UNIVERSIDADE FEDERAL DO CEARÁ

13 | CAEN PÓS-GRADUAÇÃO EM ECONOMIA

SÉRIE
ESTUDOS ECONÔMICOS CAEN

The use of Yanai's Generalized Coefficient of Determination to reduce the number of variables in DEA models

Maurício Benegas

FORTALEZA · AGOSTO · 2016

UNIVERSIDADE FEDERAL DO CEARÁ
CURSO DE PÓS-GRADUAÇÃO EM ECONOMIA - CAEN

SÉRIE ESTUDOS ECONÔMICOS – CAEN
Nº 13

The use of Yanai's Generalized Coefficient of Determination to reduce the number of variables in DEA models

FORTALEZA – CE
AGOSTO – 2016

THE USE OF YANAI'S GENERALIZED COEFFICIENT OF DETERMINATION TO REDUCE THE NUMBER OF VARIABLES IN DEA MODELS

Maurício Benegas
Graduate Program in Economics - CAEN/UFC
mauricio_benegas@caen.ufc.br

ABSTRACT

This paper proposes a novel method of reducing the number of inputs and outputs in DEA models. The method is based on Yanai's Generalized Coefficient of Determination and on the concept of pseudo-rank of a matrix. It is also proposed a rule to determine the cardinality of the subset of selected variables so as to optimize discretionary power without significant loss of information.

Keywords: DEA; Yanai's Generalized Coefficient of Determination; pseudo-rank.

JEL Codes: C6; C8

1 INTRODUCTION

DEA (Data Envelopment Analysis) model is a nonparametric method of estimating production frontiers. DEA involves solving a linear programming (LP) problem to determine a production frontier against which the technical efficiency of Decision Making Units (DMU) will be calculated. A basic DEA model was originally proposed by Charnes, Cooper and Rhodes (1978). That initial model is also known as “the CCR model”.

The basic CCR model proposes maximizing the ratio between a weighted sum of outputs and a weighted sum of inputs. The weights of these sums are chosen according to feasibility conditions and assuming a hypothesis of constant returns to scale. Charnes, Cooper and Rhodes have transformed the fractional (primal) CCR model into a linear (dual) model, usually referred to in the literature as the DEA [1] model.

The CCR model and its variants have been increasingly used since the first description by Charnes, Cooper and Rhodes in 1978. Since then, researchers worldwide have used the model as a tool to assess technical efficiency. The DEA model has become the method of choice for this type of study, especially in the absence of an explicit production function to define the relationship between inputs and outputs[2]

One of the most frequent problems associated with the CCR model is the lack of discrimination between DMUs when the number of inputs and outputs is very large in relation to the number of DMUs. A large number of variables relative to the number of observations may entail a large number of efficient DMUs in the sample – thus reducing the model’s ranking capability. This is a characteristic of the CCR model: the lower the number of DMUs, the less active the restrictions imposed on maximum efficiency multipliers.

Several alternatives have been proposed to increase the ranking capacity of the CCR model[3], including the super-efficiency (Andersen and Petersen, 1993) and cross-efficiency evaluation methods (Sexton *et al*., 1986; Doyle and Green, 1994; and Green *et al*., 1996). Other methods that use additional information, usually characterized by adding restrictions, include the cone-ratio and assurance-region approaches.

In the present work we propose a simple and objective method to address a situation in which the original inputs and outputs have been correctly selected, but the low number of observations has translated into low discrimination power. This approach proposes to reduce the dimensions of inputs and outputs and, therefore, does not require additional information. In addition, it does not require post-estimation procedures (as in the case of the super-efficiency and cross-efficiency approaches). The approach we propose relies on

[1] For details on this procedure it is recommended to consult the work of Cooper, Seiford and Tone (2006).
[2] If there is evidence that inputs and outputs can be linked through a function, then the stochastic frontier model of production can be used as an alternative to DEA. Please refer to the work of Kumbhakar and Lovell on models of stochastic frontier production models (2000).
[3] See Adler et al. (2002), Angulo-Meza and Lins (2002), Podinovski and Thanassoulis (2007) and Senra et al. (2007).

multivariate statistical techniques, notably Principle Components Analysis, as on the use of a correlation matrix. Again it should be underscored that the approach is not intended for selection of inputs or outputs. Rather, it is a support tool to increase the discriminatory power without significant loss of information – that is, discriminatory power is increased regardless of the knowledge concerning which variables are essential for the model.

