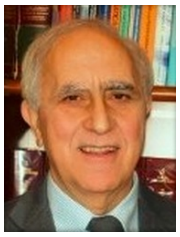
JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

PRINCIPAL COMPONENT ANALYSIS TO RANKING TECHNICAL EFFICIENCIES THROUGH STOCHASTIC FRONTIER ANALYSIS AND DEA

Sergio SCIPPACERCOLA
Associate Professor, Department of Economics, Management, Institutions
University of Naples "Federico II", Italy



E-mail: sergio.scippacercola@unina.it

Enrica SEPE
PhD Candidate, Department of Economics, Management, Institutions
University of Naples "Federico II", Italy

E-mail: enrica.sepe@unina.it

Abstract

The Stochastic Frontier Analysis permits evaluating the Technical Efficiency scores for one output variable to obtain the corresponding Technical Efficiency of n Decision-Making Units (DMU). The objective of this work is a comparison between a Stochastic Frontier Analysis, with same input and different output variables, and the Data Envelopment Analysis. You get k Technical Efficiency TE(yi) which are unified by a Principal Component Analysis and compared with the results of a DEA on the same data.

Keywords: Principal Component Analysis; Stochastic Frontier Analysis; Technical Efficiency; Data Envelopment Analysis; Secondary Schools

1. Introduction

The evaluation of Technical Efficiency (*TE*) is a fundamental tool for seeing which determinants slow down the development of production. We have two distinct approaches to evaluating Technical Efficiency, namely a parametric approach which is Stochastic Frontier Analysis (*SFA*) and a non-parametric deterministic approach which is Data Envelopment Analysis (*DEA*). DEA is an approach which uses mathematical programming to identify the efficient frontier, and does not impose functional forms (Kumbhakar, Lovell, 2003; Ray, 2004; Cooper, 2006). The main advantage of DEA is that it does not require any hypothesis about the analytical form of the production function. In DEA we have many inputs and many outputs jointly considered. DEA is based on the chosen inputs and outputs of entities that are named Decision-Making Units (DMUs). For example, all the schools (DMUs) are compared in relationship to the "best" performing schools. DEA is a non-parametric linear programming method for assessing the efficiency of (DMUs).

SFA requires strong distribution assumptions of both statistical random errors (i.e. normal distribution) and non-negative technical inefficiency random variables. SFA

considers many input variables (x_1, x_2, ..., x_k) but only one output variable (y). Our proposal is very interesting when we have more than one output variable and with SFA cannot be considered jointly as happens with DEA. Our goal is therefore to make more SFAs and unify *TEs* into a single list as for DEA. So, the main objective of this work is to obtain a single ranking of different SFAs**.**

Section 2 introduces the Stochastic production frontier methodology and Section 3 the Data Envelopment Analysis. Afterwards, in Section 4 we suggest how to organize the SFA with the same input and different outputs in order to obtain a synthetic indicator of efficiency instead of many outputs in accordance with the hypothesis of the stochastic model; an application follows of the methodology on a real case (Secondary Schools) with a brief discussion of the main findings. Finally, in Section 5, our comparison between DEA and SFA is presented.

2. Stochastic Frontier Analysis

The SFA is a *parametric approach* that hypothesizes a functional form and uses the data to econometrically estimate the parameters of this function. SFA requires functional forms on the production frontier, and assumes that units may deviate from the production frontier not only owing to technical inefficiency but also to measurement errors, statistical noise or to other non-systematic factors. In addition, the SFA requires strong distribution assumptions of both statistical random errors (i.e. normal distribution) and the non-negative technical inefficiency random variables (i.e. half-normal or truncated normal distribution) (Coelli et. al., 2005). The Stochastic Frontier Analysis searches for the production function, which represents the maximum output attainable given a certain quantity of inputs (Rao et al., 2005).

