A COMBINED PRINCIPAL COMPONENT ANALYSIS-REGRESSION ANALYSIS MODEL TO STUDY THE EFFECT ON TECHNICAL EFFICIENCY OF BAD LOANS IN BANK INDUSTRY

Omid Mashhadifarahani, Narges Rezavi, Loghman Hatami-Shirkouhi

This paper presents a combined principal component analysis-regression analysis (PCA-RA) model to study the effects of bad loans on the economic performance of banking systems. This model first applies PCA to calculate the overall technical efficiency of banks and then uses regression analysis to find the impact of bad loans on technical efficiency. Bad loans or Non-performing loans (NPLs) include: past due loans; bankrupt and quasi-bankrupt assets, and doubtful assets. The technical efficiency of banks is measured with reference to three input indicators: number of branches, deposits, and costs; and three output indicators: income, profit, and loans. Then technical efficiency is regressed on two explanatory variables of loans and bad loans. Results of a case study in governmental banks in Iran show that bad loans have negative impact on the technical efficiency of banks.

Keywords: bad loans, banking industry, non-performing loans, principal component analysis, regression analysis, technical efficiency

1 Introduction

The standard methods applied in banking are the intermediation and production approaches. Under the intermediation approach, banks use purchased funds together with physical inputs to produce various assets (measured by their value). The production approach assumes that banks use only physical inputs such as labour and capital to produce deposits and various assets (measured by the number of deposit and loan accounts at a bank, or by the number of transactions for each product). We adopt a combined approach of the intermediation approach and production approach to assess the technical efficiency of banks.

Berger and Humphrey [1] discussed several approaches of modelling the bank production process: the production approach, user-cost approach, value added approach and dual approach. Berger and Humphrey [1] suggested the intermediation approach is best suited for evaluating bank efficiency, whereas the production approach is appropriate for evaluating the efficiency of bank branches. Koutsomanoli-Filippaki et al. [2] employed the directional technology distance function approach to decompose profit efficiency into its technical and allocative components.

Traditionally, multivariate techniques are extensively used for assessing the performance of banking systems. Data envelopment analysis is one of the popular tools in this respect [3-6]. Frontier techniques such as stochastic frontier analysis (SFA) are also used for bank efficiency assessment. Perera and Skully [7] investigated the consistency of parametric stochastic frontier analysis (SFA) and nonparametric data envelopment analysis (DEA) estimates for bank efficiency. Furthermore, principal components analysis (PCA) and numerical taxonomy (NT) are used and applied to verify and validate DEA findings in Azad et al. [8]. PCA is one of the multivariate analysis techniques usually used for correlation analysis, data reduction and also efficiency assessment [9, 10]. Zhongsheng and Dong [11] analysed the operating efficiency of Chinese commercial banks during 1999 to 2003 based on PCA.

Formally, the bad loan is defined to be a debt instrument (loan) whose contractual interest and principal payments are difficult to collect. The effects of bad loans have also been addressed in the literature. Using DEA, Chen [12] evidenced that bank loss on bad loans is one of the major reasons for lower cost efficiency of privately owned banks than publicly owned banks in Taiwan. Matthews and Zhang [13] considered and incorporated non-performing loans as a bad output for calculating Malmquist total factor productivity growth index of nationwide commercial banks of China. Barros et al. [14] analysed technical efficiency of the Japanese banks based on the Non-radial directional model with undesirable output that takes into consideration not only desirable outputs but also an undesirable output that is represented by non-performing loans (NPLs).

To the best knowledge of the authors there is no study that used PCA for technical efficiency of banks with bad outputs (i.e. Non-operating or bad loans) and their effect on the overall technical efficiency. This paper presents a combined PCA-Regression analysis for efficiency assessment and analysis of the effect of bad loans on the technical efficiency of banks.