Following this Introduction, the work is divided as follows: Section 2 briefly reviews multivariate variable selection in DEA; Section 3 introduces the proposed methodology; section 4 discusses applications of the proposed method to the CCR model; and finally, Section 5 presents conclusions and suggestions for future research.

2 VARIABLE SELECTION/REDUCTION IN DEA

One of the issues relating to the CCR model is that of the dimensions of inputs and outputs. A consequence of using a large number of variables relative to the number of observations is the loss of discrimination power due to the generation of a large number of efficient DMU´s. There is no consensus on the optimal number of inputs and outputs to be used. However, Cooper, Seiford, and Tone (2006) suggest the following rule: J:J>max{M x N, 3(M + N)}, where M and N correspond to the number of outputs and inputs respectively.

One of the most common ways of selecting variables in DEA is the use of correlation matrices for inputs and outputs. When two variables (inputs or outputs) are highly correlated, one is discarded, usually on the basis of *ad hoc* criteria. However, eliminating one or the other variable could have a dramatic impact on estimated efficiency (Dyson *et al*., 2001).

In recent years, the application of multivariate statistical methods, especially Principal Components Analysis (PCA), has appeared as a satisfactory alternative for variable reduction. The formulation of a DEA model in which inputs and outputs are summarized as principal components is the focus of the work of Ueda and Hoshiai (1997) and Adler and Golany (2002). One limitation of using PCs instead of inputs and outputs is that these "new" variables may take a negative value, and therefore must be transformed for PCA. That transformation, however, may impact results. One way of overcoming the problem is to use the additive model of DEA (Ali and Seiford, 1990) that is invariant to translation of inputs and outputs. Pastor (1996) has also shown that an input-oriented BCC model (after Banker, Charnes and Cooper, 1984) is invariant to output translation.

There is also a practical difficulty. Even if the problem of negative PCs is resolved, the question remains of how to interpret the results in terms of the projection of inputs and outputs (that is, predicting quantities). The fact is that the only satisfactory way of accomplishing this is back transformation to original variables - which might require a considerable computational effort. Some authors select variables based upon their contribution to PCs.

Specifically, the variables with the largest absolute linear combination coefficient are selected. Because it is common for the first few components to explain most data variance, the result is a considerably reduced subset of variables.

Jenkins and Anderson (2003) employ the method of partial covariance analysis to identify the correlation between variables as well as the contribution of each variable to these correlations. With this technique, the authors demonstrate that the removal of variables with little contribution to the correlations does not significantly change the results.

The method we propose is based on the work of Cadima (2001) and Cadima and Jollife (2001) and combines PCA with elements of the Jenkins and Anderson (2003) method. It consists of generating a correlation matrix between two data sets – the orthogonal projection of data onto the subspace generated by a subset of PCs and the orthogonal projection of data onto the subspace generated by a subset of original variables. This matrix measure of the closeness of the two subspaces is known as Yanai's Generalized Coefficient of Determination (GCD) (Yanai, 1974). The resulting subset of PCs (which usually includes PCs that explain eighty to ninety percent of data variance) corresponds to the original variables that maximize the GCD.

The number of PCs is determined prior to the generation of the correlation matrix, estimating the pseudo-rank for the matrix, as proposed by Cadima (2001) based on the specific geometric structure of the cone of positive semidefinite matrices.

In the following section a brief discussion of the geometrical structure of the cone of positive semidefinite matrices and of PC analysis is presented. GCD is then defined.

3 THE PSEUDO-RANK OF A MATRIX AND YANAI'S GENERALIZED COEFFICIENT OF DETERMINATION

This section briefly describes basic concepts described in detail in the work of Jolliffe (1986), Cadima (2000), and Cadima and Jolliffe (2001).

Assuming C_p to be the cone of positive semidefinite matrices with dimension $p \times p$ provided with the Frobenius inner product $\langle \cdot, \cdot \rangle_F : C_p \times C_p \to \Re$ such that

$$\langle A, B \rangle_F = tr(A'B) \text{ for any } A, B \in C_p \quad (1)$$

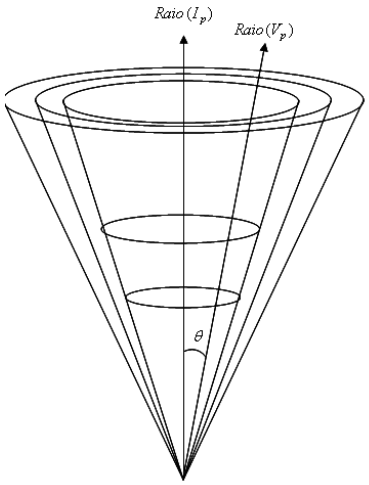
The norm induced by (1) will be denoted by $\| \ \|_F$. For any matrix $V \in C_p$, the set $Ray(V) = \{A \in C_p ; A = \lambda V, \lambda \geq 0\}$ is called the ray associated with V. The ray associated with the identity matrix of dimension p is called the central ray of C_p

