Improving Discrimination in Data Envelopment Analysis: PCA-DEA or Variable Reduction

Nicole Adler1 and Ekaterina Yazhemsy

School of Business Administration, Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel

Within the data envelopment analysis context, problems of discrimination between efficient and inefficient decision-making units often arise, particularly if there are a relatively large number of variables with respect to observations. This paper applies Monte-Carlo simulation to generalize and compare two discrimination-improving methods; principal component analysis applied to data envelopment analysis (PCA-DEA) and variable reduction based on partial covariance (VR). Performance criteria are based on the percentage of observations incorrectly classified; efficient decision-making units mistakenly defined as inefficient and inefficient units defined as efficient. A trade-off was observed with both methods improving discrimination by reducing the probability of the latter error at the expense of a small increase in the probability of the former error. A comparison of the methodologies demonstrates that PCA-DEA provides a more powerful tool than VR with consistently more accurate results. PCA-DEA is applied to all basic DEA models and guidelines for its application are presented in order to minimize misclassification and prove particularly useful when analyzing relatively small datasets, removing the need for additional preference information.

Keywords: Data Envelopment Analysis, Principal Component Analysis, discrimination, simulation

1 Corresponding author. e-mail: msnic@huji.ac.il. Address: School of Business Administration, Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel. Telephone: +972-2-5883449.
The aim of this research is to understand the extent of the misclassification of decision making units (DMUs) as efficient in the data envelopment analysis (DEA) context and the degree to which it may be possible to correct this error. Two methodologies suggested in the literature as potential paths for improving discriminatory power without requiring additional preferential information are evaluated, namely principal component analysis combined with DEA (PCA-DEA) and variable reduction (VR) based on a partial covariance analysis. Secondary aims include the application of principal component analysis to all basic DEA models and the determination of the most effective way of implementing the preferable framework. Adler and Golany (2001, 2002) suggested using principal components, a methodology that produces uncorrelated linear combinations of original inputs and outputs, to improve discrimination in DEA with minimal loss of information. This approach assumes that separation of variables representing similar themes, such as quality or environmental measures, and the removal of principal components with little or no explanatory power, improves the categorization of efficient and inefficient DMUs. Jenkins and Anderson (2003) subsequently suggested a different statistical methodology in order to identify complete variables that could be omitted from the analysis whilst minimizing information reduction. They concluded that omitting even highly correlated variables could have a major influence on the computed efficiency scores, as argued in Dyson et al. (2001), hence an analysis of simple correlation is insufficient in identifying unimportant variables. Consequently, Jenkins and Anderson (2003) promulgate the use of partial covariance analysis to choose a subset of variables that provide the majority of information contained within the original data matrices.

While Adler and Golany (2001, 2002) and Jenkins and Anderson (2003) applied their methods to datasets published in the DEA literature, this study uses a simulation technique to generalize the comparison between the two approaches. A Monte Carlo simulation is used to generate a large number of DMUs, based on various production functions, inefficiency distributions, correlation between variables and sample sizes. In addition, to ensure that the conclusions are as general as possible, various forms of misspecification of the DEA models are also analyzed. The results of the various DEA, PCA-DEA and VR approaches are compared to the ‘true’ efficiency scores. Two potentially incorrect predictions of efficiency are designated, namely efficient decision-making units defined as inefficient (under-estimation) and inefficient DMUs defined as efficient (over-estimation). Furthermore, a rule-of-thumb is introduced, which defines the percentage of retained
In this paper we have substantially extended the application of PCA to DEA by applying the idea to all basic DEA models (previously it was applied to the additive model alone (Adler and Golany (2001, 2002)). In this context, we discuss the units and translation invariance properties of each of the linear programs. In addition, we also discuss the potential repercussions of the application of PCA-DEA with respect to the local estimates of substitution, targets and efficient peers as a result of the reduction in dimensions. We demonstrate the consequences of information reduction on a simulated database in order to draw general conclusions as to the advantages and disadvantages of the methods being evaluated. Previous papers analyzed real datasets and therefore could not evaluate the effect of the data reduction with respect to the accuracy of the DEA categorization of efficient and inefficient DMUs. In addition, we discovered that DEA models including all salient variables and the correct returns-to-scale assumption, never incorrectly categorize inefficient DMUs. The variable reduction methodology was published after PCA-DEA, and it was suggested that VR is the preferable method. This paper demonstrates the opposite to be true, including a reduction in bootstrap bias and confidence intervals. A rule-of-thumb is introduced, which defines the percentage of retained information required to balance the trade-off between the two incorrect definitions of (in)efficiency. This parameter was tested over varying levels of correlation between variables, sample sizes, inefficiency distributions and production functions. In addition, to ensure that the conclusions are as general as possible, various forms of misspecification of the DEA models were analyzed and two very different experimental designs were programmed, all reaching similar conclusions.

The paper is organized as follows. Section 2 presents a literature review, develops various PCA-DEA models, describes the variable reduction technique and discusses the bootstrap method for confidence interval estimation. Section 3 describes a number of experimental designs and distributions that generate the simulated data subsequently utilized to compare the methods under
discussion. Section 4 describes the results and section 5 the conclusions and recommendations for implementing the selected approach.

2. The Data Envelopment Analysis Framework

DEA measures the relative efficiency of decision-making units with multiple inputs and outputs and assumes neither a specific functional form for the production function nor the inefficiency distribution, in contrast to parametric statistical approaches. Problems related to discrimination arise, for example, when there are a relatively large number of variables as compared to DMUs, which in extreme cases may cause the majority of observations to be defined as efficient. As shown in Section 3, this is generally due to a large number of inefficient units incorrectly classified as efficient, which is a direct result of the weak assumptions of the DEA framework. Kneip et al. (1998) and Simar and Wilson (2000) developed models to determine the statistical properties of the nonparametric estimators. In particular, they showed that the speed of convergence of DEA estimators relies on (1) the smoothness of the unknown frontier and (2) the number of inputs and outputs relative to the number of observations. If the number of variables is relatively large, estimators exhibit very low rates of convergence and the applied researcher will need large quantities of data in order to avoid substantial variance and very wide confidence interval estimates. To avoid the discrimination problem, Simar and Wilson (2000) suggested that the number of observations ought to increase exponentially with the addition of variables, but general statements on the number of observations required to achieve a given level of mean-square error are not possible since the exact convergence of the nonparametric estimators depends on unknown smoothing constants. According to the Simar and Wilson bootstrap results, even the index case with one input and one output requires at least 25 observations, and preferably more than 100 for the confidence intervals of the efficiency estimator to be almost exact. Unfortunately, large samples are generally not available in practice and researchers try to analyze relatively small multivariate datasets, hence the need for discrimination improving methodologies.

Banker (1996), Simar and Wilson (2001) and Pastor et al. (2002) propose statistical tests for measuring the relevance of inputs or outputs, as well as tests to consider potentially aggregating inputs or outputs (Kittelsen (1993)). Olesen and Petersen (1996) proposed a data aggregation procedure when the efficient frontier is not composed of complete facets, described as ill
conditioned. If there is a lack of full-dimensional efficient facets, they argue that at least one linear aggregation of inputs or outputs could be applied such that DEA-efficient DMUs either maintain an efficient score or are noted to be efficient in their subsequent two-stage procedure. Their stated aim of ensuring the existence of well defined marginal rates of substitution is very different to the aims of this current research. PCA-DEA and VR aim to rectify misclassified efficiency as a result of the analysis of small datasets, thus deliberately changing the DEA frontier.

