10. USING THE MULTIVARIATE DATA ANALYSIS TECHNIQUES ON THE INSURANCE MARKET

Vasile DEDU*
Daniel ARMEANU**
Adrian ENCIU***

Abstract

In the present financial theory, we confront with complex economic phenomena and activities which cannot be studied or analyzed profoundly because of the plurality of existing variables, ratios and information. The economic, financial and social activity carried on under crisis or economic growth conditions registered year by year a development of the products and instruments in use. The complexity of the economic area may be simplified through techniques of multi-dimensional analysis. Such a method is the analysis of the principal components which allows the decreasing of the initial causal space dimension generated by the functional links which are established among the initial explanatory variables. The dimension of this space is determined by the number of explanatory variables identified as causes of the economic phenomenon and the higher their number, the more difficult it is to analyze the initial causal space because the information volume, the complexity of calculations, the risk not to identify the contribution of each variable to the creation of the initial causal space variability and the decrease in the initial variables significance in case they would be inter-correlated grow. The simplification of the initial causal space means the determination of a change which consists in transition from a space with a large number of variables to another one of fewer dimensions, equivalent but on the conditions of keeping maximum information from the initial space and maximizing the variability of the new space (called principal space). Variables from the principal space represent the principal components, they are un-correlated and the vectors which define them have a unitary length.

Keywords: original variables, covariance matrix, eigenvalue, eigenvector, principal components, total variance, generalized variance, factor matrix, factor loadings, factor scores, classification

JEL Classification: G15, G22, C44

* Professor, Academy of Economic Studies, Bucharest.
** Professor, Academy of Economic Studies, Bucharest.
*** Lecturer, Ph.D., Academy of Economic Studies, Bucharest.

We defined the principal components **(**PC) as standardized linear combinations(SLC) of the original variables. Therefore, if $\mathbf{z_i}$ is principal component **i**, $\boldsymbol{\beta}^{(i)}$ the vector whose elements define SLC **i** and $\mathbf{x_j}$ the original variables, $j = 1,2,...,n,$ we have:

$$z_i = \beta_1^{(i)} \cdot x_1 + \beta_2^{(i)} \cdot x_2 + ... + \beta_n^{(i)} \cdot x_n = \sum_{j=1}^{n} \beta_j^{(i)} \cdot x_j \quad (1)$$

Let $z = \begin{pmatrix} z_1 & z_2 & \dots & z_n \end{pmatrix}^t$ be the vector of the ***n*** PCs, $B = \begin{pmatrix} \beta_1^{(1)} & \beta_1^{(2)} & \dots & \beta_1^{(n)} \\ \beta_2^{(1)} & \beta_2^{(2)} & \dots & \beta_2^{(n)} \\ \vdots & \vdots & \dots & \vdots \\ \beta_n^{(1)} & \beta_n^{(2)} & \dots & \beta_n^{(n)} \end{pmatrix}$ be an ***nxn*** matrix consisting of column vectors and $x = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}^t$ be the vector of the ***n*** original variables. In terms of matrix algebra, equation (2) can be written as:

$$z = B^t \cdot x \quad (2)$$

Let **z** and **β** be generic notations for a PC and a vector whose elements define a SLC, respectively. In order to determine the PCs, we must maximize the variance of each PC. This is done by solving the following system:

$$\begin{cases} z = \beta^t \cdot x \\ \max VAR(z) \end{cases} \quad (3)$$

Taking into account that **β** is a SLC, system (4) can be equivalently written as:

$$\begin{cases} \max\limits_{\beta} \beta^t \cdot \Sigma \cdot \beta \\ \beta^t \cdot \beta = 1 \end{cases} \quad (4)$$

where **Σ** is the covariance matrix of the original variables.