With the definitions above, it is possible to find the angle between the rays associated with any two matrices A and B. This angle is given by the arc whose cosine is

$$\cos(A,B) = \frac{\langle A,B \rangle_F}{\|A\|_F \|B\|_F} \tag{2}$$

In Cadima (2001), the author demonstrates that C_p has a layered structure, with containing several cones fitted inside the other:

Figure 1. - Cone of positive semidefinite matrices



Based on this observation, the author argues that the region close to the central ray contains only matrices of full rank, or at least with rank $p-1$ (which can occur at the boundary of such regions). The farther away from the central ray the lower the rank of matrices.

However, matrices of full rank are also found outside of the core. Because may have eigenvalues close to zero, they behave as low-rank matrices. The question then is how far (into the core of the cone) does one have to move to avoid such matrices? The answer to this question lies in the concept of the pseudo-rank of a matrix. According to Cadima (2001), the pseudo-rank of a $V \in C_p$ matrix is the smallest integer $k^* \leq p$ such that

$$\cos(V, I_p) \leq \sqrt{\frac{k^*}{p}} \tag{3}$$

The author shows that this value is given by

$$k^* = \left\lceil \frac{tr(V)^2}{tr(V^2)} \right\rceil \quad (4)$$

where [z] is the nearest to z greater integer. Letting V the covariance/correlation matrix of a data matrix A, with p variables and n observations, or $V = n^{-1}A'A$. In this case the pseudo-rank of V corresponds to the number of components to be used as representative of all accumulated variance associated with A.

3.1. Yanai's Generalized Coefficient of Determination

Let A represent a data matrix with dimension $(n \times p)$, where p indicates the number of variables and n the number of observations for each variable. In this context, A may refer to a matrix of discretionary/nondiscretionary outputs/inputs. It is important to keep in mind that in DEA models, the number of observations indicates the number of columns in the pertinent matrices. Thus, for the implementation of the proposed method, matrix A should be considered the transposed output/input matrix.

Given the covariance matrix/correlation of data $S = n^{-1}A'A$, let Λ and P be the diagonal matrix of eigenvalues (arranged in decreasing order) and the matrix of normalized eigenvectors of S respectively. The PCs are the columns of the matrix $(n \times p)$ given by $C = AP$ Using the spectral decomposition of S, it is easy to show that the covariance matrix of C is exactly Λ, so that the variables in C are uncorrelated. For this reason the AP transformation is sometimes called data "decorrelation."

Consider K to be a subset of indices associated with $k \leq p$ PCs arranged in decreasing order of eigenvalues (in general the first k's). Similarly, let Q be defined as a subset of indexes associated with the $q \leq p$ original variables. The sets $\mathbf{K}$ and $\mathbf{Q}$ are the subspaces generated by vectors with indices K and Q respectively. The following matrices are then defined:

- A_K is the submatrix of A in which columns with indexes in K are maintained;
- $S_K = n^{-1}A_K^{'}A_K$ is the covariance matrix associated with A_K;
- Λ_K is the matrix of the eigenvalues associated with S_K;
- P_K is the matrix of eigenvectors associated with eigenvalues in Λ_K;

Let us assume $P_{\mathbf{K}}$ to be the matrix of orthogonal projection on the subspace $\mathbf{K}$ such that

$$P_{\mathbf{K}} = n^{-1} A S_K^{-1} A'$$

where S_K^{-1} is the Moore-Penrose generalized inverse of $S_{\mathbf{K}}$. Similarly, $P_{\mathbf{Q}}$ is the matrix of orthogonal projection on the subspace $\mathbf{Q}$ defined as:

$$P_{\mathbf{Q}} = n^{-1} A I_Q S_Q^{-1} I_Q A'$$

where I_Q is the identity matrix of the submatrix obtained by selecting the q columns with indices in Q and $S_Q = n^{-1} I_Q A' A I_Q$.

