Exploratory 2- and 3-way data analysis and applications

by Sergio Camiz[1]
Departement of Mathematics, University «La Sapienza», Roma.
E-mail: camiz@mat.uniroma1.it

Abstract. This paper tries to consider the framework of scientific investigation, where quantitative methods find their natural field of application. Three different phases are identified, exploratory, confirmatory, and modelling, referring to different mathematical background and having very different rules for a correct use. The main exploratory data analysis methods, classification and factor analysis are introduced. Of factor analyses methods some mathematical background is proposed, considering in particular the two most used: Principal Components and Correspondence Analyses. Three-way data analysis methods are also considered, in particular introducing the mathematical framework of most known multiple table analysis methods. Some examples, taken from different investigations, show the ability of exploratory methods to synthesize the information contained in the considered data tables and help the scientist to identify the main structures contained in the data.

1. Methodological reflections on relationships between data analysis and investigation

1.1 What is a scientific investigation?

Once a question is asked in a scientific environment, in order to give a correct answer, one starts from the present knowledge, if necessary collecting the literature concerning the question and the way answers were given to similar questions, and checks if the answer may be obtained using the collected knowledge. If it is possible, the answer is *deduced* from the current knowledge.

As an alternative, one may start by reviewing the present knowlegde in a scientific environment, by collecting the literature related to the topic of interest and, if some open question is found, one may check if the answer may be given, based on the present knowledge. When this is possible, again the answer is *deduced* from the knowledge.

When this is not the case, one may try to build up an answer. In a scientific environment, this may be done through empyrical experience. To this purpose, one must do some observations and collect some data, that will be further analyzed. In order to be found in this way, the searched answer *must* be *contained* within the data, in the sense that the collected data must be able to give an answer to the asked question.

There is a difference between the first and the second case. In fact, in the first case, an answer may be *deduced* from other knowledge in a purely logical way. Several different levels of logic may be considered. It may be the *associative* logic, that is the logic pervading the dreams, whose connections may be own of each single person, although having some universal rules, as psychoanalysis argues. It is in many cases the specific science framework *paradygma*, that

[1] This paper was carried out during a visit of the author to the Institute of Applied Mathematics of University of Tbilisi in June 1999, in the frame of its cultural agreement with Department of Mathematics “Guido Castelnuovo” of Rome University “La Sapienza”. Both institutions grants are gratefully acknowledged.

condition the construction of most scientific theories, by grounding them to some very general hypotheses. It may be *mathematical logic*.

Mathematical logic has very strict rules. In fact, both mathematics and mathematical logic may be completely formalized, in order to provide at the same time a language for the description of most natural phoenomena, and a tool for *proving* the truth of what is expressed, once that this description is consistent with the *axioms*, some properties of the objects of the investigation that are empyrically proved as true for the considered objects. In other sciences, reasoning is more free and the deduction of scientists are less constrained: it is frequent to find some hypothesis considered reliable only by *analogy* with some other, based on the researcher experience.

As an alternative, the truth of a hypothesis may be *induced* by experimentation. This may be achieved either by setting controlled conditions to repeat observations concerning the phenomenon under study, isolating it from all possible noise and bias, or observing many times the phenomenon in the reality and connecting it with the conditions present during the observation. In both cases, once the conclusions of the experimentation are drawn, one may procede by *inference*, that is by generalizing the results to a broader *population*, that is a larger set of objects, individuals, or situations, that share with the observed ones the same characters that condition the results of the experimentation.

It may be necessary some particular caution, in order to proceed by generalization, in particular considering the causal relationships between phenomena. In fact, one may distinguish at least between two different causalities. The one usually considered in mathematical terms is *strict causality*, that is any kind of functional relationship between cause and effect, typical of most of relations described in classical mechanics. For such an identification of the relations, and for their quantification, phenomena may be isolated in the experimentation, in order to avoid the co-occurrence of other causes that may bias both the effects and their magnitude.

There is another kind of causality, namely *weak causality*. In this case, many events may favour the phenomenon under study or be co-occurrent with it. It is the typical case of observations in human sciences, in etology, and in ecology, since individual behaviour, responses, and choices are various and not always dependent by identifiable causes. In addition, an experimentation is not always possible, whereas surveys may be performed. In this context, the searched relationships cannot be expressed in functional form, since one may better describe *tendencies*, so that relationships assume a *probabilistic* character.

1.2 What is data analysis?

In the context of the experimentation, once a data collection has been done, the scientist has the need to deal with observed data, in order to extract from them the contained information, in order to draw his conclusion. For this reason, he may apply to *data analysis*. Data analysis is the set of mathematical and statistical techniques, based on computer science methods via computers use, useful for a scientist to read, synthesize, and understand the data he collected for an investigation.

Data analysis methods are many, so that it became nowadays a discipline in its own, getting an identity different from those that contribute to its methods. In fact, whereas computers are the primary tools for data analysis, computations are performed through mathematical methods, and statistics is taken into account every time one deals with populations, samples, indexes, etc., one has to recognize that data analysis, although an intersection of all, is in some respect different from all of them. In addition, since data analysis must always deal with existing empyrical data and their explanation in the frame of an existing scientific discipline, one may not forget that its use cannot be considered outside such frame.

Let us consider an investigation in any field. The relationship between the scientist and its field of investigation deals through *observations*, i.e. a series of elements noted during his observation. A set of observations is thus a set of such series of elements, that form the basis for the knowledge that the scientist aims at building. We shall call each unit of observation a *unit,* other synonymous are *statistical unit*, *individuals*, *observations*. To units, *characters* correspond, the nature of which is as various as the science fields are, since it may range from elements of reality of every possible kind, movies, pictures, or texts to coded characters, frequencies, or measures. In order to deal with such a set of *data*, it is necessary that to each unit the same set of characters is identified with the same meaning, such as weight, colour of eyes, temperature, number of bars in the town, number of cars produced, etc., and obtained with the same technique. If necessary, characters may be distinguished according to their nature. In fact, particular interest may be devoted to those data suitable of a mathematical treatment, most of which are those that can be expressed *numerically*, although recently the treatment of *textual data* has become a current practice (Lebart and Salem, 1988; 1994).

The content of the data depends on both the scientific field of reference and the way the scientist performs his observations, something strictly related to his previous *culture*. So, data are the elements of the observation taken from reality by the observer, based on his culture, whose treatment is suitable to the growth of scientific knowledge, thus to culture. One may say that at the very beginning of a study concerning a phenomenon very little of it, or nothing at all, is known, whereas at the very end, most of it, or maybe everything, should be understood, in order to take advantage of the acquired knowledge for any possible use.

Since the acquired knowledge is supposed to become a known truth, it is necessary to take special cautions, in order to proceed from the beginning to the end through a methodological path, allowing to state that the results obtained by the application of these methods to the observed data may be considered as true. A good reference may be to consider the investigation path subdivided into *three steps*, corresponding to three different different phases of the formation of the knowledge. To each phase correspond different aims and, as a consequence, a different approach is necessary and different methods and tools are supposed to be used.

It may appear as an author's obsession to stress the difference between the different steps of a research process and to suggest different approaches for each step. In fact, only the knowledge of such steps may indicate to the researcher the caveats to notice, particularly in data analysis, since different tools may be specific to a particular step, and they may not be used in the others without care. The steps, as proposed in Camiz (1993) may be defined as follows:

- *Exploratory phase*, in which one defines a frame of reference for his/her work, defines the aims of the investigation, and begins a data collection. In this phase the analysis of the data aims at searching structures and relationships, in order to formulate some hypotheses. Data are thus submitted to *exploratory data analysis*, to recover as much synthesized information as possible, in order to reveal any existing data structure and, in particular, to see whether or not the research aims are reachable on the basis of the collected data.
- *Knowledge acquisition phase*, in which the information obtained by the previous analysis is used to improve both experimental design and sampling, in order to be able to clarify relationships among measures and among observations. In this phase, *confirmatory analyses* are applied to the data, in order to figure out the assumed relationships, through tests of hypothesis. Sampling requires particular care to allow statistical inference of results to the reference population and statistical techniques are used to reveal the significance of relationships detected.
- *Modelling phase*, in which a theoretical and formalized (or mathematical) description of an observed phenomenon is given, able of giving evidence on the relationships observed. In this phase, data must be used to test the effectiveness of the *model*, so that the most different data sets are used to evaluate its application range. *Simulations* are an outcome, either to show the effectiveness of the model to depict the studied phenomenon, or to derive further information from *simulated data*, i.e. data produced by the model and not from direct observation. Another outcome is *forecasting*, i.e. the ability of the model to predict the behaviour and the evolution of the phenomenon itself.

The advantage to proceed to a modelling after a better knowledge of the data structure is the ability to tailor the model in order to fit the data in the best way. In this sense, one may roughly consider it a *marxist* approach to modeling, i.e. to derive the model from the observation, in opposition to the *hegelian* one, i.e. to decide a model and impose it to the reality. This is specially dangerous in human sciences, where the decision to apply a non-suitable model, e.g. an economic or a political one, to a population may result in terrific consequences if it does not suit sufficiently well the reality. The sad memories of either monetary policy in South-America or of socialist system in USSR may be kept as a memento for this argument.

1.3 Exploratory phase

In the exploratory phase, the scientist defines a framework for the investigation and some aims; consequently, he searches the literature and collects some data, either from literature or from observation. Data are then inspected in order to understand their content, in particular aiming at identifying some structures, able to give some general idea of the key-elements of the phoenomenon. It is the identification of structures that leads the scientist to the formulation of some hypotheses concerning the relationships among the elements composing the phoenomenon.

In order to inspect the data, rather than reading them all, a cumbersome work very difficult to carry on successfully, one may apply to some instruments able at revealing the information contained in the data in a synthesized way. The instruments should be able to reorganize the data, in order to reveal the structures that may exist. Such structures represent a way to synthesize the information contained in the data, or some kind of structure should be revealed by the synthesis that the instruments allow to organize. In fact, exploratory data analysis aims at describing the data information content through a *strong synthesis* of the data by describing the resulting

structures. In particular, this leads to describe which relationships exist among the considered characters and which resemblance can be detected among the considered units.

The synthesis is performed through a *synthetical* representation of the *important* information in graphical form. The hypothesis is that the important information contained in the data may be extracted via some mathematical techniques, based on *association measures*. Two main graphical approaches are generally used, based on two different mathematical models:

- the *objects* are represented as *clouds of points* on a plane, forming scatter diagrams; plane subsets are also used to show regions or groups of points;
- the *objects* are represented as *nodes* of a graph, whose *edges* represent *relationships*; particular graphs used are *dendrograms* and *pyramids*.

Both models are useful to reveal the structures contained in the data. In the first case, the *units* are points in an *affine space* and characters are generators of the vector space that sustains the affine space. In the second case, objects and relationships are nodes and arches of a graph, so that most of the information is easily detected. Often both models are used at the same time, by representing the graph nodes as points on the graphical plane whose coordinates depend on the scattering relative to the represented characters. In this case, edges may have different thickness, showing the different intensity of the relationship.

An exploratory data analysis attempts to synthesize the information contained into the data, by describing the existing structures and aiming at building a global structure hypothesis. This may be achieved through the identification of

- *factors*, the characters that best fit the objects diversity; they may be merely descriptive but they are useful to represent the objects position in their respect; in many cases they are also interpreted as the causes of the diversity.
- *classes*; building homogeneous classes of objects allows to consider the obtained partition as a structure of the considered set.

Summarizing, an exploratory study pathway may be outlined in the following way:

- an argument of investigation is decided;
- references in the literature are searched;
- some data are gathered:
    - by collecting them from the literature;
    - by performing observations:
        - concentrated to the situations of interest,
        - aiming at identifying interesting objects, and
        - their relations with possible causes or other elements.
- data are submitted to exploratory data analysis tools:
    - descriptive statistics, to check for consistency of data, existance of outliers, etc.
    - factor analyses, to reveal the main sources of information, to be considered as the best references for the representation of data;
    - hierarchical cluster analysis, to build a complete set of encapsulated partitions, able to reveal the structure of the objects set; this allows to search later the partitions having optimal qualities;
- strong syntheses are searched on
    - *factors* describing objects diversity,
    - *classes* of homogeneous objects.
- the results are translated in the reference framework.

An exploratory analysis has some limits, partly depending on the used tools that do not consider any statistical paradygm, but are limited to the results of mathematical procedures leading to optimal data representations. In fact, since no particular attention is usually paid in this phase to data collection, the obtained synthesis is only a synthetical description of considered data. Thus, neither it may be inferred to reference populations nor it concerns other (similar) data. It is possible that the proved capacity to help the scientist in formulating hypotheses depends on the *natural* way used for performing the synthesis. Instead, the hypotheses formulated based on this synthesis ought to be validated through other suitable methods.

It must be pointed out that in exploratory analysis data are studied without attributing to them an interpretative *a priori* model (Benzécri, 1973-82): a warrant of the possible objectivity of resulting information.

