Improving discrimination in Data Envelopment Analysis: PCA-DEA versus Variable Reduction. Which method at what cost?

Nicole Adler[1] and Ekaterina Yazhemsky[2], Hebrew University of Jerusalem

Working Paper of the Hebrew University Business School

In the data envelopment analysis context, problems related to 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 presents a comparison of two discrimination-improving methods published in the literature that do not require additional preferential information; principal component analysis applied to data envelopment analysis (PCA-DEA) and variable reduction based on partial covariance (VR). A simulation based approach was used to generalize the comparison as to which methodology was preferable under which conditions. Performance criteria were based on the percentage of observations incorrectly classified; efficient decision-making units mistakenly defined as inefficient and inefficient units defined as efficient. According to the simulation results, 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. The comparison of the two methodologies showed that PCA-DEA provides a more powerful discrimination tool than VR with consistently more accurate results when the curse of dimensionality exists. Guidelines for the PCA-DEA user are presented based on a rule-of-thumb that aims to minimize both types of error.

Keywords: Data Envelopment Analysis, Principal Component Analysis, discrimination, ranking, 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.

[2] katy1@mscc.huji.ac.il

The aim of this research is to compare two methodologies suggested in the literature as potential paths for improving the discriminatory power within data envelopment analysis (DEA) without requiring additional preferential information, namely principal component analysis combined with DEA (PCA-DEA) and variable reduction (VR) based on a partial covariance analysis. A secondary aim is to determine the most effective way of implementing the preferable framework. Lack of discrimination often defined as the curse of dimensionality, means that a large number of DMU's are incorrectly classified as efficient due to overestimation. Adler and Golany (2001, 2002) suggested using principal components, a methodology that evaluates 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, will aid in the correct categorization of efficient and inefficient decision-making units (DMU's). 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 choosing the correct variables to be removed. 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 DMU's, 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 (Error type I) and inefficient DMU's defined as efficient (Error type II). Furthermore, a rule-of-thumb named the 'optimal index' is introduced, which

defines the percentage of retained information providing the greatest possible accuracy of the PCA-DEA model. It is shown that the optimal index is necessary for effective application of the aforementioned method in practice. To further test the models' capabilities, coverage probabilities of the confidence intervals for DEA, PCA-DEA and VR radial efficiency estimators were evaluated using Simar and Wilson's bootstrapping methodology (1998, 2000). The results further demonstrate the improvements that occur when applying PCA-DEA to estimate the efficiency score parameter.

The paper is organized as follows. Section 2 presents various DEA linear programs, the two discrimination-improving models within the DEA context and 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 is a non-parametric technique of frontier estimation that determines both the relative efficiency of a number of decision-making units (DMU's) and targets for their improvement. DMU's can represent any set of organizations or departments that perform fundamentally the same task with the same set of variables. 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 DMU's, which in extreme cases may cause the majority of observations to be defined as efficient. As shown through the use of simulation, this is generally due to a large number of type II errors (i.e. inefficient units incorrectly classified as efficient), which is a direct result of the weak assumptions of the DEA framework. One of the goals of this study is to determine the levels of two potentially incorrect predictions of efficiency that occur in the DEA environment, namely efficient decision-making units defined as inefficient (Error type I) and inefficient DMU's defined as efficient (Error type II). 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 a rather large quantity of data in order to avoid substantial variance and very wide confidence interval estimates. To avoid the curse of dimensionality, 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. Furthermore, Banker (1996), Simar and Wilson (2001) and Pastor et al. (2002) suggested statistical tests for measuring the relevance of inputs or outputs, as well as tests to consider potentially aggregating inputs or outputs. Unfortunately, large samples are generally not available in practice and researchers try to handle small multivariate datasets, hence the need for discrimination improving methodologies.

In the simulation analysis, 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 optimal solution of the linear programs 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, necessary for a DMU to be defined as relatively efficient. An observation is rated as relatively efficient if, and only if, there are no output shortfalls or resource wastage at the optimal solution. The additive linear program (LP) is particularly useful in our context because this formulation corresponds to the Pareto-Koopmans (mixed) definition of technical efficiency. It also possesses a translation invariance property[3] under VRS (Pastor, 1996) and data may be non-positive without the need for transformation.

[3] An efficiency measure is independent of the linear translation of the input and output variables.

Another property that is generally considered crucial in performance analysis is units invariance.[4] Lovell and Pastor (1995) introduced a normalized, weighted, additive LP that contains both translation and units invariance properties under VRS, and units invariance only under CRS. Therefore, it was chosen for the present study in place of the standard, additive linear program. 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 appropriate variables.

2.1 Principal Component Analysis - Data Envelopment Analysis

The idea of combining DEA and PCA methodologies 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 without much loss of information. As stated in Johnson and Wichern (1982), let the random vector $X=[X_1,X_2,\ldots,X_p]$ (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.

$$
\begin{aligned}
&X_{PC_i} = l_i^t X = l_{1i}X_1 + l_{2i}X_2 + \ldots + l_{pi}X_p \\
&Var(X_{PC_i}) = l_i^t V l_i = \eta_i \qquad , \quad i=1,2,\ldots,p \\
&Cov(X_{PC_i}, X_{PC_k}) = l_i^t V l_k = 0 \qquad , \quad i=1,2,\ldots,p,\ k=1,2,\ldots,p,\quad i \neq k
\end{aligned}
\tag{1}
$$

[4] An efficiency measure is independent of the units in which the input and output variables are measured.

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. It should be noted that PCA-DEA is based on correlation rather than on covariance due to the different variable measurement units[5]. 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). PCs can be used to replace either all the inputs (outputs) simultaneously or, alternatively, groups of variables with a common theme, such as a set of environmental or transportation variables, thus linear program (2) refers both to original data and PCs in order to develop a generalized program.

$$
\begin{aligned}
&\underset{s_0, s_{pc}, \sigma_0, \sigma_{pc}, \lambda}{Max} \; w^t_{Y_0} s_o + w^t_{Ypc} s_{pc} + w^t_{Xo} \sigma_o + w^t_{Xpc} \sigma_{pc} \\
&s.t. \quad Y_o \lambda - s_o = Y^a_o \\
&\quad -X_o \lambda - \sigma_o = -X^a_o \\
&\quad Y_{pc} \lambda - L_y \, s_{pc} = Y^a_{pc} \\
&\quad -X_{pc} \lambda - L_x \, \sigma_{pc} = -X^a_{pc} \\
&\quad \sigma_{pc} \geq 0 \\
&\quad s_{pc} \geq 0 \\
&\quad s_o, \, \sigma_o, \lambda \geq 0
\end{aligned}
\qquad (2a)
$$

$$
\begin{aligned}
&\underset{V_o, \, U_o, \, V_{pc}, \, U_{pc}}{Min} \; V^t_o X^a_o + V_{pc}{}^t X^a_{pc} - U^t_o Y^a_o - U^t_{pc} Y^a_{pc} \\
&s.t. \; V^t_o X_o + V^t_{pc} X_{pc} - U^t_o Y_o - U^t_{pc} Y_{pc} \geq 0 \\
&\quad V^t_o \geq w^t_{Xo} \\
&\quad U^t_o \geq w^t_{Yo} \\
&\quad V^t_{pc} L_x \geq w^t_{Xpc} \\
&\quad U^t_{pc} L_y \geq w^t_{Ypc} \\
&\quad V_{pc} \text{ and } U_{pc} \text{ are free}
\end{aligned}
\qquad (2b)
$$

where subscript 'o' ('pc') is the index of original (principle component) variables; X_{pc} represents an mxn input matrix; Y_{pc} an rxn output matrix; X^a and Y^a input and output column vectors for DMUa respectively; λ a column n-vector of DMU weights; σ a column m-vector of input excess; s a column r-vector of output slack variables; w^t 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 variable returns-to-scale (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

[5] Performing PCA on a standardized data matrix has the same effect as performing the analysis on the correlation matrix.