Given the definitions above, Yanai's GCD between subspaces $\mathbf{Q}$ and $\mathbf{K}$ is defined as:

$$GCD(\mathbf{Q},\mathbf{K}) = \frac{\left\langle P_{\mathbf{Q}}, P_{\mathbf{K}} \right\rangle_F}{\left\| P_{\mathbf{Q}} \right\|_F \left\| P_{\mathbf{K}} \right\|_F} \tag{5}$$

Supposing that $K = \{1,2,...,k^*\}$ remains fixed, where k^* is the pseudo-rank of the covariance matrix, in this case the selection of variables exhibiting the greatest contribution to the principal components selected in K is the set of indices $\tilde{Q}$ such that

$$\tilde{Q} = \arg\max_{Q} GCD(\mathbf{Q},\mathbf{K})$$

4 REDUCED CCR MODEL[4]

A practical example of the proposed method is provided in this section. For that, the CCR model will be presented with the original variables (hereafter called the "general model" and denoted by CCR_g), and with the subset of selected variables ("reduced model," denoted by CCR_r). For the sake of simplicity, only product-oriented models will be discussed below.

Let us suppose that there are J DMUs under study, each using a vector $\mathbf{x} \in \Re_+^N$ of inputs to produce a vector $\mathbf{y} \in \Re_+^M$ of outputs with a technology defined by

[4] The use of the CCR model is an example. The method can actually be applied to any DEA model.

$$T_{CCR_g} = \{(\mathbf{x}, \mathbf{y}); \mathbf{x} \geq \mathbf{X}\lambda, \mathbf{y} \leq \mathbf{Y}\lambda, \lambda \geq 0\}$$

where $\mathbf{X}_{(N\times J)}$, $\mathbf{Y}_{(M\times J)}$ and $\lambda_{(J\times 1)}$ are input and output matrices and the vector of intensities respectively.

Given a DMU *j*, its technical efficiency in the model CCR_g, denoted by ET_j^g , is estimated by solving the following linear programming problem (the minimization of slacks is omitted for simplicity)

$$ET_j^g = \begin{cases} \max\limits_{\theta,\lambda} \theta \\ subject\ to \\ (\mathbf{x}_j, \theta\mathbf{y}_j) \in T_{CCR_g} \end{cases} \quad (6)$$

Let us now suppose that a set of variables was selected following the procedure presented in Section 3. For simplicity's sake, let us assume that only outputs were selected. Let us denote by $\mathbf{y}_j^q$ the vector product of a DMU *j = 1,2,…J* , with selection of $q < N$ outputs, and let us denote by $\mathbf{Y}^q$ the respective reduced output matrix. The technology in the CCR_r model is defined as

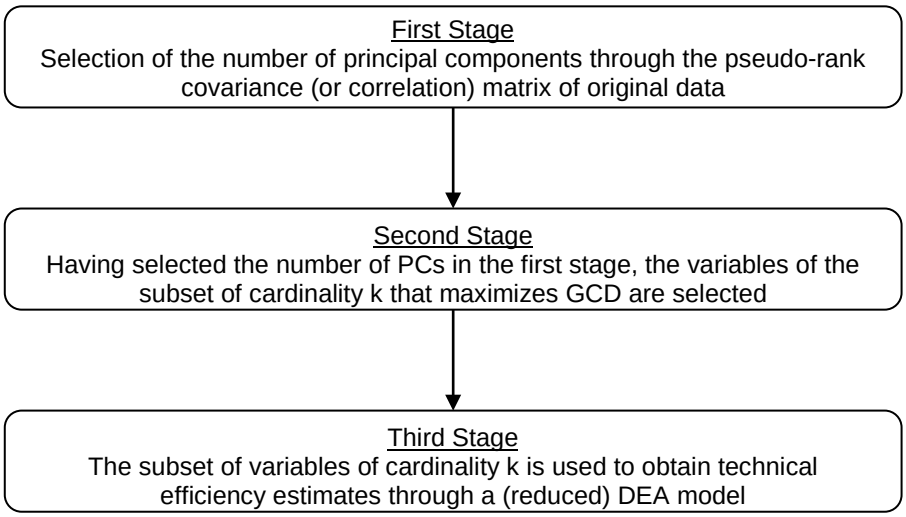
$$T_{CCR_g} = \{(\mathbf{x}, \mathbf{y}^q); \mathbf{x} \geq \mathbf{X}\lambda, \mathbf{y}^q \leq \mathbf{Y}^q\lambda, \lambda \geq 0\}$$

