Jestr

Journal of Engineering Science and Technology Review 4 (3) (2011) 256 – 260
Special Issue on Econophysics

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Research Article

Factor Data Analysis and econophysics: application in market segmentation

D. Karapistolis and G. Stalidis*

Dep. of Marketing, Alexandrian Technological Educational Institute of Thessaloniki, Greece.

Abstract

In this paper, an overview of factor data analysis methods is presented, as an alternative approach to classic statistical methods and it is shown that they are a powerful tool for analyzing economic phenomena. The principles on which data analysis methods are based are in a large degree inspired by physics, not only as general considerations but also as specific concepts, terminologies and methods. The notions of energy, entropy and inertia are matched with information theory, linear algebra and statistics to provide powerful tools for modeling and analyzing non-linear economic phenomena. Considering that any phenomenon under study is a complex open dynamic system where a large number of factors interact with each other, factor data analysis methods are able to examine such interactions as a whole, instead of a set of independent pair-wise comparisons of factors. The mechanism underlying these methods is to map the problem to a multidimensional vector space and based on the data themselves, to discover the underlying patterns, to find out how series of figures organize and which variables or group of variables are correlated. Model construction is thus not restricted to any initial assumption and is entirely driven by the data (Greenacre, 2007). In order to depict the potential of such methods in economic analysis, we present the application of multiple correspondence analysis to the market segmentation in the business plan for an internet radio venture.

Keywords: Multivariate Data Analysis, Factor Analysis, Information Entropy in Economic Phenomena, Multiple Correspondence Analysis.

1. Introduction

Economics and physics, just like the majority of sciences, are called to handle large quantities of data before they are able to establish findings on any specific research topic. Luckily enough, a tool which allowed the scientists to operate on such quantities of data has been developed, which is no other than the computer. The ability of the computer to manage and analyze data, together with a remarkable evolution in analysis and modeling methods has enabled economics and physics to study complex phenomena that could not be modeled with analytical means. At the same time, new challenges have been set to modeling efforts, especially to increase their effectiveness by reducing the need for oversimplifications and doubtful assumptions.

In the beginning of the 20th century, European psychologists attempted to utilize the evaluations of certain variables (e.g. memory, intelligence), captured from their patients through specific tests, to estimate composite variables which were not directly observable in the collected data and would interpret in the best way the human behavior. These composite variables were called "factors". These efforts of European psychologists, resulted in the work of Charles Spearman [1] and Thurstone [2] and also became the foundations of a new family of statistical methods under the name "Factor Data Analysis" or "Data Analysis". The first member of this family to appear was the Principal Component Analysis [3]. Thereafter, Hirschfeld [4] formulated thoughts based on linear algebra on the correlation between lines and columns in a contingency table. He was followed by Fischer [5], who formulated the discriminant factors, approaching in a different way the issue of revealing the relations inside a multi-dimensional matrix with data corresponding to qualitative variables. The most recent and effective method of the Data Analysis family has been developed in the '60s by the French professor J.P. Benzecri [6], with the name Factor Correspondence Analysis **(**Analyse Factorielle des Correspondances -A.F.C-). This method is applied on qualitative data in the form of multidimensional contingency matrices. In parallel with the French school, in the other side of Atlantic, appeared the American school with J.D Carrol, J. B. Kruskal, R. S Sheppard, G. Yang and others, under the name "multidimensional scaling". John Tukey [7] was the first who differentiated Data Analysis from Statistics and presented it as an independent scientific field. According to Tukey, Data Analysis includes: (a) designing of data collection tools to facilitate easier, more correct and more precise analysis (b) analysis processes and (c) interpretation techniques.

1.1 Principles of Factor Data Analysis

The most important principles of the French school of Data Analysis, as stated by its founder J.P. Benzecri are the following:

1. Statistics should not be confused with probability theory. Statistical mathematics are often structured on

* E-mail address: stalidgi@mkt.teithe.gr

the basis of numerous assumptions which are rarely satisfied in practice.

2. The model should be constructed following the data and not the other way around. By ignoring this principle, one can easily be lead to a false application of mathematics to sciences related to human behavior. The attempt to match the data to a priori models is often unsuccessful, as a primary role in human behavior is played by memory, which means that behaviors of the past are likely to be avoided in the future.
3. It is of great importance to be able to handle information in as many dimensions as possible, since multi-dimensional representations are much closer to real-world problems.

