World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:1, 2008

Joint Use of Factor Analysis (FA) and Data Envelopment Analysis (DEA) for Ranking of Data Envelopment Analysis

Reza Nadimi, Fariborz Jolai

***Abstract*—**This article combines two techniques: data envelopment analysis (DEA) and Factor analysis (FA) to data reduction in decision making units (DMU). Data envelopment analysis (DEA), a popular linear programming technique is useful to rate comparatively operational efficiency of decision making units (DMU) based on their deterministic (not necessarily stochastic) input–output data and factor analysis techniques, have been proposed as data reduction and classification technique, which can be applied in data envelopment analysis (DEA) technique for reduction input – output data. Numerical results reveal that the new approach shows a good consistency in ranking with DEA.

***Keywords*—**Effectiveness, Decision Making, Data Envelopment Analysis, Factor Analysis

I. Introduction

DATA envelopment analysis (DEA) initially proposed by Charnes et al.[1] is a non-parametric technique for measuring and evaluating the relative efficiencies of a set of entities, called decision-making units (DMUs), with the common inputs and outputs, and it is a linear programming-based technique that converts multiple input and output measures into a single comprehensive measure of productivity efficiency. DEA provides a measure by which one firm or department can compare its performance, in relative terms, to other homogeneous firms or departments. DEA is mainly utilized under two different circumstances. First, it can be used when a department from one firm wants to compare its level of efficiency performance against that of a corresponding department in other firms. Second, it can be used in a longitudinal nature by comparing the efficiency of a department or firm over time[2].

There are other combination methods in the DEA context. Adler and Golany[3], employed principle component analysis method to overcoming the difficulties that DEA encounters when the number of input and output data is excessive, they used of PCA to data reduction in inputs and outputs then applied PCA results into DEA model and compared achieved similar results. Adler and Berechman[4] adapted above approach to develop a model to determine the relative efficiency and quality of airport and showed their methodology have high relative efficiency. A new method, in multiple inputs and outputs forms have been proposed by Freidman and Sinuany_Stern[12], which have been evaluated DMUs based on Canonical Correlation Analysis and data envelopment analysis or CCA/DEA model.

This article proposes a combination of DEA and Factor Analysis (FA/DEA approach). The rest of this article is organized as follows. In Section 2, a brief description of the DEA models used in this article is presented. Section 3 gives the fundamental of factor Analysis technique. The FA/DEA approach is developed in Section 4. A numerical comparison of the FA/DEA and DEA procedures for consistency is presented in Section 5. Finally, Section 6 concludes this research.

Tehran University, Tehran, Iran (e-mail: fjolai@ut.ac.ir)

II. Data Envelopment Analysis

Data envelopment analysis (DEA), is analytical tool which first introduced by Charnes et al.[1], in 1978. It is the performance measurement technique that applies to evaluation the relative efficiency of decision-making units (DMU's) in organization such as banks, dental services, police, motor registries, hospitals etc.

Various models, used for computation efficiency of decision making units, such as CCR[1] and BCC[5]. The original fractional BCC model proposed by Banker *et al.*[5] evaluates the relative efficiency of $\boldsymbol{n}$ DMUs (j=1,..,n) , each with $\boldsymbol{m}$ inputs and $\boldsymbol{s}$ outputs denoted by $x_{1j},x_{2j},...,x_{mj}$ and $y_{1j},y_{2j},...,y_{sj}$, respectively , maximizing the ratio of weighted sum of outputs to the weighted sum of inputs, so, efficiency measure of DMU o according to the variable-returns-to-scale model is given as follows:

$$\min_{\lambda,s,\sigma} \varphi - \varepsilon(s+\sigma) \tag{1}$$

subject to

$$\boldsymbol{Y}\lambda - s = \boldsymbol{Y}^o \tag{2}$$

$$-\boldsymbol{X}\lambda - \sigma = -\phi\boldsymbol{X}^o \tag{3}$$

$$e^t\lambda = 1 \tag{4}$$

$$\lambda, s, \sigma \geq 0$$

Where λ is a vector of DMU weights which achieve from above linear program, ε^t a transposed vector of extremely small values or infinitesimal values, e^t a transposed vector of ones, σ and s are vectors of input and output slack, respectively, inhere, the column vector of input and output for DMU o have been showed by X_p and Y_p, and ϕ represented a constant which we use of ϕ to compare our approach results with results of DEA.