Given a DMU *j*, its technical efficiency in the model CCR_r, denoted by ET_j^r , is estimated by solving the following linear programming problem

$$ET_j^r = \begin{cases} \max\limits_{\theta,\lambda} \theta \\ subject\ to \\ (\mathbf{x}_j, \theta\mathbf{y}_j^q) \in T_{CCR_r} \end{cases} \quad (7)$$

The reduced model is obtained in three stages, as summarized below:

Figure 2 - Steps for Obtaining the Reduced Model



One issue not covered by the procedure presented above is that of defining the cardinality of the subset of selected variables. One suggestion would be to combine the gain in discrimination with some measure that would reveal loss of information. The gain in discrimination would be obtained by the difference between the percentage of efficient DMUs in the general model and in the reduced model. The complexity of this issue relates to what measure of informational loss should be used. One possibility is to use the Kolmogorov-Smirnov statistic, which quantifies the difference between distributions. In this context, this statistic is given by

$$KS(q) = \sup_x \left| F(x) - F_q(x) \right|$$

where F and F_q are the empirical cumulative distribution functions of technical efficiency estimated by the general and reduced models (the latter with a subset q of selected variables) respectively.

Let K^* and K_q^* be the number of efficient DMUs in the general and reduced models respectively; let $\delta_q = (K^* - K_q^*)/K$ such that δ_q represents, in proportional terms, the gain in discrimination power of the reduced model in relation to the general model. Then the optimal cardinality would be given by q^* such that

$$q^* \in \arg\max_q \left\{ \frac{\delta_q}{KS(q)}; KS(q) \le 1.36\sqrt{\frac{2}{K}} \right\} \quad (8)$$

In which $KS(q) \le 1.36\sqrt{2/K}$ is included so that optimal cardinality will depend on acceptance of the null hypothesis of the Kolmogorov-Smirnov test. The amount $1.36\sqrt{2/K}$ represents the

nullity condition of the Kolmogorov-Smirnov test with a significance level of 0.05, such that if $KS(q) > 1.36\sqrt{2/K}$ the null hypothesis of equality between the distributions is rejected. It should be noted that $1.36\sqrt{2/K}$ is generally valid for $K \geq 8$, otherwise it is necessary to consult tabulated values. [5]

If there are multiple solutions to (8) the lowest maximize q^* is selected, so that

$$q^* = \min_{\hat{q}} \left\{ \hat{q} \in \arg\max_{q} \left\{ \frac{\delta_q}{KS(q)}; KS(q) \leq 1.36\sqrt{\frac{2}{K}} \right\} \right\} \quad (9)$$

4.1. Example

This example employs real-world data previously described by Benegas and Silva (2010), who examined the technical efficiency of the public health care system in Brazil.

The database uses one input (x), represented by the annual *per capita* expenditure on health of the three levels of government, and 12 outputs (y_i, $i = 1,...12$), , representing health indicators available in the Ministry of Health's Information System DATASUS. [6]

For the present example, to reduce the power of discrimination, only 12 of the 27 states studied by Benegas and Silva (2010) were selected. The data and some of the descriptive statistics are shown in Table 1 (Appendix Table A1 provides a description of the variables used in the example).

Table 1 - Data of example

State	x	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
RO	387.5	68.2	73.8	70.9	80	0.8	1.7	111.2	106.9	107.5	108.3	47.6	70
AC	556.6	68.6	73.8	71.1	71	0.8	2	86.5	83.64	110.5	90.14	40.3	67.4
AM	513.4	68.4	74.4	71.3	78	0.9	1.5	106.3	93.88	131.1	92.36	57.1	72.8
RR	674.5	67.2	72.1	69.6	83	1.1	1.6	93.13	88.62	114.1	89.85	71.9	79.3
PA	263.8	68.8	74.7	71.7	76	0.8	1.6	116.6	112.3	144.5	119.8	55	76.9
AP	512.5	66.2	74.1	70.1	79	0.8	1.5	94.12	96.3	126.7	98.17	28.2	90.5
TO	503.8	68.8	73.3	71	78	1.1	1.8	99.57	107.2	110.1	107.4	21	69.9
MA	285.3	63.4	71.3	67.2	69	0.6	2.4	108.2	106.6	135.6	111.9	50.1	59
PI	315.8	65.6	71.7	68.6	73	0.8	2.5	101.9	102.1	106	104.6	61.7	49.6
CE	291.4	65.7	74.4	69.9	74	0.9	1.9	108.3	108.3	113.3	112.3	40.8	71.3
RN	405	66.3	74.1	70.1	69	1.2	2.3	98.87	95.45	108.2	94.61	45.1	82.3
PB	335.1	65.2	72.2	68.6	68	1.1	2.7	105.7	103.6	117.8	104	48.5	74.7
Max	674.5	68.8	74.7	71.7	83	1.2	2.7	116.6	112.3	144.5	119.8	71.9	90.5
Min	263.8	63.4	71.3	67.2	68	0.6	1.5	86.5	83.64	106	89.85	21	49.6
Mean	420.4	66.9	73.3	70	75	0.9	2	102.5	100.4	118.8	102.8	47.3	72
SE	130.1	1.73	1.18	1.33	4.8	0.2	0.4	8.529	8.772	12.64	9.732	13.9	10.6

