

---

## Variable elimination in nested DEA models: a statistical approach

---

Jirawan Jitthavech

School of Applied Statistics,  
National Institute of Development Administration,  
Bangkok 10240, Thailand  
Email: jirawan@as.nida.ac.th

**Abstract:** In this study, a new definition of a relevant variable in a DEA model is proposed for variable selection. The selection procedure is the conventional iterative backward elimination procedure with multiple statistical comparisons. The multiple tests of null hypothesis are reduced to a simple hypothesis test using either the binomial probability or the McNemar test with Bonferroni correction of significant level. From the results based on two simulation populations of moderately and lowly correlated input variables, the proposed procedure using either one of the suggested statistical tests can identify the relevant variables with high accuracy and eliminate the irrelevant variables effectively. In the dataset from a large scale experiment in the US public school education, the reduced model selected by the proposed procedure is shown to be the better approximation of the full model than the ones selected by the Pastor et al. method.

**Keywords:** data envelopment analysis; DEA; relevant variable;  $2 \times 2$  contingency table with structural zero; binomial probability test; McNemar test; Bonferroni correction.

**Reference** to this paper should be made as follows: Jitthavech, J. (2016) 'Variable elimination in nested DEA models: a statistical approach', *Int. J. Operational Research*, Vol. 27, No. 3, pp.389–410.

**Biographical notes:** Jirawan Jitthavech is currently working as an Associate Professor at School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand. She received her BSc (First Class Honour) from Prince of Songkla University in 1974, an MSc in Statistics (Honour) from National Institute of Development Administration in 1976 and a PhD in Statistics from University of Georgia, USA in 1984. Her research interests include statistical modelling, sampling design and analysis and operations research.

---

### 1 Introduction

Data envelopment analysis (DEA) is a non-parametric method for measuring a decision making unit's (DMU) relative efficiency in both multiple input and multiple output production settings. It has been applied to evaluate the performance of entities in various fields, such as, schools (Charnes et al., 1981), farms (Heidari et al., 2012; Zangeneh et al., 2010), financial institutions (Azadeh et al., 2010; Mostafa, 2009; Luo et al., 2012; Titko

et al., 2014), hotels (Pulina et al., 2010; Cheng et al., 2010; Manasakis et al., 2013; Shirouyehzad et al., 2014), etc. But DEA itself does not provide any guidelines for variable selection (Nataraja and Johnson, 2011; Madhanagopal and Chandrasekaran, 2014). The selection of input and output variables in a DEA model has been widely and extensively addressed in the DEA literature (Titko et al., 2014). At a given number of DMUs, the greater the number of variables a DEA model has, the more efficient DMUs will be (Jenkins and Anderson, 2003; Xie et al., 2014). DEA loses its power to discriminate between efficient and inefficient DMUs when excessive input and output variables are included in a DEA model (Xie et al., 2014; Nataraja and Johnson, 2011). Omission of relevant variables significantly affects the DEA efficiency scores (Nataraja and Johnson, 2011; Smith, 1997). Selection of an appropriate set of variables is, therefore, one of the most crucial tasks in DEA.

Conceptually, there are at least three approaches to variable selection in a DEA model. The first one is to select the input and output variables based on the recommendations of subject matter experts in the fields related to the problem under study (Golany and Roll, 1989; Katharakis et al., 2014). Instead of the experts' subjective recommendations, a theoretical framework which is available in the domain of application can be used to select an appropriate set of input and output variables for a DEA model (Joo et al., 2011). The second approach is statistical in nature (Banker, 1993, 1996; Lewin et al., 1982; Roll et al., 1989; Chilingerian, 1995; Lovell and Pastor, 1997; Pastor et al., 2002; Jenkins and Anderson, 2003). The third approach is a deterministic approach which iteratively eliminates the variables in a backward manner until the average of differences in efficiency scores of the same DMU between the reference model and reduced model is greater than a predetermined threshold (Wagner and Shimshak, 2007). In this paper, we focus only on the second approach.

In the earlier studies of the statistical approach, the correlation analysis between variables and efficiency scores has been used as a criterion for selecting the variables included in a DEA model (Lewin et al., 1982; Roll et al., 1989; Chilingerian, 1995). However, the variable selection process based on correlation coefficients may be insufficient and may also have to consider the causal relationship between the variable and the efficiency scores (Norman and Stoker, 1991; Sigala et al., 2004). Jenkins and Anderson (2003) improve the input variable selection for the nested model by using a partial covariance analysis. The first few principal components can be used to replace the original variables with minimum loss of information (Adler and Yazhensky, 2010). Adler and Yazhensky (2010) show via Monte Carlo simulations that the PCA-DEA methodology provides a more powerful discrimination tool than the variable reduction methodology using the partial covariance analysis with consistently more accurate results when the curse of dimensionality exists. Furthermore, the PCA-DEA methodology performs well with highly correlated inputs and even for small dataset (Nataraja and Johnson, 2011). The independent components which are generated from independent component analysis (ICA) are proposed to use as new input variables in a DEA model (Kao et al., 2011; Lin and Chiu, 2013). From the results using a simulated dataset and a real life hospital dataset, Kao et al. (2011) conclude that the ICA-DEA methodology provides a more discriminatory power than the PCA-DEA methodology. Bian (2012)

uses a subset of selected Schmidt variables via Gram-Schmidt process as new input variables in a DEA model to evaluate DMUs' efficiencies effectively when there are multiple correlations among original input variables. However, the variable transformation method generally does not reduce the number of variables but does reduce the dimensionality of the dataset. The disadvantage of the variable transformation method is the difficulty of interpreting the results in the subsequent analysis such as the measurement of the efficient level for each original variable, etc. Several statistical tests on the efficiency scores in the independent DEA models are proposed by Banker (1993, 1996). Other statistical methodologies are proposed to select the input and output variables for a DEA model, such as, the discriminant analysis using the external evaluation by three-level fractional factorial design (Morita and Avkiran, 2009), the genetic algorithm to select the 'best' subset of variables for a DEA model (Madhanagopal and Chandrasekaran, 2014).

In nested DEA models, forward selection and backward elimination of variables are the procedures suggested in the literature (Lovell and Pastor, 1997; Kittelson, 1993). Lovell and Pastor (1997) use a backward elimination procedure to determine the appropriate variables in a DEA model by deleting a group of variables because of their 'unreasonable' range of variation and further deleting another smaller subset of variables based on the same criterion until there is no possibility of deleting more variables without significantly affecting the efficiency scores. A statistical test of nested radial DEA models is proposed based on the efficiency contribution measure (Pastor et al. 2002) in which a candidate variable is considered to be relevant if more than  $p_0\%$  of DMUs have an associated efficiency score change greater than a threshold  $\bar{p}$ . A Bernoulli variable  $T_i$  is defined as follows.

$$T_i = \begin{cases} 1, & \text{if } \rho_i > \bar{p} \text{ for } i = 1, 2, \dots, n, \\ 0, & \text{otherwise} \end{cases}$$

where  $\rho_i$  is the ratio of the efficiency score of DMU  $i$  in the input-oriented full model of constant returns to scale to the efficiency score of DMU  $i$  in the reduced model,  $n$  is the number of DMUs and  $\bar{p} > 1$ . The test statistic  $T = \sum_{i=1}^n T_i$  is shown to be distributed as  $B(t; n - 1, p_0)$  (Pastor et al., 2002). The omitted variable in the reduced model is relevant and should be incorporated in the model if the null hypothesis  $H_0: p \leq p_0$ , is rejected at the significance level  $\alpha$  where  $p$  is the ratio  $T / n$ . Pastor et al. (2002) subjectively suggest using  $p_0 = 15\%$  and  $\bar{p} = 1.10$  but do not provide any guidelines for the specification of  $p_0$  and  $\bar{p}$ . Another statistical tool used in variable selection is the regression analysis which Ruggiero (2005) uses to identify relevant variables if the associated coefficients are statistically significant and have the proper signs.

