Professor Nicolae DARDAC, PhD
Assistant Iustina Alina BOITAN
The Bucharest Academy of Economic Studies

STATISTICAL MODELS FOR MONITORING THE LIKELIHOOD OF CREDIT PORTFOLIO IMPAIRMENT

Abstract. *Academic literature and the studies of international financial institutions are the field of a wide debate on the best suited financial indicators and econometric models for predicting, in real time, a wide series of adverse events (credit institutions' rating downgrade, capital adequacy, banking or currency crises). Our empirical approach consists in combining PCA, as a factor analysis technique, with binary logistic regression, in order to forecast the likelihood of a credit portfolio impairment for the whole Romanian banking system. We distinguished several types of financial indicators, related to macroeconomic climate and bank specific data, that are likely to contribute to the determination of the probability of credit portfolio quality impairment. We have applied PCA and identified three principal components. The significance of each component and its predictive power was then tested in a binary logistic model.*

Key words : *banking system; probability of credit portfolio impairment; early warning system; principal components analysis (PCA); binary logistic regression.*

JEL Classification : C 23, G 01, G 21.

Introduction

The current context, characterized by uncertainty and major turbulences on international financial markets, along with the fear of spreading of the effects of financial crisis to banking systems in Europe, have determined credit institutions in this area to be reluctant in providing external financing. Given the concern on the reduction of access to external financing and increase of the indebtness price, Romanian credit institutions should prove a greater prudence in managing the assets and liabilities portfolio. This international trend overlaps a period of aggressive expansion of lending activity, whose effects, unless being carefully managed, could materialize, in the medium term, in a major deterioration of the quality of banks' loans portfolio and of the degree of capitalization.

In the last decade, a wide body of literature tried to analyze the leading factors of episodes of vulnerability or banking crisis, by creating early warning systems to signal, at an earlier stage, any episodes of distress, inadequate capitalization or rating downgrade. The objective of our study is the testing of a model for monitoring the probability of impairment in the quality of loans portfolio to the aggregate level of the Romanian banking system. We have applied a reference technique for this kind of empirical analysis, namely binary logistic regression.

In economic literature, there are ample controversies aiming at assessing both the nature of indicators that will be predictor variables (macroeconomic, microeconomic, institutional) and the appropriate number of economic variables that can be included in the regression. Some authors (Estrella, Park, Peristiani 2000; Rojas-Suarez 2001; Jagtiani, Kolari, Lemieux, Shin 2003) favor a small number of variables, usually 2, while others propose to introduce a significant number of variables, assuming, implicitly, the presence of a major multicollinearity.

To avoid this main drawback specific to the use of a large number of indicators, and also, in order to maintain a high level of the economic significance, by keeping in the further analysis all the variables in the initial dataset, we have decided to apply a factor analysis technique, namely *Principal Component Analysis (PCA)*, to reduce redundancies in the dataset and to synthesize their influence in a small number of uncorrelated factors. This approach had been followed also by several authors (Whalen and Thomson 1988, Stock and Watson 1999, Gosselin and Tkacz 2001, Illing and Liu 2003, Lestano and Kuper 2003), the main purpose being that of reducing the number of explanatory variables included in forecasting models.

As a result of the analysis of empirical studies developed in economic literature, dedicated to the study of distress phenomena in the banking system, and of trends manifested in the Romanian banking system in the period 1997-2008, we held onto the analysis an initial set of 32 indicators, able to assess the state of financial stability, prices stability, the state of financial liberalization, external and current account dynamics, cyclic indicators and of population behavior.

After carrying out univariate tests, we kept in the analysis a number of 12 indicators, which proved to be statistically significant and to have discriminatory power. The dependent variable had been defined as the ratio of nonperforming loans to total assets, because Romanian banking system had been characterized in the mentioned period (especially until end year 2007) by a structural excess of liquidity and a high exposure to credit risk. Therefore, in the context of the actual financial crisis, it is of major importance that supervisory authorities carefully monitor and manage credit risk exposure.

The paper is structured as follows: the first section consists of a brief overview of factor analysis technique, with emphasis on the principal components analysis (PCA). We have presented the premises of our research and discussed the results obtained. Section two integrates the principal components identified in the previous stage into a binary logistic regression, in order to test their ability in predicting future episodes of credit portfolio quality impairment.