Adler et al. (2002), Angula-Meza and Lins (2002) and Podinovski and Thanassoulis (2007) review various approaches to increasing discrimination between DMUs. The most widely published approaches that do not require additional information include super- and cross-efficiency. Under cross-efficiency (Sexton et al. (1986), Doyle and Green (1994)), the efficiency score of each DMU is calculated \( n \) times, using the optimal weights evaluated in each of the \( n \) linear programs (LPs). The scores are then averaged (median, minimum etc) to produce a single cross-efficient score (Green et al. (1996)). DMU's are thus both self- and peer-evaluated. If the weights of the LP are not unique, a goal programming technique can be applied to choose between the optimal solutions. The secondary goals are generally defined as either "aggressive" or "benevolent" (Sexton et al. (1986)). However, all the remaining information that may be drawn from a DEA approach is thus lost and some DMUs may be defined as efficient and inefficient simultaneously. The concept of super-efficiency (Andersen and Petersen (1993)) leads to a ranking of the efficient DMUs through the exclusion of the unit being scored from the LP and an analysis of the resulting change in the Pareto frontier. This methodology enables an extreme efficient unit to achieve an efficiency score greater than one, i.e. super-efficiency. The method is problematic because 'specialized' DMU's may achieve a high ranking, requiring bounds on the weights (Sueyoshi (1999)) and various super efficient DEA models have proven infeasible for some extreme efficient points, subsequently requiring assurance regions (Hashimoto (1997)). While both the super- and cross-efficiency approaches are additional procedures for ranking DMUs utilizing complete information, PCA-DEA and the variable reduction approach attempt to reduce the dimensionality of the model in order to improve discrimination.

---

2 Alternative approaches, such as weight restrictions or cone-ratio analysis (Thompson (1986) and Charnes et al. (1990)) require additional preferential information prior to their application.

3 Seiford and Zhu (1999) and Zhu (2001) provide necessary and sufficient conditions for infeasibility of the super-efficiency DEA measures, reveal the relationship between the infeasibility and returns to scale classification and suggest the application of the super-efficiency approach is most appropriate within a DEA sensitivity analysis.
2.1 Principal Component Analysis - Data Envelopment Analysis

Zhu (1998) suggested principal component analysis could be applied to ‘output divided by input’ ratios as a complementary approach to DEA. The idea of combining DEA and PCA methodologies to achieve dimension reduction was developed independently by Ueda and Hoshiai (1997) and Adler and Golany (2001, 2002). In these papers it is suggested that the variables can be divided into groups, based on their logical composition with respect to the production process, and then replaced with principal components representing each group separately. Alternatively, PCA could be applied to the complete set of variables (inputs and/or outputs individually) in order to improve the discriminatory power of DEA by reducing the data to a few uncorrelated principal components, which generally describe 80-90% of the variance of the data. If most of the population variance can be attributed to the first few components, then they can replace the original variables with minimal loss of information. As stated in Johnson and Wichern (1982), let the random vector $X=[X_1,X_2,...,X_p]$ (in our case the original inputs or outputs chosen to be aggregated) possess the covariance matrix $V$ with eigenvalues $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_p \geq 0$ and normalized eigenvectors $l_1,l_2,\ldots,l_p$. Consider the linear combinations, where the superscript $t$ represents the transpose operator, as specified in equations (1). The new variables, commonly known as principal components, are weighted sums of the original data.

$$X_{PC_i} = l_i^t X = l_{i1} X_1 + l_{i2} X_2 + \ldots + l_{ip} X_p,$$

$$\text{Var}(X_{PC_i}) = l_i^t V l_i = \eta_i, \quad i=1,2,\ldots,p$$

$$\text{Cov}(X_{PC_i},X_{PC_j}) = l_i^t V l_j = 0, \quad i=1,2,\ldots,p, \quad k=1,2,\ldots,p, \quad i \neq k \quad (1)$$

The principal components, $X_{PC_1},X_{PC_2},\ldots,X_{PC_p}$, are the uncorrelated linear combinations ranked by their variances in descending order. In order to counter bias that might occur due to differences in the magnitude of the values of the original variables, the PC transformation should be applied to the correlation matrix of the normalized variables. Principal components are computed based solely on the correlation matrix and their development does not require a multivariate normal assumption. The complete set of principal components is as large as the original set of variables. $L_x$ is the matrix of all $l_i$ whose dimensions drop from $mxm$ to $hxm$, as PCs are dropped ($X_{PC}$ becomes an $h \times n$ matrix).
Following Adler and Golany (2001, 2002), we begin with adaptations of the additive DEA, constant returns-to-scale (CRS) case (Charnes et al. (1985)), which computes inefficiencies identified in both inputs and outputs simultaneously. The additive linear program is particularly useful because this formulation corresponds to the Pareto-Koopmans (mixed) definition of technical efficiency. It also possesses a translation invariance property under variable returns-to-scale (VRS) (Pastor, 1996) and data may be non-positive without the need for transformation. Another property that is generally considered crucial in performance analysis is units invariance. Lovell and Pastor (1995) introduced a normalized, weighted, additive LP that meets both translation and units invariance properties under VRS (units invariance only under CRS). Normalized, weighted, additive DEA utilizes the same constraints as basic additive but replaces the objective function of the primal additive formulation with the maximum sum of slacks weighted by the reciprocal sample standard deviations of the variables.

Principal components can be used to replace either all the inputs (outputs) simultaneously or, alternatively, groups of variables with a common theme, thus linear program (2) refers to both the original data and PCs in order to present a generalized formulation.

\[
\begin{align*}
\text{Max} & \quad w'_o s_o + w'_p s_p + w'_o \sigma_o + w'_p \sigma_p \\
\text{s.t.} & \quad Y_o \lambda = Y_o^a \\
& \quad -X_o \lambda - \sigma_o = -X_o^a \\
& \quad Y_p \lambda - L_s \sigma_p = Y_p^a \\
& \quad -X_p \lambda - L_s \sigma_p = -X_p^a \\
& \quad \sigma_p \geq 0 \\
& \quad s_p \geq 0 \\
& \quad s_o, \sigma_o, \lambda \geq 0
\end{align*}
\]

\[
\begin{align*}
\text{Min} & \quad V'_o X_o^a + V'_p X_p^a - U'_o Y_o - U'_p Y_p \\
\text{s.t.} & \quad V'_o X_o + V'_p X_p - U'_o Y_o - U'_p Y_p \geq 0 \\
& \quad V'_o \geq w'_o \\
& \quad U'_o \geq w'_p \\
& \quad V'_p L_s \geq w'_p \\
& \quad U'_p L_s \geq w'_p \\
& \quad V'_p \text{ and } U'_p \text{ are free}
\end{align*}
\]

\(4\) The optimal solution of the additive LPs is an efficiency score that measures the longest distance from the DMU being evaluated to the relative efficient production frontier. In other words, the objective function measures the maximum sum of absolute improvements measured as slacks. An observation is rated as relatively efficient if, and only if, there are no output shortfalls or resource wastage at the optimal solution. Translation invariance implies that the efficiency measure is independent of the linear translation of the input and output variables. Units invariance implies that the efficiency measure is independent of the units in which the input and output variables are measured.
where subscript ‘o’ (‘pc’) is the index of original (principle component) variables; $X_{pc}$ represents an $m \times n$ input matrix; $Y_{pc}$ an $r \times n$ output matrix; $X^a$ and $Y^a$ input and output column vectors for DMU$a$ respectively; $\lambda$ a column $n$-vector of DMU weights; $\sigma$ a column $m$-vector of input excess; $s$ a column $r$-vector of output slack variables; $w^d$ is a vector consisting of reciprocals of the sample standard deviations of the relevant variables. An additional constraint $e^t \lambda = 1$ can be added to (2a) corresponding to the VRS case (Banker et al. (1984)). (2b) is the dual version of (2a). As described in Adler and Golany (2002), by definition $V_{pc}^t X_{pc} \equiv V_{pc}^t L_x X$ where $V_{pc}^t$ represents a row vector of dual variables. Therefore $V_{pc}^t L_x$ equals the weight of the ‘original’ $X$ input matrix and the normalized, weighted, additive LP can be replaced by the algebraically equivalent linear program (2). The same is true for output matrix $Y$. The PCA-DEA formulation is exactly equivalent to the original linear program if and only if the PCs explain 100% of the variance in the original input and output matrices. Following the normalized, weighted, additive DEA model, each variable is divided by the corresponding standard deviation, the correlation matrix of standardized inputs and PCs are calculated and finally linear programs (2) are used to derive efficiency scores.

