Multiple Data Envelopment Analysis: The Blessing of Dimensionality

Borko D. Stosic
Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros s/n, Dois Irmãos,
52171-900 Recife-PE, Brazil

Ivon P. Fittipaldi
Ministério da Ciência e Tecnologia
Esplanada dos Ministérios, Bloco E, 2° andar, Sala 215
70067-900 Brasília-DF, Brazil

In this work we propose an extension of the standard Data Envelopment Analysis (DEA), termed Multiple Data Envelopment Analysis (MDEA), based on performing multiple DEA runs for different possible choices of subsets of input and output variables. The proposed procedure is suitable for the situations where a large number of input and output variables must be taken into account, and the number of Decision Making Units (DMUs) is relatively small, so that the traditional DEA looses its discriminative power (effect known as "curse of dimensionality", when a large number of the considered DMUs become efficiency standards in themselves). To deal with this problem, the proposed method invests additional computational effort to account for the large number of inputs and outputs, and their respective combinations. This approach provides efficiency spectra (frequency distributions) for each DMU, from which efficiency ranking can be extracted, together with confidence intervals. The proposed approach is tested on two controlled, artificial data sets.

1. Introduction

Data Envelopment Analysis (DEA) is a non-parametric method for estimating technical efficiency of Decision Making Units (DMUs), by application of a linear programming technique for comparative analysis of input and output variables. The method considers empirical efficiency frontiers, spanned by those DMUs that use minimal input to produce maximal output, within the observed sample. The foundations of the method were established by Farell in 1957 [1], while the term "Data Envelopment Analysis" was coined, and the method rendered operational only two decades later, by Charnes, Cooper and Rhodes in 1978 [2]. With the exponential advent of easily accessible powerful computing resources over the last decades, application of DEA has been steadily gaining momentum in diverse research areas, ranging from economics to social sciences (extensive lists of references have been compiled in [3] and [4]), while its wider acceptance for application in other disciplines may be expected in the near future. Another non-parametric method closely related to DEA is Free Disposal Hull, introduced by Deprins et al [5], which has also been gaining in popularity since it imposes weaker assumptions on the production set, although it has somewhat less discriminative power then DEA. Methods such as DEA and FDH are particularly attractive because of the fact that they do not require a priori knowledge of the functional relation between the input and output variables (technology), nor do they impose arbitrary statistical weights (relative importance) on the input and output variables.

On the other hand, there are situations for which these methods are not suitable in their usual form. One such situation is encountered when the number of observed DMUs is small, and it is essential that a large number of input and output variables be considered. In such situations most (or all) of the DMUs are represented by points on the efficiency frontier, they all receive unit efficiency scores, and the method looses its discriminative power. This effect is known as "curse of dimensionality", and it is not unique to non-parametric technical efficiency estimation methods such as DEA and FDH. Rather, it is encountered whenever the dimension of the variable space is large, and an enormous sample size is required to provide a reasonable representation of the volume of interest. In particular, techniques dealing with maximization/minimization of functions, numerical integration, etc., all suffer from this effect, when

the dimension of the variable space is large. However, along with the notion of "curse of dimensionality", it has been recently noted by Donoho [6] that the "blessing of dimensionality" effects may also exist. In particular, three concrete situations have been identified when the "curse of dimensionality" may be countered by considering the specific nature of the problem: i) when the function of interest is highly concentrated in a small part of the variable space volume, ii) when the continuum limit can be analytically treated by considering infinite dimensional variable space, and iii) when the (numerous) variables of interest represent a sample from an underlying continuous function.

