passionately involved in their field.

The contribution of **Alain Desrosières [Analyse des données et sciences humaines : comment cartographier le monde social?]** shows in a precise manner how great an impact methods of data analysis had on the methods of the social sciences in France in the 70s, an impact at the limit of infatuation, as indicates for example the publications of principal axes graphical displays in high circulation magazines (see, e.g., the colourful display in the annex of Desrosières paper). In a rather surprising way, the methodology itself had certain political overtones. Even the temple of official statistics (the "INSEE" [National Institute of Statistics and Economic Studies], to which Alain Desrosières belongs) was perturbed for a while by the frenzy of sociologists and economists who had finally discovered a tool that could measure up to the complexity of their objects. This influence has continued to make itself felt in recent years in the context of "geometrical data analysis" with the works of Henry Rouanet and Brigitte Le Roux, developed often among researchers who are disciples of the sociologist Pierre Bourdieu.

The contribution of Fionn Murtagh which has already been mentioned also brings up the particularity of the computational environment at the origin and development of contemporary data analysis but mainly it contains a very sharp analysis of the linguistic and cultural obstacles that slowed down the understanding and the diffusion of the methods. The lack of translations of the works from the Benzécri research department in the 70s led to a lack of understanding within the Anglo-Saxon world that contrasted with the adhesion of linguistically close countries like Italy (see the contribution of A. Rizzi in this collection). The misunderstanding continued later on when a simplified version of Benzecri's "Traité d'Analyse des Données" intended for practitioners and close to the manual "Pratique de l'analyse des Données" by Bastin et al. (1980), was translated into English under the title " Correspondence Analysis Handbook " (Benzécri, 1992). This work was erroneously greeted as being the translation of the treatise from 1973. "This is a translation of the Benzécri 'bible' on correspondence analysis previously available only in French" wrote David Hand (1994) in the "Journal of Classification". Despite the fact that the translated book was not intended for statisticians, D. Hand nevertheless notes some qualities of the book : "With its many examples of correspondence analysis being applied (in different ways) the book provides an excellent illustration of how sensitive and sophisticated use of a single technique can shed light on data in many different ways. It serves to support the position that a thorough grasp of a few techniques is better than a weak grasp of many. ".

2.4 Multiple correspondence analysis

Multiple correspondences analysis is a simple extension of the area of applicability of correspondence analysis to a *complete disjunctive binary table*. The properties of the method are interesting, the computational procedures and the rules of interpretation and representation obtained are simple and specific. The principles of this method can be traced back to Guttman (1941), but also to Burt (1950) or to Hayashi (1956). Other types of extensions of correspondence analysis based on generalized canonical analysis have their foundation particularly in the works of Carroll (1968), Horst (1961) and Kettenring (1971).
Multiple correspondences analysis has also been developed under the name *Homogeneity Analysis* by the research team of de J. de Leeuw since 1973 (cf. Gifi, 1981/1990) and under the name of *Dual Scaling* by Nishisato (1980). In Nakache (1973) we find an application of

correspondence analysis to a complete disjunctive table. A synthetic exposition of these various approaches is provided by Tenenhaus and Young (1985).

The contribution of **Willem Heiser** to this issue of JEHPS **[Psychometric Roots of Multidimensional Data Analysis in the Netherlands: From Gerard Heymans to John van de Geer]** contains precious information about all the developments of principal axes methods. There is a general consensus among all the authors of this special issue that psychology (and more precisely differential psychology, as both Heiser and Desrosières underline) was important for its foundation. W. Heiser describes how the two psychologists Gerard Heymans, and then John van de Geer methodically manipulated and recorded multidimensional data and how the "Dutch School of psychology" (recognized in particular by Spearman and his contemporaries) finally contributed to the birth of a "Dutch School of data analysis", based mainly in Leiden and in Groningen. The readers will also discover the identity of Gifi (the statistical "Bourbaki" of the Netherlands), one of the authors that is most frequently mentioned in this special issue.