System (5) will be solved using the Lagrangian approach. Thus, we can define:

$$L = \beta^t \cdot \Sigma \cdot \beta - \lambda \cdot \left(\beta^t \cdot \beta - 1\right) \quad (5)$$

The first-order conditions for maximization are:

$$\begin{cases} \dfrac{\partial L}{\partial \beta} = 0 \\ \dfrac{\partial L}{\partial \lambda} = 0 \end{cases} \Leftrightarrow \begin{cases} 2 \cdot \Sigma \cdot \beta - 2 \cdot \lambda \cdot \beta = 0 \\ \beta^t \cdot \beta - 1 = 0 \end{cases} \Leftrightarrow \begin{cases} \left(\Sigma - \lambda \cdot I_n\right) \cdot \beta = 0 \\ \beta^t \cdot \beta - 1 = 0 \end{cases} \quad (6)$$

where $\mathbf{I_n}$ is the identity matrix.

It can be easily proven that the normated eigenvectors of matrix **Σ** satisfy system (6). The variance of each PC, given by the bilinear form $\beta^t \cdot \Sigma \cdot \beta$, will equal the corresponding eigenvalue of matrix **Σ**. It is obvious that **Σ** has ***n*** eigenvalues

$\lambda_1, \lambda_2, ..., \lambda_n$; rearranging them so as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$, the following inequality holds:

$$VAR(z_1) = \lambda_1 \geq VAR(z_2) = \lambda_2 \geq ... \geq VAR(z_n) = \lambda_n \quad (7)$$

Consequently, increasing the number of PCs retained will result in a greater share of variance recovered from the original variable space. It is worth mentioning that the PC transformation preserves the **total variance** ($\mathbf{V_T}$) from the initial space:

$$V_T = \sum_{i=1}^{n} VAR(x_i) = \sum_{i=1}^{n} VAR(z_i) = \sum_{i=1}^{n} \lambda_i \Leftrightarrow tr(\Sigma) = tr(K) \quad (8)$$

where **K** is the covariance matrix of the PCs. **K** is a diagonal matrix[1]:

$$K = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix} \quad (9)$$

PCA also results in the preservation of generalized variance ($\mathbf{V_G}$):

$$V_G = |\Sigma| = |K| \quad (10)$$

For the reasons showed below, it is considered that the analysis of the principal components is a simplified re-expressing of the initial causal space, the simplification being made on the conditions of maximization of the information quantity from the original space.

We used the principal components method in order to analyze the activity of insurance company branches, for the purpose of determining the best branch of the company. We took into consideration eight ratios relevant for the activity of branches (subscribed net premium minus cancelled premium – PRI_NET, total income – VEN_TOT, administrative expenses – CH_ADM, acquiring expenses – CH_ACH, net indemnification – DESP_NET, number of employees – NR_ANG, premium net reserve – REZ_NE_P, net claims reserve – REZ_NE_D and the financial result – REZ_FIN).

In the beginning, we have calculated the average and the standard deviation for the considered ratios:

Table 1

Average and the standard deviation for the considered ratios

	Average	Standard deviation
PRI_NET*	3882.834	2518.938
VEN_TOT	4632.147	3080.051
CH_ADM	847.484	771.328
CH_ACH	1338.690	855.323

[1] As a matter of fact, PCA basically consists of diagonalizing a symmetrical matrix (in our case Σ). Matrix K is computed as $B^t \cdot \Sigma \cdot B$.

	Average	Standard deviation
DESP_NET*	3299.661	2419.452
NR_ANG	29	17
REZ_NE_P	1666.036	1245.263
REZ_NE_D	1347.465	1448.332
REZ_FIN	-899.070	1316.304

The higher the standard deviation is, the more different the companies' branches are. In our case, the deviation has high values for all the considered ratios and this means the branches are very different from each other, forming a space with high variability where the causal dependencies structure is very complex and very difficult to be visualized. It is worth noticing that for the financial result ratio the average is negative, and that means the activity carried on in 2007 was materialized in loss.

