HUB RESEARCH PAPER
Economics & Management

Negative Data in DEA: A Simple Proportional Distance Function Approach

Kristiaan Kerstens and Ignace Van de Woestyne

HUB RESEARCH PAPER 2009/4
APRIL 2009

partners in
Hogeschool-Universiteit Brussel

Negative Data in DEA: A Simple Proportional Distance Function Approach

Kristiaan Kerstens[*] Ignace Van de Woestyne[†]

April 2009

Abstract

The need to adapt Data Envelopment Analysis (DEA) and other frontier models in the context of negative data has been a rather neglected issue in the literature. Silva Portela, Thanassoulis, and Simpson (2004) proposed a variation on the directional distance function, a very general distance function that is dual to the profit function, to accommodate eventual negative data. In this contribution, we suggest a simple variation on the proportional distance function that can do the same job.

Keywords: Data Envelopment Analysis, Linear Programming.

1 Introduction

The seminal article of Farrell (1957) and the revised interest of Charnes, Cooper, and Rhodes (1978) have led to the development of the Data Envelopment Analysis (DEA) literature that has developed at the interface of operational research and economics.[1] This DEA literature has meanwhile become one of the success stories of the operational research area (see, e.g., Emrouznejad, Parker, and Tavares (2008)). The estimation of frontier or best practice models to determine the relative efficiency of organizations has found its way to a large variety of domains of application. In terms of empirical surveys of certain well-analyzed sectors, one could, for instance, point to banking (e.g., Harker and Zenios (2001)), education (Worthington (2001)), health care (e.g., Ozcan (2008)), insurance (Cummins and Weiss (2000)), public transit (e..g., De Borger, Kerstens, and Costa (2002)), and real estate (Anderson, Lewis, and Springer (2000)). In addition to this surge of empirical applications, there have been a vast series of methodological developments in this literature (see, e.g., the surveys in Färe, Grosskopf, and Lovell (1994) or Thanassoulis, Silva Portela, and Despić (2008)).

[*]CNRS-LEM (UMR 8179), IESEG School of Management, 3 rue de la Digue, F-59000 Lille, France. k.kerstens@ieseg.fr Corresponding author.

[†]Hogeschool Universiteit Brussel, Brussels, Belgium

[1]On the history of DEA, see for instance Førsund and Sarafoglou (2005).

While in a traditional production context inputs and outputs are assumed to be non-negative, frontier applications have also moved into areas where negative data may occur.[2] Examples include, among others, the analysis of financial statements (e.g., Smith (1990) or Feroz, Kim, and Raab (2003)) or the rating of mutual funds (see the seminal article by Murthi, Choi, and Desai (1997)). Obviously, growth rates or returns can be both negative and positive. The issue of handling negative data has attracted some research attention. For instance, proposals have been made to translate the data (e.g., by adding a number making all data positive), though in many models this may have implications on the efficiency measures, among others (see, e.g., Ali and Seiford (1990)). In fact, very few DEA models turn out to yield solutions that are invariant to such data transformations (i.e., are translation invariant). This small literature has been competently summarized in Pastor and Ruiz (2007) or Thanassoulis, Silva Portela, and Despić (2008).

The rather recently introduced directional distance function generalizes existing distance functions by accounting for both input contractions and output improvements and it is dual to the profit function (see Chambers, Chung, and Färe (1998)).[3] Furthermore, the directional distance function is flexible due to the variety of direction vectors it allows for. In the more pragmatic, managerially oriented benchmarking models allowing for negative data, Silva Portela, Thanassoulis, and Simpson (2004) suggest working with some variations of this directional distance function. In this contribution, we argue that a very simple modification of the traditionally defined proportional distance function can equally well be used to accommodate for negative data.

2 Technology and Directional Distance Function

Production technology traditionally transforms inputs $x = (x_1, \ldots, x_p) \in \mathbb{R}^p_+$ into outputs $y = (y_1, \ldots, y_q) \in \mathbb{R}^q_+$. The production possibility set or technology T summarizes the set of all feasible input and output vectors and can be defined as follows:

$$T = \{(x, y) \in \mathbb{R}^{p+q}_+;\ x \text{ can produce } y\}.$$

Throughout this contribution, technology satisfies the following standard assumptions: (T.1) no free lunch; (T.2) boundedness; (T.3) closedness; (T.4) strong disposal of inputs and outputs; and (T.5) convexity (see Färe, Grosskopf, and Lovell (1994) for details).