In this work we identify another such "blessing of dimensionality" effect relevant for methods such as DEA and FDH, which helps reduce (or overcome) the "curse of dimensionality" problem. More precisely, we present a new deterministic resampling scheme termed Multiple Data Envelopment Analysis – MDEA (or Multiple Free Disposal Hull – MFDH), which broadens the spectrum of applicability of DEA (and FDH) in the above-mentioned problematic situations. We start from the point of view that the actual choice of a particular set of input and output variables in itself represents a sort of "parameterization", since it may strongly affect the efficiency score of any given DMU within the observed group. The MDEA (MFDH) procedure removes this inconsistent intrinsic "parameterization" from the analysis, by sequentially choosing *all possible combinations* of subsets of inputs and outputs from the considered data set, and performing individual DEA (FDH) runs (for all of the DMUs) for each particular choice. The reduction of the dimensionality of the parameter space removes the degeneracy from the DEA (FDH) analysis, and simultaneously all of the DMUs get the same "fair" chance to be evaluated in all the possible different contexts (combinations of input and output variables). This procedure provides efficiency spectra (frequency distributions) for each individual DMU, from which efficiency ranking can be established together with confidence intervals.

This paper is organized as follows. In the next Section we introduce the MDEA (MFDH) approach, in the subsequent Section 3 we present an application of MDEA on two controlled, artificial datasets, and finally in Section 4 we draw the conclusions.

2. Multiple Data Envelopment Analysis

Our approach is based on reasoning that in general there are no deterministic rigorous methods for the actual choice of parameters to be used in a given problem. Rather, this is usually a somewhat subjective procedure, based on phenomenological considerations, as well as on the availability of reliable data. While one researcher may argue that a given choice of input and output variables is perfect for the problem at hand, another may disagree, defending a different choice of variables. Similarly, the parties having interests associated with the DMUs under study (e.g. the CEO of a corporation being evaluated for efficiency) may argue that a given choice of variables is more adequate (or just) than another, while in fact motivated by a better ranking of their DMU within the observed sample. To resolve such controversies and simultaneously reduce the dimensionality of the parameter space (thus augmenting the DEA/FDH discriminative power), one may consider the following procedure.

First identify the largest sets of N inputs and M outputs, considered relevant for the observed phenomenon, and then make successive choices of all the different subsets of $n \in \{1,2,\ldots,N\}$ inputs and $m \in \{1,2,\ldots,M\}$ outputs. Since there are $\binom{N}{n}$ possible ways of choosing a subset of n inputs from the total of N, there are $\sum_{n=1}^{N}\binom{N}{n} = 2^N - 1$ possible choices for the inputs, and analogously, $2^M - 1$ choices for the outputs. For each combination of input and output subsets, perform a DEA run (implement the linear programming algorithm) for all of the DMUs, storing the results (there are altogether $N_C \equiv (2^N-1)(2^M-1)$ such combinations). At the end of this process, each DMU receives N_C efficiency scores, each corresponding to a comparison with the other DMUs in a specific context (choice of input and output variables). This set of efficiency values represents probably the most "fair" form of evaluating a given DMU in comparison with the others, as it contains all the possible evaluation contexts. One can now plot

the efficiency frequency distribution (histogram) for each DMU, wherefrom the mean (or some other statistic) can be extracted and identified with the DMUs efficiency, as well as other usual statistical quantities. We term this conceptually simple, but computationally intensive procedure, Multiple Data Envelopment Analysis - MDEA (or Multiple Free Disposal Hull -MFDH).

The "blessing of dimensionality" is here manifested through the fact that the more variables there are, the more combinations of their subsets can be constructed (roughly 2^d combinations for variable space dimension d≡N+M), and each such combination represents a unique "context" for DMU efficiency evaluation. Besides countering the "curse of dimensionality", multiple evaluations for each DMU also provide efficiency frequency histograms, which may be used for extracting the confidence intervals. This is in fact another important aspect, where deterministic non-parametric methods such as DEA and FDH fail to provide necessary information by themselves, and some additional procedure (such as Bootstrap) is usually applied. Finally, by considering the dataset as a table where lines represent the DMUs, and the columns represent individual variables, one can perceive MDEA (MFDH) as a sequence of multiple Jackknife procedures (from remove-one, to remove-all-but-one) on the input and output columns (rather then on the lines, as in usual Jackknife).