The contribution of **Antoine de Falguerolles [L'analyse des données ; before and around]** gives another point of view, that of the statistician in academia (moreover a cultured and curious one) about, on the one hand, the situation of data analysis in France in the period 1965-1985, which was, as we have seen, characterized by an exceptional frenzy, and on the other hand about several historical events, some of which go far back in time (1588 for the " Felissima Armada "…) that have contributed to the examination (not yet the analysis !) of multidimensional data. A short section also discusses the introduction of probabilistic modelling in some data analysis approaches. This contribution also illustrates the renewed interest (possibly related to the accessibility of multimedia communication tools) in graphical methods and their history.

The contribution of **Alfredo Rizzi [Italian Contributions to Data Analysis]** is presented in relation to factorial methods since it attributes much importance to this theme, but it also concerns clustering techniques in which Italian researchers were extremely active. A. Rizzi begins by recalling that the first congress of the *International Statistical Institute* was held in Rome in 1887 and then recalls the works of the pioneer Corrado Gini in 1912 (which was already an elaborate descriptive approach). He then mentions authors who are all known for their publications in English but the Italian texts are either original or published earlier, which gives them an unquestionable historical interest.

2.5 Multiway arrays

There is no equivalent of the theorem of Eckart and Young for the case where the table is three dimensional. This can be expressed in the following way: the hierarchical decomposition of an element of the tensor product of two Euclidian spaces into a sum of tensorial products of pairs of vectors belonging to each space is unique. But such a decomposition is not unique in the case of an element of the tensor product of more than two Euclidian spaces (cf. Benzécri, 1973; Tome 2B, n°6). Therefore, in this case, there cannot be an exploratory approach that is as well established as in the case of tables with two entries.

Let us briefly mention some works of reference on the topic of multiway arrays. We find a synthesis and a classification of the main approaches in the work of Kroonenberg (1983). The first works on this theme were those of Tucker (1964, 1966) followed by those of Harshman

(1970), both in the context of classic factor analysis.

Generalized canonical analysis has been presented in Horst (1961), where it is mentioned in third place among four other possible generalizations of canonical analysis. Carroll (1968) and Kettenring (1971) returned to it and developed it. Let us also mention the works of Pagès et al.(1976) based on the operators defined by Robert and Escoufier (1976).

Procrustean analyses are related to a frequent concern that is common in multidimensional statistics: *n* individuals or observations are described on one hand by *p* variables (columns of X), and on the other hand by *q* other variables (columns of Z). How can the two clouds of individuals and the two systems of distance between individuals be compared? It was Tucker (1958) who originally proposed such a method for comparing two sets of tests concerning the same individuals. The technique was then studied by Cliff (1966), Schönemann (1968), Schönemann and Carroll (1970), and generalized by Gower (1975).

3 Clustering

Clustering methods lead to a grouping into classes of objects (partitioning methods), or in families of hierarchical classes for the methods of hierarchical clustering. The elements that are to be described are grouped in the least arbitrary way possible based on their vectors of description.

Clustering is a branch of data analysis that has given rise to many and varied publications. It has been developed considerably these last years to address the need for automatic extraction of hidden information or for identifying groups or classes from huge data sets. The basic historical work is probably that of Sokal and Sneath (1963) following the seminal article by Sneath (1959). The fist manuals that were published were those of Lerman (1970), Jardine and Sibson (1971), Anderberg (1973), Benzécri (1973), Bock (1974), Hartigan (1975). One of the first historical syntheses on this subject is that of Cormack (1971). A more recent work of synthesis in hierarchical classification was done by Gordon (1987).

3.1 Mobile centers and k-means

The contribution of **Hans Hermann Bock** in this number **[Origins and extensions of the k-means algorithm in cluster analysis]** constitutes an original and complete clarification of this theme and includes recent developments that go well beyond the 1980s. Obviously it is difficult to identify with certainty the first user of a method that is based on so simple principles (it could have been used without giving rise to an official publication), but the variety of sources and the diversity of the versions of the algorithm are surprising. Despite the fact that the formalism is limited and that its efficiency is to a high extent confirmed only by experimental results, the method of k-means clustering is probably the technique that is currently the best adapted to large collections of data and also the most widely used one in this type of application. The algorithm can be ascribed mainly to Forgy (1965) despite the fact that several works, sometimes earlier ones (Thorndike, 1953, is often mentioned, but as H. H. Bock remarked in his contribution, the relationship with the real k-means method is not very clear), most often later (MacQueen, 1967, Diday, 1971) have been carried out in parallel and independently introducing different versions or generalisations.