The following inputs/outputs variables are defined in this study. The inputs include three indicators: Number of Branches, Deposits, and Costs; and The outputs include...
three indicators: Income, Profit, and Loans. PCA indicators are defined as the ratio of output variables to input variables hence we have 9 PCA indicators. The notion of dividing outputs by inputs reflects an indication of efficiency from both production and intermediation points of view. PCA will find weights for outputs and inputs and use them to calculate a total efficiency score of the banks namely technical efficiency. Then technical efficiency is regressed on two explanatory variables of loans and bad loans. Finding the relationship between technical efficiency and bad loans will help analysing the effect of bad loans on the technical efficiency of banks under study.

### Table 1 Statistical characteristics of the applied variables (Billions Iranian Rials)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs. Mean</th>
<th>Std. dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inputs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number Branches</td>
<td></td>
<td>2.0271</td>
<td>825</td>
<td>973</td>
</tr>
<tr>
<td>Deposits</td>
<td>336.299</td>
<td>187.665</td>
<td>53.438</td>
<td>789.314</td>
</tr>
<tr>
<td>Costs</td>
<td>13.849</td>
<td>8.007</td>
<td>2.244</td>
<td>36.111</td>
</tr>
<tr>
<td>Output</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>15.656</td>
<td>8.844</td>
<td>2.360</td>
<td>36.544</td>
</tr>
<tr>
<td>Profit</td>
<td>1.806</td>
<td>1.867</td>
<td>1.112</td>
<td>7.776</td>
</tr>
<tr>
<td>Loans</td>
<td>185.027</td>
<td>97.436</td>
<td>36.056</td>
<td>432.850</td>
</tr>
<tr>
<td>Non-performing loans</td>
<td>55.495</td>
<td>24.070</td>
<td>4.543</td>
<td>89.228</td>
</tr>
</tbody>
</table>

**Figure 1** The working algorithm of PCA–regression

This paper is organized as follows. In Section 2, a step-by-step working algorithm of the study is presented. Section 3 introduces the case study; data and variables for technical efficiency of banks. In Section 4, the results of PCA for technical efficiency are presented. Also, the PCA procedure of this study is presented in this section. Section 5 analyses the impact of bad loans on technical efficiency. Section 6 presents the main findings and conclusion.

### 2 The working algorithm of the study

In this section the proposed working algorithm of the paper is presented. At its early stage, in the first phase, this algorithm proposes to collect data and pre-process them using PCA. This pre-processing involves the use of PCA and Eigen structure of data to find a weight for each of PCA variables to be used for integrating all PCA indicators and calculate an overall efficiency score for each bank in a year. The second phase involves the use of regression analysis as a statistical analysis tool to test the significance of the relationship between bad loans and technical efficiency of banks. The construct of regression model is such that the model uses technical efficiency as the response variable and two variables loans and bad loans as explanatory variables. Standard statistical tools such as MINITAB could be used to estimate the regression coefficient and test the statistical significance of coefficients. This working algorithm is presented in Fig. 1.

### 3 The case study: data and variables

As a case study, to show the applicability and usefulness of the working algorithm of this paper, we have considered the performance indicator of seven governmental banks in Iran. The data are related to seven performance indicators of these banks in the time period 2006–2010. Tab. 1 presents some descriptive statistics of the collected data.

### 4 Applied PCA

PCA is a multivariate technique commonly used for data reduction and performance assessment. When it is used for performance assessment of decision-making units (DMUs), first it is applied to find a weight for the variables under study and then ranking of DMUs. Generally, the DMUs utilize a variety of resources as inputs to produce several outputs.

Basically, PCA finds weights according to the Eigen value and Eigen vectors of a specific matrix describing the covariance or correlation between variables. This would eventually lead to scoring and rankings of DMUs. A precise description of procedure of ranking with PCA can be found in Zhu [15] and Premachandra [16], Slišković et al. [17].
In this section, the PCA procedure is illustrated step by step. There are three input variables and three output variables. Dividing each output by all three inputs will yield 9 output/input ratios which are considered as PCA indicators. In the case study we considered the performance data of seven governmental banks in 5 years from 2006 to 2010. So we have 35 DMUs for PCA. Hereafter we call each DMU as bank-year. PCA is performed by identifying Eigen structure of the covariance or singular value decomposition of the original data. Here, the former approach will be used.

