European Journal of Operational Research 000 (2016) 1–11

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Comparison of variable selection techniques for data envelopment analysis in a retail bank

Juha Eskelinen*

Aalto University, School of Business, Department of Information and Service Economy, P.O. Box 21220, 00076 Aalto, Helsinki, Finland.

ARTICLE INFO

Article history:
Received 8 January 2015
Accepted 2 November 2016
Available online xxx

Keywords:
Data envelopment analysis
OR in banking
Variable selection
Productive efficiency analysis

ABSTRACT

If there are too few units compared to inputs and outputs, the efficiency evaluation based upon the data envelopment analysis suffers from a lack of discrimination. The literature has proposed various statistical techniques when the value judgments do not guide the selection of the inputs/outputs. Two techniques, the variable reduction procedure of Jenkins and Anderson (2003) and the approach based upon the efficiency contribution measure of Pastor, Ruiz, and Sirvent (2002), were compared in an empirical retail bank context. The objective was to select a representative set of outputs from the services the bank provides. As the techniques take different approaches to selecting influential variables, the output sets proposed by the techniques diverged. This created some significant differences in the efficiency evaluations of the bank branches. The bank management gave feedback on the techniques and the results from a practical perspective. The techniques led to different managerial interpretations of the performance complementing each other. Thus, the techniques can be utilized to evaluate the units from multiple perspectives.

1. Introduction

Data envelopment analysis (DEA) is a method used to evaluate the efficiencies of comparable units with multiple inputs and/or outputs. It takes an optimistic perspective by choosing, for each decision-making unit (DMU), the most beneficial non-negative (i.e., those weights will maximize its efficiency score). This virtue has a consequence. Having a low number of units under assessment compared to inputs and outputs leads to finding a large portion of the units efficient or nearly efficient. The identification of many units falsely efficient and too optimistic efficiency scores is called a *lack of discrimination* (see e.g., Podinovski & Thanassoulis, 2007). Such an efficiency estimation may not serve the purpose of the evaluation.

Banks consume several resources (inputs) to produce several services or other outcomes (outputs). Even though there are large nationwide and international banks, the majority of the branch networks are small with a relatively low number of branches.[1] In addition, in the large banks, the branches are often clustered into smaller homogeneous subsets. Thus, the lack of discrimination can be an issue in bank branch efficiency evaluations (see, e.g., the discussion of Paradi & Zhu, 2013).

This essay focuses on the variable side (i.e., inputs and outputs) of the discrimination problem. The number of variables can be reduced by selecting the most important outputs/inputs to depict the activity. Numerous statistical techniques have been proposed for situations where judgmental knowledge is not available or not sufficient enough for selecting the variables. The following review illustrates the variety of techniques.[2]

Some selective techniques omit variables prior to DEA. Lewin, Morey, and Cook (1982), among others, used correlation and regression analysis to test redundant variables to be omitted. Jenkins and Anderson (2003) proposed a variable reduction procedure based on partial covariance (abbreviated as VR in this paper), where a combination of variables is selected based upon the proportion of the total variance retained. Gonzalez-Bravo (2007)

* Corresponding author.
E-mail addresses: juha.p.eskelinen@aalto.fi, juha.eskelinen@melkior.fi

[1] In the United States, the commercial banks had, on average, 13.2 branches in 2011 (data source: U.S. Federal Deposit Insurance Corporation). In the United Kingdom, the monetary financial institutions have 28.7, in Germany 18.9, in France 31.8, and in Finland 3.8 branches (data source: European Central Bank 2011, UK data from 2010). The exceptionally low figure in Finland is explained by the high proportion of small local cooperative and savings banks.

[2] Besides selective techniques, that are the focus of this paper, there is another branch of statistical techniques that deals with the lack of discrimination by determining statistically the weights for summing up the data in a reduced set of variables. Various techniques ending up in aggregation have been proposed by Sengupta (1990), Sinuany-Stern and Friedman (1998), Ueda and Hoshiai (1997), Adler and Golany (2001, 2002), Bian (2012), and Amirteimoori, Despotis, and Kordrostami (2014).

http://dx.doi.org/10.1016/j.ejor.2016.11.009
0377-2217/

proposed a procedure to help model specification by using an output/input ratio-based analysis prior to DEA.

Another sub-branch of selective techniques tests the impact of the retained variables on DEA results. Banker (1996) proposed the use of statistical tests for different inefficiency distribution assumptions, to indicate the significance of a variable in the production process (i.e., efficiency estimates). Simar and Wilson (2001) recommended bootstrapping to test whether the input or output variables are relevant. Pastor, Ruiz, and Sirvent (2002) introduced an efficiency contribution measure (ECM) to test for the significance of a variable in efficiency estimation. Fanchon (2003), Ruggiero (2005) and Sharma and Yu (2015) regressed the efficiency scores against the variables and tested the regression coefficient in dropping the insignificant variables. Wagner and Shimshak (2007) proposed a stepwise procedure that evaluated the impact of the variable set on the average efficiency scores.

There are a few published studies comparing some of the statistical variable selection techniques (see Nataraja & Johnson, 2011; Sirvent, Ruiz, Borras, & Pastor, 2005). These comparisons are based upon Monte Carlo simulations, and the efficiency estimates are compared to the known true frontier.

We take another approach in this paper. The main interest is in the implications that the use of variable selection could have in the interpretation and acceptance of the DEA results for performance management. We use data that have been gathered during the research collaboration with a Finnish retail bank. The variable selection has a specific contextual focus: to specify a model for evaluating the operational efficiencies of the branches during 2007–2010 using DEA. The techniques and their outcomes are reviewed by the branch network management.

