European Journal of Operational Research 215 (2011) 662–669

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Guidelines for using variable selection techniques in data envelopment analysis

Niranjan R. Nataraja, Andrew L. Johnson *

Department of Industrial and Systems Engineering, Texas A&M University, College Station, Texas, United States

ARTICLE INFO

Article history:
Received 10 September 2010
Accepted 30 June 2011
Available online 13 July 2011

Keywords:
Data envelopment analysis
Model specification
Efficiency estimation

ABSTRACT

Model misspecification has significant impacts on data envelopment analysis (DEA) efficiency estimates. This paper discusses the four most widely-used approaches to guide variable specification in DEA. We analyze efficiency contribution measure (ECM), principal component analysis (PCA-DEA), a regression-based test, and bootstrapping for variable selection via Monte Carlo simulations to determine each approach's advantages and disadvantages. For a three input, one output production process, we find that: PCA-DEA performs well with highly correlated inputs (greater than 0.8) and even for small data sets (less than 300 observations); both the regression and ECM approaches perform well under low correlation (less than 0.2) and relatively larger data sets (at least 300 observations); and bootstrapping performs relatively poorly. Bootstrapping requires hours of computational time whereas the three other methods require minutes. Based on the results, we offer guidelines for effectively choosing among the four selection methods.

1. Introduction

Nonparametric frontier estimation evaluates production units' relative efficiency in both multiple input and multiple output production settings. The method in widest use is data envelopment analysis (DEA), popularized by Charnes et al. (1978). DEA itself does not provide guidance for the specification of the input and output variables; rather, they are left to the user's discretion, judgment and expertise. However, several issues may arise when selecting variables, e.g., the unavailability of data, high dimensional production processes, and the inclusion of irrelevant inputs or outputs. This paper examines the latter issue, reviews eight variable selection methods to identify the relevant variables, and offers guidelines for choosing the most appropriate method.

Sexton et al. (1986) and Smith (1997) demonstrate that model misspecification has significant impacts on DEA efficiency estimates. Sexton et al. consider three different scenarios to investigate the impact of inclusion of a variable in the production function: including an additional input variable; including an additional random input; and selecting inputs based on statistical significance. They show that efficiency estimates cannot lessen when adding either more inputs or outputs, but a variable can change the shape and position of the frontier which in turn alters the ranking of efficiency estimate. Using Monte Carlo simulations, Smith analyzes model misspecification issues considering sample size, number of inputs, correlation between inputs, and variation in importance of input to the production process. He concludes that model misspecification has a more severe impact on efficiency estimates when the data set contains fewer observations. Dyson et al. (2001) show that the omission of a highly correlated variable can have a significant impact on the efficiency estimates of some production units because DEA generally is not translation invariant.[1]

Further, variable selection methods are important because DEA is a non-parametric approach and loses discriminatory power as the dimensionality of the production space increases. As the number of inputs and outputs increases, the observations in the data set are projected in an increasing number of orthogonal directions and the Euclidean distance between the observations increases. This results in many observations lying on the frontier; thus DEA loses its discriminatory power. An insightful discussion on the curse of dimensionality can be found in Fried et al. (2008).

The literature proposes several methods to address the issues of determining relevant variables. All approaches are statistical in nature. Our own literature survey identifies eight methods, four of which have already been compared against each other by Sirvent et al. (2005) and Adler and Yazhemsky (2010), and thus we omit them from this paper. The four remaining methods to be analyzed are: efficiency contribution measure by Pastor et al. (2002), PCA applied to DEA by Ueda and Hoshiai (1997) and Adler

* Corresponding author. Tel.: +1 9798459025.
E-mail address: ajohnson@tamu.edu (A.L. Johnson).