The *first step* of SFA consists in the specification and in the estimation of the stochastic frontier production function as well as in the estimation of technical inefficiency effects, assuming that these inefficiency effects are identically distributed. SFA methodology allows a functional form and the breakdown of the inefficiency error term. SFA is a parametric approach that hypothesizes a functional form and uses the data to econometrically estimate the parameters of this function. A production function ***f*** is defined as the schedule of the maximum amount of output that can be produced from a specified set of inputs, given the existing technology. The model of the Stochastic Frontier Analysis is (Rao et al., 2005):

$$\ln y_i = x_i'\beta + v_i - u_i \qquad (1)$$

where $\boldsymbol{y_i}$ is the output of the *n-th* producer (i.e. DMU), $\boldsymbol{x_i}$ is a vector of inputs, $\boldsymbol{\beta}$ is a vector of *k+1* parameters to be estimated, $v_i \approx iid\, N(0, \sigma_v^2)$ is the noise or error term or the measure of effects independent of the producer, $\boldsymbol{v_i}$ is homoskedastic; $\boldsymbol{u_i}$ is *iid*, $\boldsymbol{u_i}$ is a non-negative random variable measuring the technical inefficiency with $N^+(0, \sigma_u^2)$ (half-normal either normal-truncated model $N^+(\mu, \sigma_u^2)$ or exponential or gamma); $\boldsymbol{v_i}$ and $\boldsymbol{u_i}$ are distributed independently of each other and of the regressors. We can define the Technical Efficiency (*TE*) as the ratio of realised output to the stochastic frontier output:

$$\ln TE_i = \ln y_i - \ln y_i^* = \ln(y_i/y_i^*) = -u_i \qquad (0 \le TE \le 1) \qquad (2).$$

The parameters of stochastic frontier function are estimated by the maximum likelihood method. An estimation of stochastic frontier is the use of the γ (Battese and Corra,1977):

$$\gamma = \sigma_u^2/(\sigma_u^2 + \sigma_v^2).$$

When the parameter $\gamma = 0$ the variance of the technical inefficiency effect is zero, if γ *is close to one* it indicates the deviations from the frontier are due mostly to technical inefficiency, and if $\gamma = 1$ it indicates that one-sided error component dominates the symmetric error component.

The main hypothesis of interest of the SFA is:

$$\mathrm{H}_0 : \beta_1 = \ ... = \beta_q = 0 \quad q < K \ .$$

The omission of u_i is equivalent to imposing the restriction specified in the null hypotheses i.e.

$$\mathrm{H}_0 : \gamma = \delta_0 = \cdots = \delta_J = 0.$$

This indicates that the inefficiency effects in the frontier model are not present (no efficiency). Null hypotheses of interest are tested using the generalized likelihood ratio. The null hypothesis is $\mathrm{H}_0 : \gamma = 0$ which specifies that technical inefficiency effects are not stochastic.

We reject the null hypothesis of no technical inefficiency effects given the specifications of the stochastic frontier and inefficiency effect model. If the parameter $\gamma = 0$ we accept null hypothesis then the variance of the technical inefficiency effect is zero and so the model reduces to the traditional mean response function. Leaving a specification with parameters that can be consistently estimated using ordinary least squares.

The *second step* of SFA involves the specification of a regression model for predicted technical inefficiency effects. OLS is inappropriate and either the dependent variable must be transformed prior to estimation or a limited dependent variable estimation technique must be employed.

3. Data Envelopment Analysis

The parametric method involves the application of econometric techniques where efficiency is measured relative to a statistically estimated frontier production function. The non-parametric method revolves around mathematical programming techniques, the most commonly applied of which is generically referred to as DEA. In this case, the former body of method imposes a particular functional form, while the latter does not. Therefore, another linear programming method for assessing the efficiency and productivity units is the Data Envelopment Analysis. In particular, DEA is a non-parametric linear programming method for assessing the efficiency and productivity units called decision-making units (DMUs) because they enjoy a certain decision-making autonomy. DEA application areas have grown since it was first introduced as a managerial and performance measurement tool in the late 1970s. The DEA approach was introduced by Charnes et al. (1978) who proposed the

efficiency measurement of the DMUs for constant returns to scale (CRS), where all DMUs are operating at their optimal scale. Later Banker et al. (1984) introduced the variable returns to scale (VRS) efficiency measurement model, allowing the breakdown of efficiency into technical and scale efficiencies in DEA.

