G.U. Journal of Science
19(2): 127-133 (2006)

GUJS
www.gujs.org

The Comparison of Principal Component Analysis and Data Envelopment Analysis in Ranking of Decision Making Units

Filiz KARDİYEN, H.Hasan ÖRKCÜ[♣]

Gazi University Arts and Science Faculty, Department of Statistics
Teknikokullar, 06500, Ankara, TURKEY, hhorkcu@gazi.edu.tr

Received:05.04.2005 Accepted: 06.12.2005

ABSTRACT

In this study, Data Envelopment Analysis (DEA) and Principal Component Analysis (PCA) were compared when these two methods are used for ranking Decision Making Units (DMU) with multiple inputs and outputs. DEA, a nonstatistical technique, is a methodology using a linear programming model for evaluating and ranking DMU's performance. PCA, a multivariate statistical method, uses new measures defined by DMU's inputs and outputs. The results of both methods were applied to a real data set that indicates the economic performances of European Union member countries and also, a simulation study was done for different sample sizes and for different numbers of input-output, and the results were examined. For both applications, consistent results were obtained. Spearman's correlation test is employed to compare the rankings obtained by PCA and DEA.

Key Words: Principal Component Analysis, Data Envelopment Analysis, Ranking

1. INTRODUCTION

Data Envelopment Analysis is a methodology that uses linear programming in the evaluation of the relative efficiency of decision making units (DMU) with common inputs and outputs. DEA is not only used to determine efficient and nonefficient units but recently, it is also used to rank DMUs. CCR model, basis of DEA, was first proposed by Charnes, Cooper and Rhodes [1] and then extended by Banker [2] (Banker, Charnes and Cooper (BCC) model). These methods are called classical models and they can not be used in ranking efficient units. Andersen and Petersen [3], provided ranking of efficient units through improving these methods.

DEA, has been used commonly in a variety of fields since it was developed and its development continues through interacting with other techniques. Since the method can be applied to multiple inputs and outputs, it interacts with multivariate statistical methods. The association of DEA with PCA and Canonical Correlation Analysis (CCA) is shown in some studies. PCA is a multivariate analysis method used to destroy the independence structure between variables or to reduce the number of dimensions. Moreover, it can be used for ranking units. Therefore, DEA and PCA can be compared for ranking decision making units [4].

In this study, DEA and PCA were compared when these two methods are used for ranking DMUs with multiple inputs and outputs. Since the model used for measuring relative efficiency in DEA is based on the output/input ratio, treatments were performed by using the output/input ratio instead of the original input data in PCA in order to compare these two methods. Different to other studies which use only real data set, in this study, together with the real data set we made a simulation study comprising of a total of 20000 trials for different situations (number of DMU and number of input-output) to make our results more reliable. In section 2, DEA and its use in ranking DMUs are explained. In section 3, PCA and the use of its adaption to DEA for ranking DMUs are introduced. In section 4, these two methods are compared on a real data set and in section 5, the results of this comparison are supported by simulation results. Results of the research are presented in the last section.

2. DATA ENVELOPMENT ANALYSIS

DEA, a nonparametric efficiency method, was first proposed by Charnes, Cooper and Rhodes to measure the relative efficiency of DMUs' that are similar to each other in terms of products or services. The characteristics of this method are the ability to define measure of inefficiecy and resources in each DMU and since the efficiency value of each

[♣]Corresponding author, e-mail: hhorkcu@gazi.edu.tr

DMU is computed relative to each other, computed efficiencies are relative and they do not make functional assumption on variables [1].

Charnes, Cooper and Rhodes developed Farrell's ideas and the efficiency value obtained by dividing single output to single input was extended to multiple output/input ratio. By this means, there is an artificial output and input for each DMU and the efficiency value of DMUs can be computed by the help of these artificial inputs and outputs. Here, weights are chosen in a way that the efficiency values will not be greater than 1 [5].

In DEA, there are various models used for measuring efficiency and these models are derived from a ratio model measured by dividing the weighted sum of outputs to that of inputs [6].