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 correlation in the original input and output matrices[6]. 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.

The translation invariance property is not crucial for PCA-DEA because geometrically 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). Therefore, PCA-DEA may also be applied to the standard radial CRS and VRS DEA (Charnes et al. (1978) and Banker et al. (1984) respectively). The PCA-DEA formulation for the input oriented, CRS, radial linear program is presented in (3).

$$
\begin{aligned}
&\underset{s_0,s_{pc},\sigma_0,\sigma_{pc},\lambda,\theta}{Min}\ \theta \\
&s.t.\quad Y_o\lambda - s_o = Y_o^a \\
&\quad -X_o\lambda - \sigma_o = -\theta X_o^a \\
&\quad Y_{pc}\lambda - L_y s_{pc} = Y_{pc}^a \\
&\quad -X_{pc}\lambda - L_x\sigma_{pc} = -\theta X_{pc}^a \\
&\quad \sigma_{pc} \geq 0 \\
&\quad s_{pc} \geq 0 \\
&\quad s_o,\ \sigma_o, \lambda \geq 0
\end{aligned}
\qquad (3a)
$$

$$
\begin{aligned}
&\underset{V_o,\ U_o,\ V_{pc},\ U_{pc}}{Max}\ U_o^t Y_o^a + U_{pc}^t Y_{pc}^a \\
&s.t.\ \ V_o^t X_o^a + V_{pc}^t X_{pc}^a = 1 \\
&\quad V_o^t X_o + V_{pc}^t X_{pc} - U_o^t Y_o - U_{pc}^t Y_{pc} \geq 0 \\
&\quad V_o \geq 0 \\
&\quad U_o \geq 0 \\
&\quad V_{pc}^t L_x \geq 0 \\
&\quad U_{pc}^t L_y \geq 0 \\
&\quad V_{pc} \text{ and } U_{pc} \text{ are free}
\end{aligned}
\qquad (3b)
$$

The disadvantage of PCA-DEA is that the data must be transformed and then, once results are obtained, it must be transformed back to the original form in order to interpret the results. In DEA the results obtained with respect to each DMU reflect its position within the "production possibility set" (PPS) relative to the efficient part of the boundary of the PPS. The imposition of weights restrictions in DEA will render parts of the efficient boundary of the PPS no longer efficient. Allen et al. (1997) showed that the interpretations of the inefficiency rating, the targets

[6] 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)

and the efficient peers change under weights restrictions. Indeed, the targets and the efficient peers obtained could reflect a substantial change in the current mix of input-output levels of the inefficient DMU's. A similar phenomena occurs under the PCA-DEA formulation (as a result of the free sign in PCA). However problems related to discrimination often arise and in extreme cases, the majority of DMU's may prove efficient, which means that there is a need for a trade-off between complete DEA information and a need to improve discrimination.

2.2 Multivariate Statistical Approach for Variable Reduction

Jenkins and Anderson (2003) introduced a systematic multivariate statistical approach to reduce the number of variables, omitting those providing the least information. The variables to be omitted are chosen based on a partial correlation technique, 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. DEA is a non-parametric approach and it is unclear, particularly with a 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: $i=1,\ldots,p$ representing the variables to be omitted, and $i=p+1,\ldots,m$ 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 $i=1,\ldots,p$ V_{22} represents the variance-covariance matrix of variables $i=p+1,\ldots,m$, V_{12} (V_{21}) represents the covariance matrix of variables $i=1,\ldots,p$ and $i=p+1,\ldots,m$ (and vice versa).

iv. Calculate the trace of $V_{11.2}$ which represents the size of the remaining variance of variables $i=1,\ldots,p$ after conditioning on the retained variables $i=p+1,\ldots,m$.

v. Repeat the procedure of labeling of the $i=1,\ldots,m$ 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 variables using their original measurements.

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 DEA, PCA-DEA and VR models, to determine under which circumstances each approach proves more accurate in terms of Error types I and II, in an attempt to define an optimal 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 Intervals 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[7], 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*)*.

$$x^* = x\hat{\theta} / \theta^*, y^* = y \tag{4}$$

where values θ^* 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).

$$\Pr(-\hat{b}_a \le \hat{\theta}_b^*(x_0, y_0) - \hat{\theta}(x_0, y_0) \le -\hat{a}_a) \approx 1 - \alpha \tag{5}$$

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.

$$BIAS(\hat{\theta}(x_0, y_0)) \equiv E(\hat{\theta}(x_0, y_0)) - \theta(x_0, y_0) \tag{6}$$

[7] The data generation procedure assumes continuous density of inefficiency, without mass on the boundary.

$$BIAS_B\left(\hat{\theta}(x_0, y_0)\right) \equiv \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*(x_0, y_0)) - \hat{\theta}(x_0, y_0) \quad (7)$$

The bias estimates highlight the inaccuracies caused by using a sample set of DMU's 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 Tables 3 and 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 proving to be 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 basic DEA and the alternative models. 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 very different experimental designs were programmed, the first school of thought is 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 decision-making units of the entire population are absolute efficient (lie on the frontier). The second school, based on Kneip et al. (1998) and Simar and Wilson (1998, 2000, 2001), assume single output inefficiency and that no DMU is strictly efficient. Basic results from each of the experimental designs reach the same general conclusions.

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

experiment. The single output is used for simplicity and in order to permit the use of standard production functions to compute the output values. We assume homogeneity, namely that all DMU's operate under the same conditions, using the same production process, hence the same measures of efficiency apply equally to all DMU's (Haas and Murphy (2003)). Table 1 presents the Cobb-Douglas production functions used initially because they permit interaction among factor inputs and are relatively easy to manipulate mathematically.

Table 1 here

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)) was 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.

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

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 from the production function, the input values are calculated using the expression $x_i e^{\tau_i}$, where τ_i represents a non-negative, input-specific inefficiency (Bardhan et al. (1998)). Inefficiencies τ_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 τ_i=0. 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 larger values of inefficiency are less likely and that the relatively small variances of inefficiency are highly likely, causing a cloud of DMU's near the frontier.

3.2 Data Generation Procedure II

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

$$y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25} e^{-\tau} \qquad (9)$$

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 DMU's now possess some level of inefficiency, we assume that DMU's 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. These assumptions were also tested for various forms of the inefficiency distributions (exp(0.2231) and HN(0,0.2796)) using radial DEA. The percentage of relatively efficient units in the original population varied based on the inefficiency distribution and cut-off assumption ranging from 38% under the exp(0.2231) distribution with $e^{-\tau} > 0.9$ to 0.9% under the HN(0,1) distribution with

$e^{-\tau} > 0.99$. However, we observed exactly the same general tendencies as the results of the first experimental design.

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 and 25 observations). It should be noticed that the correlation of 0.4 or 0.8 between variables holds for the entire ideal population of 10,000 DMU's, 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.

In addition, various forms of misspecification were purposely introduced (see for example Smith (1997) and Galagedera and Silvapulle (2003)), namely one or two of the inputs were omitted, an irrelevant input or two were incorporated into the model and the incorrect assumption as to the type of returns-to-scale was made. Performance criteria based on the percentage of observations that predict the efficiency incorrectly were used to compare the findings obtained by the models. The criteria included the number of efficient units defined as inefficient (Error type I) and the number of inefficient units defined as efficient (Error type II). It should be noted that error type II is more likely to occur because there are more inefficient simulated DMU's by definition. The coverage probabilities of the confidence intervals for bootstrapped DEA, PCA-DEA and VR radial efficiency scores were also estimated.