3. Application on controlled datasets

To demonstrate de performance of MDEA on limited datasets with a large variable space, we use two well-defined DGPs to generate data where efficiency of each DMU is known exactly, and may be compared with the MDEA estimates. We start by using the recipe of Simar and Zelenyuk [7] to generate artificial DMUs with multiple inputs and a single output. More precisely, we consider situations with N inputs and a single output characterized by the Cobb-Douglas technology

$$y_k^* = \prod_{i=1}^{N} x_{ik}^{\alpha_i} \quad , \tag{1}$$

where y_k* is technically efficient output level for DMU k, given the mutually independent input levels x_{ik}, i=1,…,N, drawn from a uniform distribution on (0,1), and $0<\alpha_i<1$ are the coefficients defining the distribution: here we choose $\alpha_i=i/10(N+1)$ (this choice yields a balanced range of values for y_k*). To emulate a skewed efficiency distribution with most of the DMUs close to the efficiency frontier and a diminishing tail towards lower efficiency (see [7] for more details), the "observed" output is generated as

$$y_k = \frac{y_k^*}{1+u_k} \tag{2}$$

where u_k is a positive deviate drawn from the normal distribution with mean μ=0 and standard deviation σ=1.0, yielding for the efficiency expectation value E(θ)≈0.56.

For this type of simulated datasets it was shown by Simar and Zelenyuk [7] that DEA exhibits a pronounced "curse of dimensionality" effect for K=20 DMUs already for N=7. In what follows, we test the performance of MDEA for a rather extreme situation with K=20 and N=20 (which may be considered "hopeless" by practitioners using current state of the art tools), by comparing the MDEA efficiency estimates with the known exact values. The current choice of N=20 inputs and M=1 outputs offers a total of $N_C≡(2^{20}-1)(2^1-1)= 1048575$ subsets of inputs (there is a single possible choice for the unique output).

As a first test, we run DEA for all of the 1048575 choices, and record the percentage of cases when DEA yields unit efficiency, as a function of the maximum number of inputs used to form the variable subsets (subspace dimension), where we differentiate between true efficiency and efficiency by default (when a DMU receives a unit efficiency score but is not peer to any other DMUs). As may be expected, it turns out that the proportion of cases corresponding to efficiency by default becomes dominant with increasing dimension of the subspace, while in fact these scores do not provide any information on the true efficiency of the DMUs. The results displayed on Figure 1 demonstrate saturation of the percentage curve, where e.g. at the level of maximum nine inputs 80% of the times DEA attributes unit efficiency scores, while 40% of the times efficiency by default is observed. While in itself a large

percentage of unit scores represents a manifestation of the "curse of dimensionality", significant presence of efficiency by default may be considered an undesirable artifact of DEA methodology, which enhances failure of DEA discriminative power, and we therefore proceed by excluding cases of efficiency by default from mean efficiency calculations. We have performed calculations for several successive variable subspace dimensions, and it turns out that successive MDEA level curves demonstrate systematic shift toward higher efficiency values, where lower level curves tend to underestimate, and the higher level curves tend to overestimate the true efficiency scores. It turns out that the case of maximum four inputs, with a total of $\sum_{i=1}^{4}\binom{20}{i}=6195$ DEA runs, demonstrates the smallest standard deviation in relation to the true efficiency scores. In this case a total of 40,09% unit scores were observed, and 5,76% of efficiency by default (the last were then discarded in calculations of averages). These values may be taken as a starting point for MDEA application on real datasets, or for unknown distributions. From the results shown in Figure 2 it is seen that results lie in the correct range, the largest absolute deviations from the true score being of the order 0.2. It should be stressed that here we are dealing with an extreme situation where straightforward DEA and FDH *have no discriminative power at all*, and therefore these results may be considered rather satisfactory.