All of the above-mentioned tests are based to some degree on the efficiency scores of all DMUs in a DEA model. Given a DEA model, the DMU efficiency scores are dependent on the input and output variables. Misspecification of the model significantly affects the DEA efficiency scores (Nataraja and Johnson, 2011; Smith, 1997). It should be noted that when a variable is excluded from the current model, the new set of efficient

DMUs becomes a subset of the original set of efficient DMUs (Lopez and Dula, 2008). If the new set of efficient DMUs after omitting a variable from the reference model remains unchanged and if the identification of efficient DMUs is the objective of the study, then that variable can be definitely discarded from the model. This conclusion suggests redefining a relevant variable based on the number of efficient DMUs in a DEA model.

The objectives of the paper are three folds: to propose a new definition of a relevant variable in a DEA model, to introduce two hypothesis tests for using in the proposed iterative procedure of backward variable elimination in nested DEA models and to sequentially select an appropriate DEA model for a particular application. The remainder of the paper is organised as follows. In the next section, a new definition of a relevant variable is proposed, and the backward variable elimination procedure is described. Two statistical methods for multiple comparisons in the iterative procedure are developed for using in the proposed procedure. In Section 3, two examples are used to illustrate variable elimination by the proposed procedure. One example utilises two simulation datasets and another uses a dataset published in the literature. The DEA model selected by the proposed procedure in the second example is compared with the ones selected by the Pastor et al. method (Pastor et al., 2002), which is widely used in DEA literature (Sirvent et al., 2005; Jones, 2006; Chen and Johnson, 2010). In Section 4, the model selections are evaluated by the Komolgorov-Smirnov test and Kullback-Leibler divergence. The final section presents the conclusions.

## **2 Proposed procedure**

The first part of the section covers the role of theories and empirical studies in the domain of the application or the management preference in variable selection in a DEA model and also a new definition of a relevant variable for using in the proposed procedure of backward variable elimination. The statistical tests used in the hypothesis testing in the iterative proposed procedure are presented in the last part of the section.

### *2.1 Variable classification and the proposed procedure*

Normally, a set of predetermined input and output variables in a DEA model of the studied topic are thoroughly scrutinised from the related theories, empirical studies or management preference. A number of predetermined output and input variables are definitely to be incorporated in the model from the strong support of the theoretical or empirical causal relationship between the efficiency score and the variables, or from the management preference (Sirvent et al., 2005). In the proposed procedure, these variables are to be grouped in a set  $S_1$ . The remaining predetermined variables, referred to as candidate variables, are grouped in a set  $S_2$ . For example, in a DEA study of an output from plants in an industry, the number of labours is definitely incorporated in the model and, therefore, is in  $S_1$  but various supporting tools, which are known to be useful for increasing some levels of productivity but cannot specifically be justified, are included in  $S_2$ . By classifying the original dataset into two subsets,  $S_1$  and  $S_2$ , the dimensionality of the original dataset is reduced to the dimensionality of the dataset  $S_2$ .

The DEA model described by all predetermined variables in  $S_1$  and  $S_2$  is the initial reference model, also referred to as the full model. When some variables in the full model are omitted, the model is referred to as a reduced model. Using the notations  $S_1$  and  $S_2$ , the input-oriented CCR DEA full and reduced models can be represented as

Full model:	Reduced model:
Min $\theta$	Min $\theta$
subject to	subject to
$\theta x_{i0} \geq \sum_{k=1}^n \lambda_k x_{ik}, i \in S_{11};$	$\theta x_{i0} \geq \sum_{k=1}^n \lambda_k x_{ik}, i \in S_{11};$
$\theta x_{i0} \geq \sum_{k=1}^n \lambda_k x_{ik}, i \in S_{21};$	$\theta x_{i0} \geq \sum_{k=1}^n \lambda_k x_{ik}, i \in \text{any subset of } S_{21};$
$y_{h0} \leq \sum_{k=1}^n \lambda_k y_{hk}, h \in S_{12};$	$y_{h0} \leq \sum_{k=1}^n \lambda_k y_{hk}, h \in S_{12};$
$y_{h0} \leq \sum_{k=1}^n \lambda_k y_{hk}, h \in S_{22};$	$y_{h0} \leq \sum_{k=1}^n \lambda_k y_{hk}, h \in \text{any subset of } S_{22};$
$\lambda_k \geq 0, k = 1, 2, \dots, n.$	$\lambda_k \geq 0, k = 1, 2, \dots, n.$

where  $\theta$  is the efficiency of DMU 0,  $S_{11}$  and  $S_{12}$  are respectively the set of predetermined input and output variable vector,  $\mathbf{x}_i$  and  $\mathbf{y}_i$  which are definitely incorporated in the model,  $S_{21}$  and  $S_{22}$  are the set of input and output candidate variables respectively,  $S_1 = S_{11} \cup S_{12}$  and  $S_2 = S_{21} \cup S_{22}$ . Other radial DEA models can be represented in a similar manner.

The inefficient DMUs in the full model are still inefficient in any reduced model but some efficient DMUs may become inefficient in a reduced model (Lopez and Dula, 2008). This leads to a new definition of a relevant variable in a DEA model:

A variable in  $S_2$  is said to be relevant if its omission from the model specification significantly decreases the number of efficient DMUs.

The new definition of a relevant variable is conceptually different from the conventional definitions which depend on some degree of the efficiency scores of DMUs (Lewin et al., 1982; Roll et al., 1989; Banker, 1996; Pastor et al., 2002; Jenkins and Anderson, 2003). Since the variables in  $S_1$  are not the candidate variables but definitely to be incorporated in the model, the new definition of a relevant variable is applied only among the candidate variables in  $S_2$ . The proposed definition is appropriate if the interest is focused on the identification of the efficient DMUs for benchmarking purpose. The statistical significance tests will be discussed in Section 2.2.

By definition, both input and output variables in  $S_1$  are definitely incorporated in the DEA model. Only the variables in  $S_2$  are candidates for variable selection in this study. Consider the case that  $S_{11}$  and  $S_{12}$  consists of at least one input and one output variable respectively. In this case, the proposed procedure is of backward elimination and can be described as follows:

- 1 Initialise  $t = 1$ , and start with the full DEA model described by  $v_{in}$  input variables in  $S_{11} \cup S_{21}$  and  $v_{out}$  output variables in  $S_{12} \cup S_{22}$ . Let  $k_{in}^t$  and  $k_{out}^t$  be respectively the numbers of members in  $S_{21}$  and  $S_{22}$  at the start of iteration  $t$ .
- 2 Classify DMUs in the full model as efficient and inefficient DMUs and the number of efficient DMUs is denoted by  $m_{full}$ .
- 3 If  $k_{in}^t > 0$  ( $k_{out}^t > 0$ ), create all possible  $k_{in}^t$  ( $k_{out}^t$ ) reduced models by excluding one member at a time in  $S_{21}$  ( $S_{22}$ ). Otherwise, do not create any reduced model. The total number of reduced models in iteration  $t$  is  $k_{in}^t + k_{out}^t$  denoted by  $I_t$ .
- 4 Classify DMUs in reduced models created in Step 3 as efficient and inefficient DMUs and let  $m_{it}$  be the number of efficient DMUs in reduced model  $i$ ,  $i = 1, 2, \dots, I_t$ .
- 5 Perform the statistical tests to determine the discarded variable (see details in Section 2.2). If none of variables in  $S_2$  can be discarded, then terminate the algorithm; otherwise, go to step 6.
- 6 Update  $t = t + 1$ . If the omitted variable in step 5 is an input variable, update  $k_{in}^t = k_{in}^{t-1} - 1$  and remove the omitted variable from  $S_{21}$ . Otherwise, update  $k_{out}^t = k_{out}^{t-1} - 1$  and remove the omitted variable from  $S_{22}$ .
- 7 If  $k_{in}^t = 0$  and  $k_{out}^t = 0$ , terminate the algorithm; otherwise, return to step 3.

When the algorithm is terminated either in step 5 or 7, the final model is the appropriate DEA model for the given dataset in a sense in accordance with the null hypothesis defined in step 5.