III. FACTOR ANALYSIS (FA)

Factor analysis is a statistical method that is based on the correlation analysis of multi-variables. The main applications of factor analytic techniques are: (1) to reduce the number of variables and (2) to detect structure in the relationships between variables, that is to classify variables. Therefore, factor analysis is applied as a data reduction or structure detection method. Factors are formed by grouping the variables that have a correlation with each other.

There are mainly three stages in Factor analysis [6]:

A correlation matrix is generated for all the variables. A correlation matrix is a rectangular array of the correlation coefficients of the variables with each other. Factors are extracted from the correlation matrix based on the correlation coefficients of the variables. The factors are rotated in order to maximize the relationship between the variables and some of the factors.

Let $\boldsymbol{d}_{(n\times 1)}$ be a random vector with a mean of $\boldsymbol{\mu}$ and a covariance matrix named $\boldsymbol{\Sigma}_{(p\times p)}$., where d_i specifies efficiency or an overall performance index of the i^{th} DMU. Then a k-factor model holds for $\boldsymbol{d}$, if it can be written in the following form:

$$\boldsymbol{d} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{u} + \boldsymbol{\mu} \tag{4}$$

where $\boldsymbol{H}_{(n\times k)}$ is a matrix of constants and $\boldsymbol{f}_{(k\times 1)}$ and $\boldsymbol{u}_{(n\times 1)}$ are random vectors. The elements of $\boldsymbol{f}$ are called *common* factors and the elements of $\boldsymbol{u}$ are *specific* or *unique* factors. In this study we shall suppose that:

$$\begin{aligned} &E(\boldsymbol{f}) = \boldsymbol{0},\ Cov(\boldsymbol{f}) = \boldsymbol{I} \\ &E(\boldsymbol{u}) = \boldsymbol{0},\ Cov(u_i,u_j) = 0;\ i\neq j \\ &Cov(\boldsymbol{f},\boldsymbol{u}) = \boldsymbol{0} \end{aligned} \tag{5}$$

Thus, if (5) holds, the covariance matrix of d can be split into two parts, as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{\Psi} \tag{6}$$

where $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$ is called the *communality* and represents the variance of d_i which is shared with the other variables via the common factors and $\boldsymbol{\Psi}$=$Cov(\boldsymbol{u})$ is called the *specific* or *unique variance* and is due to the unique factors $\boldsymbol{u}$. This matrix explains the variability in each d_i that is not shared with the other variables. The main goal of FA is to apply $\boldsymbol{f}$ instead of $\boldsymbol{d}$ for assessing DMUs. The number of factors would usually be determined by considering how well the model fits the data. Often a scree-test 1 is used for this[9]. Scree test is a criterion that in it eigenvalues plotted against factors. Factors in descending order, are arranged along the abscissa with eigenvalue as the ordinate, this graph is useful for determining how many factors to retain. Because the variance that each standardized variable contributes to a factor analysis extraction is one, a factor with an eigenvalue less than 1 is not as important, from a covariance perspective, as an observed variable. Thus, inhere, scree test is used to choice the number of factors. Let's summarize and formulize the above steps as follows. In this study, we skip the rotation step.

First, the correlation matrix, namely $\boldsymbol{R}$, is computed on the basis of data due to the variables, d_{ij}:

$$\boldsymbol{R} = Corr(\boldsymbol{D}) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} \tag{7}$$

where, $\boldsymbol{D}$ is an $n\times p$ matrix of p variables for n DMU's.