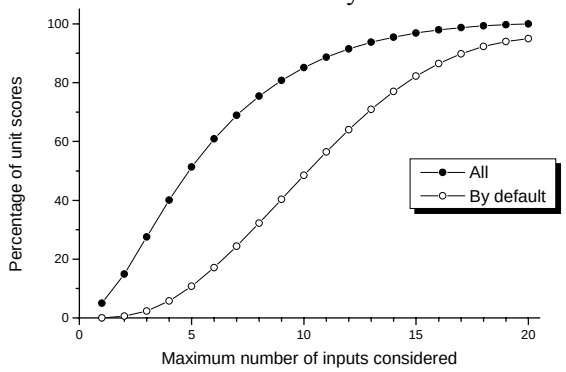

Figure 1. Percentage of unit scores obtained from DEA runs, as the number of inputs is increased from 1 to 20.

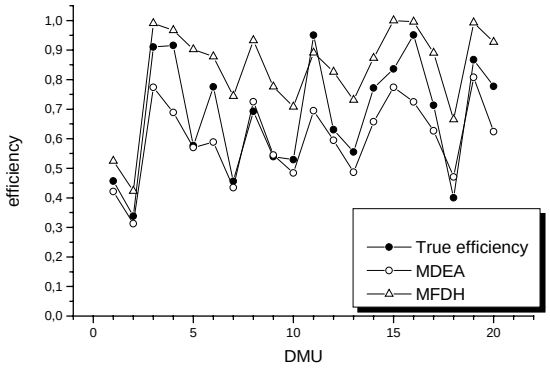

Figure 2. Average MDEA and MFDH scores for $N\leq 4$ inputs, together with true efficiency scores, for DGP I.

Finally, on Figure 3 we display a (typical) histogram of efficiency values for the first DMU from the first dataset, with true efficiency of 0,45671 and the MDEA4 estimate of 0,42. It is seen that the efficiency frequency histogram is well behaved, and may be used to extract confidence intervals.

To test MDEA on a truly multivariate case (multiple inputs and multiple outputs), we modify the procedure of Daraio and Simar [8] as follows. As before, the inputs are generated from a uniform distribution, and the total technically efficient output level y_k^* of DMU k is given by equation (1). The individual technically efficient outputs are now constructed by first generating M uniform deviates Y_{km} , m=1,…,M on (0.2,5) to fix the output slopes $y_{km}^*/y_{k1}^*=S_m\equiv Y_{km}/Y_{m1}$, and then imposing $y_{k1}^*+y_{k2}^*+\ldots+y_{kM}^*=y_k^*$. Denoting by S the normalizing factor (the sum of the output slopes) $S=S_1+S_2+\ldots+S_M$, for the individual technically efficient outputs we now have $y_{km}^*=S_m y_k^*/S$. Finally, efficiencies are generated as $\exp(-U_k)$, where U_k are drawn from an exponential with mean $\mu=1/3$, and the individual outputs are calculated as $y_{km}=\exp(-U_k)\, y_{km}^*$.

For testing the performance of MDEA we choose K=20 DMUs with N=10 inputs and M=10 outputs, giving a total of $N_C\equiv(2^{10}-1)(2^{10}-1)=1046529$ possible subsets of inputs and outputs. Again it turns out that increasing the dimension of the variable space augments both the percentage of the DMUs on the frontier and the efficiency by default, and we find that the case with maximum 4 inputs and a maximum of 4 outputs yields results closest to true efficiency levels (in terms of standard deviation). Results of our calculations are shown on Fig 4, where it is seen that level of performance of MDEA and MFDH are similar to that of the previous case.

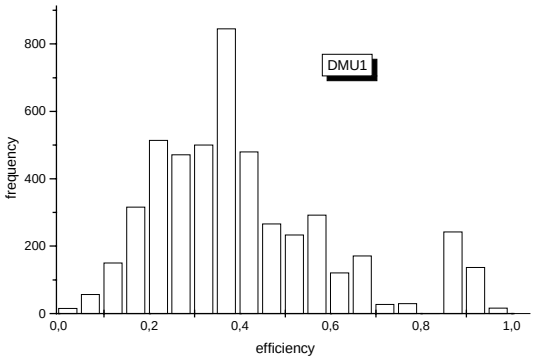

Figure 3. Histogram of efficiency values for N≤4 inputs, for the first DMU from DGP I.