A DEA model must consist of at least one input and one output variable. Therefore, in the case that  $S_{11}(S_{12})$  is an empty set, the condition  $k_{in}^t(k_{out}^t) > 0$  in step 3 and the termination condition  $k_{in}^t(k_{out}^t) = 0$  in step 7 must be changed to  $k_{in}^t(k_{out}^t) > 1$  and  $k_{in}^t(k_{out}^t) = 1$  accordingly.

The proposed procedure with a new definition of a relevant variable can be used in variable elimination in a variety of DEA models: either CCR or BCC model with either input or output orientation. The proposed procedure also provides valuable information to the decision maker for each efficient DMU as to which variable has the most influence in maintaining that efficiency.

## 2.2 Statistical testing for multiple comparisons

In this subsection, two different statistical methods are developed for hypothesis testing in order to determine the discarded variable in the proposed iterative procedure of backward variable elimination in Section 2.1.

### 2.2.1 The binomial probability test

When a sample of a fixed size  $n$  is randomly selected from a DMU population with the proportion of efficient DMUs  $p_0$  and only two outcomes: efficient DMU and inefficient DMU are considered, it can be concluded that the number of efficient DMUs,  $X$ , in a sample is a binomial random variable (Milton and Arnold, 1990). The sample randomly selected from the population consists of all predetermined variables in  $S_1$  and  $S_2$ , referred to as a full model. Generally, the proportion  $p_0$  is unknown but can be estimated by  $\hat{p}_0 = m_{full} / n$  where  $m_{full}$  is the number of efficient DMU in the full model. Let  $B(x; n, p)$  denote the binomial probability distribution with parameters  $(n, p)$ . If  $X = m_{full}$ , then

$$P(X = m_{full}) = B\left(m_{full}; n, \frac{m_{full}}{n}\right). \quad (1)$$

The null hypothesis for variable elimination in a DEA model using the binomial probability test is

$$H_0 : \text{variable } i \text{ is relevant, } i = 1, 2, \dots, I_t, \quad (2)$$

against the alternative hypothesis

$$H_a : \text{variable } i \text{ is irrelevant, for at least one value of } i\text{'s}. \quad (3)$$

If the omitted variable is relevant, it can be interpreted from the new definition of a relevant variable in Section 2.1 that the difference,  $d_{it} = m_{full} - m_{it}$ , at iteration  $t$  in the proposed procedure is significantly different from zero. The difference  $d_{it}$  can be considered as a binomial random variable taking on a value,  $m_{full} - m_{it}$ , by chance (McNemar, 1947; Mosteller, 1952). Therefore, the conditional probability of discarding variable  $i$  in the full model as an irrelevant variable at iteration  $t$ , given  $X = m_{full}$ , is equal to

$$P(\text{discarding variable } i | X = m_{full}) = \sum_{x=0}^{d_{it}} B(x; m_{full} - 1, 0.5). \quad (4)$$

Let  $p_{it}$  be the joint probability of discarding variable  $i$  in the full model as an irrelevant variable at iteration  $t$  when the number of efficient DMU in the full model is equal to  $m_{full}$ .

$$\begin{aligned} p_{it} &= P(\text{discarding variable } i, X = m_{full}) \\ &= B\left(m_{full}; n, \frac{m_{full}}{n}\right) \sum_{x=0}^{d_{it}} B(x; m_{full} - 1, 0.5). \end{aligned} \quad (5)$$

In the proposed procedure of backward variable elimination, at iteration  $t$ , only hypothesis testing between the full model and all  $I_t$  reduced models is required to be performed. The multiple tests of  $H_0$  in (2) for  $1 \leq i \leq I_t$  with significance level  $\alpha$  are equivalent to a comparison between the full model and the reduced model  $i$ ,  $1 \leq i \leq I_t$ , with Bonferroni adjusted significance level  $\alpha_t = \alpha / I_t$  (Neter et al., 1996). Let  $p_i^* = \inf_{1 \leq i \leq I_t} p_{it}$ . If  $p_i^*$  is greater than  $\alpha_t$ , the proposed procedure of variable elimination is

terminated at iteration  $t$ . No more variables in  $S_2$  can be discarded at the significance level  $\alpha$ . Otherwise, the omitted variable  $i$  in  $S_2$  with  $p_{it} = p_i^*$  can be discarded.

2.2.2 *The McNemar test*

Consider a general correlated  $2 \times 2$  contingency table where the numbers of concordant pairs are denoted by  $a$  and  $d$  and of discordant pairs are denoted by  $b$  and  $c$ . The McNemar (1947) method is commonly used to test the marginal homogeneity of the two correlated populations widely encountered in clinical trials, biological studies, epidemiological studies, social anthropology, etc. (Toyota et al., 1999; Leisenring et al., 2000; Agresti, 2002; Tang and Tang, 2002; Durkalski et al., 2003; West and Hankin, 2008; Fagerland et al., 2013). The McNemar (1947) test statistic can be written as

$$\hat{Q} = \frac{(b - c)^2}{b + c}. \tag{6}$$

The statistic  $\hat{Q}$  has a  $\chi^2$  distribution with one degree of freedom.

The results of DEA analysis in the full model and in a reduced model at iteration  $t$  are obviously correlated and can be summarised as a correlated  $2 \times 2$  contingency table as shown in Table 1 where  $m_{full}$  and  $m_{it}$  are the sample number of efficient DMUs in the full model and reduced model  $i$  at iteration  $t$ . It should be noted that the set of efficient DMUs in any reduced model at iteration  $t$  is a subset of efficient DMUs in the full model, i.e.,  $m_{it} \leq m_{full}$  for  $1 \leq i \leq I_t$  (Lopez and Dula, 2008). For a given  $n$ , any of the possible outcomes in Table 1 is controlled by  $m_{full}$  and  $m_{it}$ .

**Table 1** The correlated  $2 \times 2$  contingency table in a sample for testing the hypothesis

		<i>Reduced model <math>i</math> at iteration <math>t</math></i>		<i>Total</i>
		<i>Number of inefficient DMUs</i>	<i>Number of efficient DMUs</i>	
Full model	Number of inefficient DMUs	$n - m_{full}$	0	$n - m_{full}$
	Number of efficient DMUs	$m_{full} - m_{it}$	$m_{it}$	$m_{full}$
Total		$n - m_{it}$	$m_{it}$	$n$

Because an inefficient DMU in the full model cannot become an efficient DMU in a reduced model, the discordant cell in the first row in Table 1 is always zero. It can be concluded that Table 1 is a correlated  $2 \times 2$  contingency table with a structural zero. The null hypothesis for testing marginal homogeneity in this case can be written as

$$H_0 : \frac{M_{full} - M_{it}}{N} = 0, i = 1, 2, \dots, I_t \tag{7}$$

And the alternative hypothesis can be written as

$$H_a : \frac{M_{full} - M_{it}}{N} > 0, \text{ for at least one value of } i\text{'s} . \tag{8}$$

where  $M_{full}$  and  $M_{it}$  are the number of efficient DMUs in the full model and in the reduced model in which the variable  $i$  in  $S_2$  is omitted at iteration  $t$  and  $N$  is the number of DMUs in the population. The test statistic  $\hat{Q}$  in (6) becomes

$$\hat{Q}_{it} = m_{full} - m_{it}, 1 \leq i \leq I_t. \quad (9)$$

The statistic  $\hat{Q}_{it}$  has a  $\chi^2$  distribution with one degree of freedom for a sufficiently large  $m_{full} > 12$  (Sheskin, 2007). It is trivial that the hypotheses (7) and (8) can be rewritten as

$$H_0 : M_{it} = M_{full}, i = 1, 2, \dots, I_t \quad (10)$$

against the alternative hypothesis

$$H_a : M_{it} < M_{full}, \text{ for at least one value of } i\text{'s}. \quad (11)$$

It can be concluded that the marginal homogeneity test in Table 1 leads to the hypothesis developed from the proposed new definition of a relevant variable in Section 2.1. Let  $Q_t^* = \inf_{1 \leq i \leq I_t} Q_{it}$  where  $Q_{it} = M_{full} - M_{it}$ ,  $1 \leq i \leq I_t$ . As in the case of the binomial probability test, it is sufficient to test

$$H_0 : Q_t^* = 0 \text{ against } H_a : Q_t^* > 0, \quad (12)$$

at the Bonferroni adjusted level of significance  $\alpha_t$ . If the null hypothesis (12) is not rejected in iteration  $t$ , the omitted variable  $j$  in the reduced model in which  $Q_{jt} = Q_t^*$  can be discarded from  $S_2$ . Otherwise, the iterative procedure of variable elimination is terminated.