Technology can be characterized by the use of distance functions. To simplify notation, denote the netput vector $z = (x, y) \in T$ and the direction vector $g = (h, k) \in (-\mathbb{R}^p_+) \times \mathbb{R}^q_+$, that is partitioned in an input and an output direction vector $-h$ and k respectively. The directional distance function is seeking a simultaneous improvement in both the input and output dimensions in the direction of the vector g and is formally defined as:

[2] In a traditional production context, see, e.g., Färe, Grosskopf, and Lovell (1994) for conditions on the input and output data matrices.

[3] Luenberger (1992) introduced the benefit function as a directional representation of preferences generalizing the input distance function defined in terms of the utility function. Luenberger (1995) transposed this benefit function in a production context under the name of the shortage function. Chambers, Chung, and Färe (1998) relabel this same function as a directional distance function and this name has become its most common denomination.

Definition 2.1. For a given technology T, the directional distance function D_T is the function $D_T : T \times \left((-\mathbb{R}^p_+) \times \mathbb{R}^q_+\right) \to \mathbb{R} \cup \{+\infty\}$ with

$$D_T(z; g) = \sup_{\delta} \{\delta \in \mathbb{R} : z + \delta g \in T\}.$$

The vector $g \in (-\mathbb{R}^p_+) \times \mathbb{R}^q_+$ is called a direction vector.

Remark first that, by extending the target set $\mathbb{R}$ with $+\infty$, the directional distance function is well-defined for all possible choices of the direction vector. Indeed if $g = 0$, then clearly $D_T(z; 0) = +\infty$. Also notice that $D_T(z; g) \geq 0$ since $\delta = 0$ is always contained in the set $\{\delta \in \mathbb{R} : z + \delta g \in T\}$.[4] Second, this distance function has an interpretation as an efficiency (or better, inefficiency) measure, because it measures deviations from the boundary of technology. An efficient vector $z \in T$ yields a directional distance function value of zero.

The directional distance function has proven to be a useful tool in applied production analysis. For instance, it allows Chavas and Kim (2007) to shed new light on economies of scope from a primal viewpoint. Furthermore, it provides the defining components of the Luenberger productivity indicator (e.g., Chambers (2002)), a generalization of the very popular Malmquist productivity index.

We mention the following proposition that follows immediately from Definition 2.1.

Proposition 2.1. *For a given technology T, $z \in T$, $g \in (-\mathbb{R}^p_+) \times \mathbb{R}^q_+$ and an arbitrary norm function $\| \dots \|$, it follows that $D_T(z; g) = \delta^* = \frac{\|z^* - z\|}{\|g\|}$, with $z^* = z + \delta^* g$.*

Proof: Trivial, and therefore discarded. □

The directional distance function defined in Definition 2.1 uses a general direction vector g. However, sometimes one considers the special case: $h = -x$ and $k = y$ which gives rise to the (Farrell) proportional distance function (Briec (1997)). Axiomatic properties of these functions are studied in Briec (1997) and Chambers, Chung, and Färe (1998). Since this proportional distance function is a special case of the directional distance function, it also measures inefficiency. The proportional interpretation of the Farrell proportional distance function follows immediately from Proposition 2.1 (just take $g = (-x, y)$ with $z = (x, y)$).

Now, consider n decision making units (DMUs) $z_i = (x_i, y_i)$, $(i = 1, \dots, n)$ from which the technology T is derived. Furthermore, $z_0 = (x_0, y_0)$ denotes the DMU under observation and $g = (h, k)$ is the selected direction vector. Then, the directional distance function value $D_T(z_0; g)$ under variable returns to scale (VRS) and strong disposability

[4]Notice that in the more general case where a point may not be part of technology, the definition of the directional distance function must be adapted such that it distinguishes between the standard case where the distance is achieved and cases where there is no way to achieve the distance. This distinction is important since Briec and Kerstens (2009) have recently shown that there are always circumstances under very general production technologies for which this adapted function may not be well-defined.