The analysts who adopt the approach of the French Data Analysis school refuse to draw conclusions based on doubtful assumptions, they consider the data in their genuine form and elaborate on revealing the tendencies hidden in the data structure. An additional property is that the study of a phenomenon is performed by examining more than two variables simultaneously, overcoming a significant limitation of classic statistical methods. To this end, mathematics of the n-dimensional space is used. More specifically:

- Each phenomenon is considered by its nature complex and a result of the interaction of a large number of variables. For this reason, it is examined on the basis of the whole set of interactions among variables, instead of separate pair-wise examinations.
- Qualitative variables are handled as property vectors, avoiding the use of pseudo-variables, which, as indicated even by their name, include a logical abstraction.

A promising application field for Data Analysis is to study economic phenomena, where it has offered a novel methodological approach in economic thinking. Although based on a synthesis of mathematics – in particular statistics – and economic theory, just like econometrics, a much different viewpoint is proposed. In this approach, the economy is studied as a dynamic system, utilizing concepts borrowed from physics, such as energy and entropy.

Econometric methods, although successful in a number of goals, are still facing several problems, limiting their effectiveness. Among these problems is the instability of statistical data, the interference of human behavior and the element of unpredictable events inside the economic mechanisms, issues that compromise the certainty and precision of the results. It is worth noting that in all econometric models, a common condition is the well known "ceteris paribus". Methods like least squares, commonly applied in estimating economic information, are considered by several researchers as insufficient, since they result in reliable indicators only under strict preconditions, which are often not satisfied in practice. The above considerations are strong arguments for the potential of non-parametric statistical methods to the analysis of economic phenomena. In particular, methods in the family of Multidimensional Data Analysis are proposed as valuable tools to anyone dealing with economic research, since the factors under study are multiple, mostly qualitative and not directly measurable. This means that, although these factors can hardly be modeled according to the principles of economic theory, their structure can be revealed using the systemic approach inherent in Data Analysis.

2. Concepts and definitions from Physics introduced in Data Analysis

2.1 Mass and Distance

Another fundamental concept used in Data Analysis, and in particular in Correspondence Analysis, is the concept of mass. If our data are in the form of a contingency table T(n x p), each cell k_{ij} contains the number of individual statistical units characterized by the i^{th} category of variable I and j^{th} category of variable J. Each row and column of table T corresponds to the profile of a category that is a vector in the p-dimensional space (for rows) or n-dimensional space (for columns). The mass corresponding to the profile of a category is defined for any row or column as follows:

Mass of the i^{th} row = Marginal frequency of the i^{th} row/Grand total $=k_{i+}/k$ (1)
where

$$k_{i+} = \sum_{j=1}^{p} k_{ij} \text{ and } k = \sum_{i=1}^{p}\sum_{j=1}^{n} k_{ij}$$

Similarly,

Mass of the j^{th} column = Marginal frequency of the j^{th} column/Grand total $=k_{j+}/k$ (2)

A variant of Euclidean distance, called the Chi-square distance x^2, is used to measure the distances between profile points. The distance between two rows *i* and *i'* is given by:

$$d^2(i,i') = \sum_{j=1}^{J} \frac{1}{k_{+j}} \left[\frac{k_{ij}}{k_{i+}} - \frac{k_{i'j}}{k_{i'+}}\right]^2 \quad (3)$$

The distance between two columns is also defined in a symmetric fashion. The x^2 distance differs from the usual Euclidean distance in that each square is weighted by the inverse of the frequency corresponding to each term. The division of each squared term by the expected frequency is normalization in terms of variance and compensates for the larger variance in high frequencies. If no such normalization was performed, the differences between larger proportions would tend to be large and thus dominate the distance calculation, while the differences between the smaller proportions would tend to be swamped.

2.2 Inertia

Inertia is a term borrowed from the "moment of inertia" in mechanics. A physical object has a center of gravity (or centroid). Every particle of the object has a certain mass *m* and a certain distance *d* from the centroid. The moment of inertia of the object is the quantity md^2 summed over all the particles that constitute the object. This concept has been introduced in correspondence analysis, where there is a cloud of profile points to which we can assign masses. A centroid can be defined for this cloud (seen as a system) using the definitions of mass and x^2distance from the above paragraphs. The inertia attributed to a profile point can be computed by the following formula.