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 covariance between inputs in the entire ideal population, inefficiency distribution, sample size and returns-to-scale assumption. The value of the horizontal axis is the average percentage of Error type I in each case and the value of the vertical axis is the average percentage of Error type II. For example, for the entire simulated population of 10,000 observations and a sub-sample size

of 8 decision-making units, the average percentage error 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 Error type I is significantly less than the probability of Error type II, therefore the axes' lengths are different. The left point of the pictures coincide with the standard DEA program, without loss of information and then, each point to the right corresponds to the situation in which more and 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 set the number of PCs or variables retained such that the percentage of information remaining was *at least* equal to the level set by the program. 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 not informative simply because the removal of an entire variable has a different effect to dropping a single principal component.

Several figures have been chosen for illustrative purposes with the aim of demonstrating general results and conclusions. Figure 1 presents a Cobb-Douglas production function with equal weights and low covariance over all inputs tested on several sample sizes. Figure 2 presents VRS Cobb-Douglas functions, demonstrating the problems that arise 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. 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 error type I in this experimental design is substantially lower than the results presented in Figure 1 based on the previous experimental assumptions, hence the values on the x-axis are much smaller. Given the

assumption that there is no data cloud surrounding the efficient frontier, the trade-off between the two error types appears considerably healthier than previously supposed.

4.1 Level of Information to Retain

One of our initial goals was to find a rule-of-thumb or 'optimal index' for the preferable approach (PCA-DEA), specifying the percentage of retained information that provides the closest proximity to the efficiency classification. In general, improving discrimination comes at a price, since it increases the probability of Error type I. It was found that in some cases, when the value of the optimal index dropped below a certain level, the probability of both types of error increased (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 Error type II reaches its minimum.
2) If there are several such points, choose the one where Error type I is minimized.

The solution to step 2 represents the optimum index value per specific simulated case. The general, optimal index will be based on accumulated index values in order to provide a reasonable rule-of-thumb. The simulation study suggests that the optimal index 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 level of correlation between variables, of the distribution of inefficiencies or of the type of production function.

4.2 Results Drawing on Data Generation Procedure I

Figure 1 here

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 type II error faster and creates type I error more slowly than the VR methodology. Furthermore, it should be noted that in terms of errors, there is no significant difference between

variable reduction according to partial correlation as a measure of information and simple correlation. 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 type II error, defined as over-estimation of efficiency in Smith (1997). Third, it has become clear that there are trade-offs between the two types of error and most importantly, given the correct variables[8] and returns-to-scale assumption, the basic DEA makes no type I error i.e. no efficient DMU is ever classified incorrectly. This is the reason that all lines presented in Figure 1 begin on the y-axis. Unfortunately, this is at the expense of type II error whereby inefficient DMU's are incorrectly defined as efficient. The value of the different error types is clearly context dependent but in cases where 50% or more of the DMU's are defined as efficient, error choice ought to be considered.

4.2.1 Returns-to-scale Error

Figure 2 here

In the top graph of Figure 2, the influence of an incorrect CRS assumption on Error type I is demonstrated, particularly for highly correlated variables. This is the first instance of type I error being produced by the basic DEA identified by the fact that the left most point (100) represents the results of the weighted additive DEA and alternative models with complete information. On average, greater type I error occurs with high correlation between inputs because of the similarity between DMU's. 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 (Error type I). It should be noted that high correlation causes the opposite effect with respect to Error type II, slightly reducing this error. Pedraja-Chaparro 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, the user needs an index of

[8] All variables are relevant and measured accurately.

dimensionality that takes account of the intercorrelation among variables. The results of the simulation demonstrate that both PCA-DEA and VR models improve the problem of dimensionality by reducing the number of variables in the DEA analysis, which as a result cause the number of DMU's lying on the efficiency frontier to decline, the probability of type II error to decrease and the probability of type I error to increase. PCA-DEA is preferable to VR, particularly for relatively low correlation between variables[9] when this type of misspecification occurs (incorrect CRS assumption). Finally, when VRS is correctly assumed, the problem of discrimination is even more distinct causing a relatively large Error type II percentage, irrespective of the correlation between inputs. This can be seen in 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 average type II error of more than 40%.

Figure 3 demonstrates the reduction in type II error as a function of sample size for the translog production function. The top graph in Figure 3 demonstrates the influence of an incorrect CRS assumption on Error type I, particularly for relatively large samples (16 observations) and the lower graph points out 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 Error type I for relatively large samples, since the number of efficient observations in the larger sample is greater by definition, therefore the possibility of type I error (efficient units defined as inefficient) occuring 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 them (Smith (1997) and Pedraja-Chaparro et al. (1999)). In the bottom graph, Type II error is substantial in the standard DEA, around 40% for very small samples (8 DMU's) and around 20% for larger sets (16 DMU's), but as we reduce the unnecessary principal components, the same error reduction appears as previously demonstrated with the Cobb-Douglas production functions.

[9] In the extreme case where the data is uncorrelated we can apply neither PCA nor VR, therefore we discuss only 'low correlation between the variables'.

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 'true' simulated efficiencies than those of VR and proved easier to navigate because the data reduction did not occur in large jumps or changes, as occurs when an entire variable is removed from the analysis, particularly when samples are small and original variables show 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, in other words a reduction in type II error 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.

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 model 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 parameters. Table 2 summarizes the intervals of error of the original DEA and PCA-DEA according to the optimal index, when original inputs are highly correlated, the number of decision-making units relative to the number of variables is small, i.e. 1 output, 4 inputs and 8 DMU's[10], under various forms of the production function (Table 1) and misspecification, namely (a) one of the inputs was omitted from the model, (b) an irrelevant input was incorporated into the model, (c) an incorrect assumption of returns-to-scale was made and (d) an irrelevant input was not included in the production process computation of output, despite a correlation with the other inputs. Table 2 presents the trade-off between the two types of error, the influence of returns-to-scale assumptions on the results and 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.

[10] We chose purposefully extreme examples in order to demonstrate the effect of information reduction but we also note that DEA has been applied to real databases with very small sample sets and substantial numbers of variables as published in the literature e.g. Hokkanen and Salminen (1997a, b), Friedman and Sinuany-Stern (1997).

Table 2 here

For example, the lower bound of cells (1) and (2) is 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 bound is 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)).

The analysis carried out in this study has highlighted some additional issues within the DEA context. First, if the correct returns-to-scale and variables are determined, the standard DEA LP's never make type I error (cell (3) and lower bound of cell (1) in Table 2), however type II error can be quite substantial, particularly in the VRS case (cell (4) in Table 2). 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 reasonable to always include the VRS constraint in a DEA when the number of observations is relatively large and inputs are highly correlated (as shown in Figures 2 and 3 and Table 2). According to the Galagedera and Silvapulle (2003) study based on a large sample (200 DMU's with 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 type I error, but this is dependent on the relative importance of the variable omitted. If the weight on the variable is relatively high (x_1), error type I is higher that that of the relatively less important variable omission (x_4). Error type I is also relatively lower under variable returns-to-scale assumptions than constant returns.

Figures 4 and 5 here

Figure 5 compares the effect of omission versus the inclusion of an extraneous variable. With the latter, type II error is more likely. The inclusion of extraneous inputs influences the type II error of the basic weighted additive DEA dramatically, since it significantly complicates the process of defining inefficient DMU's. 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 determination of the truly efficient decision-making units is more important than the correct determination of the inefficient ones. Clearly, the omission of relevant variables leads to underestimation of the mean efficiency, while the inclusion of irrelevant variables leads to over-estimation.