assumptions is obtained by solving the following linear programming (LP) model:

$$\max\left\{\delta \;:\; \begin{array}{ll} \displaystyle\sum_{i=1}^{n}\lambda_i x_{ir} \le x_{0r}+\delta h_r, & (r=1,\dots,p),\\ \displaystyle\sum_{i=1}^{n}\lambda_i y_{is} \ge y_{0s}+\delta k_s, & (s=1,\dots,q),\\ \displaystyle\sum_{i=1}^{n}\lambda_i = 1, \lambda_i \ge 0, & (i=1,\dots,n)\end{array}\right\}. \tag{1}$$

From (1), it is clear that the Farrell proportional distance function value for the same technology can be computed by solving the following model:

$$\max\left\{\delta \;:\; \begin{array}{ll} \displaystyle\sum_{i=1}^{n}\lambda_i x_{ir} \le x_{0r}-\delta x_{0r}, & (r=1,\dots,p),\\ \displaystyle\sum_{i=1}^{n}\lambda_i y_{is} \ge y_{0s}+\delta y_{0s}, & (s=1,\dots,q),\\ \displaystyle\sum_{i=1}^{n}\lambda_i = 1, \lambda_i \ge 0, & (i=1,\dots,n)\end{array}\right\}. \tag{2}$$

3 Proportional Distance Function: A Reformulation for Negative Data

Assuming now that inputs and/or outputs can be negative, one must revise the notion of a technology. In fact, an element of T no longer needs to be contained in $\mathbb{R}^{p+q}_+$. Hence, we redefine the technology T as

$$T = \{(x,y) \in \mathbb{R}^{p+q};\ x \text{ can produce } y\},$$

with the standard assumptions stated before. With this adaptation, Definition 2.1 of the directional distance function, the corresponding model (1) for computing it and Proposition 2.1 remain valid. However, the Farrell proportional distance function is no longer well-defined when inputs or outputs can take negative values, since the direction vector g is not necessarily contained in $(-\mathbb{R}^p_+)\times\mathbb{R}^q_+$. Such a choice is crucial to guarantee a simultaneous increase in the output direction and a decrease in the input direction.

To circumvent this problem, Silva Portela, Thanassoulis, and Simpson (2004) propose a so-called range directional model. In this model, the direction vector $g=(-R_0,S_0)$ is chosen for a DMU $z_0=(x_0,y_0)$ with

$$\begin{aligned} R_{0r} &= x_{0r}-\min\{x_{ir}\,;\, i=1,\dots,n\}, && (r=1,\dots,p);\\ S_{0s} &= \max\{y_{is}\,;\, i=1,\dots,n\}-y_{0s}, && (s=1,\dots,q).\end{aligned}$$

This choice for the direction vector assures that the direction vector $g \in (-\mathbb{R}^p_+)\times\mathbb{R}^q_+$ under all circumstances, thereby realizing a directional distance function suitable for

negative as well as positive data. Again in the case of a technology satisfying VRS and strong disposability assumptions, the following model needs to be solved:

$$\max \left\{ \delta \; : \; \begin{array}{ll} \displaystyle\sum_{i=1}^{n} \lambda_i x_{ir} \leq x_{0r} - \delta R_{0r}, & (r = 1, \ldots, p), \\ \displaystyle\sum_{i=1}^{n} \lambda_i y_{is} \geq y_{0s} + \delta S_{0s}, & (s = 1, \ldots, q), \\ \displaystyle\sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, & (i = 1, \ldots, n) \end{array} \right\}. \tag{3}$$

An obvious problem of this proposal is that the efficiency measure resulting from the range directional model no longer has a proportional interpretation, which is a disadvantage for practitioners.[5]

However, there is another simple alternative that basically generalizes the proportional distance function to handle negative data as well. This seems to have gone unnoticed in the literature so far. Given a DMU $z_0 = (x_0, y_0)$, we propose the direction vector $g = (-|x_0|, |y_0|)$ in which $|x_0|$ denotes the input vector with components $|x_{0r}| \, (r = 0, \ldots, p)$, and similarly $|y_0|$ denotes the output vector with components $|y_{0s}| \, (s = 0, \ldots, q)$. It is immediately obvious that this choice assures that $g \in (-\mathbb{R}^p_+) \times \mathbb{R}^q_+$ for both positive and/or negative data. Moreover, in the case of positive inputs and outputs, the direction vector coincides exactly with the one defining the original proportional distance function. Therefore, the proposed solution can indeed be seen as a generalization of the proportional distance function suitable for both positive and negative data domains. We suggest calling it the generalized proportional distance function.

