Prof. G. Ruxanda Ph.D.
Academy of Economic Studies-Bucharest

CAUSALITY SIMPLIFICATION IN DATA ANALYSES

1. Causality simplification by means of main components analyses

One of the most efficient methods of multidimensional data analyses which has multiple and various possibilities of applications in the economic field are the method of ***main components analyses***.

Many problems of economical analyses require the investigation of the functional links between certain explanatory variables which can define a special ***causality space***.

The functional links between explanatory variables can be shown by defining a ***structure of dependence*** of these variables that are usually measured by means of the coefficients of covariance and correlation. The necessity of pointing out the dependence structure appears in all the analysis which have as a goal the causality investigation.

In order to obtain a causality suggestive interpretation it is necessary that dependence structure of a set of causative variables to be simple and clear expressed. A simple and clear evidence of a dependency's structure is one of the main goals of data analyses.

Most of the times, because of the complexity of the causative links and of the specific nature of the information which characterize the analysed variables, simple, clear, synthetic and unredundanted expression of the causality are very difficult and it cannot be directly obtained. The achievement of this desideratum suppose the use of some specific instruments which can be used to achieve a certain *simplification of the causative space* determined by the explanatory variables. In this resulted causative space it is easier to show the morphology of the causative links and to express an adequate structure of dependencies.

Such a simplification is equivalent to an important diminution of the causative space dimension, *a passing from space with a lot of dimensions to a space with a small number of dimensions* which provides a large percent of *conservation* of the information contained in the initial causative space.

The orientation of the axes of the smaller reduced space *has to permit a unredundanted decomposition of the variability of the initial causative space in factors of influence, an expression of the dependence's structure without any superposition of the individual contribution of the causative factors.*

It is obvious that to obtain such a simplification it cannot happen with a dimension diminution of the initial causative space made simplistic, based on the renunciation of one of the causative variables and the restraint of another causative variables, especially when the aprioristic importance of the variables seems to be approximately equal.

The only possibilities in making such a simplification under the mentioned restrictions is the *definition of new variables* less in number than the initial variables which can express in a consistent way the informational significance of the original variable and which can provide "the conservation "of the information included in the initial variables.

2. Main components definition

One of the most rigurous way of decreasing the dimension of a special determination space is the one based on the analyses of the main components.

Definitely, the analyses of the main components is based on a technique which is based on the calculation of the projections of every point from the initial space determined by the original variables which are analysed, on the axes of a space with a smaller dimension.

The analyses of the main components, *is a method of multidimensional analyses which has as a goal the determination of new variables named* ***main components*** *and expressed as linear combinations of original variables; this way these variables are characterized with maximum variability.*

The main components *are vectorial abstract variables defined as linear combination of original variables having the following two fundamental properties:*

-they are uncorrelated two by two and the squares sum of the coefficients which defined the linear combination that corresponds to the main component is equal to one; the verification of this property by the linear combinations coefficients defining the main components, determine that these coefficients, under a vectorial form, to form ***orthonormal system****;*

-the first main component is a normalized linear combination whose variability is maximum, the second main component is an uncorrelated linear combination with the first main component and it has variability as large as possible but smaller than that of the first component.

Thus, main components are orthogonal vectors which take as much as possible from the variation of the original vectorial variables: the first main component takes the possible maximum from the variation of the original variables, the second main component takes the maximum from the remaining variation(after the variation taken by the first component is eliminated).

We can say that the main component analysis is a method of re-expression of the original variables with a small number of new variables, named

main components, which are linear combinations of a maximum variation of the original variables.

3. The mathematical model of main components analysis

In order to state the mathematical model as the base of the main components analysis we will consider that the initial causative space which is investigated is determined by a number of n explanatory variables: x_1, x_2... x_n. We will also consider that the variables x_1, x_2... x_n are random variables which are normally redistributed, the average and the variation of every variable x_i being μ_i, respectively σ_i^2.

Considering that these random variables are the components of the n-dimensional random vector x and supposing that this has a various normally distribution of the average μ and of the matrix of covariance Σ:

$$x \sim N(\mu, \Sigma),$$

we can consider those n main components of that causative space as a n-dimensional vector ,named w:

$$w = \begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{pmatrix}$$

Every coordinate w_i of this vector represents a main defined component, related to the original variables, by means of the linear combinations:

$$w_i = \alpha_1^{(i)} x_1 + \alpha_2^{(i)} x_2 + \cdots + \alpha_n^{(i)} x_n \qquad i = 1,2...,n$$

The main component determination w_i requires the determinations of the coefficients $\alpha_j^{(i)}$ which define the linear combinations corresponding to this main component. The coefficient $\alpha_j^{(i)}$ is precisely the coordinate of the vectors corresponding to covariance matrix of the original variables x_1, x_2... x_n and the variation of the main component are values of this matrix

The definition of the main component says that the determination of these coefficients has to be made in a way in which the main component w_i has the maximum variation.