Denoting j th unit's i th input and r th output by $x_{ij}\ (i=1,2,\ ...\ ,m)$ and $y_{rj}\ (r=1,2,\ ...\ ,s)$ respectively, ratio form can be defined as follows:

$$\begin{aligned}&\max\ h_j=\sum_{r=1}^{s}u_r y_{ro}\Big/\sum_{i=1}^{m}v_i x_{io}\\&\sum_{r=1}^{s}u_r y_{rj}\Big/\sum_{i=1}^{m}v_i x_{ij}\le 1\quad j=1,2,\ ...\ ,n\\&u_r,v_i\ge 0\end{aligned}\qquad(1)$$

where u_r and v_i are the input and the output weights respectively. Equating measured DMU's weigthted sum of inputs ($\sum_{i=1}^{m}v_i x_{io}$) to 1, the fundamental efficiency model, CCR Model is obtained.

$$\begin{aligned}&\max\ h_j=\sum_{r=1}^{s}u_r y_{ro}\\&\sum_{r=1}^{s}u_r y_{rj}-\sum_{i=1}^{m}v_i x_{ij}\le 0\qquad j=1,2,\ ...\ ,n\\&\sum_{i=1}^{m}v_i x_{io}=1\\&u_r,v_i\ge 0\quad i=1,2,\ ...\ ,m\ ;\ r=1,2,\ ...\ ,s\end{aligned}\qquad(2)$$

As the efficiency of DMUs is measured with this model, the model is needs to be solved for each DMU i.e. n times. The optimal goal function gives the efficiency score of the corresponding DMU. Each DMU whose efficiency score equals to 1, $h_j=1$, is evaluated as efficient. Each unit whose efficiency score is less than 1 is evaluated as inefficient.

In DEA, variables need to be separated as input and output. The discrimination of variables as input and output is dependent on their effect on the unit. Retzlaff-Roberts showed that it will be more accurate to use the concept of positive effective and negative effective variables instead of input and output variables. They proposed that variables whose increase provides the better evaluation of unit are taken as positive effective, in conrast variables whose decrease provides the better evaluation of unit are taken as negative effective [7].

In DEA, DMUs are ranked according to the efficiency scores obtained at the end of the analysis. The DMU that has the highest efficiency score occurs in the first place while the DMU that has the lowest efficiency score occurs in the last place. However, since efficiency score of all DMUs that are effective in DEA are assigned as "1", it is not possible to rank effective units between each other. DEA can be used only for ranking inefficient DMUs and in order to abolish this disadvantage various methods were developed [8]. The most commonly used method developed for ranking efficient decision units is the super efficiency model proposed by Andersen and Petersen [3]. The basic idea in this model is to compare the analyzed decision making unit with the linear combinations of all the other decision making units. The decision making unit that has the highest super efficiency score occurs in the first place. The other decision making units are ranked in descending order according to their super efficiency scores.

The super efficiency model, CCR Model, for analyzed decision making unit is defined as follows:

$$\begin{aligned}&\max h_j=\sum_{r=1}^{s}u_r y_{rj}\\&\sum_{i=1}^{m}v_i x_{ij}=1\\&\sum_{i=1}^{m}u_r y_{rj}-\sum_{i=1}^{m}v_i x_{ij}\le 0, j=1,2,...,m,\ j\ne 0\\&u_r,v_i\ge 0,\forall r,i\end{aligned}\qquad(3)$$

where o denotes the analyzed decision making unit and $j\neq o$ means removing the analyzed decision making unit from the constraint group, this is the basic idea of the super efficiency model.

3. USING PRINCIPAL COMPONENT ANALYSIS FOR RANKING DECISION MAKING UNITS

PCA is a statistical method that explains the correlation structure explained by the correlated number of p variables with the uncorrelated number of k variables which the linear combinations of the original variables provide ($p>k$). Eigen values and eigen vectors of the covariance or correlation matrices are used to find the linear combinations of the p variables in the $\boldsymbol{X}$ data matrix. Let $\boldsymbol{\Sigma}$ be the covariance matrix and $\boldsymbol{\rho}$ the correlation matrix of the random vector $\boldsymbol{X'}\ X[X_{2}..X_{p}]$ and $\lambda_1\ge\lambda_2\ge\ ...\ \ge\lambda_p$ the eigen values and, $l_1,l_2,\ ...\ ,l_p$ the ortagonal eigen vectors of the correlation matrix . Linear combinations of the variables can be calculated as $PC_i = l_i' X$, $(i=1,2,\ \ldots,p)$. The explanation ratio of total variance of k. principal component is described as $\frac{\lambda_k}{\lambda_1+\ \ldots\ +\lambda_p}$ [9].

The ratio of the weighted sum of output to the weighted sum of input intended to be maximized in the DEA is also used as a variable in PCA to provide correspondence of the two methods. Thus, for each DMU$_j$ $(j=1,2,\ \ldots,n)$

$$d_{ir}^j = \frac{y_{rj}}{x_{ij}} \ , \quad (i=1,\ \ldots,m\ ;\ r=1,\ \ldots,s)$$

ratios will be our new variables. Unlike the h_j in DEA, d_{ir}^j gives the ratio between every output and every input for each DMU. Here, the greater the d_{ir}^j, the better the performance of DMU$_j$ in terms of the r th output and the i th input.