The table below is the correlation matrix for the nine original variables (considered ratios). Obviously, the elements located on the principal diagonal are equal with the unit:

Table 2

The correlation matrix for the original variables

	PRI_ NET*	VEN_ TOT	CH_ ADM	CH_ ACH	DESP_ NET*	NR_ ANG	REZ_ NE_P	REZ_ NE_D	REZ_ FIN
PRI_NET*	1.00	0.99	0.77	0.96	0.93	0.68	0.93	0.60	-0.35
VEN_TOT	0.99	1.00	0.79	0.98	0.93	0.68	0.94	0.64	-0.40
CH_ADM	0.77	0.79	1.00	0.83	0.78	0.91	0.82	0.28	-0.12
CH_ACH	0.96	0.98	0.83	1.00	0.92	0.73	0.93	0.67	-0.47
DESP_NET*	0.93	0.93	0.78	0.92	1.00	0.76	0.85	0.57	-0.35
NR_ANG	0.68	0.68	0.91	0.73	0.76	1.00	0.71	0.11	0.01
REZ_NE_P	0.93	0.94	0.82	0.93	0.85	0.71	1.00	0.57	-0.35
REZ_NE_D	0.60	0.64	0.28	0.67	0.57	0.11	0.57	1.00	-0.84
REZ_FIN	-0.35	-0.40	-0.12	-0.47	-0.35	0.01	-0.35	-0.84	1.00

Values over 0.7 (in modules) of the correlation coefficients show a deep correlation among ratios. It is noticed that deep correlations among the majority of the analyzed variables diminish the particular significance of them, on one side, and emphasizes the existence of informational redundancies, on the other side: there is a significant quantity of information dissipated among the variables.

Further on, we shall calculate own values of the correlation matrix. We mention that we are interested only in own values higher than 1, as we consider them useful to be analyzed only the principal components which have an informational content more abundant than the original variables (in standardized form, obviously). The results are shown in the table below.

The following graph shows the eight own values of the correlation matrix; one may notice that the first two are representative, while the next register values which tend to zero.

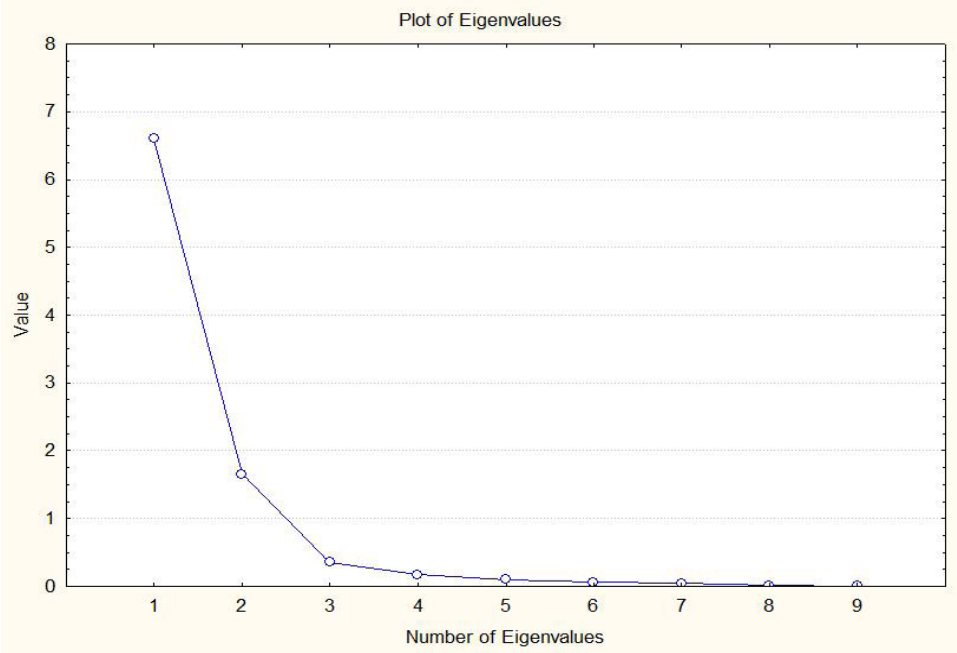
Table 3

Eigenvalues and variance

	Eigenvalues	% Total variance	Cumulated eigenvalues	% Variance cumulated
1	6.602600	73.36222	6.602600	73.36222
2	1.653816	18.37573	8.256416	91.73796