3.2 Hierarchical classification, other methods

A first history of hierarchical classification is found in a previously mentioned article by Cormack (1971), while a synthesis of more recent work is that of Gordon (1987).
Like k-means techniques, hierarchical clustering is often used to complement principal axes techniques. Thus it is treated by several authors in this number, especially by J. Gower, who retraces (section " Taxonomy " of his article) the first attempts and approaches of Sneath, Williams, Lambert, Lance which he has actually witnessed.

Let us mention an interesting historical point: It has been shown (Gower and Ross, 1969) that hierarchical classification with the single linkage criterion is equivalent to calculating the minimal spanning tree on the complete graph valued by the distances between objects. The algorithm of Kruskal (1956) for calculating this tree is the one that is most often mentioned. But the algorithm of Florek *et al.* (1951), which is older and much more efficient, allowing for manual computation, led to a real school of data analysis in Poland (the *Wroclaw taxonomy*) even before modern computational aids became available. One can consult Graham and Hell (1985) for a (fascinating) history of the minimal spanning tree algorithm whose first version known to date can in fact be traced back to Borůvka (1926).

Other agglomerative criteria, reducing the so-called "chain effect", may yield more reliable results than the single linkage criterion (see Ward, 1963 ; Wishart, 1969). Under certain conditions, these algorithms can be highly accelerated by using the concept of reciprocal neighbours introduced by McQuitty (1966).
The contribution of **Boris Mirkin** and **Ilya Muchnik [Some topics of current interest in clustering: Russian approaches 1960-1985]** has the great merit of making accessible works that are sometimes ignored by non Russian speakers (the references are sometimes not known, the authors are often known!). The themes that are treated (with modesty and humour in some of the authors' remarks) are the comparisons of clusters, the studies of consensus between classifications, the "one-clustering", the "bi-clustering", and classifications on graphs.

4 Validity of the results

The study of the validity of the results of *principal axes analysis* has lead to much research, but which has since the 1980s taken a different direction with the "*computer intensive methods*". They are mentioned by J. Gower in his contribution, but not very present in the different contributions to this special issue. Nevertheless it is an area of research that led to remarkable results from a theoretical point of view.

4.1 Numerical stability of the principal axes :

Form a purely numerical point of view, Escofier and Le Roux (1972), and Escofier (1979) have treated the stability of the axes in principal axes analysis (principal components analysis and correspondence analysis). These authors study the maximal variations of the eigenvectors and the eigenvalues when well defined modifications alter the data: removal or addition of elements to the data tables, the influence of regrouping several elements or small modifications of the values in the table, the influence of the chosen distances and weightings. Their results are based on the works of Davis and Kahan (1970). The sub-spaces corresponding to the largest eigenvalues are the most stable ones with respect to possible perturbations of the diagonalized matrix (cf. the fundamental works of Wilkinson, 1965, and of Kato, 1966).

4.2 Statistical inference and principal axes

The law of Wishart, established by Fisher (1915) in the case of 2 variables, and then more generally by Wishart (1928), concerns the distribution of an *empirical covariance matrix*. The probability density function of the eigenvalues of a Wishart matrix was expressed simultaneously by Fisher (1939), Girshick (1939), Hsu (1939) and Roy (1939), a nice example of simultaneous and practically independent discoveries. The integration of this complex density has given rise to several publications; among the main ones, that of Pillai (1965), Sugiyama (1966), Krishnaiah and Chang (1971), inspired by the works of the physicist Mehta (1960).

In the publication where he gives the expression for the distribution function of the eigenvalues of a Wishart matrix, Girshick (1939) computes the asymptotic variance and covariance (when the number of observations n tends to infinity) of the eigen-elements of the empirical covariance matrix S, in the case where all the eigenvalues of the theoretical covariance matrix S are distinct. He also gives the theoretical variance and covariance of the eigenvalues of the empirical correlation matrix in the case where all the eigenvalues of the theoretical correlation matrix R are also distinct.
Bartlett (1950) proposed a method for testing the equality of eigenvalues of the matrices S and R. Lawley (1956) studied in greater detail the case of the smallest eigenvalues of S. After his pioneering works on asymptotic distributions in 1951, Anderson (1963) generalized the results of Lawley by determining the limit distributions of the eigenvalues without assuming the corresponding theoretical values to be distinct. The confidence intervals of Anderson are still currently used by those who practice principal components analysis.