For the i^{th} row profile,

$$Inertia = m_i \sum_j \frac{(r_{ij} - \bar{r}_j)^2}{\bar{r}_j} \quad (4)$$

where

$$r_{ij} = {k_{ij}}/{k_{i+}} \text{ and } \bar{r}_j = {k_{+j}}/{k}$$

The inertia of the j^{th} column profile is computed similarly. The total inertia of the contingency table is given by:

$$\Phi^2 = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \quad (5)$$

Formally, we can write inertia as the weighted sum of the x^2 distance between each profile and the mean profile. The greater the inertia, the greater the association between row and column (the distance from the mean). Inertia can be as low as 0 (no association) and as high as the rank of the table (perfect association of each line with each column).

2.3 Information, energy and system entropy

Energy can be considered as a common denominator in a system, since each object can be expressed in terms of the quantity of energy that it includes. Therefore, any productive asset is characterised by the work it can produce, expressed as energy content, and operates by transferring energy to other objects. The entropy of a system is also a dominating concept in thermodynamics. This concept defines the energy state of a system and enables us to express the differentiation between useful exchanges of energy (work) and dissipated energy that its loss is not reversible. In the same sense, any economic activity within a system can be linked with the production of work and therefore the transfer of energy. When studying economic phenomena, an analogy can be found between and energy and information [8].

As information we define the quality variable which determines the position or state of a system within time. The information can be elementary or complex. The evolution of a phenomenon can be represented by a sequence of events, some of which are less and others more probable. The occurrence of each event is a carrier of a certain quantity of information, related to the total set of the possible events. Specifically, the quantity of information carried by the occurrence of an event, expressed as an ascending function H(x), can be defined according to equation of Hartley [9] as follows:

$$H(x) = log_2 K \quad (6)$$

H(x) determines the quantity of information that is necessary to specify an element among K elements in a set of possibilities. This quantitative definition of information, established in classic information theory, is focused on measuring the quantity of information units or the strength of a message transmission by reducing information to a binary set of symbols. Therefore, the informational content of a message is only related to its structure and not its meaning. This is in contrast to the qualitative definitions of information found in the humanistic information sciences, which are concerned with the meaning and interpretation of the information carried by a message.

Let us consider the case of an attribute I with N categories. Each category corresponds to a class E and each class E_i (i=1,...N) is a set of k_i elements. Then, the information related to the set E that is provided by the attribute I is given by

$$H(I) = \sum_i p_i \times log_2 \frac{1}{p_i} \quad (7)$$

Where

$p_{i=k_i}/_k$ and $k = \sum k_i$

The above equation corresponds to Shannon's noiseless coding theorem [10], which quantifies the resources needed to store or transmit a given body of information related to a phenomenon that consists of k independent elementary possibilities. The quantity H(I) is called entropy [9] of the partitioning of E, which does not depend on the nature of the attribute neither on the type of classes but solely on the distribution of frequencies p_i. It is also seen that it is always H(I)≥ 0 and the entropy is maximum when the elements to which the attribute I is partitioned have the minimum possible predictability. It is therefore apparent that the evolution of a phenomenon to the state of maximum entropy leads it to disorder, the latter being strongly related to the notion of non-predictability.

The information entropy defined in (7) is in close resemblance to the entropy of a thermodynamic system, which has been quantified by Boltzmann using the formula:

$$S = -k \cdot \sum p(i) \cdot log_e p(i) \quad (8)$$

where k is the Boltzman constant.

The entropy S, as introduced by Clausius [11] expresses the ability of the system to produce useful work. In a thermodynamic phenomenon, the variation of entropy dS can be written as:

$$dS = d_e S + d_i S \quad (9)$$

The term d_eS expresses the exchanges of energy between the system and the environment that correspond to reversible processes and may be positive or negative according to the direction of the exchanges. The term d_iS refers to the non-reversible processes inside the system which lead it to disorder. The variation of entropy d_iS is always positive or zero, meaning that there is a spontaneous tendency by the system to move to a state of higher entropy which means of higher disorder.