Since the null hypotheses (2) and (10), used for variable elimination by the binomial probability test and the McNemar test respectively, are not identical but have the same objective, it may be possible that the reduced models from these two statistical methods are different.

The proposed procedure of backward variable elimination is similar to the stepwise procedure using the backward approach (Wagner and Shimshak, 2007). The variable elimination criterion in Wagner and Shimshak's iterative procedure is the simple comparison between the selected index and the given threshold and the reference model is changed iteratively. The suggested criteria include the minimum of the average differences in efficiency scores exceeds some maximum level, the change in any one efficiency score exceeds some maximum level, or the number of efficient DMUs falls below some minimum number. But, the variable elimination criterion in an iteration of the proposed procedure is invariably based on the number of efficient DMUs in the full model which is the reference model only in the first iteration in Wagner and Shimshak's procedure. Furthermore, the hypothesis significance test instead of simple comparison is also suggested in the proposed procedure.

### 3 Numerical results

In this section, two examples are used for numerical illustration of variable elimination in a nested DEA model by the proposed procedure using two statistical tests in Section 2. The data in the first example are generated by simulation. The second example uses the data from program follow through (PFT), a large-scale social experiment in US public school education (Charnes et al., 1981). In the second example, the variables in the DEA models selected by the proposed procedure using two statistical tests are compared with the variables in the DEA model selected by the Pastor et al. method.

Without any loss of generality, the model in the simulation is assumed to have only one output. The output  $y$  of a DMU is generated from the following Cobb-Douglas convex production function of constant returns to scale

$$y = x_1^{0.3}x_2^{0.2}x_3^{0.4}x_4^{0.1}, \tag{13}$$

where  $x_1, x_2, x_3$  and  $x_4$  are the relevant inputs with the corresponding elasticity 0.3, 0.2, 0.4 and 0.1, respectively. The values of  $x_1, x_2, x_3$  and  $x_4$  are generated from the multivariate normal distribution  $N(\boldsymbol{\mu}, \Sigma)$  where the expectations and variances of  $x_1, x_2, x_3$  and  $x_4$  are given by  $\boldsymbol{\mu}' = [4 \ 5 \ 6 \ 3]$  and 16.0, 25.0, 36.0 and 9.0, respectively. Two simulation populations, one with moderate and another one with low correlation, are generated with the correlation coefficients among the input variables as shown in Table 2.

**Table 2** Correlation coefficients among the input variables,  $x_1, x_2, x_3$  and  $x_4$

Correlation	$\rho_{12}$	$\rho_{13}$	$\rho_{14}$	$\rho_{23}$	$\rho_{24}$	$\rho_{34}$
Moderate	0.60	0.70	0.65	0.70	0.55	0.65
Low	0.20	0.30	0.25	0.30	0.12	0.20

The output variable in the simulation is given by the translog production function of (13) multiplying by an efficient score random function  $f(w)$

$$\ln y = f(w)(0.3 \ln x_1 + 0.2 \ln x_2 + 0.4 \ln x_3 + 0.1 \ln x_4), \tag{14}$$

where  $f(w)$  is given by

$$f(w) = \begin{cases} 1 & \text{if } w \leq 0 \\ 1-w & \text{otherwise,} \end{cases} \tag{15}$$

and  $w$  is the uniform random variable in the range. Equation (15) can be interpreted that the expected proportion of efficient DMUs in the simulation population is 0.2 and the efficiency score of the remaining inefficient DMUs is uniformly distributed in the range [0.6, 1). The simulation design of 20% efficient DMUs follows the school of thought in which 5%–30% of DMUs in the entire population are efficient (Banker et al., 1993; Smith, 1997; Bardhan et al., 1998; Andor and Hesse, 2011). Another two arbitrary irrelevant and independent input variables  $x_5$  and  $x_6$  generated from the uniform  $U(30, 90)$  and  $U(50, 90)$  distributions respectively are included to form the simulation population. In summary, the full model for DEA analysis in the simulation is described by one output variable,  $\ln y$ , four relevant variables,  $\ln x_i, 1 \leq i \leq 4$  and two irrelevant variables,  $\ln x_5$  and  $\ln x_6$ .

Two independent populations of size 50,000 DMUs are generated by (13) with moderate correlation among input relevant variables referred to as case 1 and low correlation among input relevant variables referred to as case 2. The populations in these two cases are transformed by (14) and then added another two irrelevant variables,  $\ln x_5$  and  $\ln x_6$ , to form the new simulation populations. Suppose that  $x_1$  and  $x_2$  are two input variables strongly supported by theoretical studies but  $x_3$  and  $x_4$  do not have such strong supports. Under the notations in the proposed procedure,  $v_{in} = 6$ ,  $v_{out} = 1$ ,  $k_{in}^1 = 4$ ,  $k_{out}^1 = 0$ ,  $S_{11} = \{\ln x_1, \ln x_2\}$ ,  $S_{12} = \{\ln y\}$ ,  $S_{21} = \{\ln x_3, \ln x_4, \ln x_5, \ln x_6\}$  and  $S_{22} = \emptyset$ . One hundred datasets of size 100 DMUs are randomly selected without replacement from the population in each case for testing the performance of the proposed procedure. The input-oriented CCR DEA model is used to evaluate the radial DMU efficiency scores in the simulation. In iteration 1, the number of candidate variables in  $S_2$  is equal to four or  $I_1 = 4$ . For a given dataset in the simulation, the maximum possible number of iterations in the proposed procedure is equal to four when all four candidates in  $S_{21}$  can be omitted. If the null hypothesis  $H_0$  in (2) is rejected for at least one value of  $i$ 's,  $1 \leq i \leq I_t$ , by the binomial probability test or the null hypothesis  $H_0$  in (12) is rejected by the McNemar test in iteration  $t$ , the number of candidate variables in iteration  $t + 1$  decreases to  $I_t - 1$ . The Bonferroni adjusted levels of significance in the test are correspondingly equivalent to  $\alpha = .01$  and  $.05$ .

After analysing the simulation results from 100 datasets in each case of correlation, we found that the proposed procedure for variable elimination using the statistical tests as described in Section 2.2 with Bonferroni adjusted levels of significance can effectively withhold the relevant variables and eliminate the irrelevant variables. Based on the binomial probability test, only all relevant variables are incorporated, at both  $\alpha = .01$  and  $.05$ , in DEA models from 100 datasets in both cases of correlation. The McNemar test selects the variables in the DEA model correctly 94% and 95 % for moderate and low correlation respectively at  $\alpha = .01$  and 98% and 100% for moderate and low correlation respectively at  $\alpha = .05$ . The rest are misspecified. It may be concluded via Monte Carlo simulations as shown in Table 3 that the binomial probability test performs slightly better than the McNemar test.

In the second example, the CCR DEA model for evaluating program and managerial efficiency uses three output variables and five input variables and is referred to as a full model. Using their expert subjectivity, Charnes et al. (1981) selected these three output variables from a set of 11 output variables,

- $y_1$  total reading score as measured by the metropolitan achievement test
- $y_2$  total mathematical score as measured by the metropolitan achievement test
- $y_3$  Cooper-Smith self-esteem inventory, intended as a measure of self-esteem.

And another five input variables from a set of 25 input variables,

- $x_1$  Education level of mother as measured in terms of percentage of high school graduates among female parents.
- $x_2$  highest occupation of a family member according to a pre-arranged rating scale
- $x_3$  parental visit index representing the number of visits to the school site

$x_4$  Parent counselling index calculated from data on time spent with child on school-related topics.

$x_5$  number of teachers in a given site.