PCA-DEA may also be applied to the standard radial CRS and VRS DEA models (Charnes et al. (1978) and Banker et al. (1984) respectively). Using principal components in place of original data does not affect the properties of the DEA models. For instance, the input-oriented, variable returns-to-scale, radial estimators continue to be units invariant and translation invariant with respect to outputs (Pastor (1996)). Principal components (PCs) represent the selection of a new coordinate system obtained by rotating the original system with $x_1, \ldots, x_m$ as the coordinate axes. It is not the parallel translation of the coordinate system, thus PCA-DEA could be applied to all basic DEA models despite their lack of translation or units invariance.

The PCA-DEA formulation for the input oriented, CRS, radial linear program is presented in (3).

---

7 Regular DEA-solvers are not suitable for PCA-DEA, therefore we suggest utilizing free PCA-DEA software for discrimination reduction purposes (http://pluto.huji.ac.il/~msnic/PCADEA.htm). Matlab programs are also available which permits researchers to adapt the models as required.

8 Radial (technical) inefficiency means that all inputs can be simultaneously reduced without altering the proportions in which they are utilized, ignoring the presence of non-zero slacks.
\[
\begin{align*}
\text{Min} & \quad \theta \\
\text{s.t.} & \quad Y_o^a \lambda - s_o = Y_o^a \\
& \quad -X_o^a \lambda - \sigma_o = -\theta X_o^a \\
& \quad Y_{pc}^a \lambda - L_s \sigma_{pc} = Y_{pc}^a \\
& \quad -X_{pc}^a \lambda - L_s \sigma_{pc} = -\theta X_{pc}^a \\
& \quad \sigma_{pc} \geq 0 \\
& \quad s_{pc} \geq 0 \\
& \quad s_o, \sigma_o, \lambda \geq 0 
\end{align*}
\]

\[
\begin{align*}
\text{Max} & \quad U_o^Y a + U_{pc}^Y a \\
\text{s.t.} & \quad V_o^i X_o^a + V_{pc}^i X_{pc}^a = 1 \\
& \quad V_o^i X_o^a + V_{pc}^i X_{pc}^a - U_o^i Y_o - U_{pc}^i Y_{pc} \geq 0 \\
& \quad V_o \geq 0 \\
& \quad U_o \geq 0 \\
& \quad V_{pc} \geq 0 \\
& \quad U_{pc} \geq 0 \\
& \quad V_{pc} \text{ and } U_{pc} \text{ are free}
\end{align*}
\]
2.2 Multivariate Statistical Approach to Variable Reduction

Jenkins and Anderson (2003) introduced a systematic, multivariate, statistical approach to reduce the number of variables by omitting those providing the least information. The variables to be omitted are chosen based on partial correlation, in which the variance of an input or output around its mean value indicates the importance of a specific variable. If the value is constant, the variable will be incapable of distinguishing one unit from another, whereas a pronounced variation indicates an important influence. Jenkins and Anderson (2003) use partial correlation as a measure of information, instead of a simple correlation matrix (Friedman and Sinuany-Stern (1997)). However, partial correlation is based on the assumptions that the data is drawn from an approximately normal distribution and the conditional variance is homoscedastic. On the other hand, DEA is a non-parametric approach and it is unclear, particularly with a relatively small dataset, whether such conditions exist.

VR consists of the following steps:

i. Normalize the data in order to obtain zero mean and unit variance ensuring that all the variables are treated equally.

ii. Divide $m$ variables (inputs in our case) into two sets: $i=1,...,p$ representing the variables to be omitted, and $i=p+1,...,m$ the variables to be retained because they contain most of the information for all $m$ variables.

iii. Compute the partial variance-covariance matrix $V_{11,2}=V_{11}-V_{12}V_{22}^{-1}V_{21}$, where $V_{11}$ represents the variance-covariance matrix of variables $i=1,...,p$; $V_{22}$ represents the variance-covariance matrix of variables $i=p+1,...,m$; $V_{12}$ ($V_{21}$) represents the covariance matrix of variables $i=1,...,p$ and $i=p+1,...,m$ (and vice versa).

iv. Calculate the trace of $V_{11,2}$ which represents the size of the remaining variance of variables $i=1,...,p$, after conditioning on the retained variables $i=p+1,...,m$.

v. Repeat the procedure of labeling the $i=1,...,m$ variables under different partitions in order to achieve minimum variance in the first $p$ variables. That is, the number of omitted variables depends on the level of the remaining variance of variables that a user must specify exogenously.

vi. Apply DEA to the subset of original variables chosen to remain in the model.
Jenkins and Anderson (2003) applied VR to a number of published datasets and discussed the influence of the omission of variables that contain little additional information on the computed efficiency scores. They demonstrate that DEA results can vary greatly according to the variables chosen, despite the scientific or managerial justification for the inclusion or omission thereof. Hence, they advocate the use of partial covariance analysis in order to enable an objective selection of variables based on information to be considered in the subsequent analysis, leading to a more complete categorization of observations. It could be argued that VR is a private case of the PCA-DEA formulation because by removing principal components, dependent on the weights chosen, one or more variables may be dropped in their entirety.

In the following sections we will examine the performance of both radial and additive DEA, PCA-DEA and VR models, to determine under which circumstances each approach proves more accurate, in an attempt to define a reasonable implementation path. Furthermore, in order to compare radial DEA, PCA-DEA and VR models, in terms of the accuracy of the efficiency measure, we analyze the coverage probabilities of the confidence intervals.

2.3 Confidence Interval Estimation
Since efficiency is measured relative to an estimate of the frontier, estimates of DEA efficiency are subject to uncertainty due to sampling variation. Simar and Wilson (1998, 2000) proposed a bootstrapping methodology for analyzing the sampling variation and estimating confidence intervals of the radial DEA measures ($\hat{\theta}$). This research utilizes the homogeneity bootstrapping approach, presenting the input-oriented case. Simar and Wilson (1998) assumed that some underlying data generating process generates data points $(x, y)$ from the production possibility set\(^9\), observations are independent and identically distributed and the dataset is randomly sampled.

Bootstrapping occurs by repeatedly updating inputs $x^*$, as shown in equation (4), by applying DEA and comparing each DMU to the new reference set $(x^*, y^*)$.

$$x^* = x^\hat{\theta} / \theta^*, y^* = y$$

\(^9\) The data generation procedure assumes continuous density of inefficiency, without mass on the boundary.
where values $\theta^*$ are drawn from a smoothed kernel estimate of the marginal density of the original estimates of the relative efficiency ($\hat{\theta}$). The conditional density has bounded support over the interval (0,1] and is right-discontinuous at 1, hence naive bootstrapping (sampling with replacement) leads to inconsistent estimates. To address the boundary problem, Simar and Wilson (1998) draw pseudo data using the reflection method. The idea behind the bootstrap is to approximate the unknown distribution of $\hat{\theta} - \theta$, the difference between the original efficiency estimates and ‘true’ efficiency, through the distribution of $\hat{\theta}_b^* - \hat{\theta}$, the difference between the bootstrapped efficiency estimates and the original efficiency estimator, conditioned on the original data. From the empirical bootstrap distribution of the pseudo estimates, values for margins of error ($\hat{a}_a$ and $\hat{b}_a$) can be computed as presented in equation (5).

$$\Pr(-\hat{b}_a \leq \hat{\theta}_b^*(x_0, y_0) - \hat{\theta}(x_0, y_0) \leq -\hat{a}_a) \approx 1 - \alpha$$

(5)

The estimated one-sided \(\left(1 - \frac{\alpha}{2}\right)\)% confidence interval at the fixed point \((x_0, y_0)\) is then $[\hat{\theta}(x_0, y_0); \hat{\theta}(x_0, y_0) + \hat{b}_a]$. By definition, the bias of the DEA estimator is presented in equation (6) and the bootstrap bias estimate for the original estimator is presented in equation (7), where $B$ represents the number of bootstrap replications.

$$BIAS(\hat{\theta}(x_0, y_0)) = E(\hat{\theta}(x_0, y_0)) - \theta(x_0, y_0)$$

(6)

$$BIAS_b(\hat{\theta}(x_0, y_0)) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*(x_0, y_0) - \hat{\theta}(x_0, y_0)$$

(7)

The bias estimates highlight the inaccuracies caused by using a sample set of DMUs instead of the entire population, as frequently occurs in practice. This procedure was applied to several fixed points \((x_0, y_0)\) randomly sampled from the simulated data. The results of the bootstrapping procedure appear in Table 4 of the results section, after the experimental design is presented. The results clearly identify the reduced bias and increased confidence interval coverage in the
discrimination improving models, with PCA-DEA identified as systematically more effective then the VR method.