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.06.045

[1] Pastor (1996) defines a DEA envelopment model as translation invariant "if, given any problem and constructing a new (problem) by translating the original input and output values, it happens that the new problem has exactly the same optimal solution as the old one". See Pastor (1996) for a complete discussion of specific scenarios in which DEA is/is not translation invariant.

and Golany (2001), a regression-based test by Ruggiero (2005), and Bootstrapping for variable selection by Simar and Wilson (2001).

This paper is organized as follows. Section 2 presents our literature survey and why we favor four specific methods. Section 3 investigates the four methods under different scenarios via Monte Carlo simulations. Section 4 discusses the appropriateness and the limitations of the four methods as well as the severity of the limitations and their effects on DEA estimates. Section 5 develops our guidelines for choosing a variable selection method and gives our general conclusions.

2. Literature review

Our literature survey on variable selection in DEA identifies eight methods. This section gives an overview of each method and explains why we evaluate efficiency contribution measure, PCA-DEA, regression-based tests, and bootstrapping for variable selection.

1. Efficiency contribution measure (ECM): Pastor et al. (2002) develop a method for analyzing the relevance of a variable based on its contribution to efficiency. See Chen and Johnson (2010) for an application of ECM. The variable being tested is called the candidate. Two DEA formulations are considered, one with the candidate variable and one without it. A binomial statistical test determines if the effect of this variable on the efficiency measure indicates that the candidate variable is important to the production process. Two approaches are defined next: forward selection (addition of variables) and backward elimination (removal of variables). The forward selection procedure is evaluated in the Monte Carlo analysis in Section 3.
2. Principal component analysis (PCA)-DEA: Ueda and Hoshiai (1997) and Adler and Golany (2001) independently develop principal component analysis-DEA (PCA-DEA). A general statistical method used to reduce the dimensionality of the data set by expressing the variance structure of a matrix of data through a weighted linear combination of variables. Each principal component (obtained from the weighted linear combination of original variables and ordered in decreasing order of percentage variance) accounts for maximal variance while remaining uncorrelated with the preceding principal components. Adler and Golany (2002) give a separate PCA-DEA mathematical formulation to obtain the efficiency estimates in which the principal components replace the original variables. In this method, a percentage of the information is retained from each of the original variables, thus improving the discriminatory power of DEA.
3. A regression-based test: Ruggiero (2005) suggests a variable selection approach in which an initial measure of efficiency is obtained from a set of known production variables. Efficiency is then regressed against a set of candidate variables; if the coefficients in the regression are statistically significant and have the proper sign (coefficient values should be positive for inputs and negative for outputs), the variables are relevant to the production process. This analysis is repeated, identifying one variable at a time. The analysis stops when there are no further variables with significant and properly signed coefficients.
4. Bootstrapping for variable selection: Simar and Wilson (2001) discuss a statistical procedure to test the relevance of removing input and output variables as well as the potential for aggregation. Test statistics are calculated and a bootstrap estimation procedure is used to obtain the critical values for these tests.
5. Banker (1996) lists three statistical tests to indicate the significance of an input or output variable to the production process. The null hypothesis is that the variable being tested does not influence the production process. Simulation studies are conducted and the results indicate that these tests perform as well as or better than COLS-based tests (Olson et al. (1980)). This is true even when the parametric frontier form used in COLS estimation is identical to the one used to generate the simulated data.
6. Fanchon (2003) suggests a recursive method to determine the variables to be included. A five-step approach determines the variable set that best explains output behavior, followed by using DEA iteratively to analyze the increase in the number of efficient observations. To validate the included variables, two more regressions are performed, one with only efficient observations and the other with both efficient and inefficient observations. In each, a high statistical significance of regression coefficients indicates a valid input variable.
7. Jenkins and Anderson (2003) propose a variable reduction method that omits the variables containing minimum information using partial correlation as a measure of information content. Information in an input or output variable is measured as the variance over a set of production units; zero variation indicates all observed production units have the same value for that variable. The authors show that omitting highly correlated variables can have a major influence on efficiency scores, and thus the multivariate statistical approach using partial correlation measures is useful to determine the relevance of a given variable.
8. Dario and Simar (2007) aggregate highly correlated inputs and outputs to reduce the dimensionality of the production possibility space to a single input and a single output using eigenvalues.