1.4 Confirmatory phase

In this phase, the scientist is concerned with the need of establishing which hypotheses he formulated in the previous phase are acceptable and which are not. It is then a phase where *knowledge* is acquired and *decisions* are taken. For this reason, in order to achieve these tasks, it is necessary to ground most of the investigation on statistical methods, the only able to allow to test the formulated hypotheses and decide whether they are acceptable or should be rejected.

For this task, the scientist must consider the experimental design and sampling, in order to get data suitable of statistical treatment. With these data, he may build some relationships and test the hypotheses concerning these relationships; as a consequence, the results of the investigation may be inferred to a reference population, with minor risks of being in error.

Questions to be answered in order to achieve this phase are thus if stated hypotheses are acceptable, if it is possible the statistical inference of the results, and, eventually, one may wonder if it is possible to gather several hypotheses in a theoretical structure, something leading then to the modeling phase.

A confirmatory data analysis is thus the phase where one tries to build some knowledge based on empyrical observations. Following the exploratory step, one then tries to validate the hypothesized elements. As was said before, the main tool for this purpose is the *hypothesis testing* and the consequent *decision*, based on a precise statistical procedure. The procedure may be outlined as follows (Mood *et al.*, 1974, Droesbeke, 1997):

- A probability threshold α is fixed, normally 5% or 1%;
- a *null hypothesis* H_0 is set and its falsification H_1 = not H_0;
- a decision rule is set, leading to:
  - reject H_0 and consequently accept H_1, or
  - not reject H_0;
- supposing H_0 true, a random sampling is performed;
- the probability p is calculated, that the results obtained using the random sample in agreement with H_0 are random;
- if $p > \alpha$, the null hypothesis H_0 is rejected, not reject otherwise.

It must be pointed out that the procedure that currently is supposed to lead to *accept* the stated hypothesis H_0, in fact limits to *not reject* it. There is actually a *risk of error* in the procedure, corresponding to the fixed threshold, since it may happen to not reject some results, because $p < \alpha$, even if they are actually random albeit having a low probability to happen (5% or 1%). It is not useful to reduce this risk by choosing a lower α, since one may fall in the opposite error, that is to reject a *true* hypothesis H_0 because $p > \alpha$.

In order to perform a confirmatory analysis, some rules should be observed. In particular,
- the answer to the asked question should be contained in the data;
- the data used for the confirmatory analysis must be different from those that suggested the hypothesis to be tested in the exploratory phase;
- data must be suitable for the type of procedure to be used;
- if the results must be inferred to a population, units must be randomly sampled and representative of the population of reference.

There are some advantages to perform an exploratory analysis before a confirmatory phase, in addition to the fact that some hypotheses were formulated in the previous phase. In fact, the ability to hypothesize a structure of the observed phenomenon allows a better organization of the following study, in particulare allowing the following actions:
- fix errors in data input;
- withdraw outliers, that is units whose behaviour is most different from all others, or study the reasons of their difference;
- withdraw elements without relationship or influence with the phenomenon;
- check if the study aims are reachable on the basis of present data;
- modify the choice of the characters to observe, according to the factors extracted from the original data set, once both the meaning and the utility of the factors were checked;
- change the sampling according to its representativity of the population of reference;
- suggest the elements to take into account in the experimental design.

In addition to these considerations, it is interesting to consider the ability of an exploratory analysis to become a framework where the subsequent results may be included, confirming what was previously hypothesized.

Summarizing, a confirmatory study pathway may be outlined in the following way:
- The scientist has a clear aim, concerning a population;
- he designs an experiment, in order to
  - ensure that his aim may be achieved through the experiment,
  - not ignore important aspects of the phenomenon,
  - take into account every variation of factors in the sampling,
  - check compatibility among aims, data and data analysis methods;
- he performs either a random or systematic sampling, observing the rules necessary for the inference of the results;
- he submits the collected data to inferential statistical methods, such as:
  - parameter estimation,
  - regression,
  - analysis of variance,
  - discriminant analysis, etc.;
- through the tests of hypotheses he searches a confirmation concerning:

- point or interval parameters estimations,
- relations and influences among factors and characters, homogeneous classes identification, etc.

1.5 Modelling

Once that several hypotheses were confirmed, the scientist may try to tie them in a theoretical structure. The structure may be formulated in a *mathematical model*, usually a set of equations that describe the relationships between the most typical characters that appear in the phenomenon. It is clear that during the formulation of the model, attention must be paid to fit all kind of data collected, no matter how. For this reason, the previous steps are helpful, since enough knowledge of elementary relationships was built already. In any case, the model is usually continuously adjusted with a feedback process, based on tests performed on available data.

With a consistent mathematical model the phenomenon may be simulated, at least concerning its main characters. In this way one can simulate the data as they were resulting from the observation of the phenomenon in reality. It is clear that simulated data should satisfy both exploratory and confirmatory analysis results. This may be a good way for both *calibration* and *validation* of the model.

Once the model is built, well calibrated and validated, the investigation is achieved, at least as far as it concerns the structures and the relationships that were hypothesized. Nothing more may be added, apart to refine some aspects, that were ignored, taking into account some more details, previously considered as noise. In this conditions, one may repeat the three described steps, if necessary reduced in their scale, as in a fractal way.

1.6 Scientist and data analyst interaction

An investigation is first of all the work of a scientist. On the opposite, the treatment of data may be the work of a data analyst, so that an iteraction among them is necessary. This may cause problems, due to the different formation of the two characters. In addition, the treatment of data means an interference of the data analyst, since data are typically the domain of the scientist, who decides:

- which is the population he is concerned with;
- which characters to consider;
- how data must be collected;

but they belong to the data analyst domain too, since he must check:

- if the sample is representative;
- if the sample was randomly chosen;
- if characters were observed in a uniform way.

This kind of double control may be difficult to handle but it is necessary, in order to ensure the quality of the results on both points of view. It must be added that the different steps need different kind of sampling.

In the exploratory analyses, sampling rules are very weak or non-existent, depending on its aims, but nevertheless some rules should be observed:

- characters must be uniformly observed, otherwise one may find in the analysis results only the difference in observation methods;
- it is not strictly necessary a random representative sampling, but:
- if the sample is representative of a population, one may think to extend to the population the found results, although *this is not a statistical inference*;

In confirmatory analysis most attention must be paid to the sample quality. This is the reason why experimental or survey designs are so used, since they drive to optimize data collections in relation with the searched results. This would be very useful for exploratory analysis too, if one considers the inference of results to a population a mayor aim (but still, *this is not statistical inference*). Nevertheless, it must be pointed out that is not advisable to use the same sampling for both exploratory and confirmatory steps, since the test of hypotheses using the same data, on which their formulation was based, becomes trivial and unreliable. If the sample is suitable for a confirmatory analysis, it is thus advisable to partition the sample in two groups and use one group of units for the exploratory analyses and the other for the confirmatory ones.

Some observations may be useful to understand the need of a data analyst near the scientist. In fact, the knowledge of data analysis, or at least of data analysis language, is necessary to the scientist to interact with both data analysis and the data analyst, but the data analysis domain is nowadays very broad and a specialist is suitable to cover it. In addition, new data analysis methods may be developed to better fit the new problems that raise during an investigation. On the other side, the data analyst role is very delicate, since he interacts deeply with the scientist in many intimate aspects of his investigation:
- when he defines the aims of his study, to clarify them and forecast (and build if necessary) the methods to be used;
- in experimental or survey design, to optimize both survey methods (the items, the questions, etc.) and the sampling according to both aims and methods (and reduce costs);
- during data treatment and results reading, to give the scientist a reliable result.

In the current practice, the data analyst is involved only at data treatment time. His role is thus strongly limited, since usually
- items were already chosen;
- sampling was already done;
- data were already collected;
- only data may be treated in the best way according to their nature, to derive most of information.

This strongly limits both methods that may be used and the possible aims of the study (with considerable loss of money). This is frustrating for the data analyst, that is conscious that his possible positive contribution is not acknowledged, limited as it is to a technical contribution *a posteriori*, when a major contribution is no longer possible.

It may be interesting to compare the current interaction with an ideal role of data analyst, that may be near the scientist along all his investigation lasting, giving his advices in many aspects that the scientist does not necessarily control. Thus, a common interface is necessary, since the scientist, proposing his problems, should understand the methodological limits imposed by the mathematical logic underlying both data analysis and statistics methods, and, on the other side,

the data analyst should translate into his language scientist's problems, find a solution, and translate it into scientist's language.

An ideal interaction between the two parties may be thus described as follows:
- the scientist decides to work on a problem and submits it to the data analyst with a possible investigation procedure;
- the data analyst translates the problem into his language and shows his needs, concerning the characters to consider and the sample features;
- the scientist performs his experimentation or his survey;
- the data analyst treats the data, gets the results, and shows them to the scientist: they translate them together in the other language;
- the scientist draws his conclusions.

It should be clear that the data analyst role is subsidiary in all those methodological aspects that the scientist is not supposed to know so deeply as it is required for a well-grounded investigation.

2. 2-way exploratory data analysis

2.1 Exploratory data analysis

"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone - as the first step." (Tukey, 1977)

"Exploratory data analysis is a detective work." (Diggle *et al.*, 1994).

The exploratory data analysis is very synthetically described by these two sentences. The detective work may be actually considered as the beginning of the knowledge acquisition, since a quest for elements of truth is necessarily the first moment for an investigation.

The aim to describe the contents of the data through a strong synthesis, leads naturally to the graphical form of the synthetical representation of the information to be considered important. The two main graphical approaches generally used are based on different mathematical models: the *scatter diagram*, where objects are plotted on the graphical plane, and the *graph*, where objects may be nodes and their relationships are edges. They may be combined in a composite representation, but most often they remain separated, even when they concern the same sets.

Considering the data, they are represented through a *data table*, i.e. a matrix X whose n rows represent statistical *units*, also called *individuals* or *observations*, and the p columns represent *characters*, also called *variables*, relative to the considered units. So, in each row all characters detected in respect to one unit are collected and in each column all the information concerning one character in all considered units is collected. As a result, all columns are sorted according to the same sequence of units and all rows contain the information relative to characters in the same order. The following matrix is a *data table*, whose element x_{ij} is the *value* of character j on the unit i:

$$X = \begin{pmatrix} x_{11} & x_{12} & .... & x_{1p} \\ x_{21} & x_{22} & .... & x_{2p} \\ & ..... & & \\ x_{n1} & x_{n2} & .... & x_{np} \end{pmatrix}$$

So, X is an $n \times p$ table, concerning n units and p characters. The characters may be simply a sign of *presence/absence* of something (e.g. the ownship of the driving licence), a set of *different modes* relative to the characters (e.g. the colour of the eyes, the citizenship, etc.), a set of *scale* values, ordered modes (e.g. the agreement to a sentence: totally in favour, strongly in favour, in favour, neutral, against, strongly against, totally against; a school vote, etc.), a *frequency* (e.g. number of supplies of a certain category in a town, etc.), or a *measure* (e.g. a temperature, a height, a weight, etc.).

Example: Some data concerning Italian regions in 1978

Table 1 - The data table of Italian regional statistics in 1978. Data are in thousands.

Region	*Population*	*Rooms*	*Primary*	*Secondary*	*Tertiary*	*Income*	*Weddings*	*Births*
Piemonte e Valle d'Aosta	4465.500	6081.300	212	945	723	1721	24.620	50.101
Lombardia	8496.683	9861.440	164	1897	1461	1277	50.166	105.810
Trentino Alto Adige	847.226	1080.889	52	100	167	987	5.332	10.936
Veneto	4135.960	5258.836	206	703	697	1053	27.897	52.820
Friuli Venezia Giulia	1245.143	1722.637	39	178	245	1134	6.813	1.227
Liguria	1868.065	3034.254	52	211	394	1292	9.535	15.943
Emilia Romagna	3852.833	5107.119	278	642	740	1245	21.139	38.352
Toscana	3502.541	4850.555	137	589	636	1126	20.184	36.149
Umbria	773.195	960.612	45	117	141	966	4.951	9.246
Marche	1350.974	1805.806	118	244	230	989	8.553	16.826
Lazio	4764.149	5440.567	172	423	1033	1109	29.862	65.626
Abruzzo	1120.770	1473.219	100	120	198	876	7.877	15.034
Molise	299.775	397.862	51	26	41	706	2.036	3.875
Campania	4984.677	4507.638	394	456	794	816	39.921	96.683
Puglia	3498.932	3381.271	416	315	523	893	27.163	68.134
Basilicata	560.057	559.093	83	55	71	754	4.022	8.812
Calabria	1861.537	1870.928	179	148	256	710	13.550	32.409
Sicilia	4575.421	4802.599	364	370	661	842	31.908	77.726
Sardegna	1441.284	1637.154	87	127	236	954	11.374	25.799

Let us consider a small 19×8 data table, taken from the Annuario Statistico Italiano (Istat, 1978, quoted by Costa *et al.*, 1981), composed by eight characters, namely *Region population* (*Population*), *Number of rooms* (*Rooms*), *Occupied in primary sector*s (*Primary*), *Occupied in*

*secondary sector*s (*Secondary*), *Occupied in tertiary sector* (*Tertiary*), *Pro-capite Income* (*Income*), *Number of weddings* (*Weddings*), and *Number of births* (*Births*), registered in the 19 Italian regions in 1978. The data are shown in Table 1. It is evident that a complete understanding of the content of the table would require a long and painful reading of all figures even in this table, nowadays considered very small. More difficult would be to state a synthesis of the table content, showing the agreement or disagreement or dependence among characters, and the similarity among regions.