3. Design of the Experiment

The Monte Carlo simulation approach is used in this study to compare the accuracy of the different models. The ‘true’ efficiency will be calculated and then compared to the score derived from the basic DEA and the alternative LPs. This is the main advantage of using simulated data, as the ‘true’ values are known, which does not occur with a dataset collected from the real world. Two different experimental designs were programmed, the first based on Banker et al. (1993), Smith (1997) and Bardhan et al. (1998), in which inefficiencies are independently drawn for several inputs, there are “clouds” of data points surrounding the efficiency envelope (i.e. relatively small variance of inefficiency) and approximately 25% of the complete set of decision-making units of the entire population are defined as efficient (lie on the frontier). The second data generation design, based on Kneip et al. (1998) and Simar and Wilson (1998, 2000, 2001), assume single output inefficiency and no DMU is strictly efficient. Basic results from each of the experimental designs reach the same general conclusions despite the substantially different initial assumptions.

Initially, 10,000 positive observations of $X$ were randomly generated from a normal distribution with mean 10 and variance 1. The correlation of 0.4 or 0.8 (low and high levels) was applied using the Cholesky factorization in order to analyze the effects of correlation between input variables and emphasize empirical relevance, since reasonably high correlation is often found in real-world datasets. A single output and four inputs ($r = 1, m = 4$) were chosen for the first set of experiments. The single output is assumed for simplicity, permitting the employment of standard production functions to compute the output values. We assume homogeneity, namely that all DMUs operate under the same conditions and production process, hence the same measures of efficiency apply equally (Haas and Murphy (2003)). Table 1 presents the Cobb-Douglas production functions initially applied since they permit interaction among factor inputs and are relatively easy to manipulate mathematically. Parameters of functions were chosen to reflect a similar experimental design presented in Banker et al. (1993).
Table 1. Cobb-Douglas Production Functions

<table>
<thead>
<tr>
<th>Production function</th>
<th>Returns-to-scale property</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25}$</td>
<td>$\sum_{i=1}^{m} \alpha_i = 1$ (Constant Returns to Scale)</td>
</tr>
<tr>
<td>$y = x_1^{0.45} x_2^{0.3} x_3^{0.15} x_4^{0.1}$</td>
<td>$\sum_{i=1}^{m} \alpha_i = 0.7$ (Decreasing Returns to Scale)</td>
</tr>
<tr>
<td>$y = x_1^{0.35} x_2^{0.2} x_3^{0.1} x_4^{0.05}$</td>
<td>$\sum_{i=1}^{m} \alpha_i = 1.2$ (Increasing Returns to Scale)</td>
</tr>
<tr>
<td>$y = { x_1^{0.45} x_2^{0.3} x_3^{0.2} x_4^{0.1}$ for $\forall x_i &lt; 10$ $x_1^{0.4} x_2^{0.25} x_3^{0.15} x_4^{0.05}$ otherwise</td>
<td>Combined technology</td>
</tr>
</tbody>
</table>

It should be noted that the Cobb-Douglas function is restrictive in the properties it imposes upon the production structure, including a fixed returns-to-scale assumption and an elasticity of substitution equal to unity. In order to generalize the results of the experiment, a more flexible, homothetic, translog production function (Read and Thanassoulis (2000)) is also used to generate simulated data, as shown in equation (8). In this manner, we attempt to ensure that the conclusions drawn are not based on the production function assumed.

\[
\ln y = 0.25 \sum_{i=1}^{4} \ln x_i + 0.5(1.5 \sum_{i=1}^{4} \ln x_i^2 - \sum_{i=1}^{4} \sum_{j=1}^{4} \ln x_i \ln x_j) \quad \text{for } i \neq j \tag{8}
\]

The next step in producing a simulated data set is to introduce inefficiencies. We first describe the steps taken under the first experimental design and then those of the second data generation procedure.

3.1 Data Generation Procedure I

While the output values are calculated according to the production function, the input values are calculated using the expression $x_i e^{\tau_i}$, where $\tau_i$ represents a non-negative, input-specific inefficiency (Bardhan et al. (1998)). Inefficiencies $\tau_i$ for each input are independently drawn from an exponential
distribution with mean of 0.2231 or half normal distribution $HN(0,0.2796)$. Independence of input inefficiencies reflects specialization. In line with several simulation studies undertaken in the literature, such as Banker et al. (1988), approximately one quarter of the entire population of 10,000 firms were defined as 100% technically efficient, hence 25% of the randomly sampled observations from the entire population lie on the efficiency frontier, namely $\tau_i=0$. The models are applied to a small number of observations hundreds of times and the sub-samples are drawn randomly from the population, hence more points may lie on the frontier or less in each sub-sample tested. This results in the same mean inefficiency of 1.15 (or mean efficiency equivalent to 0.87) of a standard DEA model in the aggregate and are consistent with the empirical estimates reported in previous DEA studies (Banker et al. (1993)). The exponential and half-normal assumptions reflect the belief that as a result of competition, larger values of inefficiency are less likely and that the relatively small variances of inefficiency are highly likely, causing a cloud of DMUs near the frontier. The mass on the deterministic frontier reflects the assumption that within the population there exist a number of DMUs that are relatively efficient in many activities simultaneously. Since the production function parameters are unknown in practice, we only use this methodology to compare the usefulness of the PCA-DEA and VR methods. We also point out that the general trends are independent of the number of observations designated as relatively efficient.

### 3.2 Data Generation Procedure II

Under the second data generation process, following Simar and Wilson (2000), no probability mass is assumed along the frontier and a single inefficiency, $\tau_a$, was simulated for each DMU $a$, independently drawn from an inefficiency distribution e.g. $\tau_a \sim HN(0,1)$. Subsequently, the output values were calculated using function (9), where $e^{-\tau}$ represents a bound on the efficiency of [0,1].

$$y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25} e^{-\tau}$$

In this simulation design, a single inefficiency parameter was added on the output side, which permits the use of an output maximizing DEA. Under a VRS assumption, the reciprocal of an output-oriented radial estimator should be calculated. Alternatively, a radial estimation procedure under CRS may be applied which either minimizes inputs or maximizes outputs, since they produce the same efficiency score. In DEA we are searching for relative efficiency and by definition, at least
one DMU must be defined as relatively efficient. Given that all DMUs now possess some level of inefficiency, we assume that DMUs with a simulated $e^\tau$ greater than 0.9 should be deemed ‘relatively efficient’. This value was gradually increased to 0.99 in order to ensure that the general results presented in Section 4 are independent of the value assumed. We observed exactly the same general tendencies in incorrect efficiency classification as the results of the first experimental design. Furthermore, for data generation procedure II we analyzed the coverage probabilities of the confidence intervals. It should be noted that estimation of confidence intervals and tests of hypothesis in DEA are available only if the second experimental design is assumed (Simar and Wilson (1998, 2000, 2001)). The bootstrapping methodology deals with inaccuracies caused by sampling, but still suffers from the dimensionality issue and its statistical significance strongly depends on the sample size. Both PCA-DEA and VR are useful when the discrimination problem exists and statistical tests are not available, due to the relatively small sample size.