Over the last few decades, data envelopment analysis has gained considerable attention as a managerial tool for measuring the performance of organizations, and it has been used widely for assessing the efficiency of public and private sectors. This method leads to the System Selection of the optimal weights for the generic DMUs, and to the solution of a mathematical programming model in which the decision variables are represented by the weights associated with each input and output unit. DEA allows multiple inputs–outputs to be considered at the same time without any assumption on data distribution. In each case, the efficiency is measured in terms of a proportional change in inputs or outputs. A DEA model can be subdivided into an input oriented model, which minimizes inputs while satisfying at least the given output levels, and an output-oriented model, which maximizes outputs without requiring any more observed input values. The most well-known is represented by the *input oriented* CRS efficiency (Charnes, et. al., 1978), where the formulation of the linear optimization problem, for the i-th DMU, is :

$$max_{f,l}\ f, \qquad \text{subject to} \begin{cases} -fy_i + Y\lambda \geq 0 \\ -fy_i + Y\lambda \geq 0, \\ x_i - X\lambda \geq 0, \\ \lambda \geq 0, \end{cases} \qquad (3)$$

where *X* is a matrix of *kxn* input and *Y* is a matrix of *mxn* output, with *n* equal to the number of DMU, y_i and x_i are the outputs and inputs observed for the i-th DMU, *f* is a scalar $(1 \leq f \leq +\infty)$ and λ is a constant vector of *nx1*. The score of technical efficiency for the DMU is represented by the quantity *1 / f* , and varies therefore, between *0* and *1* (*f = 1* denotes a DMU that stands on the frontier of production and is therefore technically efficient).

Another goal of the input-oriented DEA model is to minimize the virtual input, relative to a given virtual output, subject to the constraint that no DMU can operate beyond the production possibility set and the constraint relating to non-negative weights. In practice, most of the available DEA programs use the dual forms as expressed in (4), which lower the calculation burden and are virtually the same as (3):

$$min_{\theta\lambda,}\ \theta, \qquad (4)$$

where λ is a semipositive vector in R^k and θ is a real variable.

In this paper we use the *input-oriented CRS* model to compare the results of the SFA; however, other variations are easily extendable and available in most DEA literature, including Coelli et al. (2005) and Cooper et al. (2006).

4. SFA with same inputs and different outputs

The Stochastic Frontier Analysis permits evaluating the technical efficiency scores for the input variables (x_1, x_2, ..., x_k) with output y_1 and to obtain a measure of the Technical Efficiency (TE_1) that can be called $TE(y_1)$ i.e. a technical efficiency that is a function of y_1. We suggest performing multiple SFA with the same group of input variables (x_1, x_2, ..., x_k) but with different output variables (y_j) *(j=2, ...,k)*. For each *i-th* SFA we have the corresponding TE(y_i) with continuous values in [0,1]. Each indicator of efficiency $TE(y_i)$

obtained by each SFA, can be transformed into values on an ordinal scale. You obtain *k* rankings each due to a specific input variable used (y_j). It becomes, therefore, a problem of ordering multivariate data of an ordinal type. In a lot of applications we are interested in a unified ranking of the DMU rather than in values of the single Technical Efficiency.

In order to obtain a single graduation, you can use a Principal Component Analysis in considering the *TE* (y_i) ($j=1,2, \ldots,k$) as variables. You may grade the DMU according to the score on the first axis, but you obtain a ranking that is dependent on the first eigenvalue. The scores on the first principal component furnish an approximate indication of the probable ranking of the DMU. However, because the first principal component maximizes the weighted sum of squares of the correlation coefficients between the original variables and the first principal component, we will use this ranking that permits obtaining a unified ranking.