Suppose \( X = (x_{11}, x_{22})_{35 \times 9} \) is a 35×9 matrix composed by \( x_{ij} \)'s defined as the value of the \( j \)th variable for the \( i \)th bank-year and therefore \( x_m = (x_{1m}, \ldots, x_{35m})^T \) \((m = 1, 2, \ldots, 9)\). Furthermore, suppose \( \hat{X} = (\hat{x}_1, \ldots, \hat{x}_m)_{35 \times 9} \) is the standardized matrix of \( X \) and therefore \( \hat{x}_m = (\hat{x}_{1m}, \ldots, \hat{x}_{3m})^T \). PCA is performed to identify new independent variables or principal components (defined as \( Y_j \) for \( j = 1, 2, \ldots, 9 \)), which are respectively different linear combination of \( \hat{x}_1, \ldots, \hat{x}_9 \). As mentioned, this is achieved by identifying Eigen structure of the covariance of the original data. The principal components are defined by a 35×9 matrix \( Y = (y_1, y_9)_{35 \times 9} \) composed by \( y_{ij} \)'s are shown by:

\[
y_m = \sum_{j=1}^{9} l_{mj} \hat{x}_j, \quad m = 1, \ldots, 9,
\]

where, \( l_{mj} \) is the coefficient of the \( m \)th variable for the \( j \)th principal component. The \( l_{mj} \)'s are estimated such that the conditions of Eqs. 6 and 7 are met. \( Y_1 \) accounts for the maximum variance in the data, \( Y_2 \) accounts for the maximum variance that has not been accounted by \( Y_1 \), and so on.

\[
i_j^2 + i_{j2}^2 = 1, \quad m = 1, \ldots, 9
\]
\[
i_{m1} \cdot i_{m1} + i_{m2} \cdot i_{m2} = 0, \quad \text{for all } m \neq n, n = 1, \ldots, 9.
\]

For obtaining the \( l_{ij} \)'s and consequently vectors \((y_{1j}, \ldots, y_{9j})\) \((j = 1, \ldots, 9)\) and PCA scores the following steps are performed:

**Step 1:** Calculate the sample mean vector \( \bar{x} \) and covariance matrix \( S \) (Eqs. 4, 5 and 6):

\[
\bar{x} = (\bar{x}_1, \ldots, \bar{x}_9)_{1 \times 9},
\]

in which,

\[
\bar{x}_j = \frac{1}{35} \sum_{i=1}^{35} x_{ij} \quad \text{for } j = 1, \ldots, 9,
\]

\[
S = (s_{qj})_{2 \times 2} = \frac{1}{35} (X - \bar{x})^T (X - \bar{x}) \quad \text{for } q = 1, \ldots, 9.
\]

**Step 2:** Calculate the sample correlation matrix.

\[
R = \left( C_i \right)^{1/2} S \cdot C_i \left( \sqrt{S} \right)^{-1} \quad \text{where } C_i \left( \sqrt{S} \right)^{-1} \text{ is a } 9 \times 9 \text{ diagonal matrix whose } j^{th} \text{ diagonal element is a } 1/\sqrt{s_{jj}} \text{ for } j = 1, \ldots, 9.
\]

**Step 3:** Solve the following equation:

\[
|R - \lambda I| = 0 \quad \text{where } I \text{ is a } 9 \times 9 \text{ identity matrix. We obtain the ordered 9 characteristic roots (eigenvalues) } \lambda_1 \geq \lambda_2 \geq \ldots \lambda_9 \text{ with } \sum_{j=1}^{9} \lambda_j = 9 \text{ and the related 9 characteristic vectors (eigenvectors) (} l_{i1}, \ldots, l_{i9} \text{) } (m = 1, \ldots, 9).
\]

Those characteristic vectors compose the principal components \( Y_j \). The components in eigenvectors are respectively the coefficients in each corresponding \( Y_j \) (Eq. 7):

\[
y_m = \sum_{j=1}^{9} l_{mj} \bar{x}_j \quad \text{for } m = 1, \ldots, 9 \text{ and } i = 1, \ldots, 35.
\]