**Table 3** Numbers of correctly, over-specified and misspecified models in case 1 and case 2 in the simulation

Statistical test	Family wise significance level	Case 1: moderate correlation: number of models, %			Case 2: low correlation number of models, %		
		Correctly specified	Over-specified	Misspecified	Correctly specified	Over-specified	Misspecified
Binomial probability	.01	100	0	0	100	0	0
	.05	100	0	0	100	0	0
McNemar	.01	94	0	6	95	0	5
	.05	98	0	2	100	0	0

The output and input data including details can be found in Charnes et al. (1981). In the following analysis, all input and output variables are considered to be candidates for variable selection. In other words, the set  $S_1$  is empty and all eight variables are in  $S_2$ . The Bonferroni adjusted levels of significance and the corresponding critical values at  $\alpha = .01$  and  $.05$  for hypothesis testing are shown in Table 4.

**Table 4** Bonferroni adjusted levels significance used in the PFT dataset.

Iteration	Number of comparisons	Bonferroni adjusted levels of significance	
		$\alpha = .01$	$\alpha = .05$
1	8	.0013	.0063
2	7	.0014	.0071
3	6	.0017	.0083
4	5	.0020	.0100
5	4	.0025	.0125
6	3	.0033	.0167
7	2	.0050	.0250

The  $p$ -values of the test statistics are tabulated in Tables 5 and 6 respectively. It is found that the proposed procedure is terminated at iteration 4 where all null hypotheses cannot be rejected when using either the test statistic  $p_{it}$  or the test statistic  $\hat{Q}_{it}$ . The DEA reduced model selected by the proposed procedure using either the binomial probability test or the McNemar test in the PFT dataset is the same model, referred as  $M_1$ , described by the remaining variables in  $S_2$  at iteration 3: two output variables,  $y_2$  and  $y_3$ , and three input variables,  $x_2$ ,  $x_4$  and  $x_5$ .

The Pastor et al. method is used to select the appropriate variables in a DEA model for the same PFT dataset in order to compare the results with the model  $M_1$  selected by the proposed procedure. The following four sets of parameters are utilised in the method:

$\bar{\rho}$	1.10	1.10	1.05	1.05
$p_0$	.15	.10	.15	.10

**Table 5** The joint probabilities  $p_{ii}$ 's in the reduced models in the second example

		Omitted variable(s)						
Iteration 1:	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$y_1$	$y_2$	$y_3$
$p_{i1}$	.0002	.0002	.0002	.0002	.0046	.0000	.0000	.0712
Iteration 2:	$y_1x_1$	$y_1x_2$	$y_1x_3$	$y_1x_4$	$y_1x_5$	$y_1y_2$	$y_1y_3$	
$p_{i2}$	.0013	.0046	.0013	.0046	.0125	.0046	.1065	
Iteration 3:	$y_1x_1x_2$	$y_1x_1x_3$	$y_1x_1x_4$	$y_1x_1x_5$	$y_1x_1y_2$	$y_1x_1y_3$		
$p_{i3}$	.0478	.0013	.0712	.0478	.0271	.1178		
Iteration 4:	$y_1x_1x_3x_2$	$y_1x_1x_3x_4$	$y_1x_1x_3x_5$	$y_1x_1x_3y_2$	$y_1x_1x_3y_3$			
$p_{i4}$	0.0920	0.1065	0.1065	0.1145	0.1178			

Note: The reduced model associated with the ital  $p_{ii}$  is the selected model.

**Table 6** The  $p$ -values of the test statistic  $\hat{Q}_{ii}$ 's in the reduced models in the second example

		Omitted variable(s)						
Iteration 1:	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$y_1$	$y_2$	$y_3$
$p_{i1}$	.1573	.1573	.1573	.1573	.0455	1.0000	.3173	.0047
Iteration 2:	$y_1x_1$	$y_1x_2$	$y_1x_3$	$y_1x_4$	$y_1x_5$	$y_1y_2$	$y_1y_3$	
$p_{i2}$	.0833	.0455	.0833	.0455	.0253	.0455	.0016	
Iteration 3:	$y_1x_1x_2$	$y_1x_1x_3$	$y_1x_1x_4$	$y_1x_1x_5$	$y_1x_1y_2$	$y_1x_1y_3$		
$p_{i3}$	.0082	.0253	.0047	.0082	.0143	.0005		
Iteration 4:	$y_1x_1x_3x_2$	$y_1x_1x_3x_4$	$y_1x_1x_3x_5$	$y_1x_1x_3y_2$	$y_1x_1x_3y_3$			
$p_{i4}$	.0009	.0016	.0016	.0016	.0002			

Note: The reduced model associated with the ital  $p$ -value is the selected model.

The first set of parameters,  $\bar{p} = 1.10$  and  $p_0 = .15$ , is suggested by Pastor et al. (2002) and the other three sets are the suggestion of Sirvent et al. (2005). Results of variable elimination are tabulated in Table 7 and can be summarised as follows. The Pastor et al. method is terminated in iteration 5 under the parameter set of  $\bar{p} = 1.05$  and  $p_0 = .10$  since the  $p$ -value of the minimum test statistic is equal to .0022 less than the corresponding Bonferroni adjusted levels of significance. The reduced model under such parameter set is the selected model at iteration 4 described by two output variables,  $y_2$  and  $y_3$ , and two input variables,  $x_2$  and  $x_3$ , referred as  $M_2$ . Similarly, under the remaining three sets of parameters, the Pastor et al. method is terminated in iteration 6 and the selected reduced model, referred to as  $M_3$ , is described by one output variable  $y_3$  and two input variables,  $x_2$  and  $x_3$ .

The DEA models selected by the proposed procedure using the binomial probability test and the McNemar test and by the Pastor et al. method are summarised in Table 8. The proposed procedure using the binomial probability test under the null hypothesis (2) or the McNemar test under the null hypothesis (10) selects the same model  $M_1$  but the model selected by the Pastor et al. method depends on the chosen values of  $\bar{p}$  and  $p_0$ . The Pastor et al. method selects the model  $M_2$  under  $\bar{p} = 1.05$  and  $p_0 = .10$  and the model  $M_3$  under the other three sets of parameters,  $\bar{p}$  and  $p_0$ . The variable elimination by the proposed procedure is more consistent than the one by the Pastor et al. method.

**Table 7** The  $p$ -values of the test statistics  $T_s$  of the reduced models in the second example

		<i>Omitted variable(s)</i>								
		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$y_1$	$y_2$	$y_3$	
<b>Iteration 1</b>										
$\bar{p} = 1.10; p_0 = .10$		.9936	.7201	.9936	.9936	.9936	.9936	.9936	.9936	.0000
$\bar{p} = 1.10; p_0 = .15$		.9996	.9428	.9996	.9996	.9996	.9996	.9996	.9996	.0000
$\bar{p} = 1.05; p_0 = .10$		.9936	.0187	.9597	.8711	.7201	.9936	.9936	.9936	.0000
$\bar{p} = 1.05; p_0 = .15$		.9996	.1740	.9961	.9817	.9428	.9996	.9996	.9996	.0000
<b>Iteration 2</b>		$y_1x_1$	$y_1x_2$	$y_1x_3$	$y_1x_4$	$y_1x_5$	$y_1y_2$	$y_1y_3$		
$\bar{p} = 1.10; p_0 = .10$		.9936	.7201	.9936	.9597	.9936	.9597	.0000		
$\bar{p} = 1.10; p_0 = .15$		.9996	.9428	.9996	.9961	.9996	.9961	.0000		
$\bar{p} = 1.05; p_0 = .10$		.9597	.0007	.5314	.7201	.7201	.5314	.0000		
$\bar{p} = 1.05; p_0 = .15$		.9961	.0218	.8656	.9428	.9428	.8656	.0000		
<b>Iteration 3</b>		$y_1x_1x_2$	$y_1x_1x_3$	$y_1x_1x_4$	$y_1x_1x_5$	$y_1x_1y_2$	$y_1x_1y_3$			
$\bar{p} = 1.10; p_0 = .10$		.0463	.9936	.9597	.9936	.8711	.0000			
$\bar{p} = 1.10; p_0 = .15$		.2880	.9996	.9961	.9996	.9817	.0000			
$\bar{p} = 1.05; p_0 = .10$		.0000	.3469	.5314	.7201	.5314	.0000			
$\bar{p} = 1.05; p_0 = .15$		.0012	.7456	.8656	.9428	.8656	.0000			

Note: The reduced models associated with the italicised  $p$ -value are the selected models.