We use the data gathered from an official survey performed by the school management of the Campania Region *(Cometa project)*. The schools surveyed by the Regional School District will be at the end of the investigation, being more than a thousand. In this work were examined only thirty-three schools that had given coherent and validate data. The survey covers attributes regarding: environment, territorial context and economic resources. We started with the model including all variables and interactions. The choice of the model is based on the Box-Cox transformation (Box and Cox, 1964), while the choice of the functional form has been carried out under the hypothesis of a parsimonious model by likelihood ratio test and AIC criteria (Akaike,1977). After significance tests, only certain variables have been kept on the list of the potential determinants of technical efficiency, that represent characteristics of the school and of the management/production. We started with the model including all variables and interactions. The choice of the functional form has been carried out under the hypothesis of a parsimonious model. The null hypothesis of absence of random technical inefficiency is rejected in the different specifications and thus the Stochastic Frontier Analysis seems appropriate for the data. After verifying the hypothesis of asymmetry present in the residuals of the OLS and after trying several models with different dependent variables, the first model of SFA (SFA1) is:

$$ln(y_{1i}) = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \beta_4\, x_{i4} + v_i - u_i \qquad (5),$$

where *i* refers to the i-th school, y_{i1} is the *number of students who have passed the average score in the national test respect to the number of students,* x_{i1} is the rate of *number of teachers who have worked for more than ten years,* x_{i2} is the *rate of use of laboratories with respect to the availability,* x_{i3} is the *rate of use gyms and sports equipment*, x_{i4} *rate of implementation of projects.* Variables v_i and u_i are defined as described in Section 2.1. In Table 1 are summarized the main results of model (5), based on data of 18 schools. The second (6) and the third model (7), SFA2 and SFA3 respectively, differ from (5) only for the output variable (y_{2i}, y_{3i}) :

$$ln(y_{2i}) = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \beta_4\, x_{i4} + v_i - u_i \qquad (6),$$

$$ln(y_{3i}) = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \beta_4\, x_{i4} + v_i - u_i \qquad (7),$$

where, in (6) y_{2i} refers to the *number of students who passed the secondary school-leaving examination with a score greater than 80/100* compared to the total number of examined students, while, in (7), y_{3i} represents the *number of regular students in the study with respect to the starting lever*.

The results (Table 1) of model (5) show that the production inputs as the rate of use of laboratories with respect to the availability and the rate of use of gyms and sports

equipment has a significant impact on the determination of the production frontier. Although positive, the presence is not significant of teachers with more than ten years' teaching experience as well as the rate of realization of projects. For reasons of space we do not report further comments on the results of model (6) and other models because we are interested mainly in the graduation of technical efficiency with respect to the stochastic frontier.

Table 1. Estimation results of Frontier Production with dependent variable being the number of students who have passed the average score in the national test with respect to the number of students

Input variables/Parameters	Coefficient	Standard Error	z	P>|z|	95% confidence interval	
Constant	2.2993660	.4785349	4.81	0.000	1.361455	3.2372770
x_{i1}	.0025374	.0026548	0.96	0.339	-.0026660	.0077408
x_{i2}	-.0084085	.0016368	-5.14	0.000	-.0116165	-.0052004
x_{i3}	-.0066347	.0037900	-1.75	0.080	-.0140631	.0007936
x_{i4}	.0068563	.0040808	1.68	0.093	-.0011420	.0148546
σ_u	.0930487	.0540731				
σ_v	.3810259	.0921572		γ =0.94		
Log likelihood = .760369 Prob > χ^2 = 0.0000						
Likelihood-ratio test of σ_u= 0 :		$\chi^2(01)$ = 2.72		Prob ≥ χ^2 =0.049		

Indeed, by means of the respective models (5), (6) and (7) were computed the Technical Efficiencies (Table 3) of individual schools (DMU) suitably codified.