In all cases, after simulating an entire population of 10,000 DMU observations, the vectors were divided into smaller, more realistic subsets (sample sizes are 8, 10, 16, 20, 25 and 50 observations). We have chosen purposefully extreme examples in order to demonstrate the effect of information reduction but we also note that DEA has been applied to real data sets with very small samples and substantial numbers of variables, as published in the literature e.g. Hokkanen and Salminen (1997a, b), Friedman and Sinuany-Stern (1997), Viitala and Hänninen (1998), Colbert et al. (2000) and Kocher et al. (2006).

Finally, it should be noted that the correlation of 0.4 or 0.8 between variables holds for the entire population of 10,000 DMUs, i.e. before the introduction of inefficiencies. Inefficiencies, $\tau_i$, for each input are independently drawn after the Cholesky factorization, therefore the correlation within the simulated population will change, as it will within the sample subsets subsequently drawn. As a result, we may only refer to low and high correlation and the ‘true’ number of relatively efficient observations in each subset will vary too.
4. **Results**

The plots presented in this section illustrate the general findings of the simulation analyses. The title of each scatter plot includes information on the simulated production function, including the level of correlation between inputs, inefficiency distribution, sample size and returns-to-scale assumption. The value of the horizontal axis is the average percentage of incorrectly defined inefficient DMUs in each case and the value of the vertical axis is the average percentage of incorrectly defined efficient DMUs. For example, for the entire simulated population of 10,000 observations and a subsample size of 8 decision-making units, the average percentage is obtained from $10,000/8 = 1,250$ samples. Since the number of inefficient observations in the entire population is at least three-fold larger than the efficient units by definition, the probability of incorrectly defined inefficiency is significantly less than the probability of incorrectly defined efficiency, therefore the axes’ lengths are different. The left point in the pictures coincide with standard DEA without loss of information and then, each point to the right corresponds to the situation in which gradually more information is reduced. The curves show convex trend lines as information is removed from 100% down to 74% in 2-percent steps. The step was chosen arbitrarily for presentation purposes. The percentage of retained information is the common parameter for both methods and determines the number of PCs or variables retained in the subsequent DEA. In other words, at each point, the program chooses the number of PCs or variables to retain such that the percentage of information remaining is at least equal to the level set by the program (tables 3 and 4 explain this point in greater details). The slopes of the curves are interpreted as the rate of error reduction, hence the steeper the slope, the more effective the approach. The gap between the curves is interpreted as the difference between the methods, as the comparison between the results at specific points is less informative simply because the removal of an entire variable may have a different effect to that of dropping principal components.

Several figures have been chosen for illustrative purposes with the aim of demonstrating general results and conclusions. Figure 1 presents the erroneous classifications resulting from a Cobb-Douglas production function with equal weights and low covariance over four inputs for several sample sizes. Figure 2 presents VRS Cobb-Douglas functions, demonstrating the problems that arise.

---

\(^{10}\) Due to space limitations not all figures are presented, but the authors would be happy to send them to interested readers.
when applying DEA with the incorrect returns-to-scale assumption. Figure 3 demonstrates the results of a translog based production function, which clearly shows the same general conclusions over different sample sizes. Figure 4 demonstrates the effects of data omission with respect to a relatively important and unimportant variable and Figure 5 presents the effects of omission and inclusion of an extraneous input in comparison to the correct and complete analysis. Figures 1 to 5 demonstrate the results of the first experimental design, Figure 6 presents the performance of the PCA-DEA and VR methods based on radial CRS DEA utilizing the second experimental design described in Section 3.2. The likelihood of incorrectly defined inefficiency in this case is substantially lower than the results presented in Figure 1, hence the values on the x-axis are much smaller. In every graph it is clear that the two approaches reduce error, although beyond a certain level of information reduction, the incorrect classification of inefficiency begins to increase.

4.1 Information Retention

One of our goals is to find a rule-of-thumb for the PCA-DEA approach, specifying the percentage of retained information that provides the closest proximity to the accurate efficiency classification. In general, improving discrimination comes at a price, since it increases the probability of underestimation. It was found that in some cases, when the value of the optimal index dropped below a certain level, efficiency misclassification is no longer reduced and inefficiency misclassification increases substantially (Figure 1), therefore it may be helpful to provide guidelines concerning the optimal choice strategy. The rule-of-thumb was determined on each graph in the following manner:

1. Search for a point on the PCA-DEA curve where the incorrect classification of efficiency reaches its minimum.
2. If there are several such points, choose the one where the incorrect definition of inefficiency is minimized.

The solution to step 2 represents the rule-of-thumb value per specific simulated case. The simulation study suggests that the rule-of-thumb for the CRS (VRS) case ought to be equal to 80 (76)%%. In other words, the data may be reduced to a few uncorrelated principal components that describe at least 80 (76)% of the variance of the original data. This value is independent of the type of experimental design, i.e. the level of correlation between variables, the distribution of inefficiencies, the type of production function or the data generation process tested. Indeed, when working with a real dataset in which the underlying production function is unknown, the most
A cautious approach would be to drop PCs one-by-one until a reasonable level of discrimination is achieved or until you have reached the rule-of-thumb of at least 80 (76)%.

4.2 Results Drawing on Data Generation Procedure I

Figure 1 demonstrates clearly three basic results. First, the figure shows that PCA-DEA is strongly preferable to VR for all sample sizes and for all levels of information retained. PCA-DEA reduces incorrectly defined efficiency faster and creates incorrectly defined inefficiency more slowly than the VR methodology. Second, the rate of discrimination improvement by PCA-DEA is highest for smaller samples. The higher the ratio of variables to observations, the lower the level of discrimination and the more likely the basic DEA will yield incorrectly defined efficiency. Third, it has become clear that there are trade-offs between the two types of erroneous classification and most importantly, given the correct variables\textsuperscript{11} and returns-to-scale assumption, no efficient DMU is ever classified incorrectly under the basic DEA model. This is the reason that all lines presented in Figure 1 begin on the y-axis. Unfortunately, this is at the expense of a substantial number of inefficient DMUs being incorrectly defined as efficient. The degree of importance of the two error types is clearly context dependent but in cases where 50% or more of the DMUs are initially defined as efficient, the lack of differentiation ought to be considered as the results will provide little assistance to any subsequent analysis.

\textsuperscript{11} The definition of a “correct” set of variables refers to the idea that those included in the DEA model are both salient and measured accurately.
4.2.1 Returns-to-Scale Assumptions

In the top graph of Figure 2, the influence of an incorrect CRS assumption on under-estimation is demonstrated, particularly for highly correlated variables. This is the first instance of false inefficiency classification being produced by the basic DEA model, as identified by the fact that the left most point (100) represents the results of the weighted additive DEA and the alternative models with complete information. On average, greater under-estimation occurs with highly correlated inputs because of the similarity between DMUs. The CRS assumption has the effect of increasing the feasible region and enveloping the data less tightly than under the VRS assumption. Therefore, if variables are highly correlated and a CRS assumption is incorrectly assumed, an efficient DMU at the extreme points may be classified as inefficient. It should be noted that high correlation causes the opposite effect with respect to over-estimation, slightly reducing this mistake. Pedraja-Chaparro

The simulation was based on production function $y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25}$, relatively low correlation between inputs and inefficiency distribution $\exp(0.2231)$. In the DEA models, constant returns-to-scale was assumed.
et al. (1999) indeed reach the conclusion that merely counting variables in a DEA is an inadequate measure of the dimensionality of the model. In addition, an index of dimensionality ought to take into account the inter-correlation among variables.