Considering that the n coefficients $\alpha_1^{(i)}, \alpha_2^{(i)}, ..., \alpha_n^{(i)}$ of the above linear combination are the coordinates of the n-dimensional vector $\alpha^{(i)}$ we can define the main component w_i as:

$$w_i = \left(\alpha^{(i)}\right)' x \qquad i = 1,2,...,n$$

where the components of the vector $\alpha^{(i)}$ maximize the variances of the main component w_i .

Because of the fact that the main component w_i is a linear transformation of the elements of the vector x supposed to be normally distributed with the average μ and the matrix of covariance Σ, it results that the main component is a random variable normally distributed with the average and variation given by these equations

$$E(w_i) = \left(\alpha^{(i)}\right)' \mu \qquad \text{Var}(w_i) = \left(\alpha^{(i)}\right)' \Sigma \alpha^{(i)}$$

that means:

$$w_i \sim N\left[\left(\alpha^{(i)}\right)' \mu, \left(\alpha^{(i)}\right)' \Sigma \alpha^{(i)}\right] \qquad i = 1,2,...,n$$

If we consider that the vectors $\alpha^{(i)}$ represent the columns of matrix U of n×n dimension as:

$$U = \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \cdots & \alpha_1^{(1)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \cdots & \alpha_2^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_n^{(1)} & \alpha_n^{(2)} & \cdots & \alpha_n^{(1)} \end{pmatrix},$$

and x is the vector whose coordinates are the original variables x_1, x_2... x_n and w is the vector whose coordinates are the main components w_1, w_2... w_n ,then, the linear combinations which define the main components will be considered as:

$$\begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{pmatrix} = \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \cdots & \alpha_1^{(1)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \cdots & \alpha_2^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_n^{(1)} & \alpha_n^{(2)} & \cdots & \alpha_n^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix},$$

and the mathematical model of the main components analysis, can be defined as:

$$\begin{cases} U'x \\ Var(w) \rightarrow \max \end{cases}$$

In what follows, we will consider that those n columns of the matrix U ,in fact, represent the normalised vectors of the matrix of covariance Σ and the maximum variation of the main components w_i is represented by the values of the same matrix.

As we mentioned earlier, the n main components of the causative space determined by the original variables $x_1, x_2 ... x_n$ are defined by the linear combinations :

$$w_i = \alpha_1^{(i)} x_1 + \alpha_2^{(i)} x_2 + ... + \alpha_n^{(i)} x_n ,$$

where i= 1,2...n whose weight $\alpha_j^{(i)}$ is determined in such a way to maximise the variation of the main components w_i.

The variation of the generic main component, w, can be expressed as:

$$\text{Var}(w) = \alpha' \sum \alpha$$

The determination of the unknowns represented by the components of the α vector in such that the variation of the main component w, Var(w) ,to be maximum,it is obviously an undetermined problem.

In order to make sure that this problem is well determined , we have to impose a constraint for the unknown vector α , the most natural constraint being the condition that the length of the α vector to be equal to 1:

$$\alpha' \alpha = 1$$

The proceeding relations allow us to reduce the problem of the determination of the α vector components , to the following extreme problem with links:

$$\begin{cases} \max\limits_{\alpha} \ \alpha' \sum \alpha \\ UR: \ \alpha^t \alpha = 1 \end{cases}$$

where the unknowns are the components of the α'.

The Lagragean problem associated to the extreme problem and used to solve indirectly this last problem has the following form:

$$L(\alpha, \lambda) = \alpha' \sum \alpha - \lambda(\alpha' \alpha - 1)$$

The necessary extreme conditions for this Lagragean function are given when the partial differentials are null with respect to the components of α vector and with λ multiplier, respectively:

$$\begin{cases} \dfrac{\partial L(\alpha, \lambda)}{\partial \alpha} \\ \dfrac{\partial L(\alpha, \lambda)}{\partial \alpha} \end{cases} \text{ adica } \begin{cases} 2\sum \alpha - 2\lambda\alpha = 0 \\ \alpha' \alpha - 1 = 0 \end{cases}$$

From the first extreme condition , which can be written under the following form:

$$\sum \alpha = \lambda \alpha ,$$

it results that the solution $\tilde{\alpha}$ of the extreme problem is one of the vectors of the matrix Σ the one associated to the $\tilde{\lambda}$ value of this matrix.