It is possible to synthesize the content of the data table by a graphical representation where characters are represented as vectors, whose agreement in behaviour is represented by the small angle between them and units are represented as point whose proximity represent their similarity, in relation to the considered characters.

Once identified two main *factors* through *Principal Components Analysis* (Benzécri *et al.*, 1973-82; Lebart *et al.*, 1977, 1979, 1984, 1995), a tool specific for this kind of synthesis, that will be outlined in the following, both characters and units may be represented on a plane having these factors as coordinate axes. In Fig. 1 the characters are represented and it is easy to see that *Population* corresponds nearly to the first factor, and all characters but two very close to it and to each other: this means that a great part of the considered characters is dependent on the *population*, all sharing a *size* effect. Instead *Pro-capite income* and *occupied in primary sector* are opposed in respect to the second factor. This opposition may be interpreted as the fact that most rich regions are poor in agriculture. One may note in addition that *weddings* and *births* are opposed to the richness, whereas *occupied in secondary* and *tertiary sectors* and *number of rooms* are on richness side.

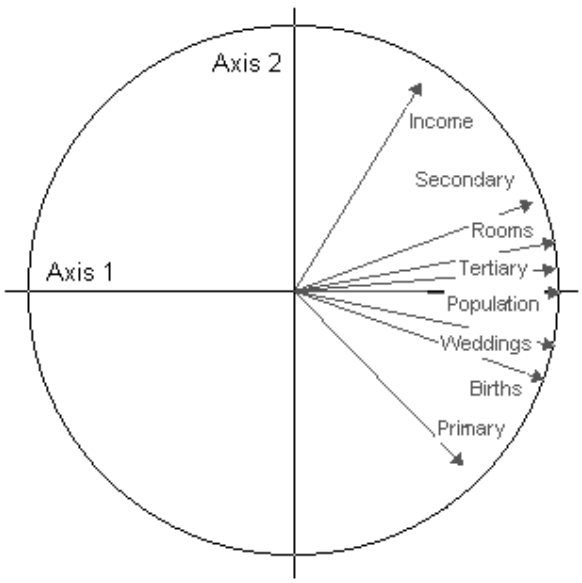

Figure 1 - Representation of characters as vectors on the plane spanned by the first two factors of Principal Components Analysis.

In analogous way, in Figure 2 is possible to see the position of regions on the plane spanned by these two factors. It is easy to note that on the right side of first axis are found the regions with highest population and on the left are found those with smaller population. Along the second

axis, *Piemonte e Valle d'Aosta*, *Lombardia*, and *Liguria*, on the upper side of the figure, are opposed to *Puglia*, *Sicilia*, and *Campania*. It is well known that the first were regions where the primary production was heavily reduced in favour of the secondary and, more recently, tertiary activities, that contributed to the raise of the income. The other three regions are those where primary productions used to have still an important share in regional economy.

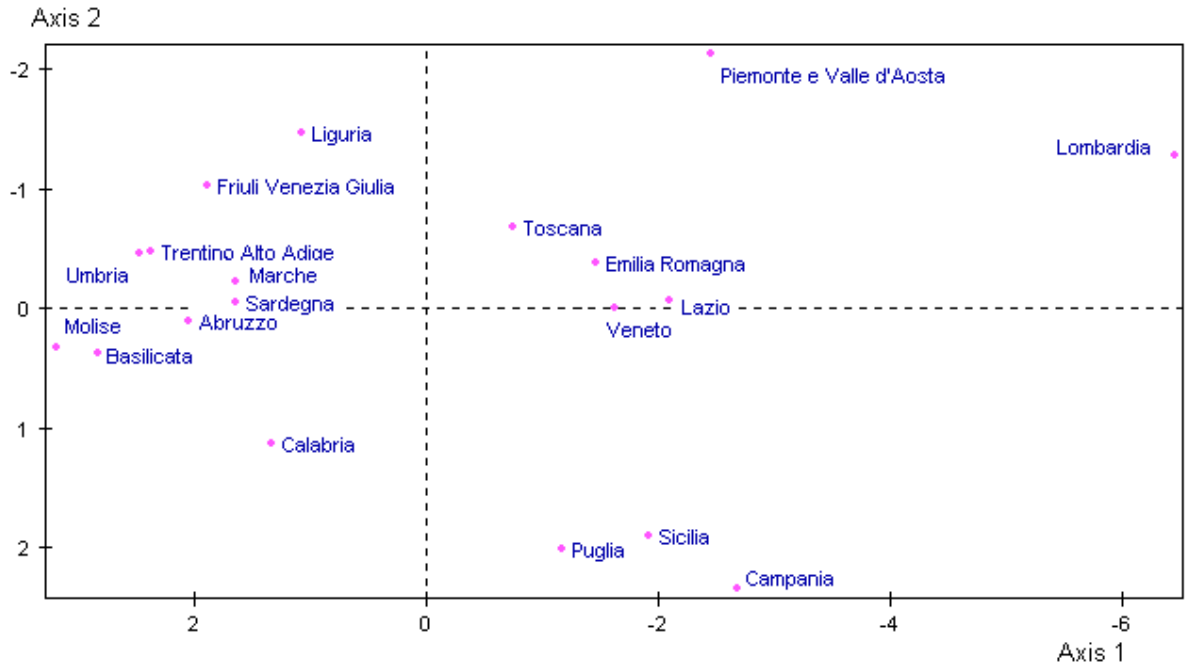

Figure 2 - Representation of regions as points on the plane spanned by the first two factors of Principal Components Analysis.

It is most evident that such a representation get very easy the understanding of the data table content. Considering that one may measure the amount of information contained in these graphics, corresponding to 93% of the total information, the power of this method is clearly outlined.

In order to go through methods of exploratory analysis, it is interesting to consider, beside the original data matrix, the so-called *association matrices*: these are squared symmetric matrices that contain information concerning mutual relationships among either units or characters. Their role is very important, since most of exploratory data analysis procedures are based on their use. Two kind of association matrices exist:

- *dissimilarity* matrices, those having the *minimum* (0) on the diagonal, so that each object is least dissimilar (0) from itself: formally, a *dissimilarity* among objects of a set E is an application

$$\begin{array}{rccc} d\colon & E \times E & \longrightarrow & \mathbb{R}^{+} \\ & (i,j) & \mapsto & d(i,j) \end{array}$$

such that

D1) $d(i,j) = d(j,i)$ $\quad \forall\, (i,j) \in E \times E$

D2) $d(i,i) = 0$ $\quad \forall\, i \in E$;

a diversity is a *semi-distance*, if holds

D3) $d(i,j) = 0 \implies i = j \qquad \forall (i,j) \in E \times E$

and is a *distance* if also
D4) $d(i,j) \leq d(i,k) + d(k,j) \qquad \forall\, i, j, k \in E$
holds.

A couple (E,d) is either a *semi-metric* or a *metric space* if d is either a semi-distance or a distance defined on E, respectively.

- *similarity* matrices those having the *maximum* (either 1 or any value) on the diagonal, since each object is most similar to itself: formally, a *similarity* among objects of a set E is an application

$$\begin{array}{ccc} s: E \times E & \longrightarrow & \mathbb{R}^+ \\ (i,j) & \mapsto & s(i,j) \end{array}$$

Such that
S1) $s(i,j) = s(j,i) \qquad \forall (i,j) \in E \times E$
S2) $s(i,i) = s(j,j) = s_{max} > s(i,j), \qquad \forall (i,j) \in E \times E$

When $E = V$ is a vector space, a similarity may be or may not be a scalar product. In a vector space V, a *scalar product* is a symmetric bilinear application

$$\begin{array}{ccc} s: V \times V & \longrightarrow & \mathbb{R} \\ (i,j) & \mapsto & \langle i,j \rangle \end{array}$$

such that
PS1) $\langle i,j \rangle = \langle j,i \rangle \qquad \forall\, i,j \in V$
PS2) $\langle ai+bj, k \rangle = a \langle i,k \rangle + b \langle j,k \rangle \quad \forall\, i,j,k \in \mathrm{V}, \forall\, a,b \in \mathbb{R}$

A scalar product is *non-degenerate* if
PS3) $\langle i,j \rangle = 0 \; \forall\, j \in V \implies i = 0$
holds, and
positive semi-definite (*psd*) if
PS4) $\langle i,i \rangle \geq 0 \qquad \forall\, i \in V$
holds. A couple (V, s) where V is a vector space and s is a scalar product is called a *pre-Hilbert space.*

Both kind of association matrices may be used for the analyses. It may be useful to use in particular the first kind for the units and the second for the characters due to some asymmetry in the role of units and characters in the analyses, although both kind of matrices may be transformed to the other without any difficulty. In particular, given a similarity s_{ij}, a dissimilarity results by setting

$$d_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$$

Here, if s_{ij} is a *psd* scalar product, d_{ij} is called *Euclidean distance*. Reciprocally, given a dissimilarity d_{ij}, a similarity is obtained by setting:

$$s_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{j}^2 + d_{..}\right)$$

here, if d_{ij} is an Euclidean distance, s_{ij} is a *psd* scalar product. The matrix $S = (s)_{ij}$ is called *Torgerson*'s (1958) matrix. A metric space whose distance is Euclidean is called *Euclidean space*.

According to the data table values, the important information that may be detected, concerns primarily the *diversity* of the considered objects, constituting the *structure* of the data table. The further explanation of the detected diversity concerns the relationships among characters, that may be sinthesized through the identification of the factors that induce the diversity. In this way, the different values of the factors relative to the different units may be used for their *ordination*. Instead, the different resemblance among the considered units leads to their *classification*, that is to partition the set of objects in homogeneous groups.

In order to represent both ordination and classification, graphical tools is used, since the synthesis of information contained in an image is much stronger than the information given by the reading of a table of numbers or a descriptive text (Bertin, 1977). In fact, all texts usually written in the exploratory analysis frame are grounded on the explanation of the graphical representations provided by the techniques used for the synthesis; nevertheless the numerical results are useful for deeper and more precise insights of the identified structures.

As already said, the graphical representations used are usually two:
- *scattering diagrams* for the search of factors: points are scattered on a plane, where straight lines represent directions of interest and points position their measure along lines.
- *graphs* for the search of classes: points (*nodes*) are tied by *arches*, representing relationships. Particular graphs are considered: dendrograms, pyramids, additive trees.

The two types of representation may be superposed, showing directly points ties on the representation plane; unfortunately, in this case it may happen to lose the accuracy of quantitative information, tied to both positions and lengths.

In the graphical representation the object are associated to the graphical elements, so that
- in scattering diagrams, units are represented as points and characters as vectors;
- in graphs, objects are nodes and relationships between two nodes are edges.

To this representation, graphical *values* may be associated, so that, at a glance, the main features may be identified:
- points proximity is usually associated to the similarity of corresponding objects;
- a small angle between vectors corresponds to an agreement between characters;
- two orthogonal vectors usually indicate that the corresponding characters are mutually independent;
- if two nodes are connected, it means that some relationship exists between the corresponding objects;
- the length of the edges indicates the weakness of the relationship between the objects corresponding to the tied nodes.

Such values are usually based on the Euclidean structure of a plane of graphical representation. It must be reminded that an *Euclidean space E* may be seen as
- an *affine space*, i.e. a triple (E, V, a), where E is a set of *points*, V a vector space and a an application

$$\begin{aligned} a:\ E \times V &\longrightarrow E \\ (P,v) &\mapsto P + v \end{aligned}$$

such that

SA1) $(P + v) + w = P + (v + w)$ $\quad \forall P \in E, \forall v,w \in V$

SA2) $(P + 0) = P$ $\quad \forall P \in E$

SA3) $\forall P,Q \in E$ there exists one and only one $v \in$ such that $Q = P + v$

If $Q = P + v$, one may write $v = Q - P$. In an affine space, terms as *alignment* and *parallelism* are meaningful.