4.3 Results Drawing on Data Generation Procedure II

Figure 6 here

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 and the definition of a "nearly efficient" unit. The figure shows that PCA-DEA is strongly preferable to VR for all levels of information retained. PCA-DEA reduces type II error faster and creates type I error more slowly than the VR methodology. Furthermore, as the definition of a 'nearly efficient' unit becomes stricter ($e^{-\tau}>0.9$), with settings ranging from 0.9 in 0.03 steps up to 0.99, error type II increases quite understandably, as the number of potential relatively efficient units decreases.

4.3.1 Confidence Intervals Estimation

Tables 3 and 4 present 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 with varied inefficiency distributions, as presented in column (1) of each table. 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 radial, radial PCA-DEA and radial VR models (column (6)). Real bias and bootstrap bias estimates (columns (4) and (5)) were calculated as shown in Equations (8) and (9). Columns (8) and (9) give the ranges of the lower and the upper bounds for estimated 97.5 % one-sided confidence intervals over 100 trials. The last part of the model name (column (2)) refers to the minimum percentage of retained information in the data. Column 10 presents information as to how many PCs or variables were retained based on the rule of thumb.

Tables 3 and 4 here

The results again indicate the greater accuracy of the PCA-DEA model over VR, in this case with respect to estimating the efficiency score[11]. As was expected, the bootstrap estimates of bias and the widths and ranges of the estimated confidence intervals decrease as sample size increases. The results are rather sensitive to the variance of the inefficiency term, for example the inefficiency distribution HN(0,1) produces especially wide ranges. 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

[11] An addition of noise in the experimental design will influence the efficiency estimator performance. Since both models are sensitive to measurement errors, the 'no noise' assumption is not crucial for the comparison between PCA-DEA and VR methods.

estimation of the 'true', absolute 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 dependant on the proportion of tied observations and we cannot neglect this problem, therefore two types of errors were utilized in this study to act as universal performance criteria.

5. Summary and Conclusions

This research has compared two methodologies previously published in the literature, both of which have the stated aim of improving the discriminatory power of DEA without the need for additional preferential information. Problems related to discrimination usually 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 been reduced to two types; Error type I whereby efficient units are defined incorrectly as inefficient and Error type II whereby inefficient units are deemed efficient, a problem that appears particularly frequently when assuming variable returns-to-scale. This study used Monte-Carlo simulation to generalize the comparison between the 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 compared to the 'true' value, permitting a computation 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 was also presented.

It was found that the PCA-DEA formulation provided consistently more accurate results that the VR technique. The results were such that PCA-DEA reduced type II error more quickly and produced type I error more slowly. The results proved robust to changes in the initial data distribution, production function, inefficiency distribution and model misspecification. This proved true under strongly different simulation assumptions, based on two schools of thought. The first school defined 25% of DMU's as strictly efficient, whereas the second school assumes

all DMU's to be inefficient to some extent, with no mass close to the frontier. The combination of VR and PCA-DEA methods (variable reduction followed by the application of principal components) proved unsuccessful due to an excessive loss of information.

In this paper we have extended the work already published in the field by applying the 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 LPs. Guidelines for the PCA-DEA user were presented based on the concept of an 'optimal index' rule-of-thumb. The optimal index considered the trade-off between the two types of error, suggesting that the data may be reduced to a few uncorrelated principal components that describe at least 80 (76)% of the variance of the original data under constant (variable) returns-to-scale assumptions. This value is independent of the level of correlation between variables, of the distribution of inefficiencies or of the type of production function. It should be noted that in some cases, when the retained information lay below this guideline, the probability of both types of error increases.

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 characteristic of a production process for relatively small samples in practice, it may be more reasonable to include a VRS constraint in a DEA 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 determination of the relatively efficient decision-making units is more important than the correct determination of the inefficient ones e.g. for benchmarking purposes.

Acknowledgements

The authors would like to thank Prof. J.S.H. Kornbluth, two associate editors of JPA and three anonymous referees for their help in substantially improving this paper. Dr. Adler would also like to thank the Recanati foundation for partial funding of this research.

References

Adler, N. and B. Golany. (2001). "Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe." European Journal of Operational Research 132 (2), 18-31.

Adler, N. and B. Golany. (2002). "Including principal component weights to improve discrimination in data envelopment analysis**."** Journal of the Operational Research Society 53, 985-991.

Allen, R., A. Athanassopoulos, R. G. Dyson and E. Thanassoulis. (1997). "Weights restrictions and value judgements in Data Envelopment Analysis: Evolution, development and future directions." Annals of Operations Research 73, 13-34.

Banker, R.D., A. Charnes and W.W. Cooper. (1984). "Models for estimating technical and returns-to-scale efficiencies in DEA." Management Science 30, 1078-1092.

Banker, R.D., A. Charnes, W.W. Cooper and A. Maindiratta. (1988). A Comparison of Data Envelopment Analysis and Translog Estimates of Production Frontiers with Simulated Observations from a Known Technology. In Dogramaci, A. and R. Fare (eds.), Applications of Modern Production Theory in Efficiency and Productivity, New York: Kluwer Academic Publishers, pp. 33-55.

Banker, R.D., V.M. Gadh and W.L. Gorr. (1993). "A Monte Carlo comparison of two production frontier estimation methods: corrected ordinary least squares and data envelopment analysis." European Journal of Operational Research 67, 332-343.

Banker, R.D. (1996). "Hypothesis tests using data envelopment analysis". Journal of Productivity Analysis 7, 139-159.

Bardhan, I.R., W.W. Cooper and S.C. Kumbhakar. (1998). "A simulation study of joint uses of data envelopment analysis and statistical regressions for production function estimation and efficiency evaluation." Journal of Productivity Analysis 9, 249-278.

Charnes, A., W.W. Cooper and E. Rhodes. (1978). "Measuring the efficiency of decision making units." European Journal of Operational Research 2, 429-444.

Charnes, A., W.W. Cooper, B. Golany, L. Seiford and J. Stutz. (1985). "Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions." Journal of Econometrics 30, 91-107.

Dyson, R., R. Allen, A.S. Camanho, V.V. Podinovski, C.S. Sarrico and E.A. Shale. (2001). "Pitfalls and protocols in DEA." European Journal of Operational Research 132, 245-259.

Friedman, L. and Z. Sinuany-Stern. (1997). "Scaling units via the canonical correlation analysis in the DEA context." European Journal of Operational Research 100 (3), 629-637.

Galagedera, D.U.A. and P. Silvapulle. (2003). "Experimental evidence on robustness of data envelopment analysis". Journal of the Operational Research Society 54, 654-660.

Haas, D.A. and F.H. Murphy. (2003). "Compensating for non-homogeneity in decision-making units in data envelopment analysis**."** European Journal of Operational Research 144, 530-544.

Hokkanen, J. and P. Salminen. (1997a). "Choosing a solid waste management system using multicriteria decision analysis". European Journal of Operational Research 98 (1), 19–36.

Hokkanen, J. and P. Salminen. (1997b). "Electre III and IV methods in an environmental problem". Journal of Multi-Criteria Analysis 6, 216-26.

Jenkins, L. and M. Anderson. (2003). "Multivariate statistical approach to reducing the number of variables in data envelopment analysis." European Journal of Operational Research 147, 51-61.

Johnson, R.A., Wichern D.W., 1982. *Applied Multivariate Analysis*, Prentice-Hall Inc., New Jersey.

Kneip, A., B. U. Park and L. Simar. (1998). "A note on the convergence of nonparametric DEA estimators for production efficiency scores." Econometric Theory 14, 783-793.

Lovell, C.A.K. and J.T. Pastor. (1995). "Units invariant and translation invariant DEA models." Operational Research Letters 18, 147-151.

Pastor, J. (1996). "Translation invariance in data envelopment analysis: a generalization." Annals of Operations Research 66, 93-102.