Other methods in our literature survey include analyzing the average change in efficiency scores, trying different model specifications, and so forth. These methods are not as statistically rigorous, and hence are not considered in this paper. Readers can refer to Lewin et al. (1982), Golany and Roll (1989), Norman and Stoker (1991), Valdmanis (1992), Sigala et al. (2004), and Wagner and Shimshak (2007).

Previous work has compared subsets of these methods. Sirvent et al. (2005) compare Pastor et al.'s efficiency contribution measure method to Banker's hypothesis tests using Monte Carlo simulation. The results show that ECM is more robust to the specification of inefficiency distribution and the type of returns to scale assumption; thus, we do not consider Banker's hypothesis test. Since Fanchon's (2003) regression-based approach is similar to Ruggiero's method and includes a variety of additional ad-hoc complications with little indication of the motivation, we do not include it in the Monte Carlo simulation comparison below. Adler and Yazhemsky (2010) show that PCA-DEA performs better than the variable reduction technique by Jenkins and Anderson (2003) in particular when analyzing small data sets. They find that PCA-DEA never produces less accurate results when compared to Jenkins and Anderson. The primary drawback of the variable reduction method, however is that it discards an entire variable whereas PCA-DEA retains a certain amount of information from all variables. Lastly, the Dario and Simar method is very similar to PCA-DEA, but requires that the final model to have only a single input and a single output; thus it is not as general as the other methods and is of little practical use. In conclusion, we choose ECM, PCA-DEA, Ruggiero's regression method (referred to as RB), and the bootstrap approach (referred to as BS) for comparison below.

3. Monte Carlo simulations

Monte Carlo simulations let us compare the performance of our four chosen variable selection methods to a known truth. We

Table 1
List of experiments and their significance.

Experiment	Correlation (ρ)	Input contribution to output	Details of the experiment
1	Independently generated	$a = b = c = 1/3$	Base case, $n = 100$
2	$\rho_{x_1x_2} = 0.8, \rho_{x_1x_3} = 0.2$	$a = b = c = 1/3$	Correlated inputs
3	$\rho_{x_1x_2} = 0.8, \rho_{x_1x_3} = 0.8$	$a = b = c = 1/3$	Highly correlated inputs
4	Independently generated	$a = 1/3$, $b = 4/9$, $c = 2/9$	Input contribution to output varied
5	$\rho_{x_1x_2} = 0.8, \rho_{x_1x_3} = 0.2$	$a = 1/3$, $b = 4/9$, $c = 2/9$	Correlated inputs and input contribution to output varied
6	$\rho_{x_1x_2} = 0.8, \rho_{x_1x_3} = 0.2$	$a = 1/3$, $b = 2/9$, $c = 4/9$	Correlated inputs + different input contribution to output
7	$\rho_{x_1x_4} = 0.8$	$a = b = c = 1/3$	Correlated input and random variable
8	$\rho_{x_1x_4} = 0.8$	$a = b = c = 1/3$	Correlated candidate input and random variable
9	Independently generated	$a = b = c = 1/3$	Small sample size, $n = 25$
10	Independently generated	$a = b = c = 1/3$	Large sample size, $n = 300$
11	Independently generated	$a = b = c = 1/4$	Base case with VRS
12	Independently generated	$a = b = c = d = 1/4$	Base case with one more relevant input x_5
13	Independently generated	$a = b = 1/2$	Base case without relevant input x_3
14	Independently generated	$a = b = c = 1/3$	Base case with exponential inefficiency distribution

calculate true efficiency from a known production function.[2] We identify one input and one output as belonging in the production function, and then we apply the four methods iteratively to test a set of relevant and irrelevant variables to determine a final set of inputs for each method. Finally, we use the set of input variables identified by each method to estimate efficiency via DEA and compare the estimates with the true efficiency levels to measure the accuracy of the variable selection methods.

