Application of the Model of Principal Components Analysis on Romanian Insurance Market

▪

Dan Armeanu
Ph.D. Senior Lecturer
Leonard Lache
Candidate Ph.D.
Academy of Economic Studies, Bucharest

Abstract. *Principal components analysis (PCA) is a multivariate data analysis technique whose main purpose is to reduce the dimension of the observations and thus simplify the analysis and interpretation of data, as well as facilitate the construction of predictive models. A rigorous definition of PCA has been given by Bishop (1995) and it states that PCA is a linear dimensionality reduction technique, which identifies orthogonal directions of maximum variance in the original data, and projects the data into a lower-dimensionality space formed of a sub-set of the highest-variance components. PCA is commonly used in economic research, as well as in other fields of activity. When faced with the complexity of economic and financial processes, researchers have to analyze a large number of variables (or indicators), fact which often proves to be troublesome because it is difficult to collect such a large amount of data and perform calculations on it. In addition, there is a good chance that the initial data is powerfully correlated; therefore, the signification of variables is seriously diminished and it is virtually impossible to establish causal relationships between variables. Researchers thus require a simple, yet powerful annalytical tool to solve these problems and perform a coherent and conclusive analysis. This tool is PCA.The essence of PCA consists of transforming the space of the initial data into another space of lower dimension while maximising the quantity of information recovered from the initial space[1]. Mathematically speaking, PCA is a method of determining a new space (called principal component space or factor space) onto which the original space of variables can be projected. The axes of the new space (called factor axes) are defined by the principal components determined as result of PCA. Principal components (PC) are standardized linear combinations (SLC) of the original variables and are uncorrelated. Theoretically, the number of PCs equals the number of initial variables, but the whole point of PCA is to extract as few factors as possible without compromising the variability of the original space. An important property of the PCs is that the first PC is extracted so as to recover the variance from the initial space to the maximum possible extent. The remaining variance is recovered by the next PCs at a declining rate: the variance of the second PC is greater than the variance of the third PC, the variance of the third PC is greater than the variance of the fourth PC and so on.*

Key words: causal space; principal components; eigen values; variance; insurance; factor matrix; generalized variance.

▪

JEL Codes: C21, G22.
REL Codes: 10B, 11C.

As we said earlier, PCA is based on a transformation that links two vector spaces of different dimensions:

$$\Psi : R^n \rightarrow R^k,\ k << n \quad (1)$$

Recall that we defined PCs as SLCs of the original variables. Therefore, if w_i is principal component i, $\alpha^{(i)}$ the vector whose elements define SLC i and x_j the original variables, $j = \overline{1,n}$, we have

$$w_i = \sum_{j=1}^{n} \alpha_j^{(i)} x_j,\ i=1,\ \dots,\ n \quad (2)$$

Let $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}$ be the vector of the *n* PCs, $A = \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \dots & \alpha_1^{(n)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \dots & \alpha_2^{(n)} \\ \dots & \dots & \dots & \dots \\ \alpha_n^{(1)} & \alpha_n^{(2)} & \dots & \alpha_n^{(n)} \end{pmatrix}$ be an *nxn* matrix consisting of column vectors and $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ be the vector of the *n* original variables. In terms of matrix algebra, equation (2) can be written as

$$w = A^t \times x \quad (3)$$

Let w and α be two generic notations for a PC and a vector whose elements define a SLC, respectively. In order to determine the PCs, we must maximize VAR(w). This is done by solving the following system:

$$\begin{cases} w = \alpha^t \times x \\ max\ VAR(w) \end{cases} \quad (4)$$

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

$$\begin{cases} max\ \alpha^t \sum \alpha \\ \alpha^t \times \alpha = 1 \end{cases} \quad (5)$$

where Σ is the covariance matrix of the original variables.