From model (2), it immediately follows that the generalized proportional distance function value for a given DMU under the same assumptions as above is computed from the following LP model:

$$\max \left\{ \delta \; : \; \begin{array}{ll} \displaystyle\sum_{i=1}^{n} \lambda_i x_{ir} \leq x_{0r} - \delta |x_{0r}|, & (r = 1, \ldots, p), \\ \displaystyle\sum_{i=1}^{n} \lambda_i y_{is} \geq y_{0s} + \delta |y_{0s}|, & (s = 1, \ldots, q), \\ \displaystyle\sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, & (i = 1, \ldots, n) \end{array} \right\}. \tag{4}$$

Remark that the generalized proportional distance function value is just like the proportional distance function a measure of inefficiency. The closer this value to zero, the more efficient the corresponding DMU.

Figure 1 illustrates the proposed direction vector on a theoretical example consisting of 65 DMUs with one input (X) and one output (Y). These DMUs are visualized by small circles. Both inputs and outputs can be negative. The DEA VRS frontier is determined

[5]This important contribution is further discussed in contrast with other proposals regarding negative data in Pastor and Ruiz (2007).

completely by five DMUs defining the vertex points of this piecewise linear frontier. These vertices have as coordinates $(-12,-6)$, $(-9,3)$, $(-4,10)$, $(8,15)$ and $(14,17)$ respectively. For four DMUs (labeled with numbers 1 to 4) the projection onto the frontier by means of the generalized proportional distance function is indicated with an arrow, whereby the direction vector is selected to be $g = (-|x_0|, |y_0|)$.

FIGURE 1 ABOUT HERE

Table 1 focuses on these four DMUs and their projections. The coordinates (x_0, y_0) of the DMUs labeled with numbers 1 to 4 are provided in columns 2 and 3. The coordinates of the direction vector $g = (g_x, g_y)$ used in the generalized proportional distance function are listed in columns 4 and 5. Consequently, the direction of the arrows in Figure 1 is determined by the absolute value of the coordinates of the position vector of the initial points. Thus, despite what Figure 1 might suggest at first sight, the direction of the arrows is not arbitrarily, but it is precisely determined by the position of the evaluated DMUs.

TABLE 1 ABOUT HERE

In Figure 1, the resulting projection points located onto the frontier are labeled with the characters A to D. Columns 6 and 7 in Table 1 represent the coordinates (x_0^*, y_0^*) of these projection points A to D. The coordinates of the difference vector $d = (d_x, d_y) = (x_0^* - x_0, y_0^* - y_0)$ connecting the initial point with the projection point (visualized in Figure 1 with an arrow) is found in columns 8 and 9. Finally, the value of the generalized proportional distance function δ for the four DMUs is found in the last column.

We remark that this value can easily be computed from the previous elements in Table 1. We illustrate this for the DMU labeled 1. It follows from Proposition 2.1 that

$$\delta_1 = \frac{\|A1\|}{\|g\|}, \tag{5}$$

with $\|A1\|$ the distance[6] from the point labeled 1 to the point labeled A and $\|g\|$ the length of the appropriate direction vector. Consequently,

$$\begin{aligned} \delta_1 = \frac{\|d\|}{\|g\|} &= \frac{\sqrt{d_x^2 + d_y^2}}{\sqrt{g_x^2 + g_y^2}} \\ &= \frac{\sqrt{(1.806-8)^2 + (12.419-7)^2}}{\sqrt{(-8)^2 + 7^2}} = \frac{\sqrt{(-6.194)^2 + 5.419^2}}{\sqrt{(-8)^2 + 7^2}} = 0.7742. \end{aligned} \tag{6}$$

The inefficiency measures for the other points can be computed in a similar fashion.