Let $d_k^j = d_{ir}^j$, with, e.g., $k=1$ corresponds to $i=1,\ r=1$ and $k=2$ corresponds to $i=1,\ r=2$, etc., where $k=1,\ \ldots,p\ ; p=m\times s$. $n\times p$ data matrix composed by d_k^j is defined as follows:

$$D=\left(\underline{d}_1,\ \ldots,\underline{d}_p\right)_{n\times p}$$

where each row represents p indivudal ratios of d_k^j for each DMU and each column represents a specific output/input ratio. That is, $\underline{d}_k = \begin{bmatrix} d_k^1 & d_k^2 & . & . & . & d_k^n \end{bmatrix}_{1xn}'$, $k=1,\ \ldots,p$.

The aim of the PCA is to find out new independent measures which are different linear combinations of $\underline{d}_1,\ \ldots,\underline{d}_p$. These measures form a weighted measure of d_k^j. To do this, principal components are represented by their eigen values, this is the basic idea of the PCA (10).

For data matrix D, PCA is processed as follows:

Step 1: Correlation matrix of sample, (R), is computed.

Step 2: Eigen value and eigen vectors of the correlation matrix of the sample are computed.. For this, solving the following equation

$$\left|R-\lambda I_p\right|=0$$

where I_p is a $p\times p$ identity matrix, p eigen values $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ \ldots\ \geq \hat{\lambda}_p$ ($\sum_{k=1}^{p}\hat{\lambda}_k = p$) and the related p eigen vectors ($\hat{l}_1, \hat{l}_2,\ \ldots,\ \hat{l}_p$) are obtained.

Step 3: Principal components are computed. Each principal component is obtained by solving the following equation $PC_k = \sqrt{\hat{\lambda}_k}\ \hat{l}_k \quad (k=1,\ \ldots,p)$.

Step 4: The first m principal components are selected satisfying $\sum_{k=1}^{m}\hat{\lambda}_k / p > 0.90$.

Step 5: $t=\sum_{k=1}^{m} w_k PC_k$ gives a linear combination weighted with the explanation ratios of m principal components selected in step 4. For determining the signs w_k s , signs of the components of the PC_k are considered.According to this:

i) If all the components of the PC_k are negative, then the weight w_k is negative , and if all the components of the PC_k are positive, then the weight w_k is positive.

ii) If more than half of the components of the PC_k is negative then w_k is negative, otherwise it becomes positive.

Step 6: To use the principal components' scores in ranking, matrix $D=\left(\underline{d}_1,\ \ldots,\underline{d}_p\right)_{n\times p}$ is standardized and matrix $D_z=\left(\underline{d}_{z1},\ \ldots,\underline{d}_{zp}\right)_{n\times p}$ is obtained.

Step 7: Principal components scores are computed with the help of the equation $PC_{Skor} = D_z.t$ and units are ranked according to values of scores.

4. APPLICATION

In this section, a real data set was used to compare the performances of DEA and PCA. This data set consists of 2 input and 3 output variables chosen from money-prices, labor force market, national income and foreign trade categories that provides an evaluation from a financial aspect of 15 countries which were members of the European Union in 2002 in a financial aspect (11). These 15 EU countries were ranked according to these variables by both PCA and DEA methods and results of these methods were compared. Input and output variables are given below.

Input 1 (x_1): Inflation rate (%)

Input 2 (x_2): Unemployment rate (%)

Output 1 (y_1): Per Capita Gross National Product ($)

Output 2 (y_2): Portion in world export

Output 3 (y_3): Portion in world import

Data set is given in Table 1.

Table 1. Output (y) and Input (x) Values of 15 European Union Countries

Countries	y_1	y_2	y_3	x_1	x_2
Belgium	21966.67	3.332	2.989	1.6	9.33
Denmark	29600	0.870	0.727	2.4	5.33
Germany	23400	9.562	7.528	1.3	9.36
Greece	10533.33	0.161	0.475	3.6	10.80
Spain	13400	1.927	2.494	3.1	18.50
France	21600	4.820	4.676	1.9	11.76
Ireland	21166.67	1.364	0.785	4.7	7.73
Italy	18566.67	3.952	3.725	2.5	11.60
Luxemburg	38733.33	0.133	0.177	2.1	2.56
Holland	22466.67	3.469	2.953	3.5	4.16
Austria	23433.33	1.134	1.096	1.8	4.56
Portugal	10033.33	0.398	0.584	3.5	5.50
Finland	22300	0.696	0.513	1.7	11.43
Sweden	24500	1.266	1.008	2.2	8.46
England	21266.67	4.312	5.116	1.6	6.46