Pastor, J.T., J.L. Ruiz and I. Sirvent. (2002). "A statistical test for nested radial DEA models." Operations Research 50, 728-735.

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

Pedraja-Chaparro, F., J. Salinas-Jimenez and P. Smith. (1999). "On the quality of the data envelopment analysis model." Journal of the Operational Research Society 50, 636-644.

Read, L.E. and E. Thanassoulis. (2000). “Improving the identification of returns to scale in data envelopment analysis”. Journal of the Operational Research Society 51, 102-110.

Simar, L. and P. W. Wilson. (1998). "Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models". Management Science 44, 49-61.

Simar, L. and P. W. Wilson. (2000). “Statistical inference in nonparametric frontier models: the state of the art.” Journal of Productivity Analysis 13, 49-78.

Simar, L. and P. W. Wilson. (2001). “Testing restrictions in nonparametric efficiency models." Communications in Statistics 30, 159-184.

Smith, P. (1997). “Model misspecification in data envelopment analysis.” Annals of Operations Research 73, 233-252.

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

Figure 1: Error percentages of weighted additive DEA (100), PCA-DEA and VR models with simulated constant returns-to-scale Cobb-Douglas production functions and varying sample sizes ($y = x_1^{0.25}x_2^{0.25}x_3^{0.25}x_4^{0.25}$, relative low correlation between inputs, eff~exp(0.2231))

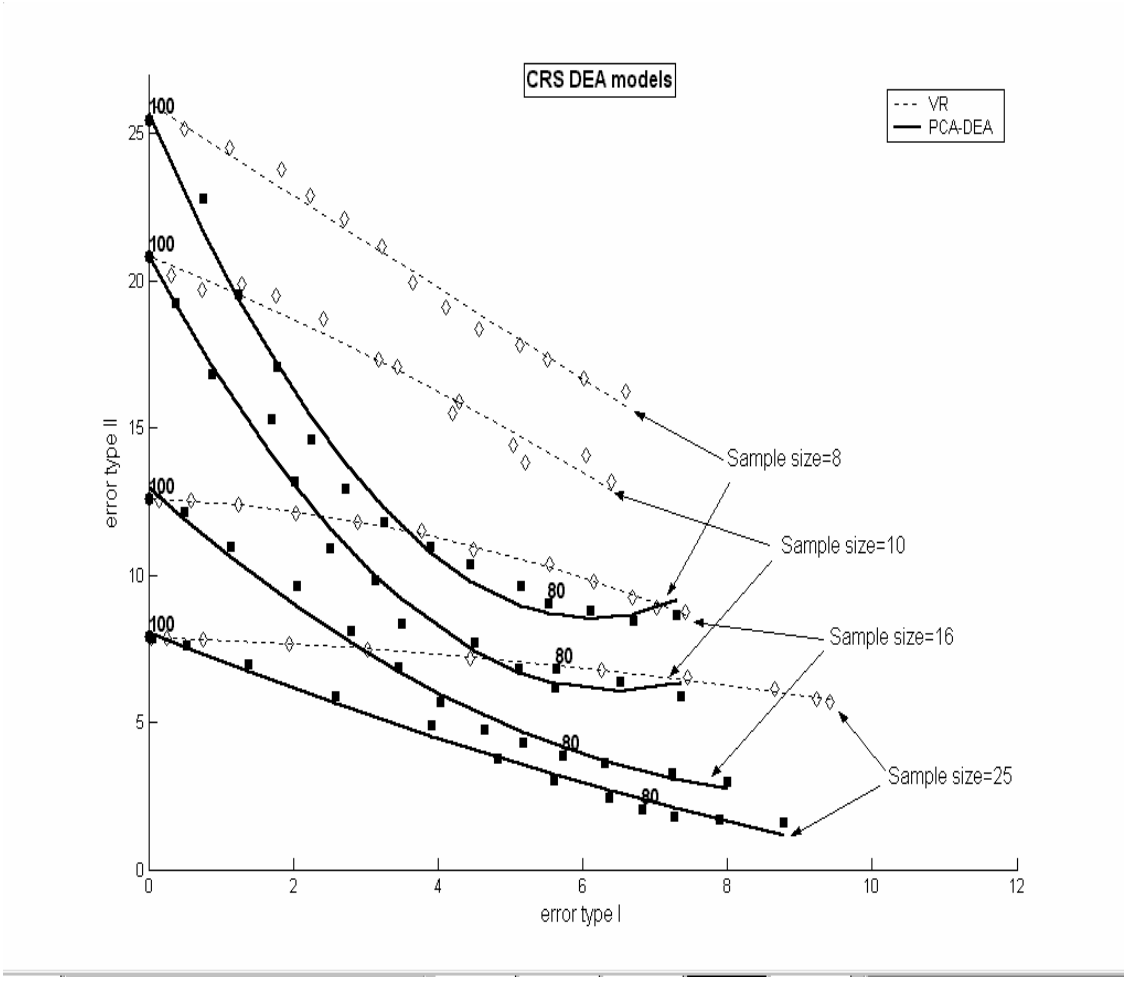

Figure 2: Error percentages of weighted additive DEA (100), PCA-DEA and VR models with simulated variable returns-to-scale Cobb-Douglas production functions and varying correlation in inputs ($y = x_1^{0.35}x_2^{0.2}x_3^{0.1}x_4^{0.05}$, sample size=8, eff~HN(0, 0.2796))