It can be easily proven that the normated eigenvectors of matrix Σ satisfy system (5). The variance of each PC, given by the bilinear form $\alpha^t \sum \alpha$, will equal the corresponding eigenvalue of matrix Σ. It is obvious that Σ has *n* eigenvalues; rearranging them so as $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, the following inequality holds:

$$VAR(w_1) \geq VAR(w_2) \geq \dots \geq VAR(w_n) \quad (6)$$

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* (V_T) from the initial space:

$$V_T = \sum_{i=1}^{n} VAR(x_i) = \sum_{i=1}^{n} \lambda_i \Leftrightarrow tr(\Sigma) = tr(\Lambda) \quad (7)$$

where Λ is the covariance matrix of the PCs. is a diagonal matrix(2):

$$\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$$

PCA also results in the preservation of generalized variance (V_G):

$$V_G = |\Sigma| = |\Lambda| \quad (8)$$

In this paper, we will use the PCA technique to determine the financial strength of the insurance companies that operated on the Romanian market in 2006. We have selected a number of eigth (8) relevant variables(3): gross written premium, net mathematical reserves, gross claims paid, net premium reserves, net claim reserves, net income, share capital and gross written premium ceded in reinsurance.

The mean and standard deviation for each variable

Table 1

	Mean	Std.Dev.
GR_WRI_PRE	143.232.114	228.320.22
NET_M_RES	43.188.939	151.739.60
GR_CL_PAID	65.569.646	125.986.40
ET_PRE_RES	41.186.950	65.331.44
NET_CL_RES	14.726.188	24.805.66
NET_INCOME	-1.711.048	22.982.14
SHARE_CAP	33.006.176	38.670.76
_WRI_PR_CED	37.477.746	93.629.54

As we can see, standard deviations are high for each variable in the model, which means that there is a large quantity of information in the original space. The high levels of standard deviations can be explained taking into account the powerful correlations between the original variables.

The correlation matrix of original variables

Table 2

	GR_WRI_PRE	NET_M_RES	GR_CL_PAID	NET_PRE_RES	NET_CL_RES	NET_INCOME	SHARE_CAP	GR_WRI_PR_CED
GR_WRI_PRE	1.00	0.32	0.95	0.94	0.84	0.22	0.47	0.82
NET_M_RES	0.32	1.00	0.06	0.21	0.01	0.30	0.10	0.01
GR_CL_PAID	0.95	0.06	1.00	0.94	0.89	0.14	0.44	0.81
NET_PRE_RES	0.94	0.21	0.94	1.00	0.88	0.19	0.50	0.68
NET_CL_RES	0.84	0.01	0.89	0.88	1.00	-0.23	0.55	0.63
NET_INCOME	0.22	0.30	0.14	0.19	-0.23	1.00	-0.15	0.21
SHARE_CAP	0.47	0.10	0.44	0.50	0.55	-0.15	1.00	0.14
GR_WRI_PR_CED	0.82	0.01	0.81	0.68	0.63	0.21	0.14	1.00

We now face the problem of information redundancy due to powerful correlations between the variables (e.g. between gross written premium and gross claims paid, between gross written premium and net premium reserves, between gross claims paid and net claim reserves etc.). Recall that we defined the PCs as uncorrelated SLCs(4) of original variables. Therefore, a primary purpose of PCA is to eliminate information redundancy, along with dimensionality reduction.

Assuming the standardization of the initial data(5), we shall now perform PCA via the correlation matrix. It is worth mentioning that only eigenvalues greater than one are of interest(6) because only the PCs with higher variance than the standardized original variables should be extracted.

The eigenvalues greater than one of the correlation matrix

Table 3

	Eigenvalue	Total variance (%)	Cumulative Eigenvalue	Cumulative (%)
1	4.648900	58.11126	4.648900	58.11126
2	1.454291	18.17864	6.103192	76.28990
3	1.014728	12.68410	7.117920	88.97400

The plot of eigenvalues below shows that the other five eigenvalues are negligible:

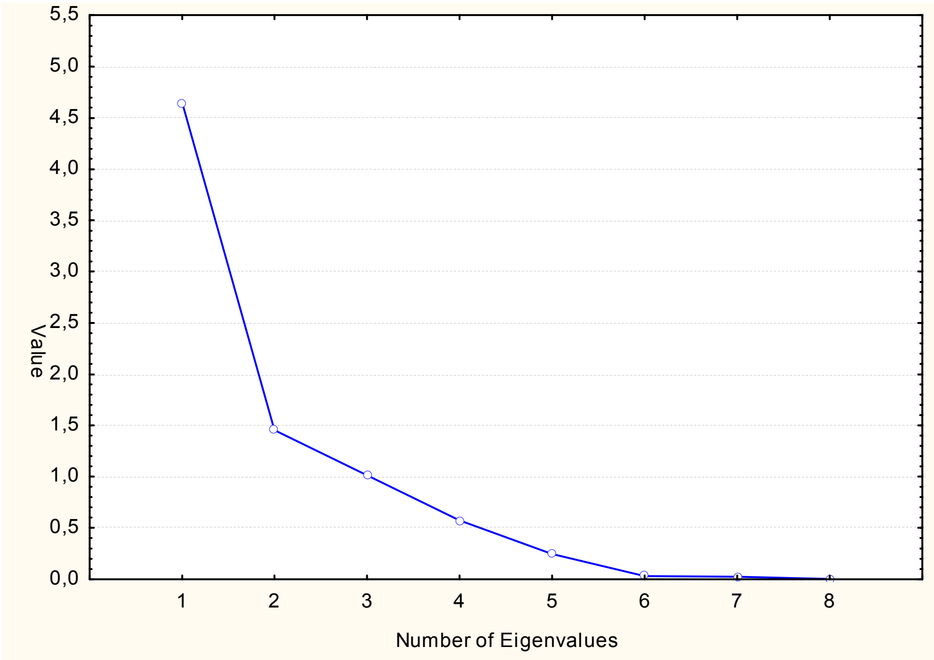

Figure 1. *Variation of correlation matrix eigenvalues*

The correlation matrix has three eigenvalues greater than one. As a result, three PCs have been retained in our analysis. They account for approximately 89% of the variability of the original space, which means that the PCA technique enabled us to transform an 8-dimensional space into a 3-dimensional space losing only 11% of the information contained in the former. It is also worth mentioning that the first PC alone explains over 58% of the variability of the original space, which means that it can be used to perform a pertinent classification of the insurance companies.

Having determined the number of principal components retained, we are now concerned with the interpretation of the PCs. Hence we computed the factor matrix (also known as factor loadings matrix), whose elements (factor loadings) represent the correlation coefficients between the original variables and the PCs. The elements of the factor matrix are computed using the formula

$$\omega_{ij} = \frac{\sqrt{\lambda_j}}{\sqrt{VAR(x_i)}} \times \alpha_i^{(j)}, \quad i = 1, 2, \ldots, n$$

$$j = 1, 2, \ldots, k \qquad (9)$$

where k is the number of retained PCs.

One common problem in PCA is that the unrotated factor matrix often provides unconclusive interpretations. To solve this problem, we determined the rotated factor matrix(7):

The rotated factor matrix

Table 4

	Factor 1	Factor 2	Factor 3
GR_WRI_PRE	0,957868	0,267247	0,032382
NET_M_RES	0,053602	0,923184	-0,044009
GR_CL_PAID	0,985009	0,014747	0,053012
NET_PRE_RES	0,936205	0,202520	0,114957
NET_CL_RES	0,877731	-0,066607	0,416310
NET_INCOME	0,165885	0,464570	-0,747579
SHARE_CAP	0,408991	0,289363	0,712994
GR_WRI_PR_CED	0,875686	-0,123486	-0,273978
Expl.Var	4,499135	1,284163	1,334622
Prp.Totl	0,562392	0,160520	0,166828

The first PC is powerfully correlated with five of the original variables (gross written premium, gross claims paid, net premium reserves, net claim reserves and gross written premium ceded in reinsurance) and recovers 56.24% of the variability from the original space (as opposed to 58% in the unrotated situation). The second and third PCs are powerfully correlated with net mathematical reserves and net income and share capital, respectively. The plot below is a graphical representation of the correlations between the PCs and the original variables.