We focus on two techniques: VR and ECM. Our prior assumption is that both techniques could be helpful in our empirical case, where the output variables represent the sales quantities of banking services that are highly positively correlated. Both techniques arise from the critique against intuitive variable exclusion, based upon high correlations between variables, and present statistical approaches for variable selection (see Jenkins & Anderson, 2003; Pastor et al., 2002)[3].

Both techniques aim to minimize information loss when omitting variables, but they use very different definitions of information loss. While the objective of VR is to minimize the reduction in data variability, ECM focuses on minimizing distortion in the distribution of efficiency scores. This paper describes what kinds of consequences the different definitions can have in the case context.

The techniques should be considered as tools to help the management when deciding on inputs and outputs for efficiency evaluation. If the purpose is solely the correct classification of efficient and inefficient DMUs in a certain context, and the purpose is not variable selection, there may be other choices to reduce dimensionality.[4]

Even though the industry is less relevant in variable selection for DEA, it is worth noting that VR has been associated as an option to improve discrimination in bank and bank branch efficiency evaluation (Paradi & Zhu, 2013; Paradi, Yang, & Zhu, 2011). The reader may also be interested in the application of ECM in the evaluation of a Spanish bank branch network (Pastor, Lovell, & Tulkens, 2006).

The rest of this paper is organized as follows. Section 2 presents the two selected techniques used in this study. Section 3 introduces the empirical case, including the data used and the initial estimation without variable reduction. Section 4 describes how the techniques were applied to construct alternative specifications for efficiency estimation. Section 5 discusses the key findings. Section 6 summarizes the management feedback to the comparison. Section 7 makes concluding remarks.

2. Variable selection techniques

First, we briefly introduce a DEA estimator that provides the baseline for this study, and then we explain both techniques applied for variable selection in this comparison.

2.1. DEA estimator

The efficiency of a DMU can be evaluated using the generalized DEA estimator presented by Banker, Charnes, and Cooper (1984). We use the output-oriented multiplier form of DEA here. Assume there are n DMUs, each consuming m inputs and producing p outputs. The inputs are denoted by $\boldsymbol{x} \in \Re^m_+$ and outputs by $\boldsymbol{y} \in \Re^p_+$. Let $\mathbf{X} \in \Re^{m\times n}_+$ and $\mathbf{Y} \in \Re^{p\times n}_+$ be the matrices of observed inputs and outputs for the DMUs. DEA finds the most favorable input weights $\boldsymbol{\nu}$ and output weights $\boldsymbol{\mu}$ for the DMU under assessment, denoted by superscript r. Scalar u_o is associated with the returns to scale (RTS) assumption. If constant returns to scale (CRS) are assumed, u_o equals zero, and the estimator is called CCR (Charnes, Cooper, & Rhodes, 1978). If variable returns to scale (VRS) are assumed, u_o is free and the estimator is called BCC (Banker et al., 1984). The generalized DEA estimator is the following:

$$\min \boldsymbol{\nu}^T \mathbf{x}^r + u_o \tag{1}$$

s.t.

$$\boldsymbol{\mu}^T \mathbf{y}^r = 1$$
$$-\boldsymbol{\mu}^T \mathbf{Y} + \boldsymbol{\nu}^T \mathbf{X} + 1^T u_o \geq \mathbf{0}^T$$
$$\boldsymbol{\mu},\ \boldsymbol{\nu} \geq \mathbf{0}$$

The objective function provides an efficiency score for the DMU, the smaller the better because of the output orientation. If the score is one, and all variable weights $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ are strictly positive, the DMU is efficient.

The first constraint normalizes the sum of weighted outputs to one. This normalization has a convenient side-product, because the weighted value of the output $\mu_i y_i^r$ depicts directly the importance of the particular variable in the efficiency evaluation of the DMU. This weighted value is called the *virtual output*, and it is an important concept in this comparison. The second constraint limits the weights in such a way that the performance of the unit is scaled to the best performance among all units.

2.2. Variable reduction based on partial covariance

As variables in the efficiency evaluation are often highly correlated, an intuitive approach is to omit some of the highly correlated variables without a significant loss of information. Jenkins and Anderson (2003) criticized this approach because interrelations between variables are less obvious and cannot be determined directly from the correlation matrix. They wanted to introduce a systematic statistical procedure to help make decisions for variable selection. The authors have not given any specific name or abbreviation to their procedure. We call it *variable reduction based on partial covariance*, similarly to Adler and Yazhemsky (2010).

VR is a multivariate statistical procedure used to find a reduced set of initial input or output variables to be included in DEA. The

[3] Even though we focus on the outputs, both techniques can be applied to inputs or both. Jenkins and Anderson (2003) present an example where they apply their approach to the entire set when there are high correlations between some of the inputs and outputs.

[4] ECM performed quite well in most tested scenarios when compared with a number of selection and aggregation-based approaches (Nataraja & Johnson, 2011). VR has been compared with a technique that combines principal component analysis (PCA) and DEA (Adler & Yazhemsky, 2010). The latter outperformed VR in the simulations. Even though the combination PCA and DEA reduces dimensionality it is an aggregation approach rather than a tool for guiding the selection of variables.

Please cite this article as: J. Eskelinen, Comparison of variable selection techniques for data envelopment analysis in a retail bank, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2016.11.009