- E is a *metric space*, since a *distance* is defined;
- V is a *pre-Hilbert space*, since a *psd scalar product* is defined, thus a (*Euclidean*) *norm*, given by the scalar product of a vector with itself;
- the defined distance is *Euclidean*, meaning that the distance between two points is the *Euclidean norm* of the difference between the corresponding vectors:

$$d^2(P,Q) = \|Q - P\| = \langle Q - P, Q - P \rangle$$

The graphical representation on a Euclidean plane suggests to consider a *Euclidean model* for the data. Nevertheless, if the considered characters are vectors spanning a vector space V and units are point of an affine space E based on V, in order to get E Euclidean an Euclidean metric is necessary. This may be obtained by defining either a distance between units or a scalar product between characters. In general this goal is achieved once a *measure of association* is given either to units or to characters, provided that this measure has suitable properties. This will be used to search for an orthonormal basis of the vector space, having optimal characteristics (e.g. it optimizes the quality of synthesized representation).

2.1 Cluster analysis

The aim of cluster analysis is to classify objects, according to their reciprocal similarity or dissimilarity, by creating a *partition* of the objects set. As a consequence, one may assume that each class may become an *idealized object*, just as we do when we give a specific name to a class of objects, to identify them among all others, based on specific characters. In this way one may fix with specific names classes of objects that share a common identity and may be treated in the same way, at least in some specific frame.

From the data analysis point of view, neither a universal nor an optimal algorithm exist for this purpose. This in fact depends too heavily from the specific configuration of the objects and from the specific aims of the classification. In addition, is is rather difficult to decide the number of groups composing the partition. The goal would be what is called a *natural partition*, meaning a partition corresponding to the data actual structure. Unfortunately, there is no possibility to describe what it is meant by natural partition in mathematical terms, so that there is no *best* method, but rather many good ones, that may suit with different capacity the particular set under study.

As an alternative, the most used method in exploratory analysis is the *hierarchical classification* (Anderberg, 1973), a technique that does not provide a single partition, but rather a *dendrogram* reflecting the whole set of association relationships among objects. A dendrogram is a particular graph, where objects represent the extreme nodes, called *leaves*, all connected to a node called *root*, through *branches*, that is edges connecting the intermediate nodes, those connected with two nodes closer to the leaves and a node closer to the root. The dendrogram is thus like a taxonomic structure, and a partition may be obtained in any number of groups by cutting the dendrogram branches at some level where they are exactly that number.

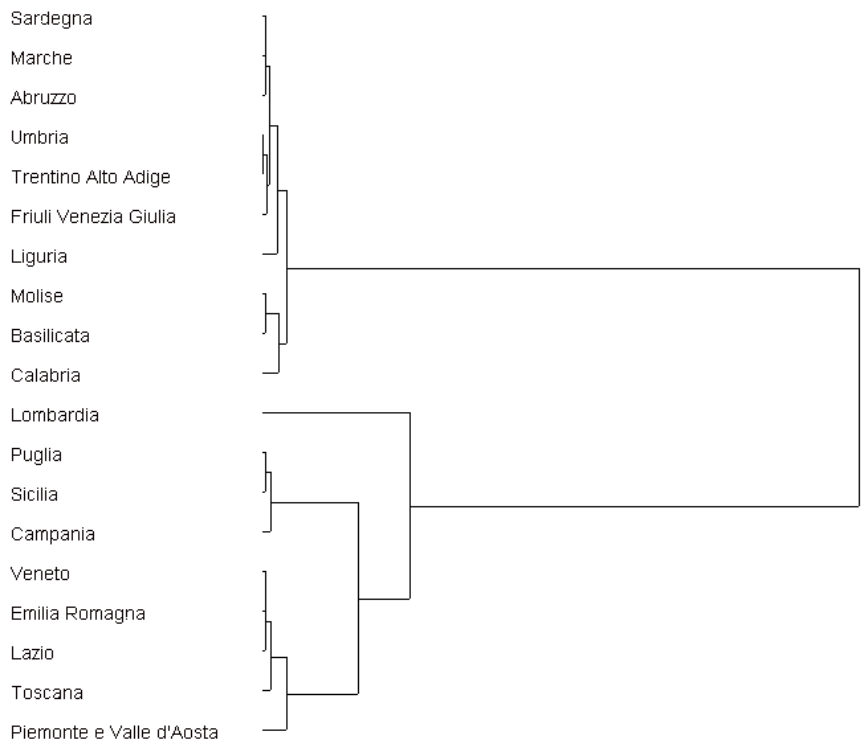

Figure 3 - The dendrogram of hierarchical classification built on the 19 Italian regions.

A dendrogram may be built in several ways, since it is sufficient to define a hierarchy to obtain a corresponding dendrogram. A typical construction is the *ascendent hierarchy*, built step by step, once that an association matrix among objects is defined. Iteratively, a couple of objects is chosen to merge into a single new object, and the association between the new object and all other objects is recalculated. The first aim is usually reached by using a choice function to be optimized, whereas the recomputation is done according to the same function. At the beginning, each object is considered a group, then, at each step:

- the couple of groups optimizing the choice function is selected,
- the two groups are merged in a new one,
- the association of the new group to all others is computed.

This procedure is repeated until only one group remains, gathering all objects.

Example: the classification of the Italian regions

As an example, let go back to the 19 Italian regions and, as association matrix, their Euclidean distances on the plane spanned by the first two axes of *PCA*: a hierarchy may be built by merging at each step the two groups that minimize the increase of within-groups variance (Ward, 1963). The resulting dendrogram is shown in Figure 3.

The simple observation of the dendrogram, considering the length of the longest branches, suggests to cut the hierarchy at the level of four branches, giving a partition in four groups, namely *Lombardia* alone, situated far away on the positive side of both *PCA* axes, a group of five regions: *Veneto, Emilia Romagna, Lazio, Toscana,* situated along the positive side of first axis, *Piemonte e Valle d'Aosta* more set on the positive side of second axis, a third group of *Puglia,*

Sicilia, and *Campania,* on the negative side of second axis with a strong agricultural tradition, and finally all other regions, on the negative side of first axis, poorer in population and all other characters. By no means, other partitions are possible and other hierarchical methods may be used, giving different partitions.

2.2 Exploratory factor analyses

The exploratory factor analyses aim at searching the factors that may explain the diversity of the objects involved. Their identification is based on the quest for subspaces of the representation space where to project the objects, with the condition that the projection optimizes some conditions fixed in advance, in order to suit some particular aspect of interest, e.g. maximize points projections on the subspace. It is interesting to note that the conditions may depend on the known structure of the data table, so that the analysis itself may be tailored accordingly. In general, the conditions may be expressed by some matrices, and the solution is usually given by their diagonalization through eigenanalysis. As a result, the extracted eigenvectors are interpreted as the searched *factors*, whereas the eigenvalues indicate the relative *importance* of the factors.

The best known exploratory factor analysis is by no means *Principal Components Analysis* (*PCA*, Benzécri *et al.*, 1973-82; Lebart *et al.*, 1977, 1979, 1984, 1995; Bry, 1994). It may be used when characters are measures and it aims at reducing the dimension of the data by minimizing the consequent loss of the information contained in the data table. It will be shown that *PCA* is the eigenanalysis of correlation matrix of characters and that resulting eigenvectors, interpreted as factors, identify the *principal components*, i.e. the orthogonal straight lines along which the scattering of the units is maximum. The amount of scattering along each axis, evaluated by the sum of square distances of the projection of the units from the origin, is given by the corresponding eigenvalue; this is a measure of the factor importance, and of reconstruction of original data. From this point of view, the most important contributions come from factors corresponding to highest eigenvalues, so that, in general, only the first few are of real interest.

PCA is a very useful exploratory technique, since it allows the synthesis of data table, through the representation of both characters and units in very reduced dimensional spaces; through this representation, it allows the identification of principal factors, thet are linear combinations of the observed characters, and easens their interpretation, based on the correlation between characters and factors. The graphical representation allows a *global* view of objects, so that the reading of the scatter diagrams graphics is a very easy task, very helpful in the construction of the searched synthesis. Such graphics are also very helpful for the communication of the analysis results.

3. Exploratory 2-way factor analyses

The aim of exploratory factor analysis is to represent in a reduced dimensional space the set of points representing the units. The problem is known as *Euclidean immersion* and it is well known that if units mutual distance is known, under some condition the solution is given:

Theorem: Given an Euclidean distance matrix (d_{ij}) between n points, they can be imbedded in an Euclidean space of dimension, at most $n-1$, equal to the rank of the Torgerson's (1958) matrix, whose generic element is

$$s_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}\right)$$

where $d^2{}_{i\cdot} = \sum_j d^2{}_{ij}/n$, $d^2{}_{\cdot j} = \sum_i d^2{}_{ij}/n$, and $d_{..} = \sum_{ij} d^2{}_{ij}/n^2$ respectively.

Proof: In fact, one point can be chosen as space origin and the others as extremes of vectors joining them to the origin. In particular, the centroid may be chosen as the origin, in which case the scalar products among vectors is given by Torgerson formula. Thus, the rank of the Torgerson matrix equals the number of the vectors linearly independent, i.e. the space dimension.

This results from the following

Proposition: Given p vectors with a scalar product, they span a space whose dimension equals the rank of the scalar product matrix.

There is a special relation between euclidean immersion and eigenanalysis. We start from the eigenproblem $AU = U\Lambda$, where A is a square matrix. We remind here some properties connected with the eigenproblem.

Proposition: The eigenvalues of a matrix A are invariant under transformation of the type

$$A \longrightarrow T^{-1}AT$$

whereas the matrix of column eigenvectors U becomes UT. If eigenvalues are different, eigenvectors are independent, so that $U^{-1}AU = \Lambda$ and $U\Lambda U^{-1} = A$.

Proposition: A symmetric real matrix A has all real eigenvalues.

Theorem (Sylvester): Given a p-dimensional symmetric real matrix A, the numbers p_1, p_2, and p_3, of the positive, null, and negative eigenvalues, summing up to p, are invariant under the same transformation.

Corollary: If a matrix A is *psd*, then $p_3 = 0$.

Definition: Given a vector space V with a scalar product, an *orthonormal basis* (v_1, v_2, ..., v_p) is a one such that

$$|\langle v_i, v_j \rangle| = \text{Kronecker's } \delta_{ij}$$

Theorem: Given a vector space V with a basis (or only generators) (v_1, v_2, ..., v_p) and a scalar product, whose matrix is $S = (s_{ij}) = (\langle v_i, v_j \rangle)$, there exist eigenvectors of S forming an orthogonal (orthonormal) basis of V.

Theorem: For any real matrix A, there exist two orthogonal matrices P and Q, such that $P'AQ$ is real diagonal with non-negative elements.

Proof: Let A an $m \times n$ real matrix, $m \geq n$, with rank $r \leq n$. The matrix AA' is symmetrical, thus with real eigenvalues, and with rank r. Let

$$P'AA'P = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}$$

be the diagonalization of AA', where P is orthogonal $m \times m$. Here, Λ is the $r \times r$ diagonal matrix of non-zero eigenvalues. Setting

$$P'A = \begin{pmatrix} D \\ E \end{pmatrix}$$

where D is $r \times n$ and E is $(m - r) \times n$, we have

$$P'AA'P = \begin{pmatrix} DD' & DE' \\ ED' & EE' \end{pmatrix} = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}$$

so that $DD' = \Lambda$ and $EE' = 0$; it follows that $E = 0$ and the elements of Λ are non-negative, since $\lambda_i = \sum_i d^2_{ij}/n$. If we set $Q' = \Lambda^{-1/2}D$, where $\Lambda^{-1/2}$ is the diagonal matrix whose elements are $(\lambda_i)^{-1/2}$, we get

$$Q'Q = I_r \quad \text{and} \quad \Lambda^{1/2}\,Q' = D \ .$$

As a result we obtain

$$P'A = \begin{pmatrix} \Lambda^{1/2}Q' \\ 0 \end{pmatrix} = \begin{pmatrix} \Lambda^{1/2} \\ 0 \end{pmatrix} Q'$$

Multiplying by Q we get

$$P'AQ = \begin{pmatrix} \Lambda^{1/2} \\ 0 \end{pmatrix}$$

and, premultipying by its transpose

$$Q'A'AQ = Q'A'PP'AQ = \begin{pmatrix} \Lambda \\ 0 \end{pmatrix}$$

so that Q and Λ are eigenvectors and eigenvalues of $A'A$ respectively. ■

From the theorem before, it results that given any A we may decompose it according to

$$A = P\Lambda^{1/2}Q'$$

where Λ is the diagonal eigenvectors matrix of AA' and P, Q are AA' and $A'A$ eigenvector matrices respectively, so that the elements of A are decomposed as follows:

$$a_{ij} = \sum_{k=1}^{r} \lambda_k^{1/2} p_{ki} q_{kj}$$

Theorem (Eckart-Young, 1936*)*: The best approximation of the matrix A of rank $s < r$ is given by

$$\tilde{a}_{ij} = \sum_{k=1}^{s} \lambda_k^{1/2} p_{ki} q_{kj}$$

once that eigenvalues (and corresponding eigenvectors) are sorted in decreasing order.