Figure 1

Eigenvalues of the correlation matrix



As one may notice, there are two own values higher than unit and, as a consequence, we shall have two principal components. This explains in a 92% proportion the variability in the initial causal space, which means that the decrease of analyzed space size from nine to two variables was realized on the conditions of an informational loss of 8%, a percentage that we can considered very good. Therewith, we can notice that the first characteristic taken individually synthesizes 73% of the information contained by the nine initial ratios and, thus, it can be subsequently used in order to make a pertinent classification of the branches.

We determined the factor matrix for the two representative components that resulted. The factor matrix is very important in our analysis since its elements (also known as factors intensities) are correlation coefficients between the original variables and the principal components. The calculation formula for an element of this matrix is:

$$\rho_{ij} = \frac{\sqrt{\lambda_j}}{\sqrt{VAR(x_i)}} \cdot \beta_{ij}, i = 1, 2, ..., n \qquad j = 1, 2, ..., k$$

where k is the number of principal components contained in the analysis.
The matrix factor is:

Table 4

The matrix factor

Indicator	Factor 1	Factor 2
PRI_NET*	-0.965660	-0.023463
VEN_TOT	-0.979685	0.013217
CH_ADM	-0.854459	-0.396520
CH_ACH	-0.990402	0.039537
DESP_NET*	-0.945573	-0.063853
NR_ANG	-0.765581	-0.540190
REZ_NE_P	-0.947269	-0.055641
REZ_NE_D	-0.649682	0.722210
REZ_FIN	0.446871	-0.820813
Variance	6.602600	1.653816
% Total variance	0.733622	0.183757

The first principal component synthesizes 73.36% of the information contained in the original space. It is strongly correlated in the negative way with the ratios subscribed net premium minus cancelled premium, total income, administrative expenses, acquiring expenses, net indemnification, number of employees and premium net reserve, providing important information about the activity volume of the company's branches.

The following table shows the coefficients of the line combinations which define the principal components (own vectors of the correlation matrix), on the basis of which we shall calculate the results of observations in the principal components space.

Table 5

Eigenvectors of the correlation matrix

Indicator	Factor 1	Factor 2
PRI_NET*	-0.146254	-0.014187
VEN_TOT	-0.148379	0.007992
CH_ADM	-0.129412	-0.239760
CH_ACH	-0.150002	0.023906
DESP_NET*	-0.143212	-0.038609
NR_ANG	-0.115951	-0.326632
REZ_NE_P	-0.143469	-0.033644
REZ_NE_D	-0.098398	0.436693
REZ_FIN	0.067681	-0.496315

We shall determine further on the coordinates of the observations in the factors space (principal components):