Figure 2: Incorrect efficiency classification under varying correlation in inputs\textsuperscript{13}

The results of the simulation demonstrate that both the PCA-DEA and the VR models partially solve the problem of dimensionality by reducing the number of variables in the DEA analysis. As the number of DMUs lying on the efficiency frontier declines, the probability of mistaken efficiency decreases and the probability of mistaken inefficiency increases. Finally, when VRS is correctly assumed, the problem of discrimination is even more distinct, causing a relatively large percentage of mistakenly efficient DMUs, irrespective of the correlation between inputs. This can be seen in

\textsuperscript{13} The simulation was based on production function $y = x_1^{0.35} x_2^{0.2} x_3^{0.1} x_4^{0.05}$, sample size $= 8$ and inefficiency distribution $HN(0, 0.2796)$. Each point on the graph represents an average percentage drawn from $10,000/8 = 400$ samples. Relative low (high) correlation means 0.4 (0.8) correlation among variables in the population set (individual subsets may have higher or lower values).
the bottom graph of Figure 2, where the left most point represents the results of the weighted additive DEA and PCA-DEA and VR with complete information, demonstrating an average greater than 40% of incorrectly defined efficiency.

Figure 3 demonstrates the reduction in incorrect classifications as a function of sample size, assuming a translog production function. The top graph in Figure 3 demonstrates the influence of an incorrect CRS assumption, particularly for relatively large samples (16 observations) and the lower graph demonstrates the serious discrimination problem that occurs when VRS are correctly assumed for relatively small samples (8 observations). In the top graph, an incorrect CRS assumption causes an undesirable increase in under-estimation for relatively large samples, since the number of efficient observations in the larger sample is greater by definition, therefore the possibility of efficient units being defined as inefficient is greater. The introduction of the VRS constraint in the bottom graph of Figure 3 demonstrates the problem of sparsity bias for relatively small samples, when a DMU consuming the lowest level of a particular input is deemed efficient, simply because there are no peers with which to compare (Smith (1997) and Pedraja-Chaparro et al. (1999)). In the bottom graph, the proportion of mistakenly efficient units is substantial in the standard DEA, around 40% for very small samples (8 DMUs) and around 20% for larger sets (16 DMUs), but as we reduce the unnecessary principal components, the same error reduction appears as previously demonstrated with the Cobb-Douglas production functions.

The comparison of the two methodologies carried out in the study identifies PCA-DEA as a more powerful discrimination tool than VR. Furthermore, PCA-DEA results were found to be closer to the simulated efficiencies than those of VR and proved easier to navigate because the data reduction did not happen in large steps, as may occur when an entire variable is removed from the analysis, particularly under small samples and low correlation. In other cases, PCA-DEA and VR results were similar, although PCA-DEA was never found to produce less accurate results. At the same time, neither of the tested techniques ensured a complete ranking, rather a significant reduction in the set of efficient units of the original DEA. The combination of variable reduction followed by use of the PCA-DEA model proved unsuccessful due to an excessive loss of information.
Figure 3: Incorrect efficiency classification under simulated translog production functions and varying sample sizes

The simulation parameters included relatively high correlation between inputs and inefficiency distribution $\exp(0.2231)$. 

---

14 The simulation parameters included relatively high correlation between inputs and inefficiency distribution $\exp(0.2231)$. 

23
4.2.2 Levels of Error and Variable Choice

Since a process of discrimination improvement within DEA begins from the common point representing the original DEA result, it may be helpful to determine the strengths and limitations of the original, weighted, additive LP, as well as PCA-DEA. For this purpose various scenarios were developed altering the simulation and DEA returns-to-scale assumptions. Table 2 summarizes the intervals of error of the original DEA and PCA-DEA, when inputs are highly correlated, the number of decision-making units relative to the number of variables is small (1 output, 4 inputs and 8 DMUs), under various forms of the production functions (Table 1) and misspecification. The types of misspecification included (a) one of the inputs was omitted from the model, (b) an irrelevant input was incorporated into the model, i.e. this input had no effect on the computed output, (c) an incorrect assumption of returns-to-scale was made. Table 2 presents the trade-off between the two types of classification error, the influence of returns-to-scale assumptions on the results and the robustness of PCA-DEA. The same tendencies were found when two inputs were omitted from the model, two irrelevant inputs were incorporated into the model and the translog production function was assumed.

Table 2. Original DEA and PCA-DEA error intervals

<table>
<thead>
<tr>
<th></th>
<th>Constant returns-to-scale</th>
<th>Variable returns-to-scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original DEA</td>
<td>PCA-DEA (at least 80% info retained)</td>
</tr>
<tr>
<td>All relevant inputs included in DEA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrectly defined</td>
<td>inefficient</td>
<td>efficient</td>
</tr>
<tr>
<td></td>
<td>0-4.5 % (1)</td>
<td>6-11 %</td>
</tr>
<tr>
<td></td>
<td>16-20 % (2)</td>
<td>5-7 %</td>
</tr>
</tbody>
</table>

|                          | Omission of one input     |                          |                               |
| Incorrectly defined       | inefficient               | efficient                 | inefficient                   | efficient                 |
|                          | 1-7 %                     | 7.5-10 %                  | 0.25-1.5 %                   | 2.5-3.9 %                 |
|                          | 12-15 %                   | 6-7 %                     | 35-40 %                      | 23-26 %                   |

|                          | Inclusion of one extraneous input |                          |                               |
| Incorrectly defined       | inefficient               | efficient                 | inefficient                   | efficient                 |
|                          | 0-3.5 %                   | 7.5-9.5 %                 | 0 %                        | 3-3.5 %                   |
|                          | 28-34 %                   | 7-10 %                    | 56-57 %                     | 25-28 %                   |

For example, the lower bounds of cells (1) and (2) in Table 2 are determined by simulated CRS Cobb-Douglas production functions \( y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25} \) with a half normal inefficiency
distribution and the upper bounds are determined by simulated VRS Cobb-Douglas production functions $y = x_1^{0.35} x_2^{0.2} x_3^{0.1} x_4^{0.05}$ with an exponential inefficiency distribution and incorrect returns-to-scale assumptions in the subsequent DEA. The lower bound of cell (4) is determined by simulated Cobb-Douglas production functions with a half normal inefficiency distribution ($HN(0,0.2796)$) and the upper bound is determined by an exponential inefficiency distribution ($exp(0.2231)$). It is clear from Table 2 that the misclassification of efficiency can at times be very severe, hence it is an issue that must be addressed. Furthermore, when moving from constant to variable returns-to-scale, the efficiency misclassification more than doubles, probably because of the extreme specialists.

The analysis carried out in this study has highlighted some additional issues within the DEA context. First, if the correct returns-to-scale assumption is applied and all salient variables are included in the analysis, standard DEA never classified inefficiency incorrectly (cell (3) and lower bound of cell (1) in Table 2), however the false classification of efficiency can be quite substantial, particularly in the VRS case (cell (4) in Table 2). In extremely small multivariate datasets, 43-45% of decision-making units proved efficient, which means that subsequent analysis and ranking is problematic, hence the need for discrimination improving methodologies. Since it is problematic in practice to determine the returns-to-scale characteristic of a production process for small samples (Read and Thanassoulis (2000)), it may be more reasonable to include the VRS constraint, particularly when the inputs are highly correlated (as shown in Figures 2 and 3). According to the Galagedera and Silvapulle (2003) study, based on a sample with 200 DMUs, 3 inputs and 1 output, the VRS specification proved to be a more accurate alternative if the DEA model does not include all relevant variables. Furthermore, when the DEA model includes irrelevant variables, they show that the true returns-to-scale assumption is crucial because of the severe over-estimation of efficiency scores. They also discuss the adverse impact of misspecification in DEA on individual DMU efficiency scores, which is more serious when salient variables are omitted as compared to the inclusion of irrelevant ones. Figure 4 shows that the omission of salient variables are undesirable and may cause substantial levels of under-estimation, but this is dependent on the relative importance of the variable omitted. If the weight on the variable is relatively high ($x_1$), the percentage of mistakenly defined inefficiency is higher than that of the relatively less important variable omission ($x_4$). This type of error occurs relatively less under variable returns-to-scale assumptions rather than constant returns.
Figure 4: Incorrect efficiency classification with the omission of one input\textsuperscript{15}

Figure 5 compares the effect of omission versus the inclusion of an extraneous variable. With the latter, over-estimation is more likely. According to previous studies (Smith (1997) and Galagedera and Silvapulle (2003)) and as demonstrated in this research through the observed trade-off tendency, it may be preferable to include an excessive number of variables in an analysis, when the correct classification of efficient DMUs is more important than that of the inefficient ones. Clearly, the omission of relevant variables leads to under-estimation of the mean efficiency, while the inclusion of irrelevant variables leads to over-estimation.

