Available online at http://ijim.srbiau.ac.ir

Int. J. Industrial Mathematics Vol. 3, No. 4 (2011) 251-258

Evaluating the Efficiency of DMUs with PCA and an Application in Real Data Set of Iranian Banks

S. Kordrostami[a]*, A. Amirteimoori [b], A. Masoumzadeh [a]

(a) *Department of Mathematics, Islamic Azad University, Lahidjan branch, Iran*

(b) *Department of Mathematics, Islamic Azad University, Rasht branch, Iran*

Received 30 November 2010; accepted 13 December 2010.

Abstract

This paper proposes a method for evaluating the efficiency of decision making units (DMUs) by using principal component analysis. This efficiency deals with undesirable outputs and simultaneously reduces the dimensionality of data set. First, we change the undesirable outputs to be desirable by reversing. Then we do PCA on the ratios of a single desirable output to a single input.

In order to reduce the dimensionality of data set, the required principal components are selected out of the generated ones according to the lowest eigenvalues. Finally these chosen principal components are treated as virtual data set into data envelopment analysis (DEA). Then the utility of proposed approach is applied to real data set of some branches of an Iranian bank.

Keywords : Data envelopment analysis (DEA), Principal component analysis (PCA), Undesirable output, Data reduction, Efficiency.

1 Introduction

Data envelopment analysis (DEA) is a methodology that uses linear programming in the evaluation of the relative efficiency of decision making units (DMUs) with multiple inputs and outputs.

Charnes, Cooper and Rhodes developed Farrell's ideas and the efficiency value that is obtained by dividing single output, to single input was extended to multiple output/input ratio and they proposed the CCR model in 1978 [5]. Banker, Charnes and Cooper also proposed another model, named BCC in 1984 [3]. These models cannot be used to rank efficient units. So Anderson and Peterson provided ranking model through improving CCR model, named AP [2].

*Corresponding author. Email address: kordrostami@guilan.au.ir

In this paper we work out a method for evaluating the efficiency and ranking DMUs by using PCA. The multivariate statistical method principal component analysis (PCA) is a data reduction technique, used to identify a small set of variables that account for a large portion of the total variance in the original variables (Bolch and Huang, 1974) [4]. PCA is also a popular ranking method in multidimensional analysis (Slottje and Scully, 1991) [11].

The idea of combining two methods (PCA & DEA) was proposed by Hoshini and Udea in 1997 and developed by Zhu in1998 [12]. He combined this statistical method with based models of DEA and proposed some new models for evaluating the efficiency of DMUs.

This paper is organized as follows: In Section 2, PCA methods are presented and used to find virtual data set from the ratio of outputs into inputs which deal with undesirable outputs and reduce the dimensionality of the data set. Section 3 presents PCA-DEA models for evaluating and ranking DMUs with virtual data set. In Section 4, an application of the proposed approach is offered to evaluate the efficiency, based upon the real data set of an Iranian bank and Section 5, the conclusion is drawn.

2 PCA Model

Principal component analysis, a multivariate statistical method, is used to reduce the dimensionality of data set without losing he main information of the problem. First we do PCA on the outputs and inputs, and then find virtual data set that is applied to evaluate the efficiency of DMUs.

Suppose we have n independent homogeneous decision making units, where each DMU_j, $j = 1, \dots, n$ consumes m inputs $x_{ij} = (x_{ij}, \dots, x_{mj}) \geq 0$ to produce k desirable outputs y_j^g and s-k undesirable outputs y_j^b:

$$\begin{aligned} y_j^g &= (y_1, \dots, y_{kj}) \\ y_j^b &= (y_{k+1j}, \dots, y_{sj}) \end{aligned}$$

We would like to produce desirable outputs as much as possible and undesirable outputs as little as possible. So the output matrix Y can be represented as follows:

$$Y = \begin{bmatrix} y^g \\ y^b \end{bmatrix} = [y_1, \dots, y_n]_{s\times n}$$

In order to do PCA on the original data set, several steps are taken as following: [6] Step1: Transform the output matrix by reversing the undesirable outputs

$$Y = \begin{bmatrix} y^g \\ -y^b \end{bmatrix} = [y'_1, \dots, y'_n]_{s\times n}$$

Step 2: Calculate the ratio matrix $D = [d_1, \dots, d_p]_{n\times p} = [d_l^j]_{n\times p}$ Where

$$\begin{aligned} d_l^j &= \frac{y'_{rj}}{x_{ij}} \\ l &= s(i-1) + r \\ p &= m \times s \end{aligned}$$

We mention that d_l^i is the ratio of single transformed output to single input. So, the bigger d_l^j cause DMUs perform better in terms of r-th transformed output and i-th input.