Proof: We search a B of rank $s < r$ that minimizes the distance

$$d^2(A,B) = tr((A-B)'(A-B)) = \langle A,A\rangle - 2\langle A,B\rangle + \langle B,B\rangle$$

that means that the derivative of the distance in respect to the elements of B must be minimum. Using the decomposition of $B = R'MS$ and deriving according to R' we get

$$0 = \langle A, XR'MS\rangle = -\langle A, XB\rangle = -\langle AB', X\rangle$$

where X is skew-symmetric and consequently AB' symmetric. Deriving according to S we get that $B'A$ is symmetric, so that A and B have the same eigenvectors and the distance becomes

$$d^2(A,B) = tr((\Lambda^{1/2} - \mathrm{M}^{1/2})'(\Lambda^{1/2} - \mathrm{M}^{1/2}))$$

In order to minimize it, it is sufficient to choose $\Lambda = \mathrm{M}$, considering only the s biggest eigenvalues. ■

Expressed in terms of least squares residuals, the distance results:

$$tr(A-\tilde{A})'(A-\tilde{A}) = \sum_{ij} \| a_{ij} - \tilde{a}_{ij} \|^2 = \sum_{ij} \left\| \sum_{k=r}^{s} \lambda_k^{1/2} p_{ki} q_{kj} \right\|^2 = \sum_{k=s}^{r} \lambda_k$$

If the scalar product matrix S is *psd*, then its approximation of lower rank is done based on its decomposition, with the same results.

Translated into vector spaces terms, the basis composed by the eigenvectors of S not only furnishes orthogonal directions but also allows to sort them according to their importance to S structure reconstruction. One may then choose a reduced dimensional representation rather freely, since each is the projection of the higher dimensional ones.

There are some limits. In fact, if the association matrix is *not psd*, negative eigenvalues occur and the Euclidean immersion is no longer possible, although an Euclidean approximation still is. In fact, considering distances or dissimilarities, one may argue that they may be written as the difference between two Euclidean distances:

$$\delta_{ij}^2 = d_{ij}^2 - e_{ij}^2$$

the first computed on the eigenvectors corresponding to positive eigenvalues and the second on those corresponding to negative ones. A good approximation may thus be obtained by considering the eigenvectors associated to positive eigenvalues, although some bias is introduced and the approximation is an overestimate of the original values (since $d > \delta$). This may be checked observing (*interpreting*) eigenvectors associated to negative eigenvalues (Gower, 1966; Mardia, 1984; Camiz, 1999).

A worst limitation depends on the fact that the solution for the best representation of a distance matrix is not

$$\sum_{ij} \left(S_{ij} - \sum_{1}^{q}{}_{\alpha} \lambda_\alpha u_{\alpha i} u_{\alpha j} \right)^2 = \text{minimum } \forall\ q < p$$

but

$$\sum_{ij} \left(D_{ij}^2 - \sum_{1}^{q}{}_{\alpha} (u_{\alpha i} - u_{\alpha j})^2 \right)^2 = \text{minimum } \forall \; q < n$$

(Le Calvé, 1976) for this, no solution in terms of matrix decomposition is possible and one may refer to optimal scaling techniques.

Once decomposed the association matrix *A*, considering the reconstruction formula

$$a_{ij} = \sum_{k=1}^{r} \lambda_k^{1/2} p_{ki} q_{kj}$$

limiting attention to the first *s* eigenvalues, the vectors-characters and the points-units are projected on the vector space spanned by the corresponding eigenvectors. In this way, *np* values are approximated using only *(n + p + 1)s* values, but the loss of information may be minimized and relative positions may be appreciated.

It must be pointed out that the selection of the number of factors to take into account is not a purely mathematical issue: one may calculate sufficiently well the amount of both individual and pooled information provided by factors, but this is not enough to decide *which factors to keep and which to discard*. According to each method, some tools may give some more detailed suggestion, but the final choice depends on the relative importance of factors, in terms of eigenvalues magnitude, and on their interpretability and interest. On the opposite, there are cases where it is important to keep either all or most of information, so that the number of factors to take into account may be very large.

3.1 Principal Components Analysis (PCA)

If a particular association matrix is not available, the Eckart-Young (1936) decomposition leads to the decomposition of the matrix *X'X*. If characters are measures with mean = 0, this is equivalent to diagonalize the covariance matrix of characters. If in addition characters have variance = 1, it is the correlation matrix to be diagonalized, since the transformation is implicitly done

$$x_{ij} \mapsto z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

In fact, the sum of squared z_{ij} is exactly the correlation amon original variables. This leads then to *PCA*, as noted before.

3.2 Correspondence Analysis

When data are frequencies, so that the matrix is a contingency table, a transformation of this kind is preferred

$$x_{ij} \mapsto k_{ij} = \frac{x_{ij}}{\sqrt{\sum_h x_{ih} \sum_k x_{kj}}} - \sqrt{\sum_h x_{ih} \sum_k x_{kj}}$$

since in this way the analysis decomposes the table chi-square statistics (Orlóci, 1978). Using the traditional notation of contingency data tables, denoting by n_{ij} the generic term of the table, whose rows, columns, and grand totals are $n_{i.}$, $n_{.j}$, and $n..$ respectively, the above transformation, proposed by Orlóci (1978) becomes

$$k_{ij} = \frac{n_{ij}}{\sqrt{n_{i.}n_{.j}}} - \frac{\sqrt{n_{i.}n_{.j}}}{n_{..}} = \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}} - \sqrt{f_{i.}f_{.j}}$$

where the terms $f_{ij} = n_{ij}/n_{..}$ and $f_{i.}f_{.j} = n_{i.}n_{.j}/n_{..}^2$ are respectively the observed relative frequency and the expected, in case of independence of rows and columns characters values. The generic term of the scalar product is thus

$$s_{ij} = \sum_k k_{ki}k_{kj} = \sum_k \left(\frac{n_{ki}}{\sqrt{n_{k.}n_{.i}}} - \frac{\sqrt{n_{k.}n_{.i}}}{n_{..}} \right) \left(\frac{n_{kj}}{\sqrt{n_{k.}n_{.j}}} - \frac{\sqrt{n_{k.}n_{.j}}}{n_{..}} \right)$$

$$= \sum_k \left(\frac{f_{ki}}{\sqrt{f_{k.}f_{.i}}} - \sqrt{f_{k.}f_{.i}} \right) \left(\frac{f_{kj}}{\sqrt{f_{k.}f_{.j}}} - \sqrt{f_{k.}f_{.j}} \right)$$

$$= \sum_k \frac{f_{ki}f_{kj}}{f_{k.}\sqrt{f_{.i}f_{.j}}} - \sum_k f_{ki}\sqrt{\frac{f_{.j}}{f_{.i}}} - \sum_k f_{kj}\sqrt{\frac{f_{.i}}{f_{.j}}} + \sum_k f_{k.}\sqrt{f_{.i}f_{.j}}$$

$$= \sum_k \frac{f_{ki}f_{kj}}{f_{k.}\sqrt{f_{.i}f_{.j}}} - \sqrt{f_{.i}f_{.j}}$$

Since in the case of independence, it results

$$s_{ij} = \sum_k \frac{f_{k.}f_{.i}f_{k.}f_{.j}}{f_{k.}\sqrt{f_{.i}f_{.j}}} - \sqrt{f_{.i}f_{.j}} = \sum_k f_{k.}\sqrt{f_{.i}f_{.j}} - \sqrt{f_{.i}f_{.j}} = 0$$

the scalar product measures the departure from independence of two columns. In addition the trace of scalar products matrix *S* is

$$tr(S) = \sum_{ij} \left(\frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}} - \sqrt{f_{i.}f_{.j}} \right)^2 = \sum_{ik} \frac{f_{ki}^2}{f_{k.}f_{.i}} - 1$$

From the definition of χ^2 we get

$$\chi^2 = \sum_{ij} \frac{(O - E)^2}{E} = \sum_{ij} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}} \right)^2}{\frac{n_{i.}n_{.j}}{n_{..}}}$$

that normalising and considering the expected values are $n_{..}f_{i.}f_{.j} = n_{i.}n_{.j}/n_{..}$, leads us to

$$\chi^2 = \sum_{ij} \frac{\left(n_{..}f_{ij} - n_{..}f_{i.}f_{.j}\right)^2}{n_{..}f_{i.}f_{.j}} = n_{..} \sum_{ij} \left(\frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}} - \sqrt{f_{i.}f_{.j}} \right)^2$$

So, trace $S = \frac{\chi^2}{n_{..}}$ and the eigenvalues of S decompose $\chi^2 = n_{..}(\lambda_1 + \lambda_2 + ... + \lambda_p)$.

Let's now consider the matrix *S** (Lebart *et al.*, 1979) whose term is

$$s*_{ij} = \sum_k \frac{1}{f_{k.}} \frac{f_{ki}f_{kj}}{\sqrt{f_{.i}f_{.j}}}$$

Since in case of independence it worths

$$s*_{ij} = \sum_k \frac{1}{f_{k.}} \frac{f_{k.}f_{.i}f_{k.}f_{.j}}{\sqrt{f_{.i}f_{.j}}} = \sum_k f_{k.}\sqrt{f_{.i}f_{.j}} = \sqrt{f_{.i}f_{.j}}$$

it is a non-centered measure and trace $S^* = \sum_{ik} \frac{f_{ki}^2}{f_{k.}f_{.i}}$.

The vector $u_p = \left(\sqrt{f_{.i}}\right)$ is an eigenvector of both matrices, in *S** corresponding to eigenvalue 1, since

$$\sum_i \sum_k \frac{f_{ki}f_{kj}}{f_{k.}\sqrt{f_{.i}f_{.j}}} \sqrt{f_{.i}} = \sum_i \sum_k \frac{f_{ki}f_{kj}}{f_{k.}\sqrt{f_{.j}}} =$$

$$\frac{1}{\sqrt{f_{.j}}} \sum_k \frac{f_{kj}\sum_i f_{ki}}{f_{k.}} = \frac{1}{\sqrt{f_{.j}}} \sum_k f_{kj} = \sqrt{f_{.j}}$$

whereas in *S* it corresponds to eigenvalue 0, since

$$\sum_i \left(\sum_k \frac{f_{ki}f_{kj}}{f_{k.}\sqrt{f_{.i}f_{.j}}} - \sqrt{f_{.i}f_{.j}} \right) \sqrt{f_{.i}} =$$

$$\sum_i \sum_k \frac{f_{ki}f_{kj}}{f_{k.}\sqrt{f_{.i}f_{.j}}} \sqrt{f_{.i}} - \sum_i \sqrt{f_{.i}f_{.j}} \sqrt{f_{.i}} = \sum_i \sum_k \frac{f_{ki}f_{kj}}{f_{k.}\sqrt{f_{.j}}} \sum_i \sqrt{f_{.i}}f_{.i} =$$

$$\frac{1}{\sqrt{f_{.j}}} \sum_k \frac{f_{kj}\sum_i f_{ki}}{f_{k.}} - \sqrt{f_{.j}} = \frac{1}{\sqrt{f_{.j}}} \sum_k f_{kj} - \sqrt{f_{.j}} = \sqrt{f_{.j}} - \sum_i \sqrt{f_{.i}}f_{.i} = \sqrt{f_{.j}} - \sqrt{f_{.j}} = 0$$

both have all other eigenvalues and eigenvectors equal, since if any u_α is an eigenvector of *S* it is orthogonal to u_p, so $\sum_i u_{\alpha i}\sqrt{f_{.i}} = 0$ and consequently (Lebart *et al.*, 1979: pp. 311-312):

$$\sum_i \left(\sum_k \frac{f_{ki} f_{kj}}{f_{k.}\sqrt{f_{.i} f_{.j}}} - \sqrt{f_{.i} f_{.j}} \right) u_{\alpha i} = \sum_i \sum_k \frac{f_{ki} f_{kj}}{f_{k.}\sqrt{f_{.i} f_{.j}}} u_{\alpha i} - \sqrt{f_{.j}} \sum_i \sqrt{f_{.i}}\, u_{\alpha i} =$$

$$= \sum_i \sum_k \frac{f_{ki} f_{kj}}{f_{k.}\sqrt{f_{.i} f_{.j}}} u_{\alpha i}$$

The first matrix has rank lower than the maximum, due to the constraint that the sum of margins equals 1. In the second matrix this corresponds to the fact that the distribution is entirely in the hyperspace $\sum_i x_i \sqrt{f_{.i}} = 1$ orthogonal to the first eigenvector, that has no interest. Then, the withdrawal of first eigenvalue and eigenvector, proposed by Lebart *et al.* (1977; 1979), is only a centering. This is a general property, as quoting Saito (1978) "Since $b_{ij} = \frac{1}{2}(\delta^2{}_{i.} + \delta^2{}_{.j} - \delta^2{}_{..} - \delta^2{}_{ij})$ is doubly centered, then exists an eigenvector *1*, whose elements are unity, associated with $\lambda_0 = 0$." Here, $\delta^2{}_{i.} = \frac{1}{n}\sum_k \delta^2{}_{ik}$ and similarly the others.