1. Factor analysis- methodology and results

Factor analysis is a statistical technique whose main purposes are the diminution of a large number of variables to a smaller number of factors, in order to handle multicollinearity, and the detection of the latent relationships between a set of variables. One of the effects generated by the presence of strong correlations between indicators is the redundancy of financial knowledge, which can bias the final results. Therefore, the aim of a factor analysis consists in reducing the initial dimension of a dataset, by combining those highly correlated variables into a single factor, without implying a significant loss of information. The accuracy of results depends, however, on the quality and validity of data used. In addition, the interpretation of results is heuristic, in the sense that the solution offered is convenient, satisfying, but not necessarily true or generally accepted.

There are several different types of factor analysis, the most frequently employed in economic literature being *principal components analysis* (PCA), which is a technique preferred for purposes of data reduction, and *common factor analysis,* which is preponderantly used for purposes of causal or confirmatory analysis.

In this paper we chose principal components analysis as a method for extracting the factors from a given dataset. As already argued, we first performed univariate tests for all the 32 initial variables selected, in order to keep in the further analysis only those financial indicators with high predictive power. Then we applied the principal components analysis to reduce this new dataset to a limited number of components, which will be used in the second part of the paper, as explanatory variables in the binary logistic model.

Financial indicators kept in the analysis, because they proved a satisfactory predictive capacity are: nonperforming loans to total loans (*cnp/ctb*); consumer price index (*cpi*); nonperforming loans to capital (*cricp*); interbank loans to total assets (*crintat*); export/gdp (*expgdp*); solvency ratio (*is*); lending/deposit interest rate ratio (*lenddep*); M2/foreign reserves (*m2forex*); M2 multiplier (*m2mult*); *openness*; return on assets (*ROA*); return on equity (*ROE*). To assess if these variables are reliable for this type of factor analysis, we performed the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy statistic test. A rule of thumb says that one can proceed with a PCA if the KMO statistic is higher than 0.6. For the variables considered we have obtained a value of 0.767 which enables us continue the analysis.

To extract the main components and to facilitate the interpretation of their scores, we opted for the varimax method of rotation, which consists in maximizing the variance of one component, while minimizing the variance around the component, such that the first component extracted captures the most possible variance of the variables in the dataset. Once the first component is extracted, it will be defined another in order to maximize the variance unenclosed in the previous one. In this fashion, subsequent components are independently one from another, and therefore uncorrelated.

Table 1. Total variance explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.120	59.335	59.335	7.120	59.335	59.335
2	1.636	13.636	72.971	1.636	13.636	72.971
3	1.127	9.393	82.364	1.127	9.393	82.364
4	0.695	5.789	88.153			
5	0.614	5.117	93.270			
6	0.251	2.095	95.365			
7	0.213	1.775	97.140			
8	0.197	1.639	98.779			
9	0.070	0.581	99.360			
10	0.052	0.431	99.790			
11	0.019	0.156	99.946			
12	0.006	0.054	100.000			

Extraction Method: Principal Component Analysis.

We observed that the first component extracted explains 59.33% of total variance of all the 12 variables, in other words, it captures to a significant share the deviation from the mean value of each variable considered. The second one explains 13.63% of deviation from the average of variables. The percentage of total explained variation decreases as the new components are extracted, the last component explaining only 0.054% of the variables dispersion around the average. Thus, by defining the principal components based on the maximum variation that they can embed, we ensured that they reflect, firstly, the extreme values recorded by the considered variables, their deviation or spread over a medium level. The closer to 100% the variance explained, the more representative is the component for the economic interpretation of the phenomenon considered, because it best surprises the characteristics of the variables' dataset, their dispersion from the average.

The number of components extracted is equal to that of initial variables. Therefore, the number of factors to use is a difficult, subjective choice. Economic literature proposes some selection criteria: the Joliffe's criterion –which consists in cutting off once the percentage of explained variance reaches a certain threshold (for example 80% so that the remaining variability be minimal); the Kaiser criterion, which keeps only those factors with eigenvalues greater than one, and the Cattell scree test, which is a graphical method in which the eigenvalues (characteristic roots) are plotted on the vertical axis and the principal components on the horizontal axis. It will be selected those factors situated on the steepest slope. It is considered that factors corresponding to the smoothest slope don't have a significant contribution.

We opted for the eigenvalue criterion because it is the most commonly used and provides the best results. Must be, however, applied with caution because, when the number of variables is very small, are extracted fewer factors than there are in fact in the data, and when the data set is very high, will be extracted several factors, to the detriment of accuracy in interpretation. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor.