4.3 Resampling

Bootstrap is nothing else than a particular simulation technique, based on the empirical distribution of the sample. In its simplest form, this technique for estimating how much confidence one should have in the estimation q* of an unknown parameter q, introduced by Efron (1979), consists in simulating m samples of the same size n as the initial sample. With the exception of the works of Gifi (1981) (see the contribution of W. Heiser) that specifically concern correspondence analysis (in fact the bootstrap principle differs significantly according to the various principal axes techniques), the firsts works where bootstrap is applied to validate results of principal components analysis are probably those of Diaconis and Efron (1983), and are thus later than the limit date that we have fixed.

5. Related topics

It is not the ambition of neither the contributions presented in this number nor of these introductory remarks to cover all the fields of a discipline that even before 1980 had developed in many directions. Inevitably, some serious gaps or omissions may still remain. For example, we could mention the exploratory methods known as *Projection Pursuit* (Friedman and Tukey, 1974) which later gave rise to a number of extensions. Another pioneering article which had much impact on principal components analysis and related methods: Wold (1966) (mentioned however by two contributors to this issue: J. Gower and A. de Falguerolles). Other descriptive methods that do not belong to the two large families studied here (principal axes and classification) will only be mentioned briefly, as is the case for the purely graphical methods intended for

representing matrices of low dimension, seriation methods and especially the methods of *multidimensional scaling* (Shepard, 1974, Kruskal and Wish, 1978) which are mentioned several times by J. Gower in his contribution, but whose history is not studied in detail.
Among the entirely visual methods, we should mention, mainly due to their historical importance, those advocated by Bertin (1967), the method of faces of Chernoff (1973), in which each face corresponds to an individual (or an observation) and each feature of the face is a variable; the curve methods of Andrews (1972), where the different parameters of the curves are the values of the variables; the method of constellations of Wakimoto and Taguri (1978). Seriation techniques (initially a method of dating in archaeology used by Petrie at the end of eighteenth century) aims to make particular structures of tables appear through simple reclassification of rows and columns. For synthetical discussions on this topic, see for example Arabie (1978). All these graphical methods tend to be used in relation to specific applications and are less adapted to the processing of large data arrays.

Bibliography

Anderberg M.R. (1973): *Cluster Analysis for Applications*. Academic Press, New York.

Anderson T. W. (1951): The asymptotic distribution of certain characteristic roots and vectors. *Proc. of the 2nd Berkeley Symp. on Math. Statist. and Prob.*, p 103-130, Univ. of California Press.

Anderson T. W., Rubin H. (1956): Statistical Inference in factor analysis. *Proc. of the 3rd Berkeley Symp. on Math. Statist.,* 5, p 111-150.

Anderson T.W. (1958): *An Introduction to Multivariate Statistical Analysis* (Second edition : 1984). J. Wiley, New York.

Anderson T. W. (1963): Asymptotic theory for principal component analysis. *Ann. Math. Statist.,* 34, p 122-148.

Andrews D. F. (1972): Plots of High-dimensional data. *Biometrics*, 28, p 125-136.

Arabie P. (1978): Constructing blockmodels : how and why. *J. of Math. Psychology,* 17, (1), p 21-63.

Bartlett M.S. (1950): Tests of significance in factor analysis. *British J. Psych. (Stat. Section),* 3, p 77-85.

Bastin C., Benzécri J.-P., Bourgarit C., Cazes P. (1980) : *Pratique de l'analyse des données*. Dunod, Paris.

Beltrami E. (1873): Sulle funzioni bilineari. *Giorn. Math. Battaglin.* 11, p 98-106.

Benzécri J.-P. (1969): Statistical analysis as a tool to make patterns emerge from clouds. In : *Methodology of Pattern Recognition* (S.Watanabe, Ed.) Academic Press, 35-74.