We assume that the three Technical Efficiencies have been collected in a data matrix ***X***, in which the rows are associated with the DMU and the columns with the three Technical Efficiencies as variables. The principal components of the three Technical Efficiencies are obtained from the PCA on ***X***. We can see (Table 2) that about 47% of the total variation is explained by the first principal component indicating that there is some conflict among the individual rankings.

The first Principal Component is expected to approximate to the common ranking quite well, therefore the scores, transformed into rank (Table 3) could be used for comparison with the results from a Data Envelopment Analysis on the same data. Thus, by considering the Pearson's correlation coefficients of ***X*** (Table 4), we note that a low positive correlation exists (0.2717, 0.2500, 0.0548) among the three Technical Efficiencies. That concordance of sign of correlation, even if low, will ensure the success of the methodology. Conversely, there is a very high correlation among the three technical efficiencies and the first principal component which reinforces the quality of the graduation carried out by the first component. Finally, the high Kendall's rank-correlation coefficient (0.8265) between the two rankings, Ist Principal Component and DEA, confirms the validity of the method shown.

Table 2. The results of Principal Component Analysis on the Technical Efficiencies

	Eigenvectors		
Variable	1st PC	2nd PC	3rd PC
TE (y_1)	0.6498	0.1374	0.7476
TE (y_2)	0.4785	-0.8381	-0.2619
TE (y_3)	0.5906	0.5279	-0.6104
Eigenvalues	1.5053	0.8722	0.6225

Table 3. Scores of Technical Efficiencies on the first Principal Component and rankings by DEA

SCHOOL CODE	TE(y_1)	TE(y_2)	TE(y_3)	I.st PC	Rank by I.st PC	Rank by DEA
S2	1.000	1.000	1.000	1.879973	1	1
S7	1.000	1.000	1.000	1.879973	1	1
S16	1.000	0.682	1.000	1.387817	3	1
S13	0.845	0.909	0.916	1.019996	4	7
S14	1.000	0.368	0.971	0.804637	5	5
S17	0.908	0.462	0.960	0.653533	6	6
S9	0.815	1.000	0.683	0.295076	7	8
S11	0.825	1.000	0.642	0.185863	8	8
S12	0.680	0.669	0.906	0.149252	9	11
S6	0.390	1.000	0.980	0.090945	10	16
S10	0.792	0.086	0.958	-0.262555	11	9
S15	0.673	0.449	0.842	-0.425536	12	12
S3	0.401	1.000	0.762	-0.608788	13	16
S8	0.854	0.504	0.505	-0.959166	14	8
S5	0.670	0.404	0.673	-1.070176	15	16
S4	0.596	0.152	0.841	-1.105909	16	16
S1	0.388	0.282	0.905	-1.277339	17	17
S18	0.232	0.097	0.716	-2.637603	18	18

Table 4. Pearson's correlation coefficients (0.05 significance level with a star)

Variable	TE(y_1)	TE(y_2)	TE(y_3)	Ist Principal Component
TE(y_1)	1.0000			
TE(y_2)	0.2717	1.0000		
TE(y_3)	0.2500	0.0548	1.0000	
Ist Principal Component	0.8045*	0.6315 *	0.5930*	1.0000

5. Discussion and Conclusions

The Data Envelopment Analysis (DEA) is a non-parametric deterministic approach that uses the mathematical programming to identify the efficient frontier, and does not impose functional forms. The main advantage of DEA is that it does not require an a priori hypothesis about the analytical form of the production function. Indeed, DEA determines the production function by applying minimization techniques on the data. Differently from regression analysis, the DEA is based on extreme observations, and this leads to the main disadvantage of DEA, i.e., that the frontier is sensitive to the extreme observations. Furthermore, DEA postulates the absence of random errors and that all deviations from the frontier denote inefficiency of the DMUs.