In order to apply a "thermodynamic" approach to study a real-world economic phenomenon, we need to consider that whenever the system under study is moved from a state of stability, it evolves by producing events (information), following a spontaneous course to a state of thermodynamic stability. When it reaches a point of maximum entropy, it becomes stationary and is characterized by inability to produce useful information but only gatherings of unstructured and unusable data, which is the analogous of heat dissipation. In such a state, the system "forgets" its history and its future course is unpredictable. Following the same thinking, in order to bring an open system (i.e. an economic phenomenon) to a state of higher ordering and enable it to produce useful information, one would need to reduce its entropy. This is possible by applying a process that exchanges energy (i.e. information) with its environment and contributes negative variation to the system entropy ($d_eS<0$). Such a process builds structure in the

system but is inevitably followed by an increase of the entropy of the environment. Whenever a system is moved from the state of maximum entropy, the introduced structuring becomes of high importance, as it is related with higher ability to provide useful information. The entropy can in general be considered as an estimation of the reduction of a system's energy and at the same time of the degree of its disorder.

What may then a researcher do in order to study an economic phenomenon? Considering the phenomenon as an open system, since it is difficult to measure directly the quantity of information available in it, a researcher is able to observe the production of information and the changes in entropy as the phenomenon evolves and to evaluate the effect caused by certain structural changes that tend to move the system from the state of maximum disorder. In this framework, it is argued that the most promising way to approach the economic reality is the systemic one, studying the structures and relations in the market, considering also human reactions and the related ordinance. Factor Data Analysis methods are applications of the above thinking, which is the study of an economic phenomenon as an open system, through the evolution of its structure and the evaluation of the effect that factors of the environment have on this system in terms of information and exchanges of entropy. The family of Factor Analysis includes several methods, the most characteristic one being Multiple Correspondence Analysis (MCA) and others such as Hierarchical Clustering, Cluster Analysis and Discriminant Analysis [12]. These methods are not based on a-priori models but seek to uncover the latent structure of a system described with a set of multidimensional qualitative or quantitative variables, by reducing the attribute space from a larger number of variables to a smaller number of factors. Another important feature of Factor Analysis, is the multivariate treatment of data through simultaneous consideration of multiple categorical variables, revealing relationships that would not be detected in a series of pair-wise comparisons.

3. Application in Market Survey for Internet Radio

3.1 Methods

In order to illustrate the potential of Factor Data analysis, in this paper Multiple Correspondence Analysis [13,14] is applied on primary survey data to capture the profile of potential customers for an original internet radio venture. The aim is to identify the characteristics of the main market segments and thus enable better product design and more accurate sales forecasting. The primary data are collected through a questionnaire-based survey, carried out through the internet in January 2011. The questionnaire contained 11 questions of closed type on the preferences and habits related to radio and internet usage of a random audience in northern Greece.

The collected data were quality checked, coded and introduced for analysis in the form of Initial Data Table in the Data Analysis software M.A.D. [15]. The initial data corresponded to 200 questionnaires, 21 categorical variables and 60 categories. From the generalized contingency table (Burt table), the relative frequencies have been calculated, the profile for each property (category) has been produced as a multidimensional vector and a mass has been assigned according to equation (1). The total information contained in the data has been estimated as the total inertia of the Burt table, according to equation (5). By applying MCA, the total behavior of the respondents has been broken down to several factorial axes, each one corresponding to a partial phenomenon that contained a measurable percentage of the total information. Each factorial axis was represented graphically and interpreted as a pattern that is expressed as a set of homogenous groups of related properties. By combining two factorial axes together, factorial planes were also produced, that allowed the identification of more complex underlying patterns.

3.2 Results

The application of MCA on the Burt Table resulted in a distribution of inertia that was not satisfactory (not high enough at the first factorial axes). Therefore, the analysis has been repeated on a non-symmetrical generalized contingency table (BurtX), where properties corresponding to population characteristics were selected as rows, while properties related to responses were selected as columns. In this case, the percentage of inertia corresponding to the first 2 axes was in total 74%. The factorial plane 1X2, which is formed by these 2 axes is shown in Fig 1. In this graph, the labeled dots mark the projection of the properties (i.e. variable categories), handled as multidimensional vectors, on the factorial plane 1X2 and are clustered manually to groups. Each group of properties corresponds, according to MCA, to a set of correlated factors, which characterize an individual population segment. Four such groups were identified graphically (Fig 1) and interpreted. The most suitable as a target market segment was clearly group C, which is interpreted as a group of mostly women, who listen to the radio many hours per day, their age is from 15-19 or 19-24, are informed about internet radio devices and are interested in buying one, listen to both frequency and internet radio, use the internet more than 12 hours per week and prefer international rock music.