Benzécri J.-P. (1973): *L'Analyse des Données*. Tome 1: *La Taxinomie*. Tome 2: *L'Analyse des Correspondances* (2de. éd. 1976). Dunod, Paris.

Benzécri J.-P. (1982): *Histoire et préhistoire de l'analyse des données*. Dunod, Paris.

Benzécri J.-P. (1992): *Correspondence Analysis Handbook*. (Translation by T. K. Gopalan). Marcel Dekker, New York.

Bertin J. (1967) : *La sémiologie graphique*. Gauthier-Villars, Paris. (*The semiology of Graphics*. University of Wisconsin Press, 1983).

Bock H. H. (1974): *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung van Daten (Cluster Analysis)*. Vandenhoeck & Ruprecht, Götingen.

Borůvka O. (1926) : O jistém problému minimálním. *Práce Mor. Přírodověd. Spol. v Brně (Acta Societ. Scient. Natur.Moravicae)*, 3:37-58, 1926.

Burt C. (1950): The factorial analysis of qualitative data. *British J. of Statist. Psychol.* 3, 3, 166-185.

Carroll J. D. (1968): Generalization of canonical correlation to three or more sets of variables. *Proc. Amer. Psychological Assoc.* 227-228.

Chernoff H. (1973): The use of faces to represent points in *k*-dimensional space graphically. *J. Amer. Statist. Assoc.*, 68, 361-368.

Cliff N. (1966): Orthogonal rotation to congruence. *Psychometrika*, 31, 33-42.

Cormack R.M. (1971): A review of classification. *J. of Royal Statist. Society*, Serie A, 134, Part. 3, 321-367.

Cox D. R. (1977): The role of significance tests. *Scandinavian Journal of Statist.*, 4, 49-70.

Davis C., Kahan W. M. (1970): The rotation of eigenvectors by a perturbation. *Journal of SIAM (Numerical Analysis)*, 7, 1-46.

De Leeuuw (1983): On the Prehistory of Correspondence Analysis. Statistica Neerlandica, vol 17, n° 4, 161-164.

Diaconis P., Efron B. (1983): Computer intensive methods in statistics. *Scientific American*, 248, (May), 116-130.

Diday E. (1971): La méthode des nuées dynamiques. *Revue Statist. Appl.* 19, n° 2, 19-34.

Eckart C., Young G. (1936): The approximation of one matrix by another of lower rank. *Psychometrika*, l, 211-218.

Efron B. (1979): Bootstrap methods : another look at the Jackknife. *Ann. Statist.*, 7, 1-26.

Escofier B. (1979 a): *Stabilité et approximation en analyse factorielle*. Thèse d'Etat, Université Pierre et Marie Curie, Paris.

Escofier B. [Cordier B.] (1965): *l'Analyse des correspondances*. Thèse, Faculté des Sciences de Rennes ; publiée en 1969 dans les *Cahiers du Bureau Universitaire de Recherche Opérationnelle*, n°13.

Escofier B., Le Roux B. (1972): Etude de trois problèmes de stabilité en analyse factorielle. *Publication de l'Institut Statistique de l'Université de Paris*, 11, 1-48

Fisher R. A. (1915): Frequency distribution of the value of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507-521.

Fisher R.A. (1939): The sampling distribution of some statistics obtained from non linear equations. *Ann. Eugen.*, 7, 179-188.

Fisher R.A. (1940): The precision of discriminant functions. *Ann. Eugen.*, 10, 422-429.

Florek K, Lukaszewicz J, Perkal J, Steinhaus H, Zubrzycki (1951): Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.*, 2, 282-285.

Forgy E. W. (1965): Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometric Society Meetings*, Riverside, California (Abstract in : *Biometrics* 21, 3, 768).

Friedman J. H., and Tukey J.W. (1974): A Projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, Ser. C, 23, 881-889.

Gabriel K.R. (1971): The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 3, 453-467.

Garnett J.-C. (1919): General ability, cleverness and purpose. *British J. of Psych.,* 9, 345-366.