**Table 7** The  $p$ -values of the test statistics  $T_s$  of the reduced models in the second example (continued)

	Omitted variable(s)					
	$y_1, x_1, x_2$	$y_1, x_1, x_3$	$y_1, x_1, x_4$	$y_1, x_1, x_2, x_3$	$y_1, x_1, x_2, x_4$	$y_1, x_1, x_3, x_4$
<b>Iteration 4</b>						
$\bar{p} = 1.10; p_0 = .10$	.0068	.9597	.9597	.9597	.8711	.0000
$\bar{p} = 1.10; p_0 = .15$	.0956	.9961	.9961	.9961	.9817	.0000
$\bar{p} = 1.05; p_0 = .10$	.0000	.0000	.2000	.0000	.0187	.0000
$\bar{p} = 1.05; p_0 = .15$	.0001	.0004	.5939	.0000	.1740	.0000
<b>Iteration 5</b>						
$\bar{p} = 1.10; p_0 = .10$	.0000	.0187	.7201	.0000	.0000	.0000
$\bar{p} = 1.10; p_0 = .15$	.0000	.1740	.9428	.0000	.0000	.0000
$\bar{p} = 1.05; p_0 = .10$	.0000	.0000	.0022	.0000	.0000	.0000
$\bar{p} = 1.05; p_0 = .15$	.0000	.0000	.0478	.0000	.0000	.0000
<b>Iteration 6</b>						
$\bar{p} = 1.10; p_0 = .10$	.0000	.0000	.0000	.0000	.0000	.0000
$\bar{p} = 1.10; p_0 = .15$	.0000	.0001	.0001	.0001	.0001	.0001
$\bar{p} = 1.05; p_0 = .10$	-	-	-	-	-	-
$\bar{p} = 1.05; p_0 = .15$	.0000	.0000	.0000	.0000	.0000	.0000

Note: The reduced models associated with the italicised  $p$ -value are the selected models.

**Table 8** The DEA models selected by the proposed procedure and the Pastor et al. method

Method	$\alpha$	Model	Selected DEA models	
			Output variables	Input variables
Proposed procedure using the binomial probability test	.01	M <sub>1</sub>	y <sub>2</sub> , y <sub>3</sub>	x <sub>2</sub> , x <sub>4</sub> , x <sub>5</sub>
	.05	M <sub>1</sub>	y <sub>2</sub> , y <sub>3</sub>	x <sub>2</sub> , x <sub>4</sub> , x <sub>5</sub>
Proposed procedure using the McNemar test	.01	M <sub>1</sub>	y <sub>2</sub> , y <sub>3</sub>	x <sub>2</sub> , x <sub>4</sub> , x <sub>5</sub>
	.05	M <sub>1</sub>	y <sub>2</sub> , y <sub>3</sub>	x <sub>2</sub> , x <sub>4</sub> , x <sub>5</sub>
Pastor et al. method with $\bar{p} = 1.10; p_0 = .10$	.01	M <sub>3</sub>	y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>
	.05	M <sub>3</sub>	y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>
Pastor et al. method with $\bar{p} = 1.10; p_0 = .15$	.01	M <sub>3</sub>	y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>
	.05	M <sub>3</sub>	y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>
Pastor et al. method with $\bar{p} = 1.05; p_0 = .10$	.01	M <sub>2</sub>	y <sub>2</sub> , y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>
	.05	M <sub>2</sub>	y <sub>2</sub> , y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>
Pastor et al. method with $\bar{p} = 1.05; p_0 = .15$	.01	M <sub>3</sub>	y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>
	.05	M <sub>3</sub>	y <sub>3</sub>	x <sub>2</sub> , x <sub>3</sub>

#### 4 Model selection

The selected reduced models in the second example using the dataset from a large scale experiment in US public school education are comprehensively evaluated by the Komolgorov-Smirnov test and the Kullback-Leibler divergence (Kullback and Leibler, 1951). In the Komolgorov-Smirnov test, the null hypothesis is tested that the efficiency scores in the full model and M<sub>j</sub>, j = 1, 2 and 3 are drawn from the same distribution. The Komolgorov-Smirnov statistic is used to measure the distance between the cumulative relative frequency distributions of DMU efficiency scores in the full model and the reduced model M<sub>j</sub> and can be written as

$$D_{M_j} = \sup_{0 \leq \theta \leq 1} |F_F(\theta) - F_{M_j}(\theta)| \tag{16}$$

where  $F_F(\theta)$  and  $F_{M_j}(\theta)$  are the cumulative relative frequency distributions of ungrouped DMU efficiency scores in the full model and the reduced model M<sub>j</sub> at efficiency score  $\theta$ . From the results, analysed by the PROC NPAR1WAY in SAS 9.3, it is found that  $D_{M_j} = 0.1429, 0.2245$  and  $0.2857$  with the  $p$ -values of two-sample Komolgorov-Smirnov statistic  $D_{M_j} = 0.6994, 0.1692$  and  $0.0366$  for  $j = 1, 2$  and  $3$  respectively. The null hypothesis is not rejected at  $\alpha = .01$  in all three cases. However, the null hypothesis is not rejected in the cases of M<sub>1</sub> and M<sub>2</sub> but is rejected in the case of M<sub>3</sub> at  $\alpha = .05$ . It can be concluded that either the reduced model selected by the proposed procedure, M<sub>1</sub>, or the reduced model M<sub>2</sub> selected by the Pastor et al. method using  $\bar{p} = 1.05$  and  $p_0 = .10$  may be used to approximate the full model.

Next, all three reduced models are further evaluated by the Kullback-Leibler (1951) divergence. By following the interpretation of the Kullback-Leibler divergence (Burnham and Anderson, 2002), the Kullback-Leibler divergence,  $D_{KL}(\text{full model} \parallel M_j)$ , is a

measure of the information lost when the relative frequency distribution of DMU efficiency scores in  $M_j$  is used to approximate the distribution of the full model and can be expressed as

$$D_{KL}(\text{full model} \parallel M_j) = \sum_{i=1}^k p_{iF} \log_2 \frac{p_{iF}}{p_{ij}}, \quad (17)$$

where  $k$  is the number of efficiency score classes,  $p_{iF}$  and  $p_{ij}$  are the relative frequencies of DMU in the full model and in the reduced model  $M_j$ , respectively, with efficiency score in class  $i$ . It can be seen that  $D_{KL}(\text{full model} \parallel M_j) = 0$  if and only if  $p_{iF} = p_{ij}$  for all  $i$ 's. Since the term  $p_{iF} \log_2(p_{iF} / p_{ij})$  is undefined for  $p_{ij} = 0$ , the DMU efficiency scores must be grouped into classes to avoid the zero frequency in any efficiency class in the reduced model. However, the data are grouped in an interval as small as possible to minimise the information loss as shown in Table 9. It can be seen that the minimum Kullback-Leibler divergence is  $D_{KL}(\text{full model} \parallel M_1)$  equal to 0.1185.

In summary, the null hypothesis is not rejected in cases of  $M_1$  and  $M_2$  by the Komolgorov-Smirnov test. But, it can be concluded that  $M_1$  is the better approximation of the full model than  $M_2$  by the Kullback-Leibler divergence.