It may be noted that, whereas the transformation operator

$$n_{ij} \mapsto k_{ij} = \frac{n_{ij}}{\sqrt{n_{i.} n_{.j}}} - \frac{\sqrt{n_{i.} n_{.j}}}{n_{..}}$$

may be wrtten in matrix form

$$K = \mathrm{diag}\left(n_{i.}^{-\frac{1}{2}} \right) (n_{ij}) \mathrm{diag}\left(n_{.j}^{-\frac{1}{2}} \right) - \mathrm{diag}\left(n_{i.}^{\frac{1}{2}} \right) n_{..} \mathrm{diag}\left(n_{.j}^{\frac{1}{2}} \right)$$

the operator

$$n_{ij} \mapsto k_{ij}^{*} = \frac{n_{ij}}{\sqrt{n_{i.} n_{.j}}}$$

becomes

$$K^{*} = \mathrm{diag}\left(n_{i.}^{-\frac{1}{2}} \right) (n_{ij}) \mathrm{diag}\left(n_{.j}^{-\frac{1}{2}} \right)$$

so that both *S* and *S** are symmetric and *psd*, thus they are scalar products, corresponding to the Euclidean distances given by

$$d^2{}_{ij} = \sum_k \frac{1}{f_{k.}} \left(\frac{f_{ki}}{\sqrt{f_{.i}}} - \frac{f_{kj}}{\sqrt{f_{.j}}} \right)^2 \quad \text{among column profiles}$$

$$d^2{}_{ij} = \sum_k \frac{1}{f_{.k}} \left(\frac{f_{ik}}{\sqrt{f_{i.}}} - \frac{f_{jk}}{\sqrt{f_{j.}}} \right)^2 \quad \text{among row profiles}$$

This may be easily checked through Torgerson's (1958) formula. Nevertheless, this is not the classical chi-squared distance proposed by the French school of data analysis (Benzécri *et coll.*, 1973-82; Lebart *et al.* 1977, 1979, 1984, 1995). In fact, the latter is a distance among *profiles*, corresponding to

$$d^2_{ij} = \sum_k \frac{1}{f_{k.}} \left(\frac{f_{ki}}{f_{.i}} - \frac{f_{kj}}{f_{.j}} \right)^2 \quad \text{among columns}$$

$$d^2_{ij} = \sum_k \frac{1}{f_{.k}} \left(\frac{f_{ik}}{f_{i.}} - \frac{f_{jk}}{f_{j.}} \right)^2 \quad \text{among rows}$$

that would lead to a scalar product of the type

$$S_{ij}^{+} = \sum_k \frac{f_{ki} f_{kj}}{f_{k.} f_{.i} f_{.j}}$$

that in matrix terms would become

$$S^{+} = \mathrm{diag}(n_{i.})(n_{ij})\,\mathrm{diag}(n_{.j})$$

differing from the previous by the diagonal terms that are the squares of the others. It is easy to see that S^* and S^+ have the same eigenvalues, whereas the eigenvectors are different.

Example: Hair and eyes of American women

This example was taken from Snee (1974): it is a contingency data table crossing the colour of the hair and of the eyes of 592 American women. The table appears as it is shown in Table 2. Applying Correspondence Analysis to the table, two factors result, the first containing over 89% and the second around 9% of the total information, so that the first factor plane summarizes 98.89% of information. The distribution of character values on the plane is shown in Figure 4.

Table 2 - The Snee (1974) contingency table of American women hair and eyes colour.

	Colour of hair				
Colour of eyes	*Black*	*Brown*	*Red*	*Blond*	*Total*
Dark Brown	68	119	26	7	220
Light Brown	15	54	14	10	93
Green	5	29	14	16	64
Blue	20	84	17	94	215
Total	108	286	71	127	592

The first axis, whose importance is much bigger than the second, distinguishes *Blond hair* from the others, in particular to *Black hair*, whereas the second opposes *Red hair* to the said, *Brown hair* remaining in a central position. Analogously, *Blue eyes* are opposed to all others, in particular to *Dark brown eyes*, whereas the second axis opposes *Green eyes* to these; *Light brown eyes* occupy an intermediate position on the second axis on this same side. The comparison among both sets of characters values suggests a strong association between *Blue eyes* and *Blond hair*, whereas other associations are not so clear: one may say that *Dark brown eyes* are common to both *Black* and *Brown hair*, *Light brown eyes* are common to both *Brown* and *Red hair*, *Green*

eyes are associate to *Red hair*, but common to *Blond hair* too. and between *Dark brown eyes* and both *Black* and *Brown hair*.

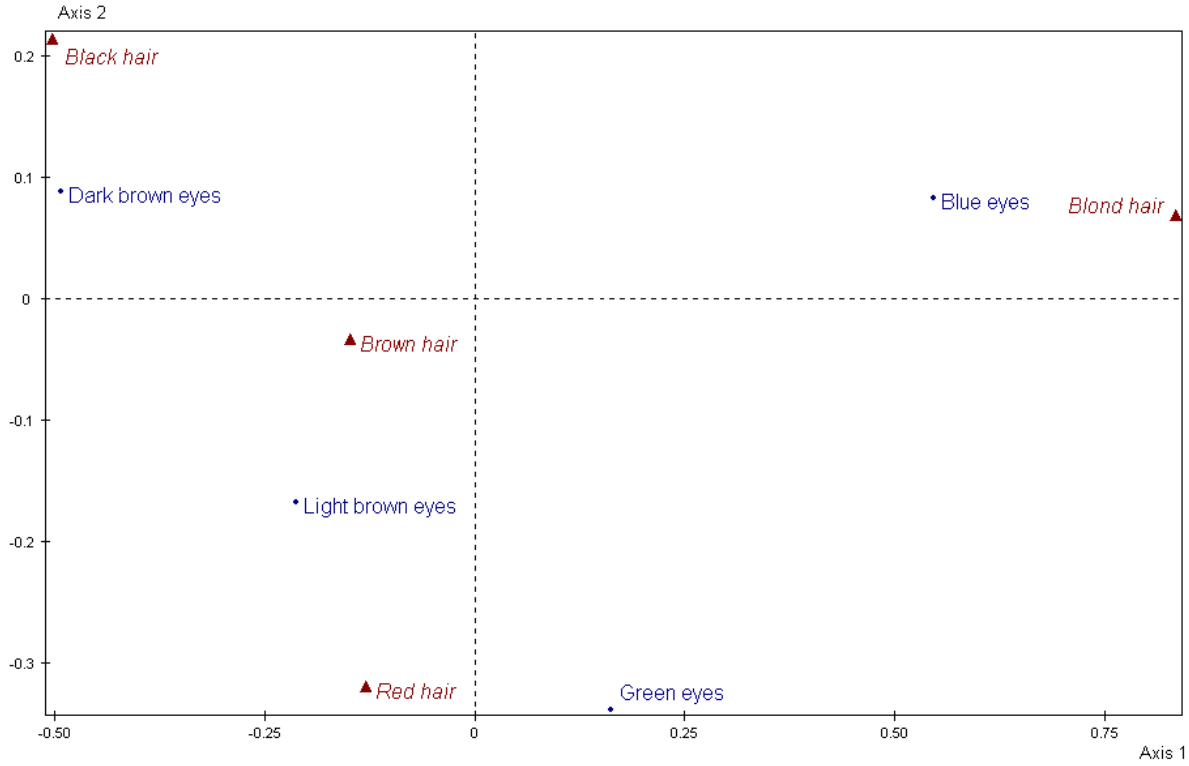

Figure 4 - The first factor plane of correspondence analysis of American women hair and eyes colour.

3.3 Supplementary elements

On factor analysis spaces, in addition to both units and characters that took part to the analysis, others objects may be represented, on condition that the relationships with the elements that took part to the analysis are known. In particular, since the position of characters is known, other units may be represented, by transforming their character values into coordinates along principal axes, using the eigenvectors matrix; in a similar way, other characters may be represented, considering the known relationships with the characters, that lead to their projection on the spaces spanned by them, thus the factorial spaces.

These elements are called *supplementary* elements, meaning that they do not participate to the identification of the representation spaces, but are simply projected onto them. They may be used for several purposes:

- characters close to the principal axes, may be used as an explanation of factors;
- if dependence relationships are suspected, the quality of the representation of a character on the space spanned by the others is a measure of their influence on it;
- supplementary units projected on factor spaces may show the ability of the factorial representation to adjust to other elements, thus the robustness or the generality of the found structure;
- if some units have a too heavy contribution to a factor, its representation as supplementary allows a more reliable representation of all others on factors that do not depend on them.

4. Exploratory 3-way factor analyses

As we said, exploratory factor analyses are among the first steps of an investigation data analysis and we just saw examples of two such analyses, different in that they were better tailored to the nature of the involved data. Even in this first step of investigation, some other knowledge may already exist concerning the data structure. For this reason, many developments of data analysis concern the introduction of the existing knowledge in the exploratory methods, in order to take advantage of the known structure. In fact, it may happen that if the known structure is not taken into account, it is likely that this will be shown in the analysis as the most important issue, a result without any interest.

Among others, the known structure may happen to have the following characters:
- there exists a causal relationships among characters already known;
- the data are taken from units belonging to populations known as different;
- the data derive from surveys repeated in different occasions, but concerning the same units;
- the characters may be partitioned into groups, since they concern different aspects of the investigation, that is better to keep apart.

The inclusion of these kind of known structure into the analysis methods leads to the so-called *multiway data analysis* (see, for reviews: Coppi and Bolasco, 1989; Bry, 1996; Dazy and Le Barzic, 1996). We shall concentrate our attention to the *3-way factor analysis* methods (Rizzi and Vichi, 1985). In this case, several data tables are considered, corresponding to different *studies* or *occasions*. According to their particular structure, they take different names. The tables may be composed by
- same units and same characters: --> *cubic data*;
- same units and different characters: --> *multiple tables*;
  - when a table explains the other: --> *instrumental variables*;
- same characters and different units: --> *multiple data sets*;
  - if indexed by time: --> *longitudinal data* (Diggle *et al.*, 1994).

The different structure deriving from these differences is of importance in order to decide the particular treatments possible.

4.1 Factor analysis on instrumental variables

Given two data tables, this analysis is considered once one wishes to explain a table structure based on the structure of the other. In this case the possible methods may be
- the projection of supplementary characters on the spaces spanned by factors extracted by the other characters (Lebart *et al.*, 1977);
- *PCA* on instrumental characters (Rao, 1984);
- Non-Symmetric Correspondence Analysis (Lauro and D'Ambra, 1984);
- *PLS* (Partial Least Squares) methods (Tenenhaus *et al.*, 1995).

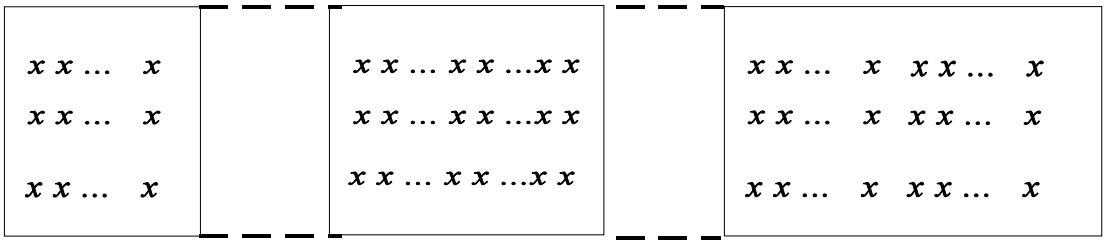

Figure 5 - Multiple tables may be aligned: rows represent the same units.

4.2 Multi-occasion data analysis

We consider as *multi-occasion* those multiple tables having either same rows or same columns, but not necessarily both. The first is the case of *multiple tables,* that may be aligned as in Figure 5, giving a pooled matrix with the same common units as rows and all tables characters as columns; the second is the case of *multiple data* that may be put on a stack as in Figure 6, giving a pooled matrix having the same common characters as columns and all tables units as rows. For such analyses, several different issues may be considered:

- syntheses concerning *occasions*, that is a simultaneous representation of the tables as a whole. This leads to the *interstructure* analysis, that is the identification of the structure that gets occasions different;
- a *virtual* occasion, able at summarizing the main features of all tables. The occasion built with this aim is called a *compromise* and its study may show some kind of common structure to all tables;
- a detailed analysis of all occasions based on the known structure, thus on the compromise. It is the *intrastructure* analysis;
- the representation in a common space of either units or characters, according to the different occasions. This allows to see how the different occasions modify the position of the elements. In particular, dealing with occasions indexed by the time, elements *trajectories* are thus revealed.