Six output/input ratios of the above two inputs and three outputs are computed:

$$d_1 = \frac{y_1}{x_1},\ d_2 = \frac{y_1}{x_2},\ d_3 = \frac{y_2}{x_1},\ d_4 = \frac{y_2}{x_2},$$

$$d_5 = \frac{y_3}{x_1},\ d_6 = \frac{y_3}{x_2}$$

Using matrix $D = \left(\underline{d}_1,\ .\ .\ .\ ,\underline{d}_6\right)_{15\times 6}$, results of the PCA computed based on the correlation matrix are summarized in Table 2 and Table 3.

Table 2. Eigen Values and Explanation Ratios of Total Variance

$\hat{\lambda}$	3.868	1.545	0.417	0.123	0.039	0.0006
Exp. Ratios	0.644	0.257	0.069	0.020	0.006	0.0001

Table 3. Coefficients of Principal Components

PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
-0.485	-0.824	-0.458	-0.568	-0.074	-0.004
-0.058	-0.558	0.404	0.486	0.078	0.013
-0.326	0.016	-0.331	0.478	-0.312	-.568
-0.477	0.065	0.440	-0.124	-0.630	0.400
-0.492	0.075	-0.281	0.306	0.492	0.579
-0.438	0.071	0.492	-0.319	0.500	-0.423

Acoording to Table 2, since first and second eigen values have an explanation ratio of %90.1 of the total variance, the principal components corresponding to these two eigen values will be used to compute the linear combination $t = \sum_{k=1}^{m} w_k PC_k$ in PCA step 5. w_k

weights' signs are determined as $w_1 = -0.644$ and $w_2 = 0.257$ according to the explanations about the w_k weights' signs in step 5. The linear combination will be as follows:

$$t = -0.644PC_1 + 0.257PC_2$$

PCA and DEA scores and ranking values computed for the data set in Table 1 are given in Table 4. The results of DEA are obtained by using model [3].

Table 4. Scores and Ranking Values Obtained by Both Two Methods

Countries	*PCA Scores*	*DEA-Super Efficiency Scores*	*PCA Ranking (U)*	*DEA Ranking (V)*
Belgium	1.386	0.751	4	6
Denmark	-1.456	0.666	9	8
Germany	6.732	2.729	1	1
Greece	-2.024	0.159	15	15
Spain	-1.916	0.237	14	14
France	1.504	1.826	3	3
Ireland	-1.674	0.286	12	13
Italy	0.390	0.533	7	10
Luxemburg	1.037	1.724	5	4
Holland	0.945	1.072	6	5
Austria	-0.784	0.708	8	7
Portugal	-1.803	0.316	13	12
Finland	-1.534	0.512	10	11
Sweden	-1.603	0.605	11	9
England	3.194	2.025	2	2

To test if there is a correlation between the ranking values obtained by both methods, Spearman's rank correlation coefficient is used (12).

$$r_s = 1 - \frac{6\sum_{i=1}^{n} dk_i^2}{n\left(n^2-1\right)},$$

$$\sum_{i=1}^{n} dk_i^2 = \sum_{i=1}^{n}\left(U-V\right)^2 \tag{4}$$

Test statistics in (4) were computed as $\left(r_s\right)_{comp} = 0.9571$, and comparison with the critical value $\left(r_s\right)_{table} = 0.7464$ shows that the H_0 hypothesis which claims that there is no correlation between the ranking values against the H_1 hypothesis which claims that there is a positive correlation between the ranking values was rejected at 0,1% level of significance.

Table 4 which shows results of both analysis are examined, Germany was found to be the most developed EU country and England and France were found to be the second and third developed EU countries, respectively, according to the ranking of EU countries for treated variables. In addition, results of both analysis showed that Greece is the least developed EU country and Spain is the second least developed EU country. Similar ranking values are observed when DEA and PCA ranking results of other EU countries are examined. These results were confirmed by testing Spearman's rank correlation coefficient. EU countries were seen to be ranked objectively and realistically, depending on the advantage of evaluation of input and output variables of DEA and PCA methods together.

In treatments performed for this data set, positive correlation was found between ranking values obtained by using DEA and PCA and it was concluded that these two methods can be used for one another when ranking DMUs.