According to the eigenvalues illustrated in table 1, we have obtained a solution composed of three principal components. The variance of the first component explains to a 59.335% the variance of the variables in the dataset considered. The second component captures 13.636% from the remaining variance, and the third one only 9.393%. Consequently, the initial dataset of 12 variables can be reduced to only three uncorrelated components, gathering a cumulative variance of 82.364%. The loss of information is of 17.636% and corresponds to the subsequent components that didn't fulfill the Kaiser criterion.

Table 2. Rotated Component Matrix[a]

	Component		
	1	2	3
CNPCTB	-0.718	0.548	-0.267
CPI	-0.280	0.869	-0.058
CRICP	-0.693	0.458	-0.093
CRINTAT	0.883	-0.245	0.241
EXPGDP	0.381	-0.105	0.880
IS	0.904	-0.107	0.211
LENDDEP	0.082	-0.863	0.203
M2FOREX	-0.383	0.768	-0.095
M2MULT	-0.496	0.452	0.436
OPENNESS	0.350	-0.269	0.837
ROA	0.852	-0.237	0.306
ROE	0.800	-0.302	0.278

Extraction Method: PCA
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Table 2 illustrates the component loadings (the correlation coefficients) established between the financial variables presented on the rows and the principal components on the columns. Depending on the intensity of correlations, we can identify those variables whose influence is best reflected by the components extracted.

As we can notice, **component 1** is related to variables expressing banking system characteristics, in terms of capital adequacy, credit portfolio quality, interbank liabilities and financial liberalization. These variables are represented by the ratio of nonperforming loans to total loans, nonperforming loans to capital, placements with and loans to other banks/total assets, solvency ratio, M2 multiplier, return on assets and return on equity. Also, these variables are included in the core set of financial soundness indicators, proposed by IMF as a prerequisite for the acquirement of the banking system's stability.

Component 2 captures variations of the variables related to prices and interest rates stability, such as: consumer price index, lending/deposit interest rate ratio and M2 to foreign reserves. **Component 3** summarizes the influence of external sector variables, represented by export to GDP and openness.

The results obtained attest the substantial role of the banking system specific variables, namely to be the foundation of any analysis on its state of health and strength. We reiterate that, by the principal components analysis we intended to keep in the study all the 12 initial variables, which have proved to be significant from a statistical viewpoint, but in a slightly modified frame, so as not to contravene the principle of the existence of a reduced multicolliniarity, and at the same time, to preserve the economic information contained therein. Therefore, their influence has been caught in various proportions, in a series of principal components, three of which have been shown to be relevant in the light of selection criteria applied. In what follows we test the predictive ability of each of the three components extracted, by including them as explanatory variables in a binary logistic regression.

2. Development of a parsimonious early warning system

As we have previously mentioned, the aim of our study is to develop a statistical monitoring tool, in order to forecast the probability of credit portfolio quality impairment. We chose to implement a binary logistic regression, estimated using panel data for the period between the IIIrd quarter 1997 and IInd quarter 2008, for financial data aggregated at the Romanian banking system level. The dependent variable had been defined as the ratio of nonperforming loans to total assets. It is a binary variable, taking value 1 for the occurrence of deterioration in the credit portfolio quality, and 0 otherwise. The vector of explanatory (independent) variables consists of the three principal components previously identified by means of PCA, and not of all the initial 12 variables.

We carried out a series of univariate tests for each component identified, only with the dependent variable and the scores related to each component, as a single predictor variable, to test which of the three components has the best predictive capability.

In order to distinguish whether a predictor variable exercises influence over the dependent variable, in order to anticipate its evolution, we applied Omnibus test (see table 3), which determines the chi-square statistics and the probability associated. Since the level of significance obtained by the univariate model for components 1 and 2 are below the critical threshold of 5%, we reject the null hypothesis and say that the model tested is statistically significant in the light of the causal relationship between the dependent variable and explanatory one. For the third component we can say, with a probability of 85%, that it does not have a good predictive power.

Table 3. Omnibus Tests of Model Coefficients

component	Chi-square	df	Sig.
1	37.316	1	0.000
2	18.166	1	0.000
3	0.036	1	0.850

To assess the extent to which the univariate models reflect the characteristics of data included in the analysis, we applied the Hosmer-Lemeshow goodness-of-fit test (see table 4). The test's null hypothesis argues that there isn't a significant distinction between the observed and estimated values. It is believed that a level of significance under the critical threshold of 5% is unsatisfactory for model's goodness-of-fit. In this case we obtained a probability that crosses this threshold, therefore we accept the null hypothesis, that the models reflect adequately the quality of the data, but without having any clues on how much of the variance of the dependent variable is explained by the models tested.