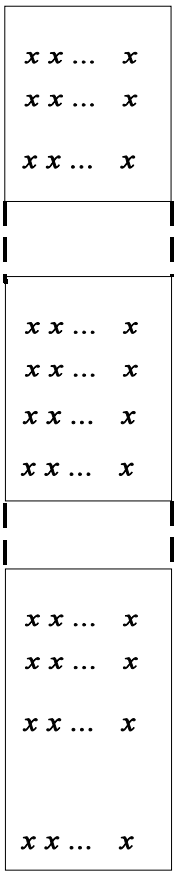

Figure 6 - Multiple data tables may be set on a stack, giving a pooled matrix whose columns are the same characters.

4.3 Compromise: intrastructure and trajectories

To build a *compromise* where to study the different tables one must distinguish among multiple tables and multiple data. Once decided the compromise, its *PCA* may be a good starting point for an intrastructure study, since on its factor spaces may be projected the vectors corresponding to each occasion characters, considered as all active characters in multiple tables case and only supplementary in multiple data sets. Units trajectories are visible only in the multiple table case, whereas characters corresponding to the different populations may be seen in the other case: a simple technique is the projection of either each *partial* unit (corresponding to the unit in each occasion), or each

partial character (corresponding to the character in each occasion) on intrastructure factor spaces.

One problem raises once that tables are pooled, since a simple *PCA* of pooled matrix may lead to some problems, tied to the different matrices structure. In fact, in multiple tables case, the different sets of characters can have different weight and a very different correlation structure, so that strongest structure tables may darken the structure of the others; in multiple data sets, both population and sample structure may be different in any occasion, once again biasing the results that are more influenced by either the largest data sets or the strongest in structure. It is thus necessary to balance the different occasions, using an appropriate set of weights.

4.4 The analysis of multiple tables

For this analysis, *PCA* of pooled data table may be considered as a possible intrastructure. Due to the said drawback, other three methods were developed in literature, in order to overcome the problem of misweighting the tables: *Generalized Canonical Analyis*, *Multiple Factor Analysis*, and *Statis*, also known as *Joint Tables Analysis* (Analyse conjointe des tableaux, Lavit, 1988): they are outlined in the following.

From now on we shall use the following notation:

- X_k ,$k = 1...K$, are the data tables in K different occasions:
  - each table has n rows, corresponding to the same units,
  - and p_k columns of standardised continuous characters (mean = 0 and standard deviation = 1), so that the pooled characters are $\sum_{k=1}^{K} p_k = p$,
- D is an $n \times n$ diagonal matrix of units weights;
- M_k , $k = 1...K$ are K association matrices among each occasion p_k characters;
- $W_k = X_k M_k X'_k$, $k = 1...K$ are symmetric *psd* $n \times n$ association matrices among units: it is a *scalar product* between units in $\mathbb{R}^n$ space.

On this basis, a pooled scalar product among units may be built, given by

$$W = \sum_{k=1}^{K} \alpha_k X_k M_k X_k^{/} = XMX^{/}$$

where M, is a $p \times p$ block diagonal matrix $\alpha_k M_k$. If for any k, the weights $\alpha_k \geq 0$, it is known that W is *psd*.

The intrastructure analysis is thus obtained by the *PCA* of matrix WD, with the constraint that $V'V = \Lambda$ (since we are dealing with units and not characters). As a consequence, the representation of supplementary elements on *PCA* factor spaces may be done as follows:

- all characters coordinates are the rows of the matrix $X' D V \Lambda^{-½}$;
- coordinates of units in all occasions are the lines of $\alpha_k X_k M_k X'_k D V \Lambda^{-1}$, $k = 1,..., K$.

The problem is thus conceptually solved. It remains to choose the weights α_k and the associations among each table characters M_k . Different choices lead to different methods.

The simplest, *PCA* of the pooled matrix, simply considers

$$\alpha_k = 1 \text{ and } m_{ii,k} = 1 / \sigma_{ii}^2$$

so, no particular adjustment is made, the *M* matrix being simply means the standardization of characters, already supposed centered. In this situation data tables having different structure may have different influence in the construction of the compromise. The following methods are possible solutions, based on different sets of weights.

The *Generalized Canonical Analysis* (Carrol, 1968) aims at finding a compromise space whose factors are the most correlated to each data table characters. For this aim, weights are equal

$$\alpha_k = 1 / K \text{ and } M_k = (X_k' D X_k)^{-1}$$

Using as association matrix the inverse of the correlation matrix of each table means that the distribution of each table becomes spherical, so that no particular influence to the compromise factors comes from a particular distribution of some table. As a drawback, it may happen that the directions identified in the tables as the most correlated with the compromise have no relation with the table structure, so that their interest may be reduced.

The *Multiple Factor Analysis* (Escofier and Pagès, 1990) was introduced aiming at reducing the influence of each table to the compromise. For this reason, the pooled *PCA* was only modified in the choice of weights, taking for each table the inverse of the largest eigenvalue:

$$\alpha_k = 1 / \lambda_k \text{ and } m_{ii,k} = 1 / \sigma_{ii}^2$$

In this way, each occasion weight is balanced, in particular tables having a stronger correlation structure among characters have the same influence as the others. As a drawback, limiting the weighting only to the first eigenvalue may seem arbitrary.

The *Statis* method (Escoufier, 1973, 1980; L'Hermier des Plantes, 1976; Lavit, 1988) has a different approach, since the compromise is built as the first principal component of an eigenanalysis performed considering the occasion as units. So, the weights are the coordinate of each occasion on the first eigenvector of this analysis. Analitically, it results

$$\alpha_k = \frac{\text{trace}\,(W_k D W D)}{\nu_1 \sqrt{\sum_i \lambda_{ik}^2}} \text{ and } m_{ii,k} = 1 / \sigma_{ii}^2$$

so that each occasion weights according to its correlation to the compromise. In this way, the compromise is the first principal component of all occasions *PCA*. For this reason, if some occasions are not well represented on the first principal component, will not be sufficiently represented in the compromise and their elements projection on it will not be sufficient for a good analysis.

4.5 Interstructure analyis of multiple tables

The three quoted methods consider the problem of the interstructure, namely the ability to give information concerning the relationships among the different occasions, considered as a whole. For this purpose

- *Generalized canonical analysis*: proposes to consider both eigenvalues and eigenvectors of the compromise analysis, with the following meaning:
  - the eigenvectors coefficients are the multiple correlation coefficients among the different characters and the compromise factors;
  - the eigenvalues are the mean of multiple correlation coefficients, thus representing the correlation of eigenvectors with each table.
- *Multiple factor analysis* aims at fixing the problem of generalized canonical analysis to reveal most correlated directions in the occasions not always meaningful ori interesting, since the scattering of units may be very poor in these directions. For this reason it suggests, as a measure of relationships, the scalar product $< W_k D,\ u_i u_i^{/} >$ of each study with the u_i , i.e. the *i*-th intrastucture eigenvector. Its meaning is the amount of inertia of *k*-th study along *i*-th axis, a measure of the importance of the axis in the description of the occasion structure. This is considered advantageus, since the intrastructure axes reflect the most important factors common to all occasions.
- *Statis* interstructure analysis is the most complete. It is the eigenanalysis of *RV* association measure between $W_k D$, a kind of correlation measure among occasions, given by

$$RV(W_k D,\ W_h D) = \frac{\mathrm{Tr}(W_k D,\ W_h D)}{\sqrt{\mathrm{Tr}(W_k D)^2\ \mathrm{Tr}(W_h D)^2}}$$

  The eigenanalysis of *RV* gives results analogous to a normal *PCA*; its interpretation is easy and meaningful. In addition, the first eigenvector may be interpreted as a *compromise occasion*, so that it may be used for the intrastructure analysis, as it was stated before. Unfortunately, no relation exists between these factors and those of the compromise, so that no mutual help in the interpretation is possible.

Example: Elections in the province of Brescia, Italy, from 1948 to 1994

This example is taken from Camiz (1996). For all 206 municipalities of Brescia province, the results of 12 Chamber of Deputies elections from 1948 through 1994 were considered, adding the other three 1994 elections, namely the majority quota of Chamber of Deputies election, the Senate (again majority) and European elections (only proportional). For these 15 tables, percentages of votes of every list were taken into account, as well as percentages of voting electors and of empty or null cards. The data may be considered as a 15-layers matrix (one for each election), 206 rows (the number of presently existing municipalities in the province), and a variable number of columns according to each election number of challenging lists.

In the following tables and figures, the various elections are labeled with a capital letter and the two last digits of the year, the 12 elections thus ranging from *A48* for 1948 through *L94* for1994[2]. In addition, regarding the other three 1994 elections, *M94* identifies the majority part of Chamber of Deputies, *N94* the Senate, and *O94* identifies the European elections.

[2] The 12 elections are *A48 = 1948, B53 = 1953, C58 = 1958, D63 = 1963, E68 = 1968, F72 = 1972, G76 = 1976, H79 = 1979, I83 = 1983, J87 = 1987, K92 = 1992, L94 = 1994*.

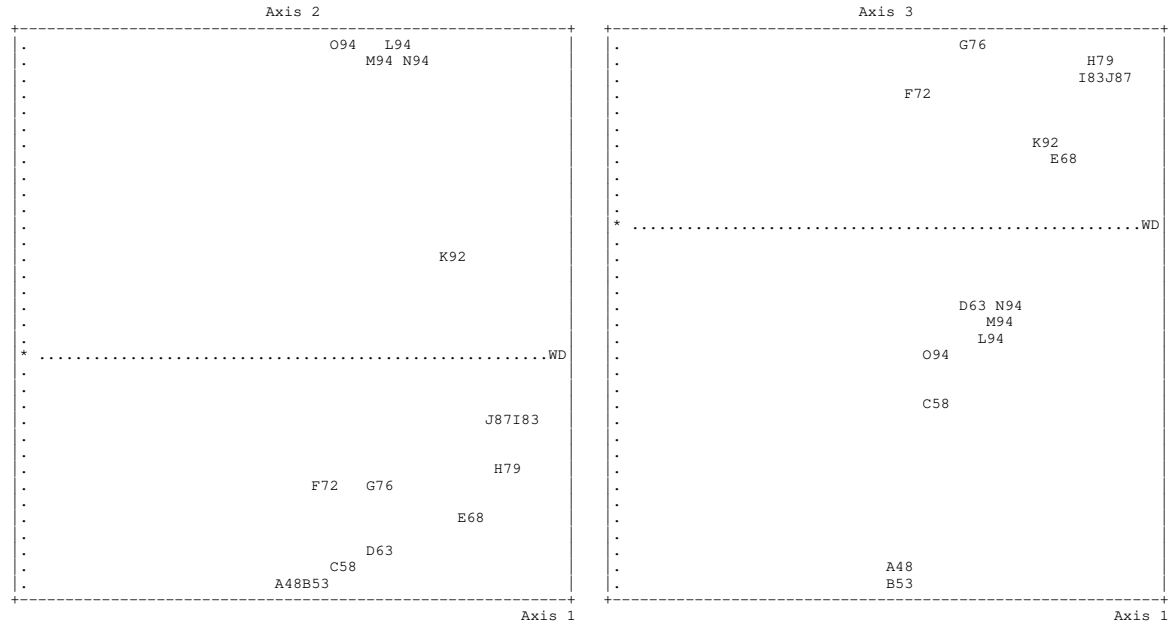

Figure 7 - Representation of the elections on the planes of the interstructure spanned by axes 1 and 2 and 1 and 3, respectively.

The lists are labeled according to the party initial: it must be noted that the label was kept unchanged even if the parties names changed along time. So, *PDP* stands for *Partito Socialista di Unità Proletaria, Partito Democratico d'Unità Proletaria*, and *Democrazia Proletaria*, extreme left parties that alternated from 1968 through 1987; *RIF* is the present *Rifondazione Comunista*; *PCI* identifies the *Communist Party*, including the 1948 *Fronte Popolare* and more recent *PDS*, *Partito Democratico della Sinistra*; *PRO* refers to the *Progressisti* candidates, the coalition of center-left, present in majority elections of 1994; *VER* stands for the *Green Federation* (in brief, *Verdi*), *RET* is *La Rete*; *PSI* is the *Socialist Party*, *PSD* the *Partito Socialdemocratico Italiano*, *PSU* the unified *Partito Socialista*, an attempt of unity between *PSI* and *PSDI* that was present only in 1968; *PR* is *Partito Radicale*, later become *Lista Pannella*; *PRI* is *Republican Party* (joint with *Radical Party* in 1958), *DC* is both ancient *Democrazia Cristiana* and present *PPI*, *Partito Popolare Italiano*, *PAT* is *Patto Segni* (joint with *Popolari* in 1994 majority), *PLI* is *Partito Liberale Italiano*, *LEG* is *Lega Nord*, *FOI* is Berlusconi's *Forza Italia*, *LFO* the coalition *Lega-Forza Italia* in 1994 majority, *MSI* is neo-fascist *Movimento Sociale*, nowadays *AN*, *Alleanza Nazionale*. In addition, percentages of voters (*V*), null (*N*) and blank cards (*X*) were added.