We can also notice that the maximum value of the square form $\alpha' \sum \alpha$ is for the point of extreme $\tilde{\alpha}$,equal to $\tilde{\lambda}$:

$$Var(\tilde{w}) = \tilde{\alpha}' \sum \tilde{\alpha} = \tilde{\lambda}$$

These results show that the determination of linear combinations coefficients that define the main component w, in the conditions of maximising the variation of this main component, is equivalent to choosing from those n values of the matrix Σ the bigger one and determining the component of the vector with α weights which define that main component by means of vector calculation associated to that value included in Σ.

Right now we can see that for each proper value λ_i of those n proper values of Σ we have one solution for the maximum problem above , namely one vector $\alpha^{(i)}$ and hence one main component w_i .

Supposing that the n proper values of Σ are ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, the first main component w_1 which have the maximum variation λ_1, is given by the linear combination :

$$w_1 = \left(\alpha^{(1)}\right)' x$$

where $\alpha^{(1)}$ is a vector corresponding to the value λ_1, so it fulfils the relations below:

$$(\Sigma - \lambda_1 I)\alpha^{(1)} = 0 \qquad \text{and} \qquad \left(\alpha^{(1)}\right)' \alpha^{(1)} = 1$$

λ_1 is the root of the characteristic equation:

$$|\Sigma - \lambda I| = 0$$

where I is a notation for the unit matrix.

In this way the main component w_1 has some properties expressed by the following relations:

$$E(w_1) = \left(\alpha^1\right)' \mu \quad \text{and} \quad \text{Var}(w_1) = \lambda_1$$

The determination of main component w_1 will be followed by another one based on a similar logics in order to obtain the second main component . This one has to fulfil the following properties : to have a maximum variation and not to be correlated to the first main component wi . In a similar way, we can determine the other main components w_3, w_4... w_n

References

[1] **Benzecri J. P.**, ***L'analyse des données***, Dunod, Paris, 1979

[2] **Boudon R., Lazarsfeld P.**, ***L'analyse empirique de la causalité*** Ed. Mouton, Paris, La Haye, 1969

[3] **Céhessat R. (ed)**, ***Exercices commentés de statistique et informatique appliquées***, Dunod, Paris, 1976

[4] **Chow G. C.**, ***Econometrics***, McGraw-Hill, New York, 1984

[5] **Desabie J.**, ***Théorie et practique des sondages***, Dunod, Paris, 1966

[6] **Dhrymes Phoebus J.**, ***Econometrics: Statistical Foundation and Applications***, Springer-Verlag, New York, 1974

[7] **Droesbeke J. J., Fichet B.**, ***Modčles pour l'analyse des données multidimensionnelles***, Economica, Paris, 1992

[8] **Harman H. H.**, ***Modern Factor Analysis***, University of Chicago Press, Chicago, Illinois, 1967

[9] **Ruxanda Gh.**, ***Limbajul EMI, Manual de utilizare***, Academia de Studii Economice, Bucureşti, 1996

[10] **Ruxanda Gh.**, ***Analiza multidimensională a datelor***, Editura ALL, Bucureşti, 1998 (în curs de apariţie)

[11] **Rysin J. V. (ed)**, ***Classification and clustering***, Academic Press, New York, 1977

[12] **Saporta G.**, ***Probabilités, analyse des données et statistique***, Ed. Technip, Paris, 1990

[13] **Saporta G.,Ştefănescu V.**, ***Analiza datelor şi informatică*** , Ed. Economică , Bucureşti, 1996

[14] **Stănăşilă O., Neagoe V.**, ***Teoria recunoaşterii formelor***, Editura Academiei, Bucureşti, 1992

[15] **Wonnacott R. J., Wonnacott T. H.**, ***Econometrics***, Wiley, New York, 1970

[16] *** ***SPSS for Windows, Release 6.0, SPSS Inc., USA, 1993***

[17] *** ***Statistica for Windows,*** Release 4.3, I, II and III Vol., StatSoft Inc., Tulsa, OK, 1991