This matrix can be decomposed to a product of three matrices:

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}} \tag{8}$$

where, $\boldsymbol{V}$ is the $p\times p$ matrix of eigenvectors and $\boldsymbol{L} = Diag([\lambda_1, \ldots, \lambda_p])$ is a diagonal matrix of the eigenvalues, assorted descendingly. Suppose (9) is rewritten as follows:

$$\boldsymbol{R} = \left(\boldsymbol{V}\sqrt{\boldsymbol{L}}\right)\left(\sqrt{\boldsymbol{L}}\,\boldsymbol{V}^{\boldsymbol{T}}\right) \tag{9}$$

Equation(10) is frequently called the fundamental equation for FA. It represents the assertion that the correlation matrix is a product of the factor loading matrix, $\boldsymbol{A} = \left(\boldsymbol{V}\sqrt{\boldsymbol{L}}\right)$, and its transpose[9]. It can be shown that an estimate of the unique or specific variance matrix, $\boldsymbol{\Psi}$, in (7) is:

$$\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} \tag{10}$$

where $\boldsymbol{I}_{(p\times p)}$ is the identity matrix. So far our study of the factor model has been concerned with the way in which the observed variables are functions of the (unknown) factors, $\boldsymbol{f}$. Instead, factor scores can be estimated by the following pseudo-inverse method:

$$\boldsymbol{W}^{\mathrm{T}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B}^{-1} \tag{11}$$

$$\boldsymbol{F} = \boldsymbol{D}\boldsymbol{W} \tag{12}$$

where $\boldsymbol{F}$ is a $n\times p$ matrix, each row of which corresponds to a DMU. The estimate in (12) is known as Bartlett's factor score, and $\boldsymbol{W}$ is called the *factor score* coefficient matrix.

Generally, FA is used to data reduction (for more detail see [11-14]). In this paper, we use the FA technique to evaluate DMUs by reducing inputs and outputs whilst minimizing the loss of information. This will be introduced in the next section.

IV. NEW APPROACH: FA/DEA METHOD

In fact, we hereby want to apply equation (13) to change DEA input and output variable to less factors enabling interpret and analysis linear combinational inputs (outputs). The important mater is the reduction of inputs/outputs through factors. In order to reduce inputs/output, we primarily normalize input/output data matrix to execute factor analysis method. Using eigenvalue and eigenvector, number of important factors will be possible to find. So, after execute FA method on inputs($\boldsymbol{X}$) and outputs($\boldsymbol{Y}$) data, $\boldsymbol{X}_{\mathrm{FA}}$ and $\boldsymbol{Y}_{\mathrm{FA}}$ have been derived, respectively, so that, $\boldsymbol{X}_{\mathrm{FA}}$ =$\mathbf{X}\mathbf{W}_{\mathrm{FAx}}$ and $\boldsymbol{Y}_{\mathrm{FA}}$=$\mathbf{Y}$ $\mathbf{W}_{\mathrm{Fay}}$. Here, we can replace $\boldsymbol{Y}_{\mathrm{FA}}$ and $\boldsymbol{X}_{\mathrm{FA}}$ in (2), (3) respectively. Thus (2), (3) can be rewrite as following:

$$\boldsymbol{Y}_{\boldsymbol{FA}}\lambda - s_{FA} = \boldsymbol{Y}_{\boldsymbol{FA}}^{\boldsymbol{\theta}} \tag{13}$$

$$-\boldsymbol{X}_{\boldsymbol{FA}}\lambda - \sigma_{FA} = -\phi\boldsymbol{X}_{\boldsymbol{FA}}^{\boldsymbol{\theta}} \tag{14}$$