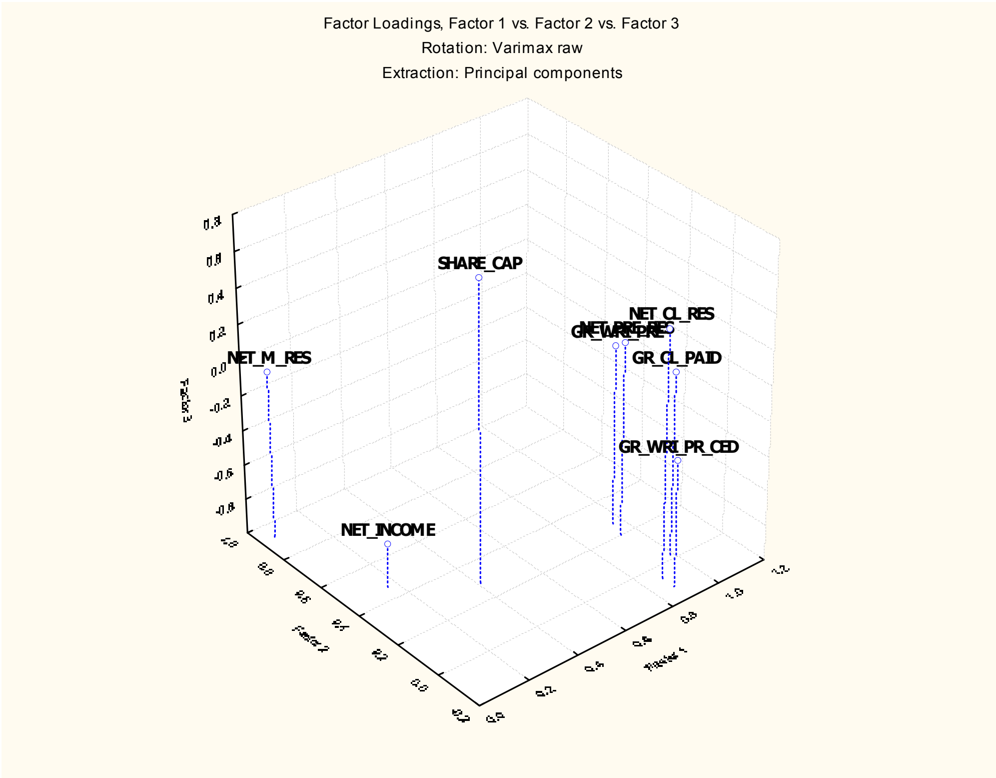

Figure 2. *Correlation between PCs and original variables*

Besides reducing dimensionality and eliminating information redundancy, PCA emphasizes the influence of common, latent (unobserved) factors on the original variables. According to Simar (2003), it is natural to believe that *most of the relevant information carried by a multivariate variable can be summarized by a limited number of latent factors.* In terms of factor analysis, the behaviour of each original variable (also known as indicator or test variable) is guided by three types of factors: *common factors* (which influence all the variables in a model), *individual factors* and *residual factors*. The last two categories influence each variable in an individual manner, thus exerting a specific influence. Equation (10) illustrates the decomposition of the values of indicators (original variables):

$$x_i = \underbrace{\sum_{j=1}^{k} a_{ij} \times f_j}_{\text{common factors' influence}} + \underbrace{a_i \times u_i}_{\text{individual factor's influence}} + \underbrace{\varepsilon_i}_{\text{residual factor's influence}}, \; i = 1,2, \ldots,8 \quad (10)$$

where x_i are the indicators and a_{ij} are the factor loadings (computed as elements of the factor matrix). Therefore, the variance of the original variables can be written as

$$\underbrace{VAR(x_i)}_{\text{total variance of original variable i}} = \underbrace{Communality}_{\text{variance explained by common (latent) factors}} + \underbrace{Individuality}_{\text{variance explained by the individual factor}} + \underbrace{Residuality}_{\text{variance explained by the residual factor}} \quad (11)$$

It should, however, be pointed out that it is difficult to distinguish between the individual and the residual factors. Equation (11) can hence be written as

$$\underbrace{VAR(x_i)}_{\text{total variance of original variable i}} = \underbrace{Communality}_{\text{variance explained by common (latent) factors}} + \underbrace{Specificity}_{\text{variance explained by specific factors (individual and residual factors)}} \quad (12)$$

We can now determine the communalities for each original variable as the sum of squares of the elements on the corresponding line of the factor matrix (that is, the factor loadings). The specificities have been computed as 1 - Communality:

The specificities of the factor matrix

Table 5

Indicator	Communality	Specificity
GR_WRI_PRE	0.989980	0.010020
NET_M_RES	0.857079	0.142921
GR_CL_PAID	0.973271	0.026729
NET_PRE_RES	0.930709	0.069291
NET_CL_RES	0.948162	0.051838
NET_INCOME	0.802217	0.197783
SHARE_CAP	0.759365	0.240635
GR_WRI_PR_CED	0.857138	0.142862

The table shows that the common factors exert a significant influence on the original variables (although this influence cannot be measured directly). In all but two of the cases the communality is above 0.85, the exceptions being net income and share capital. The specific factors account for 20% or more of the variance of these indicators because the respective variables have a somewhat subjective determination, as they can be heavily influenced by management decisions and stockholders' options. The highest communalities pertain to the indicators gross written premium, gross claims paid and net claim reserves. This happens because the behaviour of such variables is much more likely to be determined by general factors (like economic development and the general need for insurance within the economy) than by specific factors.