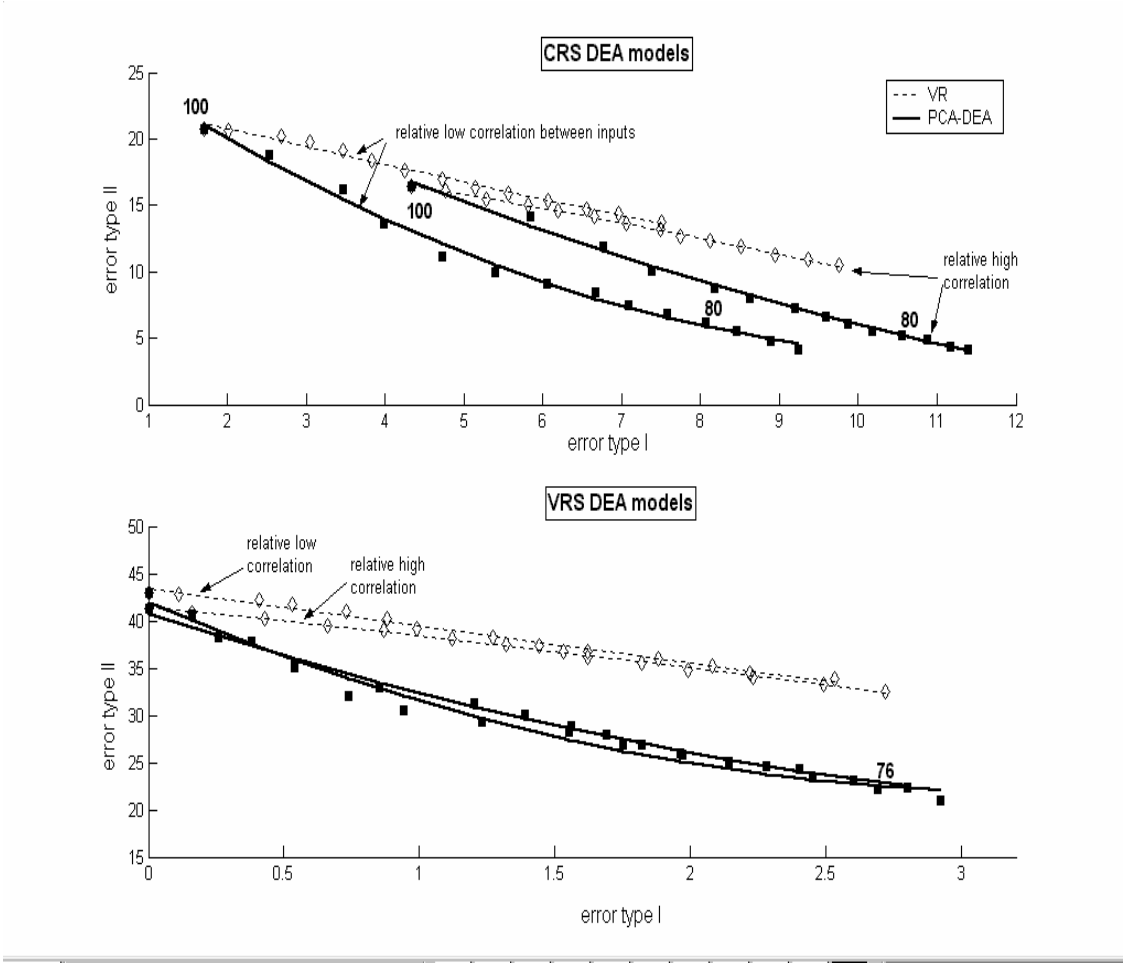

Figure 3: Error percentages of weighted additive DEA, PCA-DEA and VR models with simulated translog production functions, varying sample size (relative high correlation between inputs, eff~exp(0.2231))

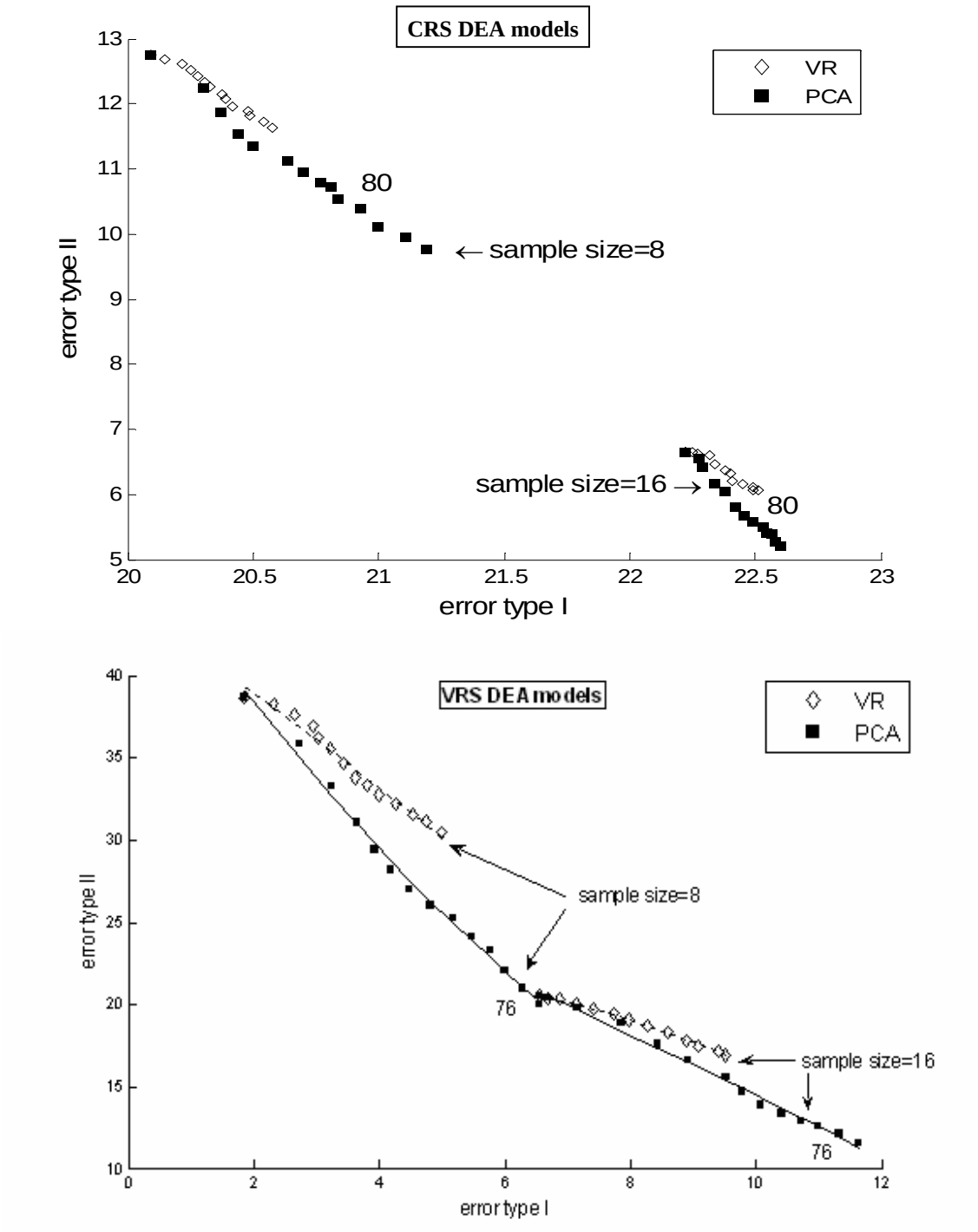

Figure 4: Error percentages of weighted additive DEA, PCA-DEA and VR models with simulated constant returns-to-scale Cobb-Douglas production functions, varying input importance and the omission of one input ($y = x_1^{0.45}x_2^{0.3}x_3^{0.15}x_4^{0.1}$, relative high correlation between inputs, eff~exp(0.2231), sample size=8)