The slacks in objective function change and convert to ε (s_{FA} +σ_{FA}) that is given in follow:

$$\text{Min } \varphi - \varepsilon\left(W_{\mathrm{FAy}}^{-1}\left(s_{\mathrm{FA}}^{+} - s_{FA}^{-}\right) + W_{\mathrm{FAx}}^{-1}\left(\sigma_{\mathrm{FA}}^{+} - \sigma_{FA}^{-}\right)\right) \tag{15}$$

Where $W_{FAy}^{-1}\left(W_{FAx}^{-1}\right)$ represents the inverse matrix of output (input) weights which achieve of (12). As well, two constraints must be added to original formula DEA which is given in below:

$$W_{\mathrm{FAy}}^{-1}\left(\mathrm{s}_{\mathrm{FA}}^{+}-s_{FA}^{-}\right)\geq 0 \tag{16}$$

$$W_{\mathrm{FAx}}^{-1}\left(\sigma_{\mathrm{FA}}^{+}-\sigma_{FA}^{-}\right)\geq 0 \tag{17}$$

So above mention the changed formula can be rewritten as follows:

$$\begin{aligned}&\text{Min } \varphi-\varepsilon\left(W_{\mathrm{FAy}}^{-1}\left(\mathrm{s}_{\mathrm{FA}}^{+}-s_{FA}^{-}\right)+W_{\mathrm{FAx}}^{-1}\left(\sigma_{\mathrm{FA}}^{+}-\sigma_{FA}^{-}\right)\right)\\&\text{subject to}\end{aligned} \tag{18}$$

$$\boldsymbol{Y_{FA}}\lambda-s_{FA}=\boldsymbol{Y_{FA}^{\theta}} \tag{19}$$

$$-\boldsymbol{X_{FA}}\lambda-\sigma_{FA}=-\phi\boldsymbol{X_{FA}^{\theta}} \tag{20}$$

$$e^{t}\lambda=1 \tag{21}$$

$$W_{\mathrm{FAy}}^{-1}\left(\mathrm{s}_{\mathrm{FA}}^{+}-s_{FA}^{-}\right)\geq 0 \tag{22}$$

$$W_{\mathrm{FAx}}^{-1}\left(\sigma_{\mathrm{FA}}^{+}-\sigma_{FA}^{-}\right)\geq 0 \tag{23}$$

$$\lambda,\mathrm{s}_{\mathrm{FA}}^{+},s_{FA}^{-},\sigma_{FA}^{+},\sigma_{FA}^{-}\geq 0$$

If we have non positive factor f_{ij} ($\boldsymbol{F}$= {$\boldsymbol{f1}$, $\boldsymbol{f}_2$...} = {fij} i=1,2,...,n, j=1,2,...,p) in the output/input data, we must set the following rule.

We have the fine related minimum $\boldsymbol{f}_j$ of every column and continue as follows:

$$\begin{aligned}&f_{ij}=f_{ij}-l_j+1,\\&\text{where } l_j=\min\{f_{ij}\}\quad j=1,2,...,p\end{aligned} \tag{24}$$

In this article we use of the MAPE criteria to compare our approach with Adler method, which it is given in follows:

$$MSE=\frac{1}{n}\sum_{t=1}^{n}\left(effic._{*}-effic._{DEA}\right)^{2} \tag{25}$$

Where *effic*. is abbreviate of efficiency and $*$ in $effic._{*}$ has been used to FA or PCA efficiency values. Here, above criteria which explained in before section, is applied to compare two approaches (our approach and Adler method), of course we used above criteria for input and output, also. In fact, square of eigenvalues is used to determination efficiency with high accuracy while there is no this item in Adler models.

V. Numerical Example

A.*Example 1:* Adler et al.[3] applied the BCC model on data from 10 different banks to evaluate the relative efficiencies (Table II). Data is analyzed by the DEA&PCA (Adler), DEA (original) and DEA&FA (proposed approach) approaches to evaluate the relative efficiencies.