We generate three inputs (x_1,x_2 and x_3) and an inefficiency term (u) and use them in a Cobb–Douglas production function to generate an output (y). Eq. (1) is the production function used in the data generating process. A single output production function allows us to compare the four methods. The values for the inputs are independently generated from a uniform distribution on the interval (10,20). The inefficiency term is half-normal with mean zero and the variance (σ^2) which we vary to obtain an average efficiency score of 85%. Exponents a, b and c define both the returns to scale (RTS) specification and the contribution of each input to output. We independently generate a random variable (x_4) from a uniform distribution in the interval (10,20) to maintain symmetry with the other inputs. Note that this variable is not part of the production process and therefore is irrelevant.

$$y = x_1^a x_2^b x_3^c e^{-u}. \tag{1}$$

To establish the correlation between inputs we adopt the following equation from Wang and Schmidt (2002).

$$x_i = \rho x_j + w\sqrt{1-\rho^2} \quad i = 2,3,4 \quad j = 1,2,3 \quad i \neq j. \tag{2}$$

Here, ρ is the correlation between x_i and x_j; w is a random variable generated from a uniform distribution in the interval (10,20).

To perform the simulation analysis, we define a base case scenario with $a = b = c = 1/3$ and independently generate the input variables. The values of a, b, and c represent their contributions to output and are set equal to each other to impose symmetry. When the sum of the exponents on the inputs is one, this indicates a constant returns to scale (CRS) production process. The number of observations equals 100. We test the basic variable set consisting of (y,x_1,u) and, x_2, x_3 and x_4 (herein referred to as candidate variables) using the four methods to determine the model specification. We also derive thirteen other experimental scenarios from the base case. These include varying: the correlation among inputs, the correlation between random variable and the inputs, the number of observations in the dataset, the RTS specification, the dimensionality changes and a different inefficiency distribution. Table 1 summarizes the experiments.

[2] Also termed the data generation process.

Table 2
Parameter values used in the experiments.

Method	Parameter value
ECM	$p_0 = 0.15, \bar{\rho} = 1.10, \alpha = 0.05$
PCA-DEA	Information retained from the variables = 80%
RB	Confidential Interval = 90%
BS	$\alpha = 0.05$

Table 2 reports the values of the parameters for the four methods. Recall that we select the values based on the literature originally introducing each method. We keep these same values throughout the Monte Carlo simulations.[3]

Having specified the input variables via the four methods, we use the output-oriented CRS measure in (3) to obtain the results presented in Section 4.

$$\max_{\theta,\lambda} \theta : -y_i + Y\lambda \geqslant 0, \quad \theta x_i - X\lambda \geqslant 0, \quad \lambda \geqslant 0, \quad \theta \text{ free}. \tag{3}$$

In experiment 11 we generate the data from a VRS production function. Thus (3) is augmented with the additional convexity constraint, and the results of the augmented model are reported, see Banker et al. (1984).

3.1. Efficiency contribution measure (ECM)

For ECM, a candidate variable is considered relevant to the production process if more than P_0% of the production processes have an associated efficiency change greater than $\bar{\rho}$. Therefore, we select $p_0 = 15\%$ and $\bar{\rho} = 10\%$ following the recommendation of Pastor et al. (2002). ECM is formulated as a hypothesis test with a binomial test statistic. Following Pastor et al., a significance level (α) of 5% is set. If the test statistic is less than the null hypothesis is rejected and a candidate variable is considered to be part of the production process. A forward selection procedure is used and initially input x_1 and output y are included in the production function, all candidate variables are tested, and the variable with the lowest test statistic below the α value is added to the production model. The ECM is repeated on the new candidate set with one less variable. The process stops when all candidate variables have a test statistic larger than α, or no variables remain in the candidate set.