Recall that we defined the PCs as linear combinations of the original variables. As we saw earlier, the coefficients of the SLCs are given by the normated eigenvectors of the covariance (or correlation) matrix[8]:

The normed eigenvectors of the correlation matrix

Table 6

	Factor 1	Factor 2	Factor 3
GR_WRI_PRE	0.202495	-0.036051	0.119201
NET_M_RES	-0.088258	0.080710	0.764878
GR_CL_PAID	0.236082	-0.055489	-0.093669
NET_PRE_RES	0.196206	0.023689	0.078077
NET_CL_RES	0.182116	0.235296	-0.102060
NET_INCOME	0.063400	-0.553745	0.273611
SHARE_CAP	-0.004514	0.566914	0.290264
GR_WRI_PR_CED	0.258101	-0.324141	-0.240419

The factor scores

Table 7

	Factor 1	Factor 2	Factor 3
ABC ASIGURĂRI	-0.528918	-0.33091	-0.28589
AGRAS	-0.467755	-0.37092	-0.29948
AIG LIFE	-0.248735	-0.55629	1.05069
AIG ROMÂNIA	-0.251086	-0.67712	-0.43519
ALLIANZ-ȚIRIAC	4.453070	-2.14199	-0.54004
ARDAF	0.260885	3.98877	-1.57900
ASIBAN	1.471385	0.89075	0.34927
ASIMED	-0.547944	-0.49811	-0.34818
ASIROM	1.782933	0.36140	0.99269
ASIROM CONCORDIA	-0.548484	-0.40579	-0.30114
ASITO KAPITAL	-0.417567	-0.48417	-0.34874
ASITRANS	-0.274011	-0.35407	-0.36907
ASTRA	0.374542	2.42955	1.07949
ATE INSURANCE	-0.556508	-0.19879	-0.22494
AVIVA	-0.542741	1.23577	0.60924
BCR ASIGURĂRI	1.565687	-0.89718	-0.51044
BCR ASIGURĂRI DE VIAȚĂ	-0.424433	-0.31347	0.08715
BT ASIGURĂRI TRANSILVANIA	0.326596	0.73905	0.00527
CARPATICA ASIG	-0.078883	-0.26261	-0.35684
CERTASIG	-0.550829	-0.41410	-0.33531
CITY INSURANCE	-0.547583	-0.38413	-0.31154
CLAL ROMÂNIA	-0.570951	-0.18724	-0.35681
DELTA	-0.495821	-0.64218	-0.00517
DELTA ADDENDUM	-0.537575	-0.37856	-0.31518
EUROASIG	-0.541163	-0.37547	-0.31967
FATA ASIGURĂRI AGRICOLE	-0.536367	-0.39092	-0.31551
GARANTA	-0.087410	-0.70769	-0.22368
GENERALI	0.624774	0.68381	-0.54874
GERROMA	-0.544834	-0.45527	-0.32081
GRAWE	-0.415959	-0.17728	0.22579
ING ASIGURĂRI DE VIAȚĂ	-0.351754	-0.22964	5.42741
INTERAMERICAN	-0.391006	0.44912	-0.04643
IRASIG	-0.553515	-0.40953	-0.34051
KD LIFE ASIGURĂRI	-0.559469	-0.15693	-0.28189
NBG INSURANCE	-0.496560	-0.34410	-0.31078
OMNIASIG	1.912550	0.87197	0.38104
OMNIASIG VIAȚĂ	-0.378050	-0.07020	0.01006
OTP GARANCIA ASIGURĂRI	-0.525552	-0.07187	-0.22618
RAI	-0.467720	-0.45225	-0.31294
UNITA	0.666760	1.68859	-0.04802

As mentioned previously, we can perform a classification of the insurance companies based on the values (scores) of the first PC. This variable captures 56.24% of the variance of the original space and is powerfully correlated with the following indicators(9): gross written premium, gross claims paid, net premium reserves, net claim reserves and gross written premium ceded in reinsurance.