Source: Benegas and Silva (2010)

[5] See Gibbons and Chakraborti (2003) for details.
[6] See site www2.datasus.gov.br.

In this example, the Kolmogorov-Smirnov null hypothesis is accepted with a significance level of 0.05 for a value up to 0.5552, such that the condition to select the cardinality of the subset of selected outputs becomes

$$q^* = \min_{\hat{q}} \left\{ \hat{q} \in \arg\max_{q} \left\{ \frac{\delta_q}{KS(q)}; KS(q) \leq 0.5552 \right\} \right\}$$

Table 2 shows the results obtained by applying the proposed variable selection method to the output matrix. Variables were selected into subsets with cardinality from 1 to 10. Using the pseudo-rank of the output covariance matrix (eq. (4)), the first four PCs were used. These for PCs accounted for approximately 84% of the total variability of the sample.

Table 2 - Results of general and reduced models *

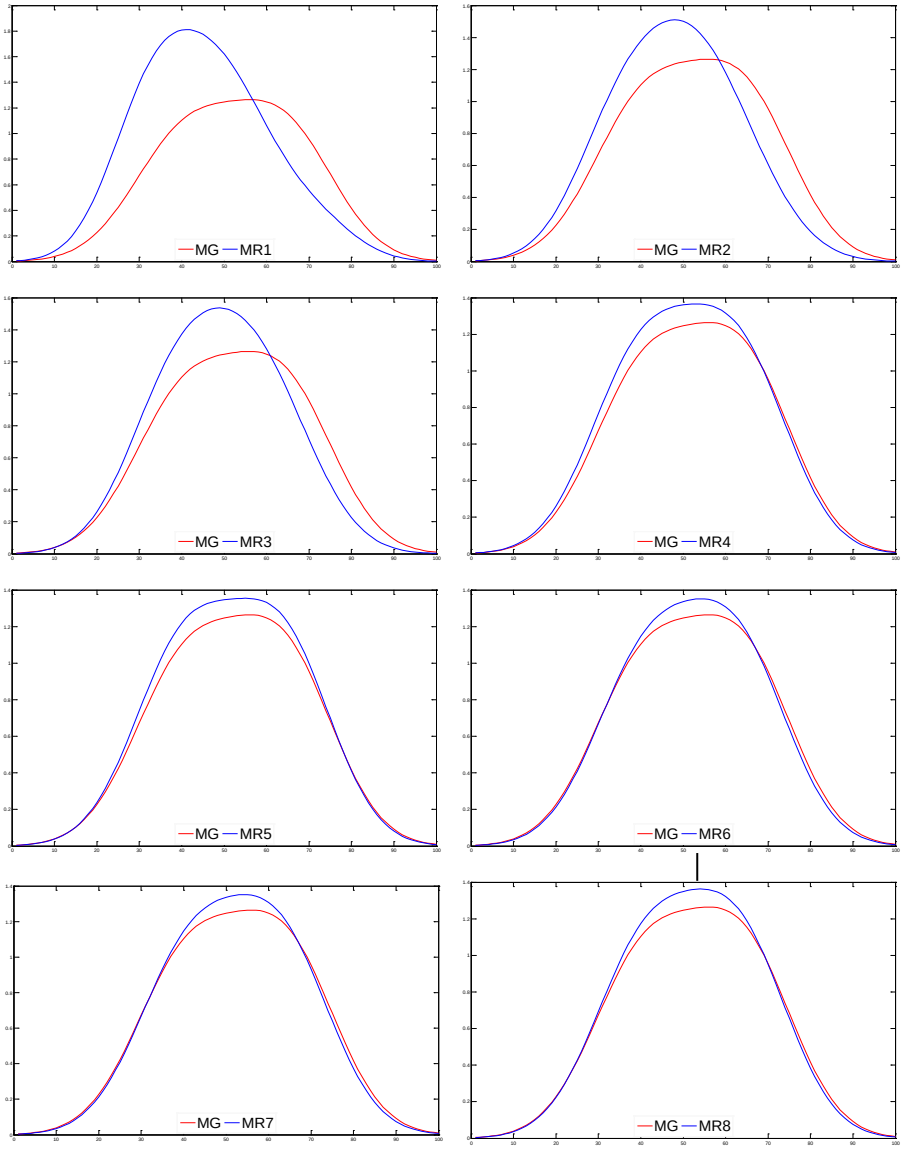
State	GM	RM1	RM2	RM3	RM4	RM5	RM6	RM7	RM8	RM9	RM10
RO	0.71	0.62	0.65	0.67	0.69	0.69	0.71	0.71	0.71	0.71	0.71
AC	0.51	0.42	0.42	0.47	0.49	0.49	0.49	0.49	0.49	0.49	0.49
AM	0.59	0.49	0.49	0.51	0.57	0.59	0.59	0.59	0.59	0.59	0.59
RR	0.53	0.40	0.40	0.40	0.48	0.53	0.53	0.53	0.53	0.53	0.53
PA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AP	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
TO	0.64	0.48	0.48	0.52	0.63	0.63	0.64	0.64	0.64	0.64	0.64
MA	1.00	0.71	0.86	0.87	0.87	0.87	0.88	0.88	0.87	0.87	0.87
PI	1.00	0.54	0.73	0.80	0.85	0.94	0.94	0.94	0.94	0.94	0.94
CE	1.00	0.84	0.84	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RN	0.88	0.70	0.70	0.70	0.88	0.88	0.88	0.88	0.88	0.88	0.88
PB	1.00	0.76	0.76	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
δ_q	-	0.33	0.33	0.33	0.17	0.17	0.17	0.17	0.17	0.17	0.17
KS(q)	-	0.42	0.42	0.42	0.17	0.17	0.17	0.17	0.17	0.17	0.17
δ_q / *KS(q)*	-	0.80	0.80	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Source: Author estimates
*The initials GM and RMq refer to general and reduced models with cardinality q