We first recall that Proposition 2.1 guarantees a proportional interpretation of the inefficiency measure. Its value, however, can be larger than one as can be observed for

[6]We remark that using a distance notion requires an appropriate norm function. As indicated in Proposition 2.1, this choice of norm function does not influence the result. Therefore, we consider here the commonly used Euclidean norm for computing distances.

the DMUs labeled 2, 3 and 4. This means that an improvement of more than 100% can be achieved in certain cases. For instance, for DMU 2 the efficiency measure amounts to 1.8571, or 185.71%. This means that the performance of this DMU can almost be doubled with respect to its original position by moving it to the location labeled B. In Figure 1, this can be observed by the fact that the origin almost halves the distance from the point labeled 2 to the point labeled B. Obviously, the closer a point is situated to the frontier, the smaller is the numerator of (5) leading to smaller inefficiency values and therefore more efficient units.[7]

Furthermore, also notice that in the case of one input and one output, all DMUs positioned in the second and fourth quadrant are projected in a direction whose support line passes the origin. This follows immediately from the choice of direction vector. This phenomenon can be observed for the points labeled 2 and 3 in Figure 1.

4 Concluding Comments

The fast growing DEA literature has for a long time neglected the issues surrounding the use of negative data in managerially oriented benchmarking models. The timely work of Silva Portela, Thanassoulis, and Simpson (2004) suggest a variation on the directional distance function, a general distance function compatible with profit maximization that has recently gained some popularity. This contribution has argued that a very simple modification of the traditionally defined proportional distance function can alternatively be employed in this context whenever a proportional interpretation is an asset.

References

Ali, A., and L. Seiford (1990): "Translation Invariance in Data Envelopment Analysis," *Operations Research Letters*, 9(6), 403–405.

Anderson, R., D. Lewis, and T. Springer (2000): "Operating Efficiencies in Real Estate: A Critical Review of the Literature," *Journal of Real Estate Literature*, 8(1), 1–18.

Briec, W. (1997): "A Graph-Type Extension of Farrell Technical Efficiency Measure," *Journal of Productivity Analysis*, 8(1), 95–110.

Briec, W., and K. Kerstens (2009): "Infeasibilities and Directional Distance Functions with Application to the Determinateness of the Luenberger Productivity Indicator," *Journal of Optimization Theory and Applications*, 141(1), 55–73.

Chambers, R. (2002): "Exact Nonradial Input, Output, and Productivity Measurement," *Economic Theory*, 20(4), 751–765.

[7]However, for points closer to the origin, the denominator of (5) decreases, consequently leading to larger inefficiencies. If a DMU would be located in the origin, then its inefficiency would measure $+\infty$.

Chambers, R., Y. Chung, and R. Färe (1998): "Profit, Directional Distance Functions, and Nerlovian Efficiency," *Journal of Optimization Theory and Applications*, 98(2), 351–364.

Charnes, A., W. Cooper, and E. Rhodes (1978): "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, 2(6), 429–444.

Chavas, J.-P., and K. Kim (2007): "Measurement and Sources of Economies of Scope: A Primal Approach," *Journal of Institutional and Theoretical Economics*, 163(3), 411–427.

Cummins, D., and M. Weiss (2000): "Analyzing Firm Performance in the Insurance Industry Using Frontier Efficiency and Productivity Methods," in *Handbook of Insurance*, ed. by G. Dionne, pp. 767–829. Kluwer, Boston.

De Borger, B., K. Kerstens, and A. Costa (2002): "Public Transit Performance: What Does One Learn From Frontier Studies?," *Transport Reviews*, 22(1), 1–38.

Emrouznejad, A., B. Parker, and G. Tavares (2008): "Evaluation of Research in Efficiency and Productivity: A Survey and Analysis of the First 30 Years of Scholarly Literature in DEA," *Socio-Economic Planning Sciences*, 42(3), 151–157.

Färe, R., S. Grosskopf, and C. Lovell (1994): *Production Frontiers*. Cambridge University Press, Cambridge.

Farrell, M. (1957): "The Measurement of Productive Efficiency," *Journal of the Royal Statistical Society Series A: General*, 120(3), 253–281.

Feroz, E., S. Kim, and R. Raab (2003): "Financial Statement Analysis: A Data Envelopment Analysis Approach," *Journal of the Operational Research Society*, 54(1), 48–58.