\textsuperscript{15} The simulation parameters included a production function $y = x_1^{0.45} x_2^{0.3} x_3^{0.15} x_4^{0.1}$, relatively high correlation between inputs, sample size=8 and inefficiency distribution $\exp(0.2231)$. 

26
Figure 5: Incorrect efficiency classification given the omission or inclusion of one input

4.3 Results Drawing on Data Generation Procedure II

Figure 6, based on the second simulation undertaken (DGP II), again demonstrates the same tendencies as Figure 1 under substantially different assumptions. The difference between Figure 1 and Figure 6 in the axes’ lengths is a direct result of the modifications in the experimental design including the computation of a single inefficiency distribution, the definition of a “nearly efficient” unit and the assumption that there is no data cloud surrounding the efficient frontier. According to the continuous efficiency distribution without mass on the frontier, we assume that DMUs with a simulated $e^2$ greater than 0.9 should be deemed ‘relatively efficient’. Thus only 8% of the DMUs are ‘relatively efficient’ and 43% of samples with 10 DMUs do not include any efficient DMUs. As a result, the incorrect categorization of inefficiency on the x axis is extremely low. Consequently, if this data generation process is more likely to reflect the real process being tested, incorrectly defined inefficiency may be almost ignored, suggesting that the discrimination improving methodologies presented here are both very useful and have relatively little “downsides”.

The simulation parameters included a production function $y = x_1^{0.25}x_2^{0.25}x_3^{0.25}x_4^{0.25}$, relatively high correlation between inputs, sample size=8 and inefficiency distribution $exp(0.223I)$.
Table 3 presents the DEA results for larger samples with 50 DMUs and 8 inputs identifying the same error trends as previously demonstrated with graphs for smaller samples. In addition, Table 3 demonstrates the influence of the percentage of retained information (from 100% down to 74% in 2-percent steps) on the number of inputs (out of the original 8) remaining in the subsequent analysis. It is clear that the two approaches reduce errors, although beyond a certain level of information reduction, the incorrect classification of inefficiency begins to increase.

\[ y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25}, \] relatively low correlation between inputs, sample size=10 and inefficiency distribution $HN(0, 1)$. 

---

17 The simulation parameters included a production function $y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25}$, relatively low correlation between inputs, sample size=10 and inefficiency distribution $HN(0, 1)$. 
Table 3. Incorrect efficiency classification under varying dimensions reduction

<table>
<thead>
<tr>
<th>The production function</th>
<th>$y = x_1^{0.25}x_2^{0.2}(x_3x_4x_5x_6x_7x_8)^{0.09}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Returns-to-scale</th>
<th>CRS</th>
<th>VRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect classification</td>
<td>Incorrectly defined inefficient</td>
<td>Incorrectly defined efficient</td>
</tr>
<tr>
<td></td>
<td>PCA-DEA</td>
<td>VR</td>
</tr>
<tr>
<td>100%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>98%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>96%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>94%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>92%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>88%</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>86%</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>84%</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>82%</td>
<td>0.06</td>
<td>0.09</td>
</tr>
<tr>
<td>80%</td>
<td>0.08</td>
<td>0.31</td>
</tr>
<tr>
<td>78%</td>
<td>0.08</td>
<td>0.36</td>
</tr>
<tr>
<td>76%</td>
<td>0.10</td>
<td>0.53</td>
</tr>
<tr>
<td>74%</td>
<td>0.10</td>
<td>0.67</td>
</tr>
</tbody>
</table>

The number of samples containing $q$ inputs as a function of information retained

<table>
<thead>
<tr>
<th>Method</th>
<th>PCA-DEA</th>
<th>VR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q$</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q$</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>200</td>
<td>2</td>
<td>14</td>
<td>1</td>
<td>39</td>
<td>1</td>
<td>39</td>
<td>1</td>
</tr>
<tr>
<td>98%</td>
<td>200</td>
<td>8</td>
<td>153</td>
<td>1</td>
<td>36</td>
<td>1</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>96%</td>
<td>19</td>
<td>145</td>
<td>36</td>
<td>1</td>
<td>23</td>
<td>1</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>94%</td>
<td>9</td>
<td>89</td>
<td>90</td>
<td>12</td>
<td>21</td>
<td>1</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>92%</td>
<td>58</td>
<td>119</td>
<td>23</td>
<td>7</td>
<td>23</td>
<td>1</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>90%</td>
<td>9</td>
<td>89</td>
<td>90</td>
<td>7</td>
<td>21</td>
<td>1</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>88%</td>
<td>1</td>
<td>29</td>
<td>118</td>
<td>24</td>
<td>31</td>
<td>1</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>86%</td>
<td>9</td>
<td>72</td>
<td>97</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>84%</td>
<td>2</td>
<td>30</td>
<td>116</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>82%</td>
<td>12</td>
<td>81</td>
<td>107</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>80%</td>
<td>4</td>
<td>47</td>
<td>149</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>78%</td>
<td>2</td>
<td>21</td>
<td>177</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>76%</td>
<td>1</td>
<td>8</td>
<td>191</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>74%</td>
<td>5</td>
<td>195</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 4 presents the results of Monte Carlo experiments to measure the performance of the dimension-reduction methods using radial DEA estimators for the randomly chosen fixed points according to data generating procedure II over a variety of production functions and sample sizes, as presented in rows (1)-(3). The last part of the model name (row (4)) refers to the minimum percentage of retained information in the data and returns-to-scale assumed. Each Monte Carlo experiment involved 100 trials and each trial evaluated 2000 bootstrap replications. Confidence intervals were estimated for a specific inefficient unit, the randomly chosen fixed point. The coverage of one-sided 97.5% estimated confidence intervals were used as performance criteria for basic DEA, PCA-DEA and VR radial models (row (8)). Real bias and bootstrap bias estimates (rows (6) and (7)) were calculated as shown in Equations (6) and (7) in Section 2.3. Rows (10) and (11) present the ranges of the lower and the upper bounds for estimated 97.5 % one-sided confidence intervals over 100 trials. Row 12 presents information as to how many PCs or variables were retained based on the requirement to include at least 80% of the information in the case of constant returns-to-scale and 76% in the variable returns-to-scale case. As was expected, the bootstrap estimates of bias and the widths and ranges of the estimated confidence intervals decrease as sample size increases. It is also notable that the discrimination improving models reduce the level of bias that exists in the standard DEA.