Table 6

Principal components

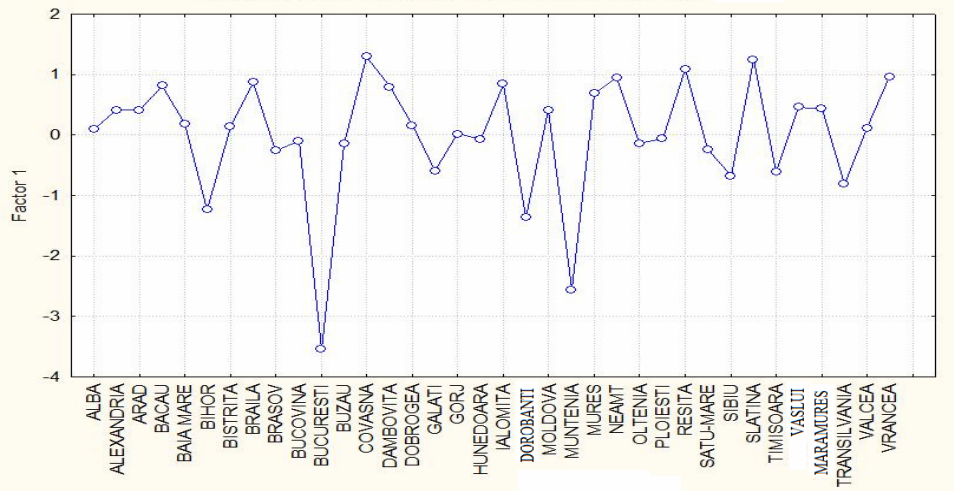
BRANCH	Factor 1	Factor 2
ALBA	0.09344	-0.09959
ALEXANDRIA	0.40922	-0.27503
ARAD	0.41135	-0.02610
BACAU	0.81536	-0.13890
BAIA MARE	0.18777	-0.46544
BIHOR	-1.23334	1.12144
BISTRITA	0.14206	0.29295
BRAILA	0.87456	0.10597
BRASOV	-0.26066	0.08114
BUCOVINA	-0.09538	-0.35187
BUCHAREST	-3.54107	-2.25638
BUZAU	-0.14688	-0.91079
COVASNA	1.29346	-0.01787
DAMBOVITA	0.79434	-0.20405
DOBROGEA	0.15797	-0.13023
GALATI	-0.59249	0.05738
GORJ	0.01358	-0.59256
HUNEDOARA	-0.06593	-0.30123
IALOMITA	0.84471	-0.00995
DOROBANTI	-1.36629	5.04952
MOLDOVA	0.41066	-0.20579
MUNTENIA	-2.55893	-0.69534
MURES	0.69032	-0.04918
NEAMT	0.95073	0.00248
OLTENIA	-0.13759	0.34542
PLOIESTI	-0.05952	-0.18522
RESITA	1.09052	-0.25960
SATU MARE	-0.23877	0.16928
SIBIU	-0.67635	-0.04648
SLATINA	1.24570	-0.06519
TIMISOARA	-0.61038	-0.09203
VASLUI	0.45854	-0.23640
MARAMURES	0.43325	0.06636
TRANSILVANIA	-0.81292	0.61249
VALCEA	0.11693	-0.15092
VRANCEA	0.96206	-0.13828

As we could see below, the first principal component synthesizes the majority of the initial ratios and offers important information about the volume activity of the branches.

Therefore, we made a classification of the branches according to this new characteristic. The following graph synthesizes the classification results.

Figure 2

The classification of branches by the first principal component



We point out that the first principal component is strongly negatively correlated with the ratios of subscribed net premium minus cancelled premium, total income, administrative expenses, acquiring expenses, net indemnification, number of employees and premium net reserve. This fact is possible since, as we could see also in the correlation matrix, there are strong correlations between these ratios; intuitively, it is logical that a high volume of net subscribed premium to be followed by high acquiring expenses or high paid indemnification.

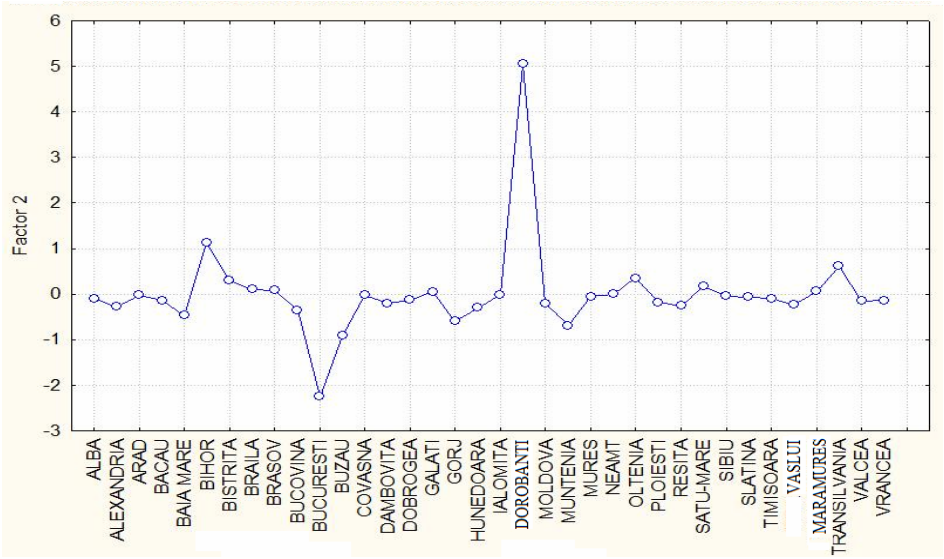
The classification by the first component is very important, since it allows easy appreciation of the turnover of each branch. The strong opposite correlation means that those branches that will register low results as to this factor, will have the highest volume of activity and, reciprocally, the branches with lower values at the first factor will have a low volume of turnover. Thus, we noticed the branches of Bucharest, Muntenia, Dorobanți and Bihor which have the highest values of the above mentioned ratios. The branches with a medium volume of turnover are Alba, Bistrița, Braşov, Bucovina, Galați, Piteşti, Timişoara, Transilvania, etc., while the branches of Bacău, Covasna, Slatina, Ialomița, Neamțt, etc., registered the lowest volume of activity in 2007.