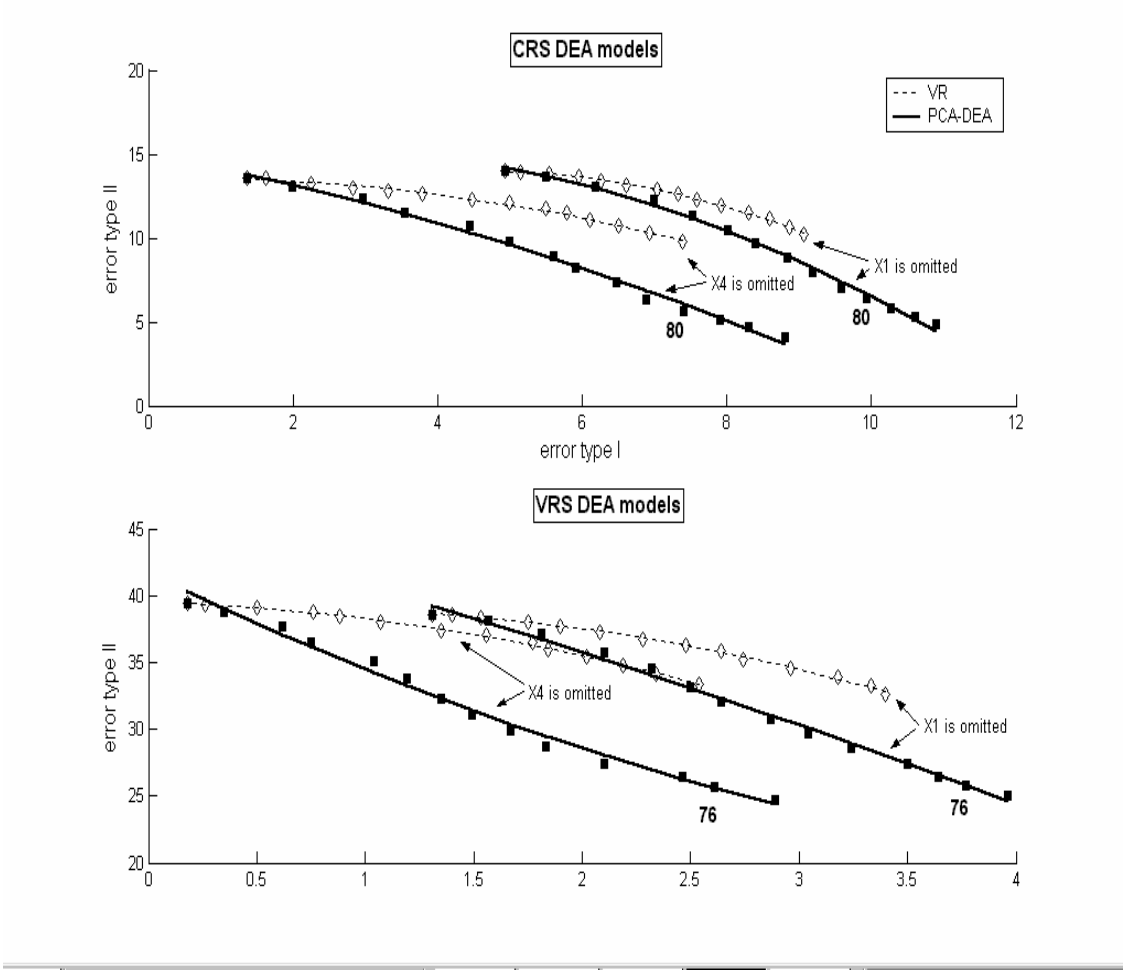

Figure 5: Error percentages of weighted additive DEA, PCA-DEA and VR models with simulated constant returns-to-scale Cobb-Douglas production functions, omission or inclusion of one input ($y = x_1^{0.25}x_2^{0.25}x_3^{0.25}x_4^{0.25}$, relative high correlation between inputs, eff~exp(0.2231), sample size=8)

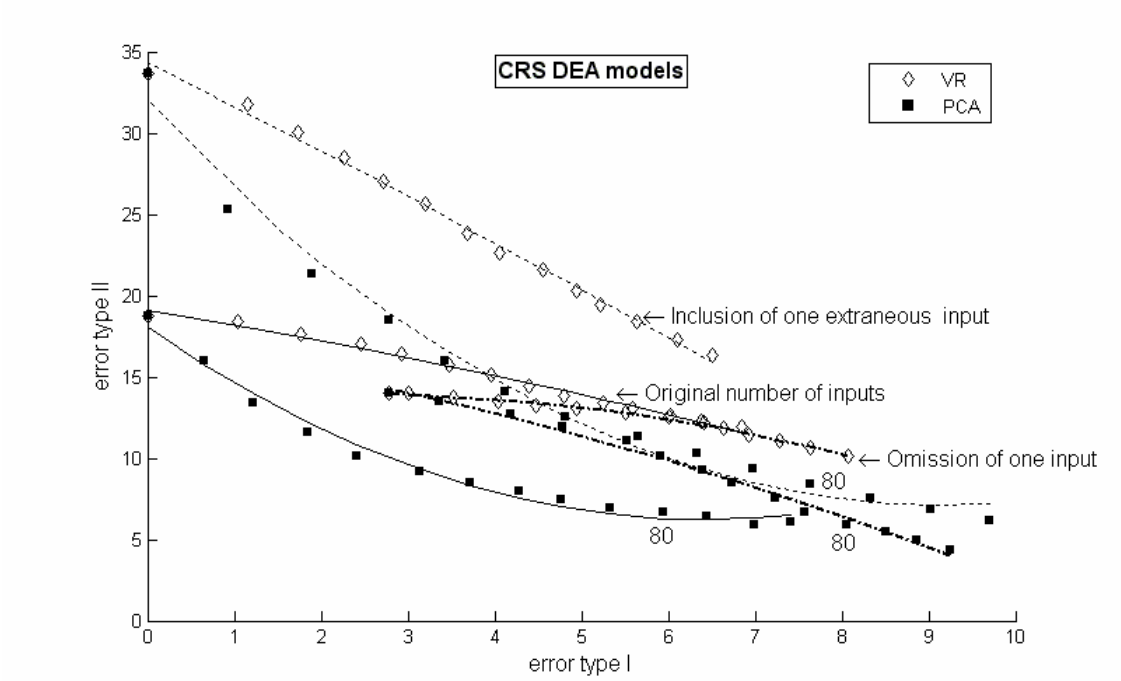

Figure 6: Error percentages of radial CRS DEA, PCA-DEA and VR models with simulated constant returns-to-scale Cobb-Douglas production functions and varying efficiency definition ($y = x_1^{0.25}x_2^{0.25}x_3^{0.25}x_4^{0.25}$, relative low correlation between inputs, eff~exp(0.2231) , sample size=16)

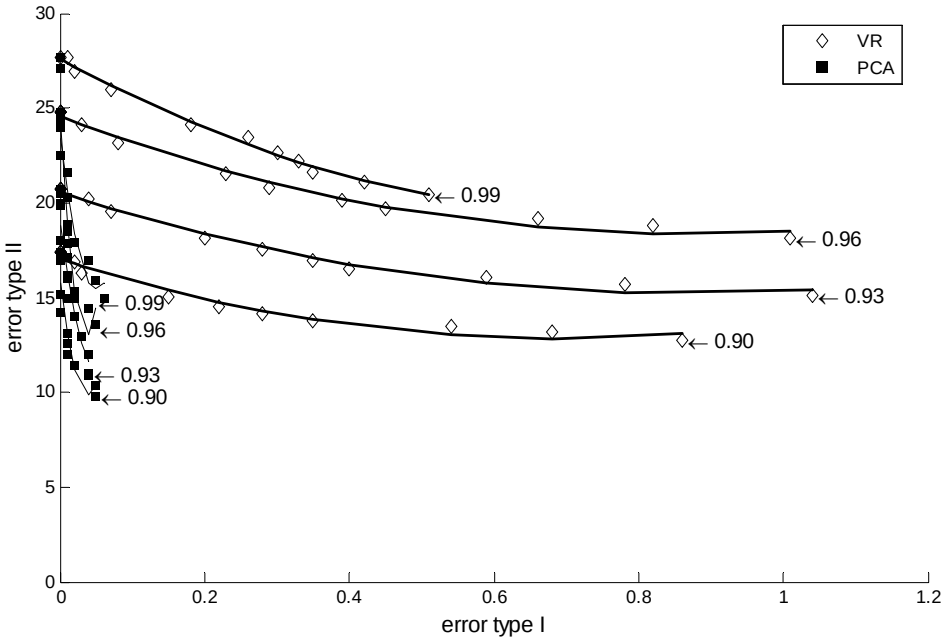

Table 1. Cobb-Douglas Production Functions

Production function	***Returns-to-scale property***
$y = x_1^{0.25} x_2^{0.25} x_3^{0.25} x_4^{0.25}$	$\sum_{i=1}^{m} \alpha_i = 1$ (Constant Returns to Scale)
$y = x_1^{0.45} x_2^{0.3} x_3^{0.15} x_4^{0.1}$	
$y = x_1^{0.35} x_2^{0.2} x_3^{0.1} x_4^{0.05}$	$\sum_{i=1}^{m} \alpha_i = 0.7$ (Decreasing Returns to Scale)
$y = x_1^{0.3} x_2^{0.3} x_3^{0.3} x_4^{0.3}$	$\sum_{i=1}^{m} \alpha_i = 1.2$ (Increasing Returns to Scale)
$y = \begin{cases} x_1^{0.45} x_2^{0.3} x_3^{0.2} x_4^{0.1} & \text{for } \forall x_i < 10 \\ x_1^{0.4} x_2^{0.25} x_3^{0.15} x_4^{0.05} & \text{otherwise} \end{cases}$	Combined technology