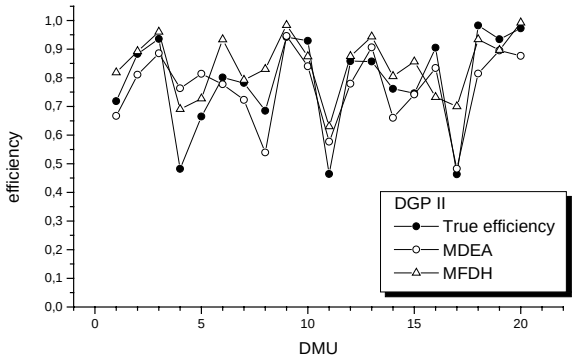

Figure 4. Average MDEA scores for N≤4 and M≤4, together with true efficiency scores, for DGP II.

4. Conclusion

In conclusion, MDEA resampling scheme represents a new tool that may help reduce the impact of the "curse of dimensionality" problem, when the observed phenomenon requires use of many variables, and the number of DMUs is limited. While it is computationally rather intensive, it can be argued that powerful computer resources have already become a rather accessible asset, and that this fact does not represent a serious impediment for its use. In the current case of 20 DMUs, 20 inputs and a single output, running all of the 1048575 possible DEA runs takes several minutes on a 1.6Ghz Celeron machine with a specialized program written in C programming language, while MDEA4 with 6195 combinations takes only a couple of seconds. However, before applying MDEA in practice, further theoretical and experimental investigations are needed, in order to establish the extent of its applicability. In particular, the question of interpretation of the efficiency histograms (is average the correct statistic, how confidence intervals can be extracted, etc.) should be addressed. Another question that should be addressed both theoretically and experimentally is whether 40% frontier and 5-10% default efficiency represent optimum levels for MDEA application, or should these be regarded as functions of sample size and the dimension of the variable space. Finally, this procedure should be tested on different artificial (other DGP's) and real datasets.

References

[1] M. J. Farrell, "The Measurement of Productive Efficiency", *Journal of the Royal Statistical Society*, vol. 120(3), pp 253-290, 1957.

[2] A. Charnes, W.W. Cooper and E. Rhodes, "Measuring the efficiency of decision making units", *European Journal of Operational Research*, vol. 2, pp 429-444, 1978.

[3] G.A. Tavares, "A bibliography of data envelopment analysis (1978–2001)", RRR 01-2002, *RUTCOR—Rutgers Center for Operations Research*, Rutgers University 2002. http://rutcor.rutgers.edu/pub/rrr/reports2002/1_2002.pdf

[4] A. Emrouznejad, "An Extensive Bibliography of Data Envelopment Analysis (DEA), Volume I - V ", *Business School, University of Warwick 2001*. Coventry CV4 7AL, England. http://www.deazone.com/bibliography

[5] Deprins, D., L. Simar, and H. Tulkens. (1984). "Measuring Labor Efficiency in Post Offices." In M. Marchand, P. Pestieau and H. Tulkens (eds) *The Performance of Public Enterprises: Concepts and Measurements*. Elsevier, 345-367.

[6] D.L. Donoho, "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality". *Aide-Memoire of the invited lecture at the conference "Mathematical Challenges of the 21st Century"*, American Mathematical Society, August 2000, Los Angeles, USA. http://www-stat.stanford.edu/~donoho/lectures.html

[7] L. Simar and V. Zelenyuk "On Testing Equality ff Distributions of Technical Efficiency Scores". *Discussion Paper #0434 of Institute of Statistics, University Catholique de Louvain*, Belgium. 2004. http://www.stat.ucl.ac.be/ISpub

[8] C. Daraio and L. Simar "Introducing Environmental Variables in Nonparametric frontier Models: A Probabilistic Approach". *Technical reportr #0318 of Institute of Statistics, University Catholique de Louvain*, Belgium. 2003. http://www.stat.ucl.ac.be/ISpub