Gifi A. (1981): *Non Linear Multivariate Analysis*, Department of Data theory, University of Leiden. Cf. also: Gifi A. (1990): *Non Linear Multivariate Analysis*, Wiley, Chichester.

Girshick M.A. (1939): On the sampling theory of roots of determinantal equations. *Ann. Math . Statist.*, 1, 10, 203-224.

Gordon A.D. (1987): A review of hierarchical classification, *J.R.Statist.Soc.*, A, 150, Part 2, 119-137.

Gower J. C. (1966): Some distance properties of latent and vector methods used in multivariate analysis. *Biometrika*, 53, 325-328.

Gower J. C. (1975): Generalized Procrustes Analysis. *Psychometrika*, 40, (1), 33-51.

Gower J. C., Ross G. (1969): Minimum spanning trees and single linkage cluster analysis. *Appl. Statistics*, 18, 54-64.

Graham R. L. and Hell P. (1985): On the history of the minimum spanning tree problem. *Ann. Hist. Comput.* 7, 43-57.

Guttman L. (1941): The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst P., ed.) 251 -264, SSCR New York.

Hand D. (1994): Review of the book: Correspondence Analysis Handbook (J.P. Benzécri). *Journal of Classification*, vol. 11, n°2, 289-290.

Harman H.H. (1967): *Modern Factor Analysis* (2nd ed.). Chicago University Press, Chicago.

Hartigan J. A. (1975): *Clustering Algorithms*. J. Wiley, New York.

Hayashi C.(1956): Theory and examples of quantification. (II) *Proc. of the Institute of Statist. Math.* 4 (2), 19-30.

Hill M.O. (1974): Correspondence analysis: a neglected multivariate method. *Appl. Statist.* 3, 340-354.

Hirschfeld H.D. (1935): A Connection between correlation and contingency. *Proc. Camb. Phil.*

Soc. 31, 520-524.

Horst P. (1961): Relation among m sets of measures. *Psychometrika,* 26, 129-149.

Horst P. (1965): *Factor Analysis of Data Matrices*. Holt, Rinehart, Winston, New York.

Hotelling H. (1933): Analysis of a complex of statistical variables into principal components. *J. Educ. Psy*. 24, 417-441, 498-520.

Hsu P. L. (1939): On the distribution of the roots of certain determinantal equations. *Ann. Eugen*. 9, 250-258.

Jardine N. Sibson R. (1971): *Principle of Mathematical Taxonomy*. Wiley, New York.

Jordan C. (1874): Mémoire sur les formes bilinéaires. *J. Math. Pures et Appliquées*. 19, 35-54.

Kato T. (1966): *Pertubation Theory for Linear Operators*. Springer, New York.

Kendall M. G. (1966): Discrimination and classification. In : *Proc. Symp. Mult. Analysis*. Dayton, Ohio, (Krishnaiah P. R. (ed.), Academic Press, New York, 165-185.

Kendall M.G., Stuart A. (1961): *The Advanced Theory of Statistics*. Charles Grifin, London.

Kettenring R.J. (1971): Canonical analysis of several sets of variables. *Biometrika,* 58, (3), 433-450.

Krishnaiah P.R., Chang T.C. (1971): On the exact distribution of the extreme roots of the Wishart and MANOVA matrix. *J. of Multivariate Anal.,* 1, (1), 108-116.

Kroonenberg P. (1983): *Three-Mode Principal Component Analysis*. DSWO Press, Leiden.

Kruskal J. B. (1956): On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.* ,7, 48-50.

Kruskal J. B., Wish M. (1978): *Multidimensional Scaling*. Sage University Paper, 11, Sage, Beverly Hills.

Lawley D. N. (1956): Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika,* 43, 128-136.

Lawley D. N., Maxwell A. E. (1963): *Factor Analysis as a Statistical Method*. Methuen, London.

Lerman I. C. (1970): *Les Bases de la Classification Automatique*. Gauthier-Villars, Paris.

MacQueen J. B. (1967): Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th),* Berkeley, 1, 281-297, Univ. of Calif. Press, Berkeley.

McQuitty L.L. (1966): Single and multiple classification by reciprocal pairs and rank order type. *Educational Psychology Measurements*. 26, 253-265.