The second principal component synthesizes the ratios of claims net reserve (positive correlation) and the financial result (negative correlation), providing information about the branches profitability as well as the level of the claims reserve. In relation to the results registered for this factor, we distinguish the profitable branches (Buzău), branches with outstanding loss and claims reserves (Dorobanți), branches with

considerable losses and high claims reserve (Braşov, Bucharest, Bihor, Transilvania, Oltenia, Timişoara, etc.), branches with high loss and/or high claims reserve (Dâmbovița, Dobrogea, Galați, Gorj, Muntenia, Brăila, Mureş, etc.) as well as branches with low loss and relatively low levels of reserves (Covasna, Reşița). The classification by the second principal component is:

Figure 3

The classification by the second principal component



In conclusion, we saw that the analysis of the principal components can be a very powerful and useful analytical instrument, which allows us not only to reduce the dimensionality and to eliminate the informational redundancies, but also to visualize more clearlz such causal complex dependencies. Based on this method and techniques specific to factor analysis, we succeed to analyze the branches of an insurance company and to obtain outstanding results. Since the first principal component contains 73.36% of the information included in the original space and it is strongly correlated in a negative way with seven from the nine indicators taken into consideration, we succeed to create a hierarchy of the branches of an insurance company in relation with their financial powers. Thus, through this data analysis method we created a classification of the branches in relation to nine indicators specific to insurance field which reveals their financial power without losing too much information generated by these indicators. This was not possible to create based on initial data. Also, the analysis of the principal components allowed us to make the graphic representation of the observations in a space significantly reduced toward the initial causal space, which simplifies a lot the analysis.

References

Armeanu, Dan, Balu, Florentina, (2007). "Utilizarea analizei componentelor principale pentru identificarea variantei optime de creditare", in *Studii şi cercetări de calcul economic şi cibernetică economică*, 41(4): 123-133.

Anghelache, Constantin, Armeanu, Dan, (2008). "Application of Discriminant Analysis on Romanian Insurance Market"/"Aplicarea analizei discriminante pe piaţta românească a asigurărilor"*, Theoretical and Applied Economics/Economie teoretică şi aplicată*, 11(528): 51-62.

Diamantaras, K. I., (2002). *Neural Networks and Principal Component Analysis*, CRC Press.

Gheorghe, Ruxanda, (2005). "Analiza multidimensionala a datelor", Master Baze de Date – Suport pentru Afaceri.

Gheorghe, Ruxanda, (2001). *Analiza Datelor*, Editura ASE, Bucureşti.

Simar, Leopold, (2004). *Applied Multivariate Statistical Analysis*, Springer.

Spircu, Liliana, (2006). *Analiza Datelor: Aplicatii Economice*, Editura ASE, Bucureşti.

Vintilă, Georgeta, Armeanu, Dan, (2009). "Using the models of data analysis in financial evaluation of companies performances ", in *Metalurgica International*, XIV(7): 195-202.