TABLE I FA RESULTS OF THE 10 BANKS EXAMPLE

Eigen values and factor scores							
	Input			Output			
Eigen Values	1.7497	0.2503		2.2419	1.2777	0.3282	0.1523
factor score	$\boldsymbol{w}_{\text{inp.1}}$	$\boldsymbol{w}_{\text{inp.2}}$		$\boldsymbol{w}_{\text{out.1}}$	$\boldsymbol{w}_{\text{out.2}}$	$\boldsymbol{w}_{\text{out.3}}$	$\boldsymbol{w}_{\text{out.4}}$
	0.94	-0.35		0.82	-0.49	-0.14	0.26
	0.94	0.35		-0.03	-0.98	0.13	-0.18
				0.86	0.24	0.45	-0.02
				0.91	0.18	-0.29	-0.23

Then the results of proposed approach Adler method compare with together. Table I shows that there are one and two eigenvalues greater than one for input and output, respectively. Thus we use of one input and two outputs to run DEA&FA models and with this same number of inputs and outputs execute Adler's models. The results of two approaches and DEA model are given in Table II.

MSE criteria is calculated for two approaches and its value is 0.80, 3.11, for FA&DEA and PCA&DEA, respectively. This example shows our and DEA method's efficiency is higher than Adler approach. Thus our approach could be good replacement approach of PCA&DEA in comparison of DMU's.

Example 2: In order to illustrate our view point, here, we apply data used by Zhu [7]. This data sets are about economic performance of 18 china cites. The following contents are defined [7]:

x_1: Investment in fixed assets by state owned enterprises
x_2: Foreign funds actually used
y_1: Total industrial output
y_2: Total value of retail sales
y_3: Handling capacity of coastal ports

Obviously, x_1 and x_2 can be assumed as two inputs and y_1, y_2, and y_3 as three outputs, data of which are presented in Table IV. Eigenvalues and factor score coefficient are summarized in Table III. In this example, two and three eigenvalues have been demonstrated for input and output data, respectively. MSE criteria to FA&DEA and PCA&DEA is 0.04, 0.38, respectively. This example is showed that MSE criteria for our method is lower than Adler approach, so, proposed method's performance is good consistency with DEA method.

TABLE II DEA INPUT AND OUTPUT DATA AND RESULT WITH OUR AND ADLER APPROACH

Banks	Original data						Result of DEA		
	Input 1	Input 2	Output 1	Output 2	Output 3	Output 4	DEA (Original)	DEA (Adler)	DEA (Proposed)
1	170	70	45	6	11	5	0.96	0.96	1
2	155	85	53	11	9	7	1	1	1
3	183	92	48	23	4	2	1	1	1
4	143	62	28	7	3	1.8	0.83	0.44	1
5	202	88	60	17	5	3	1	1	1
6	117	49	35	12	4	1.7	1	1	1
7	143	44	27	8	3	1	1	0.66	1
8	155	61	33	17	6	2	1	0.80	0.77
9	139	53	42	8	7	3	1	1	1
10	183	63	52	12	15	4	1	1	1

TABLE III FA RESULTS OF THE 18 CHINA CITES EXAMPLE

	Input			Output		
Eigen Values	1.4028	0.1791		1.6719	0.4315	0.1367
factor score	$w_{inp.1}$	$w_{inp.2}$		$w_{out.1}$	$w_{out.2}$	$w_{out.3}$
	0.99	-0.13		0.98	0.16	0.10
	0.99	0.13		0.98	0.18	-0.09
				0.93	-0.36	-0.01

VI. Conclusion

The current article presents alternative approach to evaluate DMUs which have multiple outputs and multiple inputs. The DEA -non statistical method– use linear programming technique to obtain a ration between weighted output and weighted input. Our approach is combination data envelopment analysis with factor analysis to evaluate efficiency DMUs. Factor analysis is a multivariate statistical method that uses information obtained from eignvalue to combine different ratio measures defined by every input and every output. Numerical experimental results showed that there is difference between high correlations between two diverse methods. Thus, we can use from DEA&FA to evaluate efficiency DMUs instead original DEA and without lose of information.