Theorem 2.1. *In CCR model, If DMU_j is the only DMU with this property $\frac{y_{rj}}{x_{ij}} = \max_k \frac{y_{rk}}{x_{ij}}$. Then DMU_j is efficient.*

Also some elements of a ratio matrix may be negative, however it does not stop PCA process fortunately.
Step 3: Normalize the ratio matrix

$$\begin{aligned}
&\widetilde{D} = [\widetilde{d}_1, \ldots, \widetilde{d}_p]_{n\times p} = [\widetilde{d}_i^j]_{n\times p} \\
&\widetilde{d}_l^j = \frac{d_l^j - \overline{d}_l}{\sqrt{s_{ll}}} \\
&\overline{d}_l = \frac{1}{n}\sum_{j=1}^{n} d_l^j \\
&S_{ll} = \sum_{j=1}^{n}(d_l^j - \overline{d}_l)^2 \Big/ (n-1)
\end{aligned}$$

Step 4: Calculate the correlation matrix

$$\begin{aligned}
&R = [r_{li}]_{p\times p} \\
&r_{li} = S_{li} \Big/ \sqrt{S_{ll}S_{ii}} \\
&S_{li} = \frac{1}{n-1}\sum_{j=1}^{n}(d_l^j - \overline{d}_l)(d_i^j - \overline{d}_i)
\end{aligned}$$

Step 5: Compute eignvalues and corresponding normalized orthonormal eignvectors $l_1, l_2, \ldots, l_p$.

$$\begin{aligned}
&\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \\
&\sum_{i=1}^{p} \lambda_i = p
\end{aligned}$$

Step 6: Compute and select the principal components

$$pc = \widetilde{D}[l_1, \ldots, l_p] = [pc_1, \ldots, pc_p]_{n\times p}$$

Now, it can be noticed that any two different principal components are uncorrelated with each other, which shows that there is no information superposition between them.
The correlation variance of any two principal components is

$$Cov(pc_i, pc_j) = \begin{cases} var(pc_i) = \lambda & i = j \\ 0 & i \neq j \end{cases}$$

In order to reduce the dimensionality of data set, the required principal components are selected out of the generated ones according to the lowest eigenvalues. Although, Outputs of original DEA models need to be strictly positive, the elements of the chosen principal components can be negative. So a linear data transformation is made which can change the negative result of PCA into positive ones.

$$z_{;j} = pc_l^j + Q$$
$$Q = -\min_{1\leq l\leq m}\min_{1\leq j\leq n}\{pc_l^j\} + 1$$

It is a common choice to ensure that all transformed values are positive.

3 Evaluating the efficiency

To evaluate the efficiency of DMU_0, a simplified model is proposed by combining PCA and input - oriented CCR model as follows: [9, 8]

$$\begin{aligned}
&Max \quad \textstyle\sum_{l=1}^{m} p_l z_{lo} \\
&s.t: \\
&\quad \textstyle\sum_{l=1}^{m} p_l z_{ij} \leq 1, && j = 1,\dots,n \\
&\quad p_l - p_{l+1} \geq \varepsilon_l, && l = 1,\dots,m-1 \\
&\quad \varepsilon_l = \begin{cases} 0, & \lambda_l = \lambda_{l+1} \\ \varepsilon > 0, & \lambda_l > \lambda_{l+1} \end{cases} \\
&\quad p_l \geq 0, && l = 1,\dots,m
\end{aligned}$$