5. SIMULATION

In section 4, a simulation study has been performed to test if we get the same successful results for different numbers of DMU and different numbers of output-input variable as we did for the real data set. In this study, observation values of output and input variables used to evaluate DMUs were chosen from uniform distribution (0,100) randomly. The study was repeated 1000 times for each situation by using different numbers of output-input variables for $n = 10,\ 20,\ 35$ and 50 sample sizes. For each situation, Spearman's correlation rank coefficient was used to test if there was a correlation between the ranking values obtained by DEA and PCA. The simulation study was performed with the by MATLAB 7 programme and the results are summarized in Table 5.

Table 5. Simulation Results

n	Num.Input Input	Num.Output	rhs	n	Num.Input	Num.Output	rhs
10	1	1	897	35	2	3	902
	1	2	912		2	4	905
	1	3	819		3	3	917
	2	1	902		3	4	924
	2	2	925		3	5	841
20	1	2	905	50	3	3	850
	2	2	913		3	4	897
	2	3	892		3	5	888
	3	2	878		4	2	855
	3	3	896		4	4	900

In the rhs column, the rejection number of the hypothesis which claims that there is not a correlation between the ranking values is given for each situation at each 1000 repetitions. For example, when there are 10 DMU' s, 1 input and 1 output, for 897 ($rhs = 897$) of 1000 trials the hypothesis which claims that there is not a correlation between the scores computed by the PCA and the DEA is rejected, so we concluded that there is a positive correlation between these scores.

6. CONCLUSION

In this study, two different methods were used to solve the problem "ranking decision units", which have a wide usage today. The first of these methods, DEA, uses a linear programming technique to obtain weighted input and weighted output ratio. The second method, PCA, is a multivariate statistical method which combines different ratios defined by each input and output with the use of eigen value and eigen vector information. The application of these two methods to a real data set from EU member states and comparison of PCA and DEA yielded consistent and valuable results. Spearman's rank correlation test showed that there is a high correlation between PCA and DEA ranking values for this data set. We carried out a simulation study that contains 1000 repetitions for different numbers of DMUs and different number of input-output variables; in order to be able to interpret the results without depending on a single data set. The results of the simulation supported the results of the real data set from EU member states; and in both cases, the hypothesis that there is no relation between ranking values of the PCA and the DEA has been rejected many times.

The application and simulation studies showed that PCA and DEA methods produce similar solutions to the DMU's ranking problem. Both methods yield comparable results in cases with multiple input and output. In this sense, DEA, a statistical method, and PCA, a non-statistical method, can be used for one another.

REFERENCES

[1] A. Charnes, W.W. Cooper, E. Rhodes, "The efficiency of decision making units", ***European Journal of Operational Research***, Vol. 2, pp.429-444, 1978.

[2] R.D. Banker, A. Charnes, W.W. Cooper, "Some models for estimating technical and scale inefficiencies in data envelopment analysis, ***Management Science***, Vol. 30, no. 19, 1078-1092 , 1984.

[3] P. Andersen, N. Petersen, "A procedure for ranking efficient units in Data Envelopment Analysis", ***Management Science***, Vol. 39, no. 10, pp. 1261-1264, 1993.

[4] I.M. Premachandra, "A note on DEA vs principal component analysis: An improvement to Joe Zhu's approach", ***European Journal of Operational Research***, vol. 132, pp. 553-560, 2001.

[5] W.W. Cooper, L.M. Seiford, K. Tone, "Data Envelopment Analysis", ***Kluwer Academic Publishers***, Boston USA, 100-175, 2000.

[6] A. Boussofiane, R.G. Dyson, E. Thanassoulis, "Applied data envelopment analysis", ***European Journal of Operational Research***, Vol. 52, pp. 1-15, 1991.

[7] D.L. Retzlaff-Roberts, "Relating discriminant analysis and data envelopment analysis to one another", ***European Journal of Operational Research*** , Vol. 23, pp. 311-322, 1996.

[8] N. Adler, L. Friedman, Z. Sinuany-Stern, "Review of ranking methods in the data envelopment analysis context", ***European Journal of Operational Research***, Vol. 140, pp. 249–265, 2002.

[9] H. Tatlıdil, "Uygulamalı Çok Değişkenli İstatistiksel Analiz", ***Cem Web Ofset***, Ankara, 1996.

[10] J. Zhu, “Data Envelopment Analysis vs principal component analysis: An illustrative study of economic performance of Chinese cities”, ***European Journal of Operational Research***, Vol. 111, pp. 50-61, 1998.

[11] www.foreigntrade.gov.tr.

[12] H. Gamgam, “Parametrik Olmayan İstatistiksel Teknikler”, ***Gazi Üniversitesi Yayınları***, Ankara, 1998.