TABLE IV DEA INPUT AND OUTPUT DATA AND RESULT WITH OUR AND ADLER APPROACH

China cites	Original data					Result of DEA		
	Input 1	Input 2	Output 1	Output 2	Output 3	DEA (Original)	DEA (Adler)	DEA (Proposed)
1	2874.8	16738	160.89	80800	5092	1.00	0.08	0.87
2	946.3	691	21.14	18172	6563	1.00	0.29	1.00
3	6854	43024	375.25	144530	2437	0.87	0.06	0.66
4	2305.1	10815	176.68	70318	3145	0.94	0.11	0.79
5	1010.3	2099	102.12	55419	1225	1.00	0.35	0.89
6	282.3	757	59.17	27422	246	1.00	0.52	1.00
7	17478.6	116900	1029.09	351390	14604	1.00	1.00	1.00
8	661.8	2024	30.07	23550	1126	0.51	0.18	0.89
9	1544.2	3218	160.58	59406	2230	1.00	0.25	0.83
10	428.4	574	53.69	47504	430	1.00	0.93	1.00
11	6228.1	29842	258.09	151356	4649	1.00	0.08	0.81
12	697.7	3394	38.02	45336	1555	0.92	0.22	0.89
13	106.4	367	7.07	8236	121	1.00	0.34	1.00
14	5439.3	45809	116.46	56135	956	0.21	0.02	0.30
15	957.8	16947	29.2	17554	231	0.20	0.02	0.58
16	1209.2	15741	65.36	62341	618	1.00	0.07	0.62
17	972.4	23822	54.52	25203	513	0.31	0.02	0.52
18	2192	10943	25.24	40267	895	0.20	0.06	0.58

REFERENCES

[1] A. Charnes, W.W. Cooper, E. Rhodes, "Measuring the efficiency of decision making units", *European Journal of Operations Research* 2 (1978) 429 -444.
[2] L. Easton, D.J. Murphy, J.N. Pearson, "Purchasing performance evaluation: with data envelopment analysis", *European Journal of Purchasing & Supply Management* 8 (2002) 123–134.
[3] N. Adler, B. Golany, "Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe", *European Journal of Operations Research* 132, (2001) 260-273.
[4] N. Adler, J. Berechman, "Measuring airport quality from the airlines' viewpoint: an application of data envelopment analysis", *Transport Policy* 8 (2001) 171-181.
[5] R.D.Banker, A.Charnes,W.W.Cooper, "Some models for estimating technical and scale inefficiencies in data envelopment analysis", *Management Science* 30(9)(1984)1079-1092
[6] How to perform and interpret Factor analysis using SPSS, www.ncl.ac.Uk/iss/statistics/docs/Factoranalysis.html, 2002.
[7] J.Zhu, "Data envelopment analysis vs principal component analysis : An illustrative study of economic performance of Chinese cities", *European Journal of Operation Research* 111,(1998) 50-61.
[8] M.K. Epstein, J.C. Henderson, "Data envelopment analysis for managerial control and diagnosis", *Decision Science* 20, (1989) 90-119.
[9] B.S. Everitt & G. Dunn, "Applied Multivariate Data Analysis", Edward Arnold, *London*, pp304 (1991).
[10] T. Hastie, R. Tibshirani, "Discriminant analysis by Gaussian mixtures", J. Roy. Statist. Soc, B 58 (1996) 155–176.
[11] J.D. Banfield, A.E. Raftery, "Model-based Gaussian and non-Gaussian clustering", Biometrics, 49 (1993) 803–821.
[12] J.D. Banfield, A.E. Raftery, "Model-based clustering, discriminant analysis, and density estimation", J. Amer. Statist. Assoc, 97 (2002) 611–631.
[13] D.G. Calò, "Gaussian mixture model classification: a projection pursuit approach", Comput. Statist. Data Anal., 52 (2007) 471–482.