Where p_l is the weight attached to the virtual outputs $z_{ij} : j = 1,\dots,n$ and the weight constraints $p_l - p_{l+1} \geq \varepsilon$ represent the facts that l-th principal component carries the total dispersion more than $(l+1)$-th one does. Since any two different principal components are uncorrelated with each other, that shows there is no information between them.
As a result, the measure of efficiency dosen't change any more, after ignoring some principal components with lowest eignvalues. Therefore, this model is useful for evaluating the large DMUs and reducing the dimensionality of data set.
IF we distinguish the efficient DMUs based upon CCR model, we use AP model (Anderson & Petersen, 1993). Then, for ranking the DMUs we develop a model by combining PCA and AP model which we mention as follows:

$$\begin{aligned}
&Max \quad \textstyle\sum_{l=1}^{m} p_l z_{lo} \\
&s.t: \\
&\quad \textstyle\sum_{l=1}^{m} p_l z_{ij} \leq 1, && j = 1,...,n, j \neq 0 \\
&\quad p_l - p_{l+1} \geq \varepsilon_l, && l = 1,\dots,m-1 \\
&\quad \varepsilon_l = \begin{cases} 0, & \lambda_l = \lambda_{l+1} \\ \varepsilon > 0, & \lambda_l > \lambda_{l+1} \end{cases} \\
&\quad p_l \geq 0, && l = 1,\dots,m
\end{aligned}$$

Where p_l is the weight attached to the virtual outputs $z_{lj} : j = 1,\dots,n$.

4 Application in real data set of some Iranian banks

In this section, we evaluate the efficiency of 25 branches of Guilan Saderat bank. This data set consists of two input and four output variables.
Input and output variables are given as below:
Input 1 (I_1): employee: It consists of the manager and clerks of each branches.
Input 2 (I_2): current cost: It consists of the cost of administrative, personal and energy.
Output 1 (O_1): resources: It consists of value of customer's accounts.
Output 2 (O_2): wage: That is the money that bank receives for its services.
Output 3 (O_3): income
Output 4 (O_4): loans: It consists of any kind of long and short term loans that bank gives to its customers.
There are some limitations that we can't represent the amount of inputs and outputs. The results of CCR and AP models are given in table 1.

Table 1
Efficiency and ranks of branches with DEA models.

DMU	Branch's number	Rank	Efficiency
1	49	13	0.3821
2	185	23	0.2483
3	485	4	0.8352
4	548	8	0.4792
5	549	1	1.0000
6	760	6	0.5381
7	875	7	0.4978
8	1681	22	0.2611
9	1694	24	0.2295
10	1695	15	0.3391
11	1966	18	0.2905
12	1967	21	0.2726
13	1969	9	0.4636
14	1970	14	0.3417
15	2107	17	0.3117
16	2908	20	0.2765
17	3115	19	0.2889
18	3144	16	0.3303
19	3549	25	0.2179
20	3595	12	0.4027
21	4146	10	0.4587
22	4354	11	0.4104
23	4355	2	1.0000
24	4366	3	1.0000
25	4525	5	0.5989

Then, we take PCA steps on this data set and in order to reduce the dimensionality of data set we ignore the principal components with the lowest eignvalues. According to the illustration below, the first two principal components have an explanation ratio of 80% of total variance. The rest of components have a little information that can be ignored.

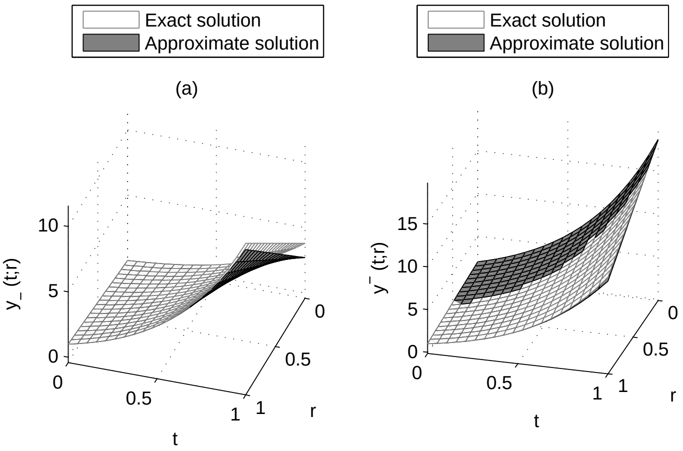

Fig. 1. Relative Importance of Principal Components

The results of new PCA-DEA models are given in table 2.

