Feature Selection in Data Envelopment Analysis: A Mathematical Optimization approach

Sandra Benítez Peña, Peter Bogetoft and Dolores Romero Morales

Abstract

This paper proposes an integrative approach to feature (input and output) selection in Data Envelopment Analysis (DEA). The DEA model is enriched with zero-one decision variables modelling the selection of features, yielding a Mixed Integer Linear Programming formulation. This single-model approach can handle different objective functions as well as constraints to incorporate desirable properties from the real-world application. Our approach is illustrated on the benchmarking of electricity Distribution System Operators (DSOs). The numerical results highlight the advantages of our single-model approach provide to the user, in terms of making the choice of the number of features, as well as modeling their costs and their nature.

Keywords: Benchmarking; Data Envelopment Analysis; Feature Selection; Mixed Integer Linear Programming

1 Introduction

Organisations need to know whether they are using the best practices to produce their products and services, and to do so they benchmark their performance with that of others. There are many documented examples of the use of benchmarking in the literature from both the private and the public sector, such as, airlines, banks, hospitals, universities, manufacturers, schools, and municipalities, see [2], and references therein.

Within benchmarking, Data Envelopment Analysis (DEA) is one of the most widely used tools, [4, 7, 8, 11, 14, 15]. It aims at benchmarking the performance of decision marking units (DMUs), which use the same types of inputs and produce the same types of outputs, against each other. DEA calculates an efficiency score for each of the DMUs, so that DMUs with a score equal to one are in the so-called efficient frontier. DMUs outside the efficient frontier are deemed as underperforming, and a further analysis gives insights as to what they can do to improve their efficiency. The efficiency of DMUs in

DEA is measured as the weighted summation of the outputs divided by the weighted summation of the inputs, and the weights are found solving a Linear Programming problem for each DMU. DEA model specification, in the form of feature (where the term feature is used to refer to either outputs, inputs or environmental variables) selection, has a significant impact on the shape of the efficient frontier in DEA as well as the insights given to the inefficient DMUs [6]. Moreover, it is known to improve the discriminatory power of DEA models [3]. Our paper proposes and investigates a mathematical optimization approach for feature selection in DEA.

2 State of the art

The complexity of the model specification phase partially explains the lack of enough guidance in the literature at this respect, [5, 12], and most of the effort goes into the analysis and interpretation of a given DEA model. With the strand of literature on feature selection, the most common approach is to use a priori rules based on Statistical Analysis (such as correlations, dimensionality reduction techniques, and regression), and Information Theory (such as AIC or Shannon entropy). Alternatively, an ex-post analysis of the sensitivity of the efficient frontier to additional features can be run to detect whether relevant features have been left out. See [1, 10, 13, 16], and references therein. Recently, there have been attempts to use LASSO techniques from Statistical Learning to build sparse benchmarking models, i.e., models using just a few features, [9].

3 Main Contributions

In this paper, the DEA Linear Programming formulation is enriched with zero-one decision variables modelling the selection of features for different objective functions, such as the average efficiency or the squared distance to the ideal point where all DMUs are efficient, and for different set of constraints that incorporate knowledge from the industrial application, such as bounds on the weights as well as costs on the features. This yields either a Mixed Integer Linear Programming (MILP) problem, or a Mixed Integer Quadratic Programming (MIQP) one. Thus, in contrast to the existing literature, that tends to combine statistical analysis with the mathematical programming based DEA, we propose an approach that is entirely driven by mathematical optimization. We illustrate our models in the benchmarking of electricity Distribution System Operators (DSOs), where there is a pool of 100 potential outputs.

The contributions of our approach are threefold. First, our single-model mathematical approach can guide better the selection of features: it controls directly the number of chosen features, as opposed to techniques based on seeking sparsity, being thus able to quantify the added value of additional features; works directly with the original features, as opposed to dimensionality reduction techniques, which create

artificial features that are difficult to interpret; and can derive a collection of models by shaping in alternative ways the distribution of the efficiencies, using different objective functions that focus on different groups of DMUs, which can be combined through, for instance, Shannon entropy. Second, while the previous literature has focused on the choice of variables from a small set of candidates, e.g., [9], in the era of Big Data, the set of alternatives to choose from is expanding at a fast pace, and the challenge is often not the lack of data, but the abundance of data, [17]. In the numerical section, we show how our MILP/MIQP approach is able to make the selection from a large pool of outputs. Third, we introduce an element of game theory when selecting features.

References

[1] N. Adler and E. Yazhemsky. Improving discrimination in data envelopment analysis: PCA–DEA or variable reduction. *European Journal of Operational Research*, 202(1):273–284, 2010.

[2] P. Bogetoft. *Performance benchmarking: Measuring and managing performance*. Springer Science & Business Media, 2013.

[3] P. Bogetoft and L. Otto. *Benchmarking with Dea, Sfa, and R*, volume 157. Springer Science & Business Media, 2010.

[4] W.D. Cook, N. Ramón, J.L. Ruiz, I. Sirvent, and J. Zhu. DEA-based benchmarking for performance evaluation in pay-for-performance incentive plans. *Omega*, 84:45 – 54, 2019.

[5] W.D. Cook, K. Tone, and J. Zhu. Data envelopment analysis: Prior to choosing a model. *Omega*, 44:1–4, 2014.

[6] B. Golany and Y. Roll. An application procedure for DEA. *Omega*, 17(3):237–250, 1989.

[7] C. Jiang and W. Lin. DEARank: a data-envelopment-analysis-based ranking method. *Machine Learning*, 101(1–3):415–435, 2015.

[8] M. Landete, J.F. Monge, and J.L. Ruiz. Robust DEA efficiency scores: A probabilistic/combinatorial approach. *Expert Systems with Applications*, 86:145–154, 2017.

[9] C.-Y. Lee and J.-Y. Cai. LASSO variable selection in data envelopment analysis with small datasets. Forthcoming in *Omega*, 2018.

[10] Y. Li, X. Shi, M. Yang, and L. Liang. Variable selection in data envelopment analysis via akaike's information criteria. *Annals of Operations Research*, 253(1):453–476, 2017.

[11] Z. Li, J. Crook, and G. Andreeva. Dynamic prediction of financial distress using Malmquist DEA. *Expert Systems with Applications*, 80:94–106, 2017.

[12] Y. Luo, G. Bi, and L. Liang. Input/output indicator selection for DEA efficiency evaluation: An empirical study of Chinese commercial banks. *Expert Systems with Applications*, 39(1):1118–1123, 2012.

[13] N.R. Nataraja and A.L. Johnson. Guidelines for using variable selection techniques in Data Envelopment Analysis. *European Journal of Operational Research*, 215(3):662–669, 2011.

[14] N.C. Petersen. Directional Distance Functions in DEA with Optimal Endogenous Directions. *Operations Research*, 66(4):1068–1085, 2018.

[15] J.L. Ruiz and I. Sirvent. Performance evaluation through DEA benchmarking adjusted to goals. Forthcoming in *Omega*, 2018.

[16] I. Sirvent, J.L. Ruiz, F. Borrás, and J.T. Pastor. A Monte Carlo evaluation of several tests for the selection of variables in DEA models. *International Journal of Information Technology & Decision Making*, 4(03):325–343, 2005.

[17] Q. Zhu, J. Wu, and M. Song. Efficiency evaluation based on data envelopment analysis in the big data context. *Computers & Operations Research*, 98:291–300, 2018.