Table 2. Original DEA and PCA-DEA error intervals according to the optimal index and DEA returns-to-scale assumptions

All relevant inputs included in DEA	**Constant returns-to-scale**		**Variable returns-to-scale**	
	Original DEA	***PCA-DEA***	***Original DEA***	***PCA-DEA***
Error type I	0-4.5 % [(1)]	6-11 %	0 % [(3)]	2-2.8 %
Error type II	16-20 % [(2)]	5-7 %	43-45 % [(4)]	23-25 %
Omission of one input	**Constant returns-to-scale**		**Variable returns-to-scale**	
	Original DEA	***PCA-DEA***	***Original DEA***	***PCA-DEA***
Error type I	1-7 %	7.5-10 %	0.25-1.5 %	2.5-3.9 %
Error type II	12-15 %	6-7 %	35-40 %	23-26 %
Inclusion of one extraneous input	**Constant returns-to-scale**		**Variable returns-to-scale**	
	Original DEA	***PCA-DEA***	***Original DEA***	***PCA-DEA***
Error type I	0-3.5 %	7.5-9.5 %	0 %	3-3.5 %
Error type II	28-34 %	7-10 %	56-57 %	25–28 %

Table 3. Monte Carlo estimates of confidence intervals for radial CRS, radial CRS-PCA and radial CRS-VR estimators with simulated constant returns-to-scale Cobb-Douglas production functions and varying inefficiency distributions (relative low correlation between inputs)

$y = x_1^{0.45}x_2^{0.3}x_3^{0.15}x_4^{0.1}$, **sample size=10**

Inefficiency distribution / True efficiency in the fixed point	DEA method	Average of relative efficiency estimators	Average of real bias	Bootstrap bias estimates	Estimate of Confidence Interval Coverages	Average width of estimated confidence intervals	Range of lower limits of estimated confidence intervals	Range of upper limits of estimated confidence intervals	Number of trials when m inputs were included
									m=4 m=3 m=2 m=1
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Basic DEA	0.8205	0.0576	0.0537	0.91	0.1189	0.5826 0.7909	0.7812 0.8904	100
exp(0.2231)	PCA_0.8	0.7981	0.0351	0.0487	0.95	0.1096	0.5971 0.7860	0.7494 0.8779	45 52 3
0.7630	VR_0.8	0.8081	0.0451	0.0514	0.87	0.1160	0.5365 0.7830	0.7343 0.8780	1 79 20
	PCA_0.9	0.8078	0.0449	0.0501	0.93	0.1110	0.6048 0.7852	0.7494 0.8877	9 79 12
	VR_0.9	0.8168	0.0538	0.0527	0.89	0.1173	0.5741 0.7938	0.7505 0.8823	46 53 1
	Basic DEA	0.7669	0.0642	0.0508	0.83	0.1123	0.5243 0.7789	0.7160 0.8787	100
HN(0,0.2796)	PCA_0.8	0.7440	0.0413	0.0474	0.88	0.1067	0.5202 0.7795	0.6760 0.8373	45 52 3
0.7027	VR_0.8	0.7534	0.0507	0.0497	0.81	0.1122	0.5306 0.7658	0.6720 0.8556	1 79 20
	Basic DEA	0.6180	0.1234	0.1120	0.87	0.2672	0.0902 0.6084	0.5195 0.8607	100
HN(0,1)	PCA_0.8	0.5942	0.0995	0.0988	0.92	0.2513	0.0879 0.6111	0.4839 0.8496	45 52 3
0.4947	VR_0.8	0.6031	0.1121	0.1067	0.86	0.2618	0.0735 0.6031	0.4736 0.8607	1 79 20

$y = x_1^{0.25}x_2^{0.25}x_3^{0.25}x_4^{0.25}$, **sample size=20**

Inefficiency distribution / True efficiency in the fixed point	DEA method	Average of relative efficiency estimators	Average of real bias	Bootstrap bias estimates	Estimate of Confidence Interval Coverages	Average width of estimated confidence intervals	Range of lower limits of estimated confidence intervals	Range of upper limits of estimated confidence intervals	Number of trials when m inputs were included
									m=4 m=3 m=2 m=1
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Basic DEA	0.8027	0.0398	0.0440	0.97	0.0844	0.6317 0.7980	0.7666 0.8496	100
exp(0.2231)	PCA_0.8	0.7876	0.0246	0.0382	0.97	0.0744	0.6469 0.7779	0.7630 0.8459	63 37
0.7630	VR_0.8	0.7909	0.0279	0.0419	0.84	0.0820	0.6188 0.7981	0.7351 0.8386	2 95 3
	Basic DEA	0.7502	0.0475	0.0423	0.88	0.0811	0.6039 0.7487	0.7187 0.8011	100
HN(0,0.2796)	PCA_0.8	0.7350	0.0323	0.0377	0.95	0.0733	0.6015 0.7464	0.7039 0.7929	63 37
0.7027	VR_0.8	0.7376	0.0349	0.0415	0.85	0.0808	0.5827 0.7174	0.6867 0.7882	2 95 3
	Basic DEA	0.5734	0.0787	0.0874	0.92	0.1761	0.1805 0.5809	0.5039 0.7048	100
HN(0,1)	PCA_0.8	0.5571	0.0624	0.0772	0.93	0.1619	0.1652 0.5848	0.4977 0.6819	63 37
0.4947	VR_0.8	0.5669	0.0722	0.0819	0.89	0.1701	0.2029 0.5932	0.4898 0.7048	2 95 3

Table 4. Monte Carlo estimates of confidence intervals for reciprocal of output-oriented radial VRS, radial VRS-PCA and radial VRS-VR estimators with simulated variable returns-to-scale Cobb-Douglas production functions and varying inefficiency distributions (relative low correlation between inputs)

$y = x_1^{0.3}x_2^{0.3}x_3^{0.3}x_4^{0.3}$, **sample size=25**

Inefficiency distribution / True efficiency in the fixed point	DEA method	Average of relative efficiency estimators	Average of real bias	Bootstrap bias estimates	Estimate of Confidence Interval Coverages	Average width of estimated confidence intervals	Range of lower limits of estimated confidence intervals	Range of upper limits of estimated confidence intervals	Number of trials when m inputs were included
									m=4 m=3 m=2 m=1
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Basic DEA	0.9065	0.0520	0.0402	0.84	0.0800	0.7582 0.9125	0.8694 0.9672	100
exp(0.2231)	PCA_0.76	0.8917	0.0371	0.0395	0.92	0.0755	0.7585 0.9048	0.8604 0.9617	32 68
0.8546	VR_0.76	0.8936	0.0390	0.0416	0.87	0.0810	0.7091 0.9024	0.8259 0.9618	91 9
	Basic DEA	0.9139	0.0378	0.0406	0.90	0.0809	0.7465 0.9102	0.8883 0.9655	100
HN(0,0.2796)	PCA_0.76	0.8998	0.0238	0.0394	0.98	0.0758	0.7608 0.8890	0.8807 0.9449	32 68
0.8761	VR_0.76	0.8984	0.0223	0.0425	0.82	0.0830	0.7123 0.8913	0.8470 0.9597	91 9
	Basic DEA	0.5957	0.0920	0.0869	0.90	0.1778	0.2831 0.5610	0.5263 0.7378	100
HN(0,1)	PCA_0.76	0.5734	0.0660	0.0761	0.93	0.1546	0.2616 0.5900	0.5042 0.7324	32 68
0.5037	VR_0.76	0.5843	0.0806	0.0821	0.92	0.1669	0.2987 0.5742	0.5141 0.7378	91 9