Table 2
Efficiency and ranks of branches with PCA-DEA models.

DMU	Branch's number	Rank	Efficiency
1	49	13	0.5142
2	185	21	0.4643
3	485	4	0.6271
4	548	12	0.5364
5	549	1	1.0000
6	760	5	0.5944
7	875	9	0.5463
8	1681	22	0.4602
9	1694	25	0.4419
10	1695	16	0.4831
11	1966	18	0.4732
12	1967	20	0.4667
13	1969	7	0.5545
14	1970	15	0.5049
15	2107	14	0.5103
16	2908	17	0.4805
17	3115	23	0.4549
18	3144	19	0.4673
19	3549	24	0.4472
20	3595	10	0.5457
21	4146	8	0.5475
22	4354	11	0.5446
23	4355	2	1.0000
24	4366	3	0.9493
25	4525	6	0.5892

The conclusion of both analysis shows that the reduction of dimensionality does not have much effect on the efficiency and ranking of the branches.

For example branch 549, in both methods, has the first rank and also the branches of 4355, 4366 and 485 have the second to fourth ranks and their efficiency are so close. Branch 4366 in DEA model is efficient, with the efficiencies of 1, but in DEA-PCA model, it has the efficiency of 0.9493 , though both methods have the third rank. Therefore we can reduce the dimensionality of data set and have the same result.

5 Conclusion

In this paper, in order to evaluate the efficiency of DMUs with undesirable outputs, we have taken the steps of PCA model on inputs and outputs. Then we have used an adjusted PCA-DEA model to evaluate DMU's performance, based on the chosen principal components. It is clear that the results of two models are similar and there are no significant differences between them. Finally, the propose approach has been applied to real data sets of Guilan Saderat Bank. We have ranked 25 branches of this bank, using the presented models and consequently it has been shown that the rank of branches is so similar.

References

[1] A.R. Amirteimoori, S. Kordrostami, A. Masoumzadeh, Attractiveness and Progress of DMUs with Context Model, Journal of Applied Mathematics, Azad University of Lahijan 8 (2006) 51-60.

[2] P. Anderson, N.C. Peterson, A Procedure for Ranking Efficient Unit in Data Envelopment Analysis. Management Science 39 (1993) 1261-1264.

[3] R.D. Banker, A. Charnes, W.W. Cooper, Models for Estimation of Technical and Scale Inefficiencies in Data Envelopment Analysis, Management Science 30 (1984) 1078-1092.

[4] B.W. Bloch, C.T. Huang, Multivariate Statistical Methods for Business and Economics. Prentice-Hall Englewood Cliffs, New Jersey. 1974.

[5] A. Charnes, W.W. Cooper, E. Rhodes , Measuring the Efficiency of Decision Making Units. European Journal of Opearational Research 2 (1978) 429-444.

[6] A. Charnes, W.W. Cooper,, Programming With Llinear Fractional. Navel Research Logistics Quarterly 15 (1992) 333-334.

[7] C.S. Cinca, M. Molinero, Selecting DEA Specifications and Ranking Units Via PCA, Journal of the Operational Research Society 55 (2004) 521-528.

[8] S. Kordrostami, A.R. Amirteimoori, A. Masoumzadeh, Evaluate The Efficiency of DMUs with Undesirable Outputs And Reduce The Dimensionality Of Data sets With PCA , Journal of Applied Mathematics, Azad University of Lahijan. Article in press.

[9] L. Liang, L. Yongjun, L. Shibing, Increasing The Discriminatory Power of DEA in The Presence of Undesirable Outputs and Larg Dimensionality of Data sets with PCA.Expert Systems With Applications, 2009, 5895-5899.

[10] I.M. Premachandra, A Note on DEA Vs Principal Component Analysis: An Improvement to Joe Zhu 's Approach 132 (2001) 553-560.

[11] D. Slottje, G.W. Scully, J.G. Ltirschberg, K.J. Hayes, Measuring the Quality of Life Across Countries: A Multidimen Sional Analysis, Westivew Press, Boulder, Co. 1991.

[12] J. Zhu, Data Envelopment Analysis Vs Principal Component Analysis: An Illustrative Study of Economic Performance of Chinese Cities, European Journal of Operational Research 111 (1998) 50-61.