Førsund, F., and N. Sarafoglou (2005): "The Tale of Two Research Communities: The Diffusion of Research on Productive Efficiency," *International Journal of Production Economics*, 98(1), 17–40.

Harker, P., and S. Zenios (2001): "What Drives the Performance of Financial Institutions?," in *Performance of Financial Institutions: Efficiency, Innovation, Regulation*, ed. by P. Harker, and S. Zenios, pp. 3–31. Cambridge University Press, Cambridge.

Luenberger, D. (1992): "New Optimality Principles for Economic Efficiency and Equilibrium," *Journal of Optimization Theory and Applications*, 75(2), 221–264.

——— (1995): *Microeconomic Theory*. McGraw-Hill, Boston.

Murthi, B., Y. Choi, and P. Desai (1997): "Efficiency of Mutual Funds and Portfolio Performance Measurement: A Non-Parametric Approach," *European Journal of Operational Research*, 98(2), 408–418.

Ozcan, Y. (2008): *Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis (DEA)*. Springer, Berlin.

Pastor, J., and J. Ruiz (2007): "Variables with Negative Values in DEA," in *Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis*, ed. by J. Zhu, and W. Cook, pp. 63–84. Springer, Berling.

Silva Portela, M., E. Thanassoulis, and G. Simpson (2004): "Negative Data in DEA: A Directional Distance Approach Applied to Bank Branches," *Journal of the Operational Research Society*, 55(10), 1111–1121.

Smith, P. (1990): "Data Envelopment Analysis Applied to Financial Statements," *Omega*, 18(2), 131–138.

Thanassoulis, E., M. Silva Portela, and O. Despić (2008): "DEA - The Mathematical Programming Approach to Efficiency Analysis," in *The Measurement of Productive Efficiency and Productivity Change*, ed. by H. Fried, C. Lovell, and S. Schmidt, pp. 251–420. Oxford University Press, New York.

Worthington, A. (2001): "An Empirical Survey of Frontier Efficiency Measurement Techniques in Education," *Education Economics*, 9(3), 245–268.

From-To	x_0	y_0	g_x	g_y	x_0^*	y_0^*	d_x	d_y	δ
$1-A$	8	7	-8	7	1.806	12.419	-6.194	5.419	0.7742
$2-B$	8	-7	-8	7	-6.857	6.000	-14.857	13.000	1.8571
$3-C$	-4	2	-4	2	-8.211	4.105	-4.211	2.105	1.0526
$4-D$	-4	-2	-4	2	-9.714	0.857	-5.714	2.857	1.4286

Table 1: Numerical Example with Four DMUs

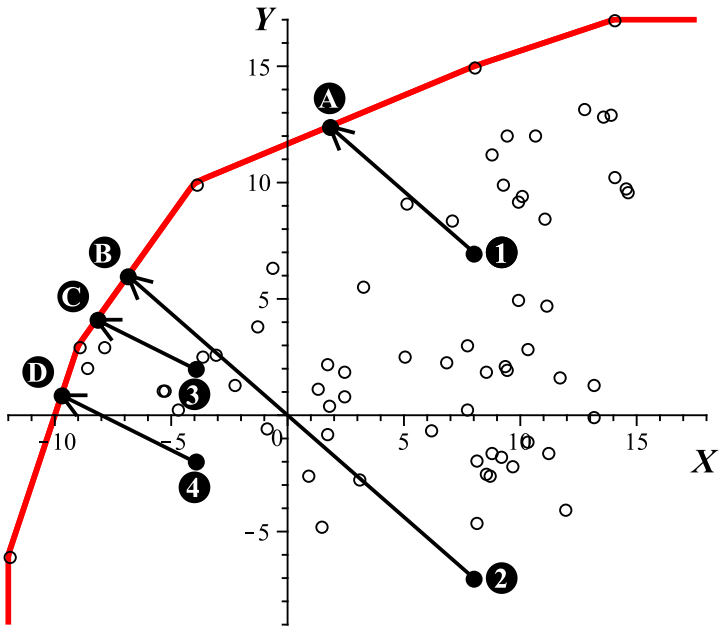

Figure 1: DEA VRS Frontier: Projections for Four Inefficient DMUs