Vice versa, the SFA, is a parametric approach that hypothesizes a functional form and uses the data to econometrically estimate the parameters of this function. The SFA requires functional forms on the production frontier, and assumes that units may deviate from the production frontier not only due to technical inefficiency but also to measurement errors, statistical noise or to other non-systematic factors. In addition, the SFA requires strong distribution assumptions of both statistical random errors (i.e., normal distribution) and the non-negative technical inefficiency random variables (i.e., half-normal or truncated normal distribution) (Coelli et. al., 2005).

With SFA the determinants of efficiency are directly obtained by estimating the production function. With SFA you can use various models changing the response variable

every time and can eventually identify the model which has greater relevance in terms of acceptance.

The method described in this work is suitable for the evaluation of efficiency. Moreover, even our partial data, the method and the results achieved already provide a useful interpretation of the efficiency frontier for the evaluation of schools. Indeed, the efficiency estimates obtained have been utilized to rank the schools according to the common efficiency index.

The comparison with the results obtained through the Stochastic Frontier Analysis with same input and different outputs correlates very well (0.8265) with the results of the DEA. This result confirms the quality of the alternative method proposed in this paper. The rankings obtained by the Stochastic Frontier, however, are more robust than those of the DEA for the very closely tested hypothesis.

Acknowledgements

The present work has been partially funded under an agreement with Nidge University (Resp. Prof. Luigi D'Ambra). We thank the School Superintendent of Campania, and in particular Angela Orabona of the "Polo Qualità di Napoli" for providing us with the data for this study.

Bibliography

1. Abdi, H. and Lynne, J. W. **Principal component Analysis, vol. 2,** John Wiley & Sons, 2010
2. Alden, H. W. **Genetic Algorithms for Real Parameter Optimization,** Foundations of Genetic Algorithms, Edited by Gregory J. E. Rawlins, Morgan Kaufman, 1991
3. Akaike, H. **Factor analysis and AIC,** Psychometrika, Vol. 52,no.3, 1987, pp. 317-332
4. Banker, R. D., Charnes, A. and Cooper, W. W. **Some models for estimating technical and scale inefficiencies in data envelopment analysis,** Management Science, Vol. 30, 1984, pp. 1078–1092
5. Battese G. E. and Corra G. S. **Estimation of a production frontier model: with application to the pastoral zone of eastern Australia,** Australian Journal of Agricultural and Resource Economics, Vol. 21, 1977, pp. 169–179
6. Box G. E. and Cox D. R. **An analysis of transformations,** Journal of the Royal Statistical Society, Series B (Methodological), 21, 1964
7. Charnes, A., Cooper, W. W. and Rhodes, E. **Measuring the efficiency of decision making units,** European Journal of Operational Research, Vol. 2, 1978, pp. 429–444
8. Coelli, T. J., Rao, D. S. P., O'Donnell, C. J. and Battese, G. E **An introduction to efficiency and productivity analysis,** Springer, 2005
9. Cooper, W. W., Seiford, L. M. and Tone, K. **Data envelopment analysis: a comprehensive text with models applications references and DEA-solver software,** Springer, 2006
10. Holland, J. **Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence,** Bradford Books, 1992
11. JGAP, **Java Genetic Algorithms Package,** http://jgap.sourceforge.net/

12. Korhonen , P. and Siljamäki, A. **Ordinal principal component analysis theory and an application,** Computational Statistics & Data Analysis, Vol. 26, No. 4, 1998, pp. 411-424
13. Kumbhakar, S.C. and Lovell, C.K. **Stochastic frontier analysis,** Cambridge University Press, 2003
14. Mizala, A., Romaguera, P. and Farren, D. **The technical efficiency of schools in chile,** Applied Economics, vol. 34, 2002, pp. 1533–1552
15. Rao, D. P., O'Donnell, C. J., Battese, G. E. and Coelli, T. J. **An introduction to efficiency and productivity analysis,** Springer, 2005
16. Ray, S. C. **Data envelopment analysis: theory and techniques for economics and operations research,** Cambridge University Press, 2004