When various forms of misspecification were purposely introduced, namely one input was omitted from the model, an irrelevant input was incorporated into the model and the incorrect assumption as to the type of returns-to-scale was made, the performance criteria for radial CRS-PCA and radial CRS-VR were quite similar. When VRS was assumed incorrectly, performance criteria indicated over-estimation for both radial VRS-PCA and radial VRS-VR models, and estimation of the efficiency score for small samples (10-20 observations) proved extremely problematic. Further examination of the accuracy of DEA and PCA-DEA for small samples would require the use of rank nonparametric statistics of relative efficiency instead of the absolute efficiency categorization. Such statistical tests are highly dependent on the proportion of tied observations, which is extremely problematic in the current context, therefore two types of errors were utilized in this study to act as universal performance criteria.
### Table 4. Confidence intervals for basic, PCA-DEA and VR radial CRS efficiency estimators

<table>
<thead>
<tr>
<th>The production function</th>
<th>Sample size (2)</th>
<th>True efficiency in the fixed point (3)</th>
<th>DEA method (4)</th>
<th>Average of relative efficiency estimators (5)</th>
<th>Average of real bias (6)</th>
<th>Bootstrap bias estimates (7)</th>
<th>Estimate of confidence intervals coverage (8)</th>
<th>Average width of estimated confidence intervals (9)</th>
<th>Range of lower limits of estimated confidence intervals (10)</th>
<th>Range of upper limits of estimated confidence intervals (11)</th>
<th>Number of trials in which $q$ inputs included in analysis (12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>($y = x_1^{0.45}x_2^{0.3}x_3^{0.15}x_4^{0.1}$)</td>
<td>10</td>
<td>0.4947</td>
<td>Basic CRS</td>
<td>0.6180</td>
<td>0.1234</td>
<td>0.1120</td>
<td>0.87</td>
<td>0.2672</td>
<td>0.5195</td>
<td>q=4</td>
<td>100</td>
</tr>
<tr>
<td>($y = x_1^{0.25}x_2^{0.25}x_3^{0.25}x_4^{0.25}$)</td>
<td>20</td>
<td>0.4947</td>
<td>Basic CRS</td>
<td>0.5734</td>
<td>0.0787</td>
<td>0.0874</td>
<td>0.92</td>
<td>0.1761</td>
<td>0.5039</td>
<td>q=3</td>
<td>45</td>
</tr>
<tr>
<td>($y = x_1^{0.3}x_2^{0.3}x_3^{0.3}x_4^{0.3}$)</td>
<td>25</td>
<td>0.5037</td>
<td>Basic CRS</td>
<td>0.5571</td>
<td>0.0624</td>
<td>0.0772</td>
<td>0.93</td>
<td>0.1619</td>
<td>0.4736</td>
<td>q=2</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCA CRS</td>
<td>0.5669</td>
<td>0.0722</td>
<td>0.0819</td>
<td>0.89</td>
<td>0.1701</td>
<td>0.5039</td>
<td>q=1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VR CRS</td>
<td>0.5734</td>
<td>0.0920</td>
<td>0.0869</td>
<td>0.90</td>
<td>0.1778</td>
<td>0.4736</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Basic VRS</td>
<td>0.5734</td>
<td>0.0660</td>
<td>0.0761</td>
<td>0.93</td>
<td>0.1546</td>
<td>0.4839</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCA VRS</td>
<td>0.5843</td>
<td>0.0806</td>
<td>0.0821</td>
<td>0.92</td>
<td>0.1669</td>
<td>0.4839</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VR VRS</td>
<td>0.5039</td>
<td>0.2616</td>
<td>0.0920</td>
<td>0.2987</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Summary and Conclusions

This research has analyzed two methodologies with the stated aim of improving the discriminatory power of DEA without the need for additional preferential information. Problems related to discrimination often arise when there are a relatively large number of variables as compared to decision-making units. In extreme cases, the majority of decision-making units may prove efficient, which means that subsequent analysis and ranking is problematic. Problems of discrimination have

---

18 The simulation assumed relatively low correlation between inputs.
been reduced to two types; efficient units that are defined incorrectly as inefficient and inefficient units that are deemed efficient. The latter occurs particularly frequently with small datasets under the assumption of variable returns-to-scale. This study used Monte-Carlo simulation to generalize a comparison between two approaches, namely PCA-DEA and variable reduction (VR). The Monte Carlo simulation generated a large dataset, from which small subsets were drawn and the DEA efficiency classification was subsequently compared to the ‘true’ value, permitting a computation of the size of the two error types. Furthermore, a bootstrapping approach in which the effects of the two approaches on the reduction of bias in the efficiency score estimates over the standard DEA linear programs is also presented.

The PCA-DEA formulation provided consistently more accurate results than that of the VR technique. The results show that PCA-DEA reduces over-estimation more quickly and produces under-estimation more slowly. The results are robust to changes in the initial data distribution, production function, inefficiency distribution and model misspecification. This proved true under different simulation assumptions, based on two experimental designs. The first data generation process defines 25% of the population’s observations as relatively efficient, whereas the second assumes all DMUs to be inefficient to some extent, with no mass on the frontier and continuous density in the inefficiencies.

In this paper we extend work already published in the field by applying PCA to all basic DEA models (previously it was applied to the additive model alone). In this context, we discuss the units and translation invariance properties of each of the linear programs. Guidelines for the application of PCA-DEA are presented based on the concept of a rule-of-thumb which considers the trade-off between the two types of erroneous classification. As the information reduction increases, the likelihood of incorrectly classifying inefficiency rises, therefore it is worthwhile gradually reducing the number of principal components remaining in the analysis and checking the subsequent results. As soon as the level of discrimination is sufficient, the information reduction process should be stopped. In general, The rule-of-thumb suggests that it is never worthwhile reducing the information below 80 (76)% of the variance of the original data under constant (variable) returns-to-scale assumptions, since at this point the marginal decrease in incorrectly defined efficient DMUs is at the expense of an increase in the inefficiency categorization. This value is independent of the level of
correlation between variables, of the distribution of inefficiencies or of the type of production function.

The analysis carried out in this study highlights some other issues within the DEA context. Since it is problematic to determine the returns-to-scale characteristics of a production process for relatively small samples in practice, it may be more reasonable to include the VRS constraint, particularly when the inputs are highly correlated. This choice, alongside the use of the PCA-DEA model, should result in reasonable levels of discrimination. With respect to the omission or addition of salient variables, according to previous studies (Smith (1997) and Galagedera and Silvapulle (2003)) and the observed trade-off tendency demonstrated in this research, it would appear to be preferable to include all potentially relevant variables for reasons of accuracy, particularly if the determination of relatively efficient decision-making units is more important than the correct determination of the inefficient ones e.g. for benchmarking purposes.

References

Adler N, Friedman L, Sinuany-Stern Z. Review of ranking methods in the data envelopment analysis context. European Journal of Operational Research 2002; 140 (2); 249-265
Andersen P, Petersen NC. A procedure for ranking efficient units in data envelopment analysis. Management science 1993; 39 (10); 1261-1294.

Banker RD, Charnes A, Cooper WW. Models for estimating technical and returns-to-scale efficiencies in DEA. Management Science 1984; 30; 1078-1092.


Bardhan IR, Cooper WW, Kumbhakar SC. A simulation study of joint uses of data envelopment analysis and statistical regressions for production function estimation and efficiency evaluation. Journal of Productivity Analysis 1998; 9; 249-278.


Colbert A, Levary RR, Shaner MC. Determining the relative efficiency of MBA programs using DEA. European Journal of Operational Research 2000; 125 (3); 656-669.


Friedman L, Sinuany-Stern Z. Scaling units via the canonical correlation analysis in the DEA context. European Journal of Operational Research 1997; 100 (3); 629-637.
Haas DA, Murphy FH. Compensating for non-homogeneity in decision-making units in data envelopment analysis. European Journal of Operational Research 2003; 144; 530-544.
Hokkanen J, Salminen P. Choosing a solid waste management system using multicriteria decision analysis. European Journal of Operational Research 1997a; 98 (1); 19–36.
Kocher MG, Luptacik M, Sutter M. Measuring productivity of research in economics: a cross-country study using DEA. Socio-Economic Planning Sciences 2006; 40; 314-332.
Olesen OB, Petersen NC. Indicators of Ill-conditioned Data Sets and Model Misspecification in Data Envelopment Analysis: An Extended Facet Approach. Management Science 1996; 42 (2); 205-219.

PCA-DEA program: http://pluto.huji.ac.il/~msnic/PCADEA.htm.


Ueda T, Hoshiai Y. Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. Journal of the Operational Research Society of Japan 1997; 40; 466-478.