Note that in this example, $\arg\max_{q} \{\delta_q / KS(q); KS(q) \leq 0.5552\} = \{4,5,6,7,8,9,10\}$ and therefore $q^* = 4$. The last part of the example appears in Figure 3, which shows the estimated densities for the general and reduced models with cardinality from 1 to 8. The procedure to estimate the densities uses the Gaussian kernel. The bandwidth was selected minimizing the mean integrated square error.[7]

[7] See Silverman, B. W. (1986) for details.

Figure 3 - Estimated densities for TE in general and reduced models



The results shown in Figure 3 suggest that in the presence of four selected variables, the inclusion of an additional variable does not have a significant impact on the comparison between the general model and the reduced model. This observation confirms the conclusions obtained by applying the method of subset cardinality selection suggested by equation (9).

5 CONCLUSIONS AND FURTHER RESEARCH

This paper proposes a method for reducing the dimension of input/output matrices used to estimate production frontiers through the CCR model (or its variants). The method is based on Yanai's Generalized Coefficient of Determination (GCD) and on the concept of pseudo-rank of a matrix. Additionally, a rule is suggested to support the choice of cardinality of the subset of

selected variables. This rule seeks to combine maximum gain in discriminatory power with minimal loss of information.

Through an example that employs real-world data, it was found that the pseudo-rank of the output correlation matrix indicates that the first four PCs should be maintained. The GCD for output subsets with cardinality from 1 to 10 was then calculated. The cardinality rule indicated the subset with four of the 12 original outputs. Finally, the estimation of densities of the general and reduced models suggested that the cardinality decision rule can support decisions concerning the number of variables required in the model to obtain maximum discrimination with minimal loss of information.

Further research is suggested on the cardinality decision rule taking into consideration various measures of loss of information. Also warranted are studies comparing the proposed method with other methods of summarization and selection.

References

ADLER, N. & GOLANY, B. (2001). Evaluation of deregulated airline networks using Data Envelopment Analysis combined with principal component Analysis with an Application to Western Europe. **European Journal of Operational Research**, 132, p. 260-273.