In Figure 7 all elections layers are represented on two *Statis* interstructure planes, spanned by the axes 1 and 2 and 1 and 3, respectively. It is easy to see that a kind of continuous evolution may be noted, that develops along a curve line in three dimensions, and a major change occurred in 1994, corresponding to an earthquake in political landscape.

Concerning the intrastructure analysis, in Figure 8 parties in each election are represented on the plane of the first two axes. On the first axis, explaining around 17% of total variation, there is a clear opposition among *PCI/PDS* and *RC* in all elections on the negative side, and *DC/PPI* in all elections on the positive side. In addition, voting electors percentages are on the negative size, as well as other lists, like most list of *PSI*, *MSI/AN*, Greens, Rete, *PR*, *PRI*, and *PLI*, more and more close to the center. On the positive side, closer to the center *PSDI* is set, as well as the null

votes and the other small lists. Considering that the plane represents the political space as understood by the electors, it is interesting to note that *PCI* and *DC* keep a stable position in electors feeling, *DC* representing the government and *PCI* the opposition, all other parties seem to be wandering in quest of a place in the same political space, in order to develop an authonomous policy. An important example in this sense is *PSI*, that in 1948 was joint with *PCI* (*PCIA*) on bottom left of the plane, in 1968 was joint with *PSDI* (*PSUE*) in top right and in the other elections fluctuated in the space in-between.

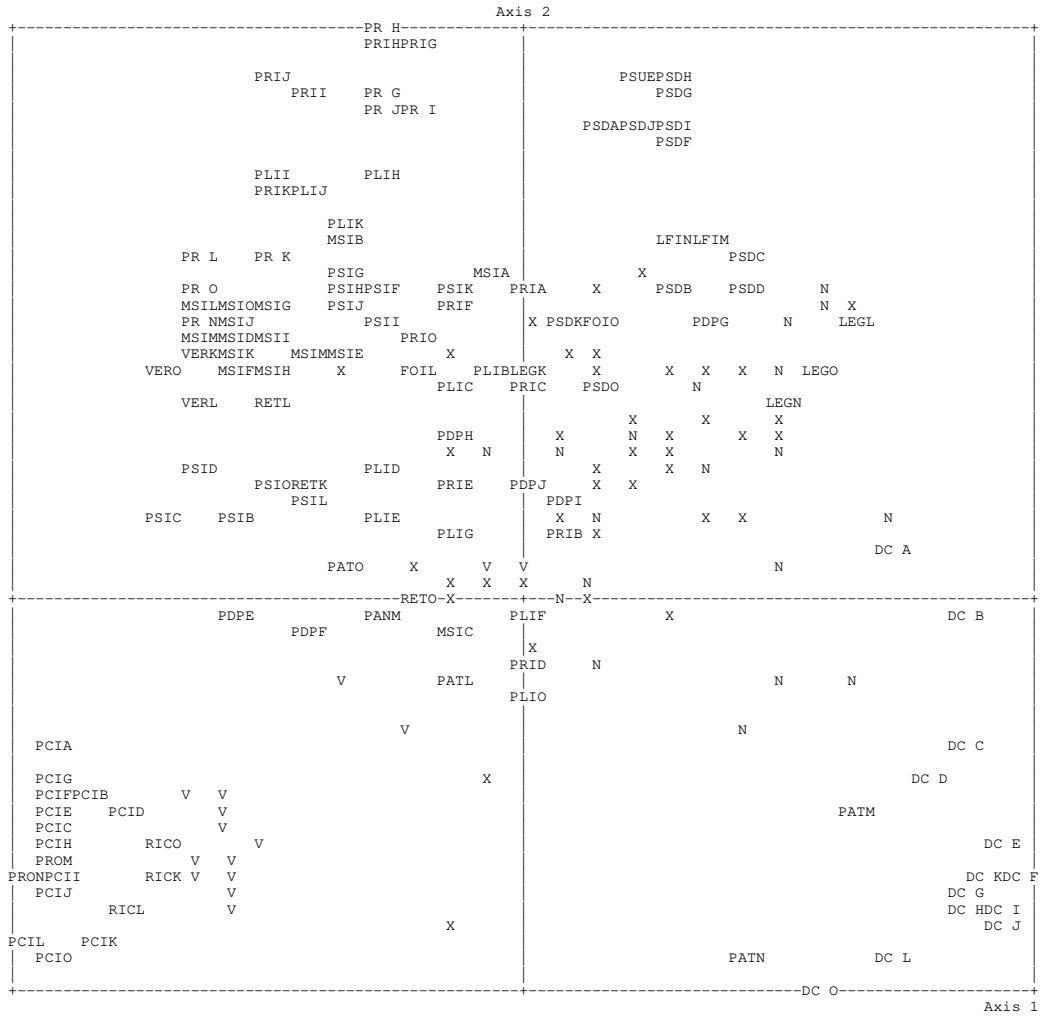

Figure 8 - Representation of lists in each election on the plane of first two axes of PCA of intrastructure analysis.

Another point of interest is the position of the parties that gathered in 1994 in the right side cohalition headed by Berlusconi: each one occupies the position previously owned by some old party that disappearede in the meantime. In addition, *Lega nord* occupies the space of *blank* and *null cards*, a position occupied by people little or not at all concerned with politics (that explains the difficulty to organize a realistic political program). It was known at that time that an attempt was done to gather both *Lista Pannella* and *Patto Segni* in the cohalition. They had to cover the extreme left and the Christian Democrat positions, but Pannella and Segni refused. This suggests that the campaign was prepared at that time using all marketing techniques, including these data analysis methods: Berlusconi's (or his collaborators) idea was to occupy *all political space* with parties belonging to the cohalition and all positions were necessary to avoid that the opponent

could have some authonomous space. Luckily this was not the case, but we assist now to the same attempt.

5. Conclusion

The ability to represent in a very synthesized way the main structures contained in a data table is the main power of exploratory analyses. It had a tremendous success in those disciplines that developed more recently, or that at least from quantitative techniques received a new boost. In fact, the use of computers was a tremendous opportunity to start using exploratory techniques otherwise impossible to exploit suitably. This enlarged the analysis abilities to very large data tables: the new *data mining*, that is the use of exploratory data analysis methods for the collection of information in very large data bases is the most recent issue, but other issues are being developed, in particular the attempt to approach the confirmatory phase with techniques closer to exploratory analysis.

It is an author's opinion that the exploratory methods, presently considered as very specialised, should become more familiar to the scientist than the tests of hypotheses and all modelling. In fact, both latter steps require particular cautions, in order to prevent misuses, not necessary in exploratory analysis, whose role is only for helping the scientist to get ideas and syntheses concerning the general framework where the other phases of the investigation are grounded.

Such uses together with the interaction between scientists and data analysts may be very useful for both sciences and data analysis, since the first may derive a more rigorous methodology as theoretical background and the second, getting in touch with different disciplines, will discover new frameworks of application where specific problems may suggest interesting developments.

References

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

Benzécri, J.P. *et coll*., 1973-82. *L'analyse des données*. 2 vol., Paris, Dunod.

Bertin, J., 1977. *La graphique et le traitement statistique de l'information*. Paris, Flammarion.

Bry, X., 1994. *Analyses factoriales simples*. Paris, Economica.

Bry, X., 1996. *Analyses factoriales multiples*. Paris, Economica.

Camiz, S., 1993. «Statis Ordination vs. Juhász-Nagy Models: the Predictability of an Exploratory Tool». *Abstracta Botanica*, 17(1-2): 29-36.

Camiz, S., 1996. *Elections in the Province of Brescia: Concurrent Use of Three-way Data Analysis Techniques.* Università di Roma «La Sapienza», Dipartimento di Matematica «Guido Castelnuovo», preprint n. 24/96.

Camiz, S., 1999. «Comparison of Euclidean Approximatins of non-Euclidean Distances». In: M. Vichi and O. Opitz (eds.), *Classification and Data Analysis - Theory and Application*. Berlin, Springer, Studies in Classification, Data Analysis, and Knowledge Organization: pp. 139-146.

Carroll, J.D., 1968. «Generalization of Canonical Analysis to three or more Sets of Variables». *Proceedings of 76th Annual convention of the the American Psychology Association*: pp. 227-228.

Coppi, R., and S. Bolasco (eds.), 1989. *Multiway Data Analysis*. Amsterdam, North-Holland.

Costa, P., M. Maravalle, M.P. Salani, and B. Simeone, 1981. *Introduzione all'analisi dei dati multidimensionali*. Università di Roma, Centro di Calcolo Interfacoltà. Dispensa Didattica.
Dazy, F. and L.F. Le Barzic, 1996. *L'analyse des données évolutives*. Paris, Technip.
Diggle, P.J., Liang, K.-Y., and S.L. Zieger, 1994. *Analysis of Longitudinal Data*. Oxford, Clarendon Press.
Droesbeke, J.J., 1997. *Éléments de statistique*. Éditions de l'Université de Bruxelles, Paris, Ellipses.
Eckart, C. and G. Young, 1936. «Approxiamtion of one matrix by another of lower rank». *Psychometrika,* 1: pp. 211-218.
Escofier, B. and J. Pagès, 1990. *Analyses factorielles simples et multiples*. Paris, Dunod.
Escoufier, Y., 1973. «Le traitement des variables vectorielles». *Biometrics*, 29: pp. 751-760.
Escoufier, Y., 1980. «Exploratory Data Analysis when Data are Matrices». In: K. Matusita (ed.), *Recent Development in Statistical Inference and Data Analysis*. Amsterdam, North-Holland: pp. 45-53.
Gower, J.C., 1966. «Some Distance Properties of Latent Root and Vector Methods used in Multivariate Analysis». *Biometrika*, 53: pp. 325-338.
Hill, M.O., 1974. «Correspondence analysis: a neglected multivariate method». *Applied Statistics*, 23: pp. 340-354.
ISTAT, 1978. *Annuario Statistico Italiano*. Roma, Istat.
Lauro, N. and L. D'Ambra, 1984. «L'analyse non symétrique des correspondances». In: E. Diday *et al.*, *Data Analysis and Informatics, III*. Amsterdam, North Holland.
Lavit, C., 1988. *Analyse conjointe de plusieurs matrices*. Paris, Masson.
Lebart, L., A. Morineau, and J.P. Fénelon, 1979. *Traitement des données statistiques*. Paris, Dunod.
Lebart, L., A. Morineau, and M. Piron, 1995. *Statistique exploratoire multidimensionnelle*. Paris, Dunod.
Lebart, L., A. Morineau and N. Tabard, 1977. *Techniques de la Description Statistique*. Paris, Dunod.
Lebart, L., A. Morineau, and A. Warwick, 1984. *Multivariate Descriptive Statistical Analysis*. New York, J. Wiley and Sons.
Lebart, L. and A. Salem, 1988. *Analyse Statistique des Données Textuelles*. Paris, Dunod.
Lebart, L. and A. Salem, 1994. *Statistique Textuelle*. Paris, Dunod.
Le Calvé, G., 1976. *Quelques remarques sur certains aspects de l'analyse factorielle*. Lab. Analyse des Données, Université de Rennes II, Cahier n. 2.
L'Hermier des Plantes, H., 1976. *Structuration des tableaux à trois indices de la statistique*. Université de Montpellier II, Thesis.
Mardia, K.V., 1978. «Some Properties of Classical Multi-Dimensional Scaling». *Comm. in Statistics - Theory and Methods*. Series A, 7(13): pp. 1233-1241.
Mood, A.M, F.A. Graybill, and D.C. Boes, 1974. *Introduction to the Theory of Statistics*. New York, McGraw-Hill.
Orlóci, L., 1978. *Multivariate Analysis in Vegetation Research*. Den Haag, Junk.
Rao, C.R., 1984. «The use and interpretation of principal component analysis in applied research». *Sankhya*, ser. A, 26: pp.329-359.
Rizzi, A., and M. Vichi, 1995. «Three-Way Data Set Analysis». In A. Rizzi (ed.) *Some Relations Between Matrices and Structures of Multidimensional Data Analysis*, C.N.R. Comitato Nazionale per le Scienze Matematiche, Applied Mathematics Monographs, Pisa, Giardini Editori: pp. 93-166.

Saito, T., 1978. «The Problem of the Additive Constant and Eigenvalues in Metric Multidimensional Scaling». *Psychometrika*, 43(2): pp. 193-201.

Snee, R.D., 1974. «Graphical displays of two-ways contingency tables». *The American Statistician*, 28: pp. 9-12.

Tenenhaus, M., J.P. Gauchy, and C. Ménardo, 1995. «Régression PLS et applications». *Revue de Statistique Appliiquée*, 48(1): pp. 7-63.

Torgerson, W.S., 1958. *Theory and Methods of Scaling*. NewYork, J.Wiley & Sons.

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

Ward, J.H., 1963. «Hierarchical grouping to optimize an objective function». *Journal of the American Statistical Association*, 58: pp. 236-244.