Table 4. Hosmer and Lemeshow Test

component	Chi-square	df	Sig.
1	1.324	7	0.988
2	8.242	7	0.312
3	10.038	7	0.186

Analyzing comparatively the outputs of the logistic regressions, we concluded that the first and second components adequately fit the data, being significant at the 5% level, meanwhile the third component obtained a p-value for Wald statistics above the critical threshold of 5%, thus we accepted the null hypothesis that its coefficient is equal to 0. As such, it isn't reliable for making forecasts.

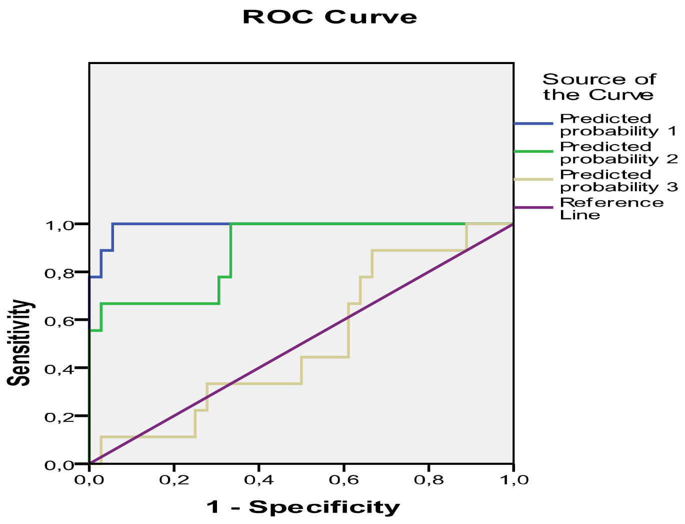
In order to conclude which principal component is best suited for predicting the probability of credit portfolio impairment, we proceeded to a comparison according to two criteria. The *first one* relies on type 1 and type 2 errors. As the fundamental aim of an early warning system consists in signaling, in an incipient stage, that a bank or the banking system as a whole is going to fall in distress, we will put more emphasis on type 1 errors, which reflect the probability that the model fails in warning the supervisory authorities about the imminence of a distress event. In this fashion, a statistical monitoring system is considered to have a good, accurate predictive ability if type 1 error is low. The results presented in table 5 show that the lowest type 1 error is committed under model 1 (11.1%), composed by the dependent variable and first component from the PCA, while the predictor variable in model 3 (the third principal component from PCA) wasn't able to identify any event of credit portfolio deterioration.

Table 5. Models' accuracy

Regression model	Type 1 error	Type 2 error	Overall accuracy percentage
Model 1	11.1 %	2.7 %	95.6 %
Model 2	33.3 %	2.7 %	91.1 %
Model 3	100 %	0 %	80 %

We preferred the model that has the least amount of Type I error, because the fundamental purpose of a forecasting model is the correct, ex ante reporting, of the potential distress of banking activity. Once again, model 1 seems to be the best suited for carrying out forecasts.

The *second criterion* is represented by the ROC *(Receiver Operating Characteristic)* curve and its curvature indicator, AUROC, which assesses the ability of discrimination of the variables included in logistic regression. ROC curve allows graphic visualization of the rate of false alarms compared to the percentage of correct predictions for all the possible values of the probability threshold. If we choose randomly a point on the ROC curve, with coordinates (*x, y*), we can show which is the *x* percentage of false alarms reported by the model, which allows a proportion of *y* correct predictions.



From the ROC curve graph it can be observed that model 1 is the closest to the perfect model, through its ability to correctly forecast the adverse events. The area under ROC curve, called AUROC, quantifies the ROC test's power. In other words, it reflects the marginal contribution of each variable / model in anticipating the likelihood of achieving the adverse event considered. An AUROC indicator equal to 1 suggests that the model tested discriminates perfectly the events of deterioration from those of sound periods, while a value equal to 0 implies the absence of any predictive capability of the model tested. In practice, it is recommended that AUROC value be above the threshold of 0.8. Table 6 illustrates, comparatively, the corresponding AUROC indicator for each component under review.