**Table 9** The Kullback-Leibler divergences of  $M_j$  with respect to the full model,  $j = 1, 2$  and 3 in the second example

Efficiency score	Relative frequency distribution			
	Full model	$M_1$	$M_2$	$M_2$
< 0.83	0.0204	0.0408	0.0204	0.0408
0.83–0.84	0.0408	0.0408	0.0612	0.0408
0.85–0.86	0.0408	0.1020	0.0816	0.0816
0.87–0.88	0.0408	0.0612	0.0816	0.1020
0.89–0.90	0.1020	0.1429	0.0816	0.1224
0.91–0.92	0.1020	0.0612	0.1837	0.2041
0.93–0.94	0.0816	0.1224	0.1224	0.1224
0.95–0.96	0.1224	0.0408	0.0816	0.0612
0.97–0.98	0.0816	0.0816	0.0816	0.0612
0.99	0.0204	0.0204	0.0408	0.0816
1.00	0.3469	0.2857	0.1633	0.0816
$D_{KL}(\text{full model} \parallel M_j)$		0.1185	0.1536	0.3798

## 5 Conclusions

An iterative backward procedure for variable elimination is proposed using a new definition of a relevant variable based on the number of efficient DMUs and not based on the efficiency scores of the DMUs. Two statistical tests, the binomial probability test and the McNemar test are used for hypothesis testing in variable elimination in the proposed procedure. In the binomial test, the joint probability of discarding a variable in the model

when the number of efficient DMUs is  $m_{full}$  is derived. It is shown in the McNemar test that the null hypothesis for testing the marginal homogeneity is equivalent to the null hypothesis that the number of efficient DMUs in the full model and the reduced model are equal.

From the simulation of 100 replications in two datasets with moderate and low correlation among input variables, it can be concluded that the proposed iterative procedure can identify the relevant variables in the input-oriented CCR model with high accuracy and eliminate the irrelevant variables effectively. The proposed procedure using the binomial probability test performs slightly better than using the McNemar test.

In the dataset from a large scale experiment in US public school education, Charnes et al. (1981) use their expertise to carefully select three output variables from 11 variables and five input variables from 25 variables to form the DEA model referred to as the full model. The proposed procedure using the binomial test and the McNemar test selects the same model  $M_1$ . The Pastor et al. method with  $\bar{p} = 1.05$  and  $p_0 = .10$  selects the model  $M_2$  and with other three sets of parameters of  $\bar{p}$  and  $p_0$  selects the same model  $M_3$ . The Komolgorov-Smirnov test cannot reject the null hypothesis in the cases of  $M_1$  and  $M_2$  but can reject in the case of  $M_3$  at  $\alpha = .05$ . However, the Kullback-Leibler divergence suggests that  $M_1$  is the better approximation of the full model than  $M_2$ .

### 5.1 *Unique contributions*

The proposed procedure of backward variable elimination combines two approaches: theoretical or empirical studies or managerial preference and statistical. The predetermined variables are classified into two subsets:  $S_1$  for the variables definitely incorporated in the DEA model and  $S_2$  for the candidate variables to be eliminated by the proposed procedure. The classification reduces the dimensionality of the original dataset to the dimensionality of the dataset of variables in  $S_2$ . The proposed procedure eliminates variables in  $S_2$  from the model specification using a new definition of a relevant variable and two statistical tests: the binomial probability test and the McNemar test. It is shown analytically that these two tests can be used in testing the hypothesis for eliminating a variable from the DEA model specification. The objective of the proposed procedure with the new definition of a relevant variable is to eliminate those variables in  $S_2$  that have the least impact on the set of efficient DMUs which define the efficiency frontier. Lastly, the Komolgorov-Smirnov test and the Kullback-Leibler divergence are introduced for sequentially selecting the appropriate DEA model for a particular application.

### 5.2 *Managerial implications*

This article recognises the importance of the managerial participation in developing a decision making model. The managerial preference of variables may be used to reduce the dimensionality of the DEA dataset. The proposed procedure also provides valuable managerial information for each efficient DMU as to which variable has the most influence in maintaining that efficiency. The decision maker should be aware of the consistency of variable elimination by various methods as illustrated in the dataset from a large scale experiment in US public school education in Section 3. Finally, the loss of information when selecting the reduced model is provided for additional guidance in managerial consideration.

### 5.3 Limitations of the research and future research directions

The proposed procedure and the statistical tests developed in Section 2 can be used for variable elimination in any type of DEA models. But the conclusions in this article are based on the Monte Carlo simulations using only one input-oriented CCR model and one Cobb-Douglas convex production function of constant returns to scale in generating the dataset. For the future research, the proposed procedure should be evaluated in other types of DEA model under more production functions of constant returns to scale and variable returns to scale with different parameters. This would be able to generalise the conclusions. Some other advanced non-parametric tests for multiple comparisons (Garcia et al., 2010) and other statistical methods of model selection should be also investigated.

### Acknowledgements

The author would like to sincerely thank the anonymous reviewers for their constructive suggestions and comments which significantly improved the presentation of this paper.

### References

- Adler, N. and Yazhemsky, E. (2010) 'Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction', *European Journal of Operational Research*, Vol. 202, No. 1, pp.273–284.
- Agresti, A. (2002) *Categorical Data Analysis*, 2nd ed., John Wiley & Sons, New York, NY.
- Andor, M. and Hesse, F. (2011) *A Monte Carlo Simulation Comparing DEA, SFA and Two Simple Approaches to Combine Efficiency Estimates*, CAWM Discussion paper No. 51, University of Germany.
- Azadeh, A., Ghaderi, S.F., Mirjalili, M. and Moghaddam, M. (2010) 'A DEA approach for ranking and optimization of technical and management efficiency of a large bank based on financial indicators', *International Journal of Operational Research*, Vol. 9, No. 2, pp.160–187.
- Banker, R.D. (1993) 'Maximum likelihood, consistency and data envelopment analysis: a statistical foundation', *Management Science*, Vol. 39, No. 10, pp.1265–1273.
- Banker, R.D. (1996) 'Hypothesis tests using data envelopment analysis', *Journal of Productivity Analysis*, Vol. 7, No. 2, pp.139–159.
- Banker, R.D., Gadh, V.M. and Gorr, W.L. (1993) 'A Monte Carlo comparison of two production frontier estimation methods: corrected ordinary least squares and data envelopment analysis', *European Journal of Operational Research*, Vol. 67, No. 3, pp.332–343.
- Bardhan, I.R., Cooper, W.W. and Kumbhakar, S.C. (1998) 'A simulation study of joint uses of data envelopment analysis and statistical regression for production frontier estimation and efficiency evaluation', *Journal of Productivity Analysis*, Vol. 9, No. 3, pp.249–278.
- Bian, Y. (2012) 'A Gram-Schmidt process based approach for improving DEA discrimination in the presence of large dimensionality of data set', *Expert Systems with Applications*, Vol. 39, No. 3, pp.3793–3799.
- Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd ed., Springer, New York, NY.
- Charnes, A.W, Cooper, W. and Rhodes, E. (1981) 'Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through', *Management Science*, Vol. 27, No. 6, pp.668–697.