The graph shows that ALLIANZ-ȚIRIAC is the leader in the insurance market, followed by OMNIASIG, ASIROM, BCR ASIGURĂRI, ASIBAN, UNITA, GENERALI, ASTRA, BT ASIGURĂRI TRANSILVANIA and ARDAF, while companies like CERTASIG, IRASIG, ATE INSURANCE, KD LIFE ASIGURĂRI, CLAL ROMANIA and others lag behind. This classification gives us a clear picture of the insurance market, as opposed to the classification provided by CSA, which is almost exclusively based on market share. Why does this happen? Because the first PC takes into account crucial indicators, such as technical reserves and reinsurance policy of insurance companies. The latter is particularly important because an adequate reinsurance policy can greatly improve the financial stability of insurance companies. For instance, according to the classification provided by CSA, ING ASIGURĂRI DE VIAȚĂ is one of the market leaders. However, the score of the first PC does not distinguish this company from others, because ING's gross written premium ceded in reinsurance is negligible (compared to its gross written premium) and this is a threat to the company.

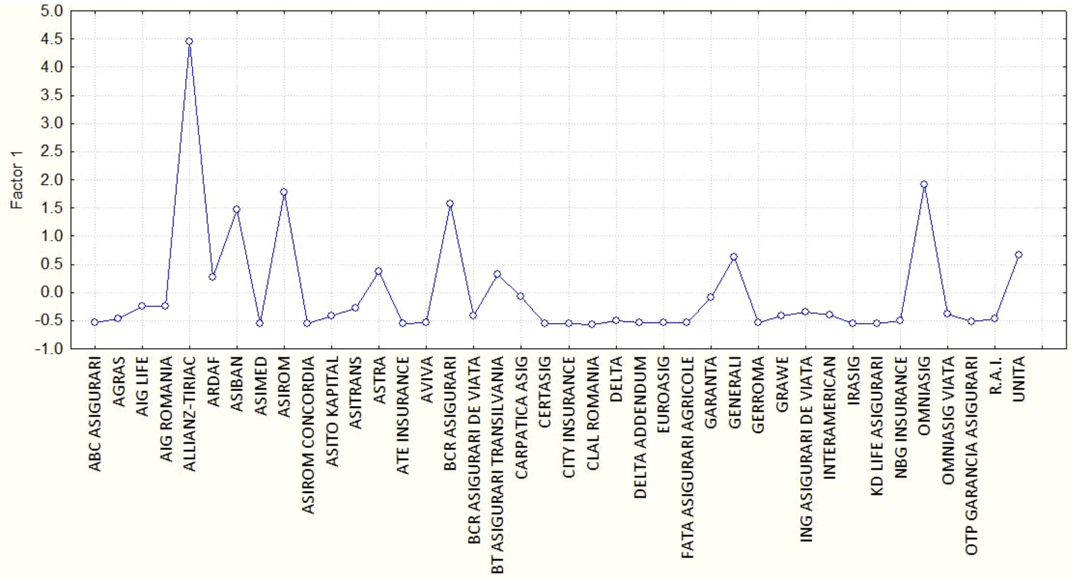

Figure 3. *Plot of observations against factor 1*

Similarly, we can perform a classification of the insurance companies based on the first two PCs:

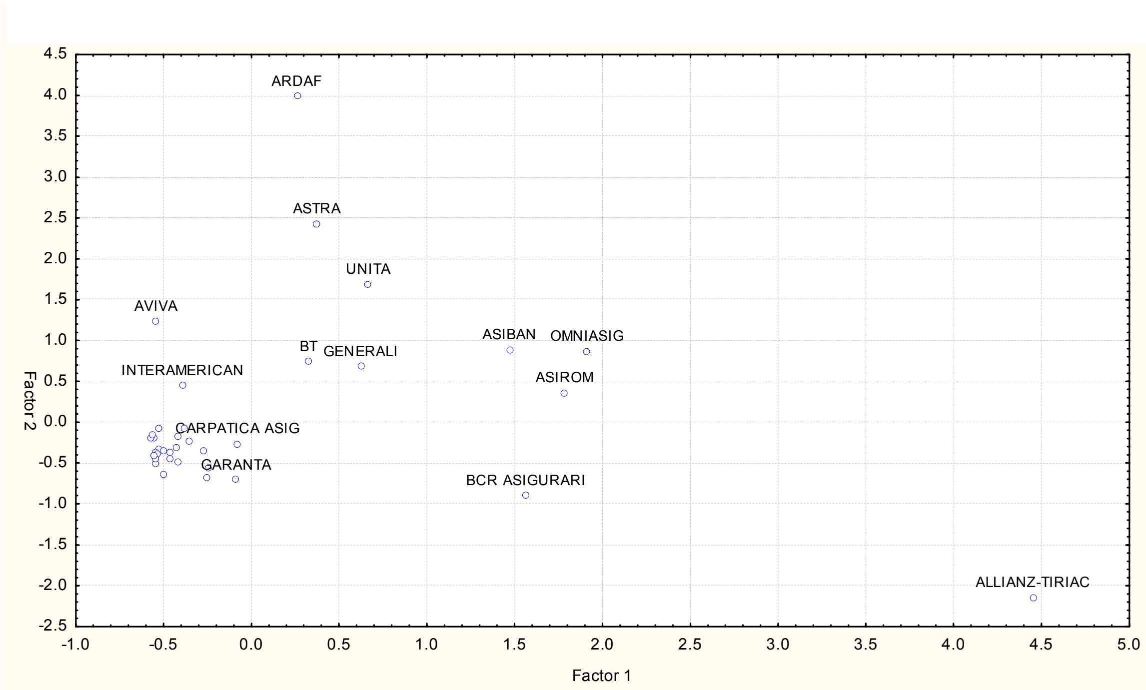

Figure 4. *Scatterplot of observations*

Recall that the second PC occurs in powerful correlation with net income and share capital (positive correlation with the former and negative correlation with the latter). Therefore, observations with negative values of the second PC make profits and have important return on equity. This is the case of ALLIANZ-TIRIAC, BCR ASIGURARI and GARANTA. Other companies (observations with positive scores of the second PC) make negligible profits or even run on losses: ASIROM, INTERAMERICAN, GENERALI, BT ASIGURARI TRANSILVANIA, OMNIASIG, ASIBAN, AVIVA, UNITA, ASTRA and ARDAF. Some of these companies (such as ARDAF, GENERALI and AVIVA) encounter serious financial problems.

It is also worth mentioning that the second PC provides a good image of the capital adequacy of insurance companies. Companies like ALLIANZ-TIRIAC, BCR ASIGURARI and GARANTA have an optimally dimensioned capital and very good levels of return on capital, while INTERAMERICAN, AVIVA, UNITA and ASTRA have too high capital-to-activity ratios. As for the other observations (companies with small market share and relatively low levels of profitability), the graph shows that they tend to form clusters (GARANTA, AIG ROMANIA, DELTA, AIG LIFE, BCR ASIGURARI DE VIATA and CARPATICA ASIG).

Notes

(1) This is measured by the generalized variance or the total variance.

(2) As a matter of fact, PCA basically consists of diagonalizing a symmetrical matrix (in our case Σ). Matrix Λ is computed as $A^t\Sigma A$.

(3) These are the original variables.

(4) Geometrically, the PCs form an orthonormal vector system.

(5) If the initial data are standardized, the covariance matrix will be nothing else but the correlation matrix.

(6) According to the Kaiser-Guttman criterion.

(7) Using the Varimax Raw rotation method.

(8) Here we provide the rotated solution.

(9) With correlation coefficients greater than 0.875.

References

Ruxanda, Gh. „Analiza multidimensionala a datelor”, Master Baze de Date – Suport pentru Afaceri

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

Spircu, Liliana (2006). *Analiza Datelor: Aplicații Economice*, Editura ASE, București

Ruxanda, Gh. (2001). *Analiza Datelor*, Editura ASE, București

Shlens, J. „A Tutorial on Principal Components Analysis”, http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf

„Factor Analysis: Statnotes, from North Carolina State University”, http://www2.chass.ncsu.edu/garson/pa765/factor.htm

„Principal Components and Factor Analysis”, http://www.uta.edu/faculty/sawasthi/Statistics/stfacan.htm

„Neural Networks and Principal Component Analysis”, K. I. Diamantaras, CRC Press, 2002