Table 6. Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Model 1	0.991	0.010	0.000	0.000	1.000
Model 2	0.889	0.058	0.000	0.771	1.000
Model 3	0.503	0.101	0.977	0.304	0.702

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

One can observe that model 1 meets the requirements of a good predictive ability, as the AUROC value is 0.991, followed by model 2, with 0.889.

Logit model assigns a probability of achieving the event of impairment to each time interval considered. To anticipate its imminence, it is necessary to define a probability threshold (cut-off) under which the periods for which the estimated probability exceeds the critical threshold will be considered to reflect an increase in the share of bad loans in total assets, while a lower estimated probability denotes a normal situation, a satisfactory quality of the loans portfolio. Although the approach is arbitrary and subjective, the choice of an optimal threshold should be done taking into account the existing compromise between type I error and type II error: a low Type I error involves a high level of type II error. Table 7 illustrates the rate of success (hit rate), and false alarms rate provided by each of the three models, for a discrimination threshold whose value was set at 21.97%.

Table 7. Comparative analysis of the discrimination power

Logistic regression	Discrimination threshold	Proportion of distress events correctly identified (sensitivity)	Proportion of normal periods incorrectly identified (1-specificity)
Model 1	21.97%	0.889	0.056
Model 2	21.64%	0.667	0.278
Model 3	21.64%	0.111	0.056

For the selected threshold, model 1 recorded the best predictive performance, the proportion of episodes of increase in the share of outstanding and doubtful assets in total assets, which were correctly identified, being 88.9% compared with 66.7% for model 2 and only 11.1% for model 3. In addition, the rate of false alarms is 5.6%, lower than that reported for model 2 (27.8%).

Conclusions

A common feature of all the statistical models designed for purposes of monitoring and early warning is the use of a wide range of financial indicators from the domestic and external sector as explanatory variables. In order to eliminate the inherent multicollinearity in a broader set of variables, we chose, for this study, the application of a technique for reducing the size of the variables dataset, namely principal components analysis (PCA). By concentrating the variability of the 12 variables considered in a reduced number of components, we have identified three main components relevant in terms of the selection criteria applied. Note that each component has its own economic significance. These were subsequently implemented in a logistic regression, to assess the predictive ability concerning the substantial increase in the share of overdue and doubtful assets to total assets.

Given the statistically and economically relevance, as well as the satisfactory predictive power, we considered that the first principal component, as a proxy variable, which summarizes the aggregated influence of financial variables that reflect the Romanian banking system's characteristics, may be subject of a statistical tool for quantifying the likelihood of credit portfolio's quality deterioration, for the whole banking system.

In addition, the results obtained recommend the application of this technique in case one wishes to test the simultaneous influence of several variables on the dependent variable considered.

REFERENCES

[1]**Brooks, C**. **(2002),** ***Introductory Econometrics for Finance***, the ISMA Centre, University of Reading, pp. 220-225;

[2] **Dardac, N., Moinescu, B. (2009),** ***The Third Wave of the Financial Crisis and Its Ripple Effects on the Deterioration Risk of Romanian Banking Sector Performance***, Economic Computation and Economic Cybernetics Studies and Research , no. 1, ASE Publishing House, Bucharest;

[3]**Demirguc-Kunt, A., Detragiache, E**. **(1999),** ***Monitoring Banking Sector Fragility: A Multivariate Logit Approach,*** IMF working paper, WP/99/147;

[4]**Garson, G.D. (2007),** ***Factor Analysis***, from Statnotes: Topics in Multivariate Analysis, http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.;

[5]**Garson, G.D. (2007),** ***Logistic Regression***, from Statnotes: Topics in Multivariate Analysis. http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.;

[6]**Gosselin, M.,A., Tkacz, G. (2001),** ***Evaluating Factor Models: An Application to Forecasting Inflation in Canada.*** Bank of Canada Working Paper No. 2001-18;

[7]**Lestano, J. J., Kuper G.H. (2003),** ***Indicators of Financial Crises Do Work! An Early-warning System for Six Asian Countries***, Department of Economics, University of Groningen;

[8]**Illing, M., Liu, Y.(2003),** ***An Index of Financial Stress for Canada***; Bank of Canada working paper 2003-14;

[9]**Stock, J.H. , Watson M.W. (1999),** ***Forecasting Inflation***. *Journal of Monetary Economics* 44: 293–335.

[10]**Whalen, G., Thomson, B**. **(1988),** ***Using Financial Data to Identify Changes in Bank Condition,*** http://clevelandfed.org/research/review/