Mehta M.L. (1960): On the statistica1 properties of the leve1 spacing in nuclear spectra. *Nucl. Phys*. 18, 395-419.

Mulaik S. A. (1972): *The Foundation of Factor Analysis*. McGraw Hill, New York.

Nakache J.P. (1973): Influence du codage des données en analyse factorielle des correspondances. Etude d'un exemple pratique médical. *Revue Statist. Appl*., 21, (2).

Nishisato S.(1980): *Analysis of Categorical Data. Dual Scaling and its Application*. Univ. of Toronto Press.

Pagès J.-P., Escoufier Y., Cazes P. (1976): Opérateurs et analyse de tableaux à plus de deux dimensions. *Cahiers du BURO,* ISUP, Paris, 61-89

Pearson K. (1901): On lines and planes of closest fit to systems of points in space. *Phil. Mag.* 2, n°ll, 559-572.

Pillai K.C.S. (1965): On the distribution of the largest root of a matrix in multivariate analysis. *Biometrika,* 52, 405-414.

Rao C.R. (1964): The use and interpretation of principal component analysis in applied research. *Sankhya* serie A, 26, 329-357.

Richardson M., Kuder G. F. (1933): Making a rating scale that measures. *Personnel Journal.*, 12, 71-75.

Robert P., Escoufier Y. (1976): A unifying tool for linear multivariate methods : the Rv coefficient. *Applied Statistics,* 25, (3), 257-265.

Roy S.N. (1939): *p*-Statistics or some generalisations of analysis of variance appropriate to multivariate problems. *Sankhya,* 4, 381-396.

Schönemann P.H. (1968): On two-sided orthogonal procrustes problems. *Psychometrika,* 33, 19-33.

Schönemann P. H., Carroll R. M. (1970): Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35, 245-255.

Shepard R. N. (1974): Representation of structure in similarity data: problems and prospects. *Psychometrika*, 39, (4), 373-421.

Sneath P. H. A. (1957): The Application of computers to taxonomy. *J. General Microbiology*, 17, 201-226.

Sokal R. R., Sneath P. H. A. (1963): *Principles of Numerical Taxonomy*, Freeman and co., San-Francisco.

Spearman C. (1904): General intelligence, objectively determined and measured. *Amer. Journal of Psychology,* 15, 201-293.

Sugiyama T. (1966): On the distribution of the largest latent root and the corresponding latent vector for principal component analysis. *Ann. Math. Statist.* 37, 995-1001.

Sylvester J.J. (1889): *Messenger of Mathematics* (cité par Eckart, Young, 1939). 19, n°42.

Tenenhaus M., Young F.W. (1985): An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 91-119.

Thorndike R.L. (1953): Who belongs in the family. *Psychometrika*, 18, 267-276.

Thurstone L. L. (1947): *Multiple Factor Analysis*. The Univ. of Chicago Press, Chicago.

Tucker L. R. (1958): An inter-battery method of factor analysis. *Psychometrika,* 23, (2).

Tucker L. R. (1964): The extension of factor analysis to three-dimensional matrices. In: *Contribution to Mathematical Psychology*, Harris C. W. (ed.), Univ. of Wisconsin Press, Madison, 109-127.

Tucker L. R. (1966): Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279-311.

Tukey J. W. (1977): *Exploratory Data Analysis*. Addison Wesley, Reading, Mass.

Wakimoto K., Taguri M. (1978): Constellation graphical methods for representing multidimensional data. *Ann. of the Inst. of Statist. Math.*, 30, (1), 97-104.

Ward J.H. (1963): Hierarchical grouping to optimize an objective function. *J. of Amer. Statist. Assoc.*, 58, 236-244.

Wilkinson J. H. (1965): *The algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

Wishart D. (1969): Mode analysis : a generalization of nearest neighbour which reduces chaining effects. *Numerical Taxonomy* (A.J. Cole ed.) 282-311, Academic Press, London.

Wishart J. (1928): The generalized product-moment distribution in samples from a normal multivariate population. *Biometrika*, 20A, 32-43.

Wold H. (1966): Estimation of principal components and related models by iterative least squares, in *Multivariate analysis*, Krishnaiah P.R. (Ed.), Academic Press, New York, pp391-420.