ADLER, N & GOLANY, B. (2002). Including principal component weights to improve discrimination in Data Envelopment Analysis. **Journal of the Operational Research Society**, 53, p. 985-991

ADLER, N & YAZHEMSKY, E. (2010). Improving discrimination in Data Envelopment Analysis: PCA–DEA or variable reduction. **European Journal of Operational Research**, 202(1), p. 273-84.

ALI, A.I. & SEIFORD, L.M. (1990). Translation invariance in Data Envelopment Analysis. **Operations Research Letters**, 9, p. 403-405.

ÂNGULO-MEZA, L. & LINS, M.P.E. (2002). Review of methods for increasing discrimination in Data Envelopment Analysis. **Annals of Operations Research**, 116, p. 225-42.

BENEGAS, M. & SILVA, F.G. (2010). Estimação da eficiência técnica do SUS nos Estados Brasileiros na presença de variáveis contextuais. **Texto para Discussão, CAEN-UFC**.

CADIMA, J.F.L. & JOLLIFE, I.T. (2001). Variable selection and the interpretation of principal subspaces. **Journal of Agricultural, Biological and Environmental Statistics**, 6(1), p. 62-79.

CADIMA, J.F.L. (2001). Redução de dimensionalidade através duma análise em componentes principais: um critério para o número de componentes principais a reter. **Revista de Estatística (INE)**, 1, p. 37-49.

CHARNES, A.; COOPER, W.W. & RHODES, E. (1978). Measuring the efficiency of decision making units. **European Journal of Operational Research**, 2(6) November, p. 429-444.

DYSON, R.G.; ALLEN, R.; CAMANHO, A.S.; PODINOVSKI, V.V.; SARRICO, C.S. & SHALE, E.A. (2001). Pitfalls and protocols in DEA. **European Journal of Operational Research**, 132, p. 245-259.

GIBBONS, J.D. & CHAKRABORTI, S. (2003). **Nonparametric Statistical Inference**, 4th Ed., CRC Press, London.

JENKINS, L. & ANDERSON, M. (2003). A multivariate statistical approach to reducing the number of variables in Data Envelopment Analysis. **European Journal of Operational Research**, 147, p. 51-61

JOLLIFE, I.T. (1986). **Principal Component Analysis**, Springer-Verlag, New York, USA.

KUMBHAKAR, S.C. & LOVELL, C.A.K. (2000). **Stochastic Frontier Analysis**. Cambridge University Press, Cambridge.

PASTOR, J. (1996). Translation invariance in Data Envelopment Analysis: a generalization. **Annals of Operations Research**, 66, p. 93-102.

SENRA, L.F.A.C.; NANCI, L.C.; SOARES DE MELO, J.C.B. & ANGULO-MEZA, L. (2007). Estudo sobre métodos de seleção de variáveis em DEA. **Pesquisa Operacional**, 27(2), p. 191-207

SILVERMAN, B.W. (1986). **Density Estimation for Statistics and Data Analysis**, Chapman and Hall, London.

UEDA, T. & HOSHIAI, Y. (1997). Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. **Journal of Operational Research Society of Japan**, 40, p. 466-478.

YANAI, H. (1974). Unification of various techniques of multivariate analysis by means of Generalized Coefficient of Determination (GCD), **Journal of Behaviormetrics**, 1, p. 45-54.

Appendix

Table A1: Input and outputs used in example

Product	Description
x	Public spending per capita on health
y1	Male life expectancy
y2	Female life expectancy
y3	Combined life expectancy
y4	Survival rate (%)
y5	Physicians per 1000 inhabitants
y6	Hospital beds per 1000 inhabitants
y7	Coverage MMR (%)
y8	Tetravalent vaccine coverage (%)
y9	BCG vaccination coverage (%)
y10	Polio vaccine coverage (%)
y11	Sanitation coverage (%)
y12	Garbage collection coverage (%)

Source: Benegas and Silva (2010)

This paper proposes a novel method of reducing the number of inputs and outputs in DEA models. The method is based on Yanai's Generalized Coefficient of Determination and on the concept of pseudo-rank of a matrix. It is also proposed a rule to determine the cardinality of the subset of selected variables so as to optimize discretionary power without significative loss of information.

Av. da Universidade, 2700 • Benfica • Fone/Fax:(085) 3366.7751
Cep: 60.020-181 • Fortaleza - CE - Brasil
http://www.caen.ufc.br • e-mail: caen@caen.ufc.br