- Chen, W.C. and Johnson, A.L. (2010) 'The dynamics of performance space of major league baseball pitchers 1871–2006', *Annals of Operations Research*, Vol. 181, No. 1 pp.287–302.
- Cheng, H., Lu, Y. and Chung J. (2010) 'Assurance region context-dependent DEA with an application to Taiwanese hotel industry', *International Journal of Operational Research*, Vol. 8, No. 3, pp.292–312.
- Chilingerian, J.A. (1995) 'Evaluating physician efficiency in hospitals: a multivariable analysis of best practices', *European Journal of Operational Research*, Vol. 80, No. 3, pp.548–574.
- Durkalski, V.L., Palesch, Y.Y., Lipsitz, S., Philip, R.F. and Rust, P.F. (2003) 'Analysis of clustered matched-pair data', *Statistics in Medicine*, Vol. 22, No. 15, pp.2417–2428.
- Fagerland, M.W., Lyderson, S. and Laake, P. (2013) 'The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional', *BMC Medical Research Methodology*, Vol. 13, Article No. 91 [online] <http://www.biomedcentral.com/1471-2288/13/91> (accessed 18 October 2014).
- Garcia, S., Fernandez, A. Luengo, J. and Herrera, F. (2010) 'Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power', *Information Sciences*, Vol. 180, No. 10, pp.2044–2064.
- Golany, B. and Roll, Y. (1989) 'An application procedure for DEA', *Omega*, Vol. 17, No. 3, pp.237–250.
- Heidari, M.D., Omid, M. and Mohammadi, A. (2012) 'Measuring productive efficiency of horticultural greenhouse in Iran: a data envelopment analysis approach', *Expert Systems with Applications*, Vol. 39, No. 1, pp.1040–1045.
- Jenkins, L. and Anderson, M. (2003) 'A multivariate statistical approach to reducing the number of variables in data envelopment analysis', *European Journal of Operational Research*, Vol. 147, No. 1, pp.51–61.
- Jones, J. (2006) 'Data envelopment analysis and its application to the measurement of efficiency in higher education', *Economics of Education Review*, Vol. 25, No. 3, pp.273–288.
- Joo, S.J., Nixon, D. and Stoeberl, P.A. (2011) 'Benchmarking with data envelopment analysis: a return on asset perspective', *Benchmarking: An International Journal*, Vol. 18, No. 4, pp.529–542.
- Kao, L.J., Lu, C.J. and Chiu, C.C. (2011) 'Efficiency measurement using independent component analysis and data envelopment analysis', *European Journal of Operational Research*, Vol. 210, No. 2, pp.310–317.
- Katharakis, G., Katharaki, M. and Katostaras, T. (2014) 'An empirical study of comparing DEA and SFA methods to measure hospital units', *International Journal of Operational Research*, Vol. 21, No. 3, pp.341–364.
- Kittlelson, S.A.C. (1993) *Stepwise DEA: Choosing Variables for Measuring Technical Efficiency in Norwegian Electricity Distribution*, Memo No. 06/93, Department of Economics, University of Oslo, Norway.
- Kullback, S. and Leibler, R.A. (1951) 'On information and sufficiency', *Annals of American Statistics*, Vol. 22, No. 1, pp.79–86.
- Leisenring, W., Alonzo, T., Salmaso, L. and Pesarin, F. (2000) 'Comparisons of predictive values of binary medical diagnostic tests for paired designs', *Biometrics*, Vol. 56, No. 2, pp.345–351.
- Lewin, A.Y., Morey, R.C. and Cook, T.J. (1982) 'Evaluating the administrative efficiency of courts', *Omega*, Vol. 10, No. 4, pp.401–411.
- Lin, T.Y. and Chiu, S.H. (2013) 'Using independent component analysis and network DEA to improve bank performance evaluation', *Economic Modelling*, Vol. 32, pp.608–616.
- Lopez, F.J. and Dula, J.H. (2008) 'Adding and removing an attribute in a DEA model: theory and processing', *Journal of Operational Research Society*, Vol. 59, No. 12, pp.1674–1684.
- Lovell, C.A.K. and Pastor, J.T. (1997) 'Target setting: an application to a bank branch network', *European Journal of Operational Research*, Vol. 98, No. 2, pp.290–299.

- Luo, Y., Gongbing, B. and Liang, L. (2012) 'Input/output indicator selection for DEA efficiency evaluation: an empirical study of Chinese commercial banks', *Expert Systems with Applications*, Vol. 39, No. 1, pp.1118–1123.
- Madhanagopal, R. and Chandrasekaran, R. (2014) 'Selecting appropriate variables for DEA using genetic algorithm (GA) search procedure', *International Journal of Data Envelopment Analysis and Operations Research*, Vol. 1, No. 2, pp.28–33.
- Manasakis, C., Apostolakis, A. and Datsaris, G. (2013) 'Using data envelopment analysis to measure hotel efficiency in Crete', *International Journal of Contemporary Hospitality Management*, Vol. 25, No. 4, pp.510–535.
- McNemar, Q. (1947) 'Note on the sampling error of the difference between correlated proportions or percentages', *Psychometrika*, Vol. 12, No. 2, pp.153–157.
- Milton, J.S. and Arnold, J.C. (1990) *Introduction to Probability and Statistics*, 2nd ed., McGraw-Hill, New York, NY.
- Morita, H. and Avkiran, N.K. (2009) 'Selecting inputs and outputs in data envelopment analysis by designing statistical experiments', *Journal of the Operations Research Society of Japan*, Vol. 52, No. 2, pp.163–173.
- Mostafa, M.M. (2009) 'Modeling the efficiency of top Arab banks: a DEA neural network approach', *Expert Systems with Applications*, Vol. 36, No. 1, pp.309–320.
- Mosteller, F. (1952) 'Some statistical problems in measuring the subjective response to drugs', *Biometrics*, Vol. 8, No. 3, pp.220–226.
- Nataraja, N.R. and Johnson, A.L. (2011) 'Guidelines for using variable selection techniques in data envelopment analysis', *European Journal of Operational Research*, Vol. 215, No. 3, pp.662–669.
- Neter, L., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996) *Applied Linear Statistical Models*, 4th ed., McGraw-Hill, Boston, MA.
- Norman, M. and Stoker, B. (1991) *Data Envelopment Analysis: The Assessment of Performance*, Wiley, Chichester, England.
- Pastor, J.T., Ruiz, J.L., and Sirvent, I. (2002) 'A statistical test for nested radial DEA models', *Operations Research*, Vol. 50, No. 4, pp.728–735.
- Pulina, M., Detotto, C. and Paba, A. (2010) 'An investigation into the relationship between size and efficiency of the Italian hospitality sector: a window DEA approach', *European Journal of Operational Research*, Vol. 204, No. 3, pp.653–620.
- Roll, Y., Golany, B. and Seroussy, D. (1989) 'Measuring the efficiency of maintenance units in Israel air force', *European Journal of Operational Research*, Vol. 43, No. 2, pp.136–142.
- Ruggiero, J. (2005) 'Impact assessment of input omission on DEA', *International Journal of Information Technology and Decision Making*, Vol. 4, No. 3, pp.359–368.
- Sheskin, D.J. (2007) *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd ed., Chapman and Hall, Boca Raton, FL.
- Shirouyehzad, H., Lotfi, F.H. and Dabestani, R. (2014) 'A modified model for sensitivity analysis of inefficient DMUs in DEA: a case study in hotel industry', *International Journal of Operational Research*, Vol. 19, No. 2, pp.186–197.
- Sigala, M., Airey, D. Jones, P. and Lockwood, A. (2004) 'ICT paradox lost? A stepwise DEA methodology to evaluate technology investments in tourism settings', *Journal of Travel Research*, Vol. 43, No. 2, pp.180–192.
- Sirvent, I., Ruiz, J.L., Borrás, F. and Pastor, J.T. (2005) 'A Monte-Carlo evaluation of several tests for the selection of variables in DEA models', *International Journal of Information Technology and Decision Making*, Vol. 4, No. 3, pp.325–343.
- Smith, P. (1997) 'Model misspecification in data envelopment analysis', *Annals of Operations Research*, Vol. 73, pp.233–252.
- Tang, N.S. and Tang, M.L. (2002) 'Exact unconditional inference for risk ratio in a correlated table with structural zero', *Biometrics*, Vol. 58, No. 4, pp.972–980.

- Titko, J., Stankeviciene, J. and Lace, N. (2014) 'Measuring bank efficiency: DEA application', *Technological and Economic Development of Economy*, Vol. 20, No. 4, pp.739–757.
- Toyota, M., Kudo, K., Sumiya, M. and Kobori, O. (1999) 'High frequency of individuals with string reaction to tuberculosis among clinical trainees', *Japanese Journal of Infections Disease*, Vol. 52, No. 3, pp.128–129.
- Wagner, J.M. and Shimshak, D.G. (2007) 'Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives', *European Journal of Operational Research*, Vol. 180, No. 1, pp.57–67.
- West, L.J. and Hankin, R.K.S. (2008) 'Exact tests for two-way contingency tables with structural zeros', *Journal of Statistical Software*, Vol. 28, No. 1, pp.1–19.
- Xie, Q., Li, Y. and Jiang, A. (2014) 'Increasing the discriminatory power of DEA using Shannon's entropy', *Entropy*, Vol. 16, No. 3, pp.1571–1585.
- Zangeneh, M., Omid, M. and Akram, A. (2010) 'A comparative study on energy and cost analysis of potato production under different farming technologies in Hamadan province of Iran', *Energy*, Vol. 35, No. 7, pp.2927–2933.