**Step 4:** Calculate the weights \( w_j \) of the principal components and PCA scores \((Z_i)\) of each country-year \(i = 1, \ldots, 35\). Furthermore, the \( Z \) vector \((Z_{1}, \ldots, Z_{35})\) where \( Z_i \) shows the score of each bank-year, is given by Eqs. 8 and 9:

\[
w_j = \frac{\lambda_j}{\sum_{j=1}^{9} \lambda_j} = \frac{\lambda_j}{9}, \quad j = 1, \ldots, 9.
\]

\[
z_i = \sum_{j=1}^{9} w_j Y_j, \quad i = 1, \ldots, 35.
\]

The PCA efficiency scores and ranking for each bank-year are presented in Tab. 2.

The results of full ranking in Tab. 2 show that Refah K. Bank in years 2006 and 2007 has the highest technical efficiency and is the best performing bank. This bank also performed very well in years 2008 and 2009 due to its ranks of 4 and 5 between all 35 DMUs (bank-year). In the table of ranking, Refah K. Bank as the best is followed by Bank Saderat Iran. This bank has got ranks 3, 7, 8, and 10 in years 2006, 2007, 2008, and 2009, respectively. Moreover, Fig. 2 shows the time trend of technical efficiency of banks. This figure shows that technical efficiency in the governmental banks has decreased during 2006 ÷ 2010.

5 Regression analysis – the impact of bad loans

In this section, to test the significance of the relationship between bad loans and technical efficiency, a linear regression model is constructed. The model uses technical efficiency as the response variable and two variables loans and bad loans as explanatory variables. Data are presented in the last three columns of Tab. 2. The MINITAB software is used to estimate the regression coefficient and test the statistical significance of coefficients. Results are presented in Tab. 3.
The null hypothesis is that regression coefficients are not values which are the risk of rejecting the null hypothesis. Because these risks are very low, we reject the null hypothesis and conclude that the coefficients are statistically significant. Furthermore, as the final finding, the negative sign of coefficient of bad loans (−0.00000364) suggests that this variable has a negative impact on the technical efficiency of banks.

### 6 Conclusion

A combined principal component analysis-regression analysis model to study the effects of bad loans on the economic performance of banking systems was presented in this paper. An overall working algorithm was presented which is capable of calculating the overall cost efficiency of banks and analysis of the impact of bad loans on technical efficiency. The applicability of the algorithm is observed via a case study. Bad loans or Non-performing loans (NPLs) include past due loans; bankrupt and quasi-bankrupt assets, and doubtful assets are suspected to have a negative effect on technical efficiency in banking industry. The technical efficiency of banks was measured with reference to three input indicators: Number of Branches, Deposits, and Costs; and three output indicators: Income, Profit, and Loans. Then technical efficiency was regressed on two explanatory variables of loans and bad loans. Results of the case study in governmental banks in Iran show that bad loans have a negative impact on the technical efficiency of banks. For future research, other forecasting and performance assessment methods in certain and fuzzy environment to assess the impact of bad loans on technical efficiency of banks can be applied. These methods are adaptive network-based fuzzy inference system (ANFIS) [18, 19], artificial neural networks (ANN), etc. The results of these methods can be also compared with the results of current study.

### 7 References


[3] Azadeh, A.; Ghaderi, S. F.; Mirjalili, M.; Moghaddam, M. Integration of analytic hierarchy process and data envelopment analysis for assessment and optimization of


Authors’ address

Omid Mashhadifarhahani
Department of Management
Shahre Qods Branch,
Islamic Azad University
Tehran, Iran

Narges Rezavi
Department of Industrial Engineering
Khatam Institute of Higher Education
Tehran, Iran

Loghman Hatami-Shirkouhi
Corresponding Author
Roudbar Branch
Islamic Azad University
Roudbar, Iran
E-mail: loghmanhatami@yahoo.com
Tel.: (+98)132 6224135