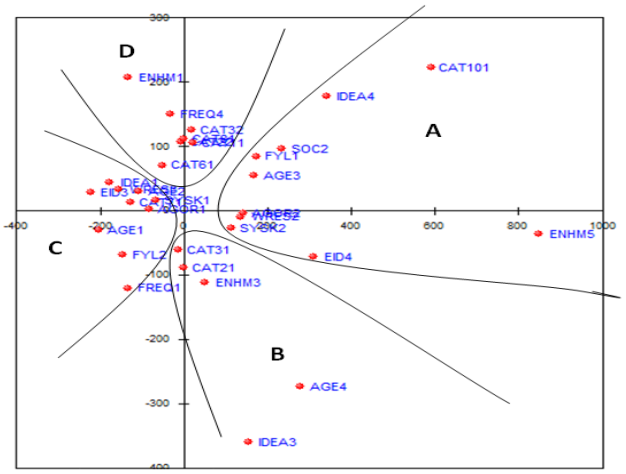

Fig. 1. Factorial plane 1 X 2, where 4 groups of properties are identified.

On the factorial axis 3 (Fig 2), corresponding to 10.2% of inertia, the one segment that is characterized by the properties projected on the negative side of the axis is interpreted as people who find the idea of interactive internet radio interesting and would buy an internet radio device, use the internet up to 4 hours per week, listen to the radio less

frequently and mostly listen to various Greek music, as opposed to people who – according to the properties projected on the positive side - are not willing to buy a radio device, who listen to the radio more frequently and prefer international rock music. After identifying the profile of the target market segment, the percentage of the respondents with profile similar to the one of the selected segment has been estimated by applying hierarchical clustering on the statistical units. The result was 24% of the sample, which can be used as a basis for sales forecasting.

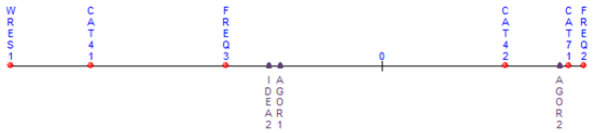

Fig. 2. Factorial axis 3. Properties in each side of zero characterize 2 opposed population segments.

4. Conclusion

The methods of Factor Data Analysis and in particular Multiple Correspondence Analysis are valuable tools in the hands of a researcher trying to analyze any economic phenomenon and they provide a powerful alternative to classic statistics and model-based econometrics. Their main advantages are that they can handle complex phenomena by breaking them to their structural elements and that they are not based on a priori assumptions or models that could restrict their effectiveness. The application of MCA to a market segmentation problem illustrates the ability of the method to handle qualitative data and to identify complex relations among variables. The result is a complete profiling of compact groups that can be addressed as target market and also the quantitative estimation of the market segment size.

References

1. C. Spearman, The proof and measurement of Association between two Things, Amer Journal Psychology, 15 72 (1904).
2. L. L. Thurstone, Multiple Factor Analysis, Chicago University Press (1947).
3. H. Hottelling, Analysis of a Complex of Statistical Variables into Principal Components", J. Educ. Psy., 24 417 (1933).
4. H. O. Hirschfeld, A connection between correlation and contingency, Cambridge Philosophical Soc. Proc. 31 520 (1935).
5. R. A. Fischer, The precision of discriminant functions, Annals of Eugenics, 10 429 (1940).
6. J.-P. Benzécri, L'analyse des correspondances. Paris: Dunod, (1973).
7. J. Tukey, The Future of Data Analysis, Ann. Math. Statist., 33 (1) 1 (1962).
8. R. Passet, , L' economie et le vivant, 2me editions Economica, Paris (1996).
9. M. Volle, Analyse des Donnes, Economica Paris (1985).
10. C. E. Shannon, The Mathematical Theory of Communication, Bell System Technical Journal, 47 379 (1948).
11. I. Prigogine, Introduction a la theromodynamique des phenomenes irreversibles, Dunod Paris (1967).
12. D. Karapistolis, Data Analysis – Statistics without models, Vol I, Thessaloniki, Anikoula (in Greek) (2008).
13. M. Greenacre, Correspondence Analysis in Practice, Chapman & Hall, (2007).
14. J.-P. Benzecri, Correspondence Analysis Handbook. New-York: Dekker, P., (1992).
15. D. Karapistolis, The MAD software (in Greek), Data Analysis Bulletin, 2 133 (2002).