[3] We observe that methods to determine values of the parameters based on the characteristics of the data set are not currently available in the literature; this is perhaps a useful area for further research.

3.2. Principle component analysis (PCA-DEA)

As mentioned, we consider the PCA-DEA formulation developed by Adler and Golany (2002). The principal components (PCs) explain the population variance, which makes it possible to replace the original variables with minimum loss of information, thereby reducing the dimensionality of the production function. Irrespective of the correlation between variables, the inefficiency distribution, and the type of production process, Adler and Yazhemsky (2010) suggest that 80% (76%) of retained information for the CRS (VRS) case provides a good approximation to the efficiency classification. After setting the PCs in decreasing order of percentage variance, we select the set describing 80% of the variation. The experiments in Table 1 assume that x_1 is a known input in the production process. Hence, we keep x_1's original values, and include the candidate variables in PCA which retains 80% of the information.

3.3. Regression-based test (RB)

For the regression-based test, Ruggiero (2005) states: if a potential input is omitted from the DEA model then that input will be positively correlated with the measured efficiency, in this case rerun the DEA model with that variable included. We implement this rule using the regression model:

$$TE = \alpha + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m + \varepsilon, \tag{4}$$

where TE is the technical efficiency as given by DEA including only y and x_1, and x_2 through x_m are the potential inputs that could have been included in the production function. Thus, we obtain TE without the x_i variables; only if the parameters β_i are greater than zero and statistically significant at a given level of significance is x_i added to the model. Ruggiero suggests that a 90% confidence interval leads to best results and hence we employ a 90% significance level in the Monte Carlo analysis in Section 4. We are cautious since Ruggiero notes this method may not perform well in the presence of correlation among the input candidate variables.[4] We regress the efficiency estimates from the basic set of variables, input x_i and output y, against the candidate variables. The procedure stops if no variable is found significant. If more than one significant variable is found, then all variables identified are added. The new efficiency score obtained with the inclusion of the new input variable (s) in the DEA are calculated and the candidate set, which is now smaller, is tested. We repeat the process until all candidate variables are either found irrelevant or included and no variable remains to be tested.

3.4. Bootstrapping for variable selection (BS)

The BS method indicates if a variable significantly contributes to the output level and thus influences the estimates of efficiency. Simar and Wilson suggest several test statistics; we use (5):

$$\gamma_{12}(s_n) = \sum_{i=1}^{n} \left[\frac{\delta_{1i}}{D_0(x_i', y_i)}\right]^2 \geqslant 0, \tag{5}$$

$$\delta_{1i} = D_0(x_i, y_i) - D_0(x_i', y_i), \tag{6}$$

where D_0 is the distance function obtained as the inverse of the output-oriented DEA formulation with (x,y) as the input–output set. Set (x_i,y_i) refers to the complete set of inputs and outputs including the candidate variable, and (x_i', y_i) refers to the reduced set without the candidate variable. We test the hypothesis that the production function does not contain the candidate variable in the production process. The test statistic is calculated using the full population n production processes. We use the bootstrap to draw independent and identically distributed (iid) samples of n observations by sampling with replacement from the population n and repeatedly estimating (6) and (5) for each sample of n to construct an empirical estimate of the distribution of $\gamma_{12}(s_n)$. With this distribution we can identify the p-values associated with the 5% significance level (α). A test statistic less than the p-value indicates both a rejection of the null hypothesis and the inclusion of the tested variable in the production function. We implement the heterogeneous bootstrap algorithm proposed in Simar and Wilson (2000) to generate the bootstrap efficiency estimates. Simar and Wilson (2001) develop six similar test statistics and find they all work equally well; we choose one that resembles minimizing the variance in efficiency estimates (see Simar and Wilson (2001) for a complete set of test statistics and a detailed explanation). A forward selection procedure is used and initially input x_1 and output y are included in the production function and all candidate variables are tested. If an iteration of BS selects more than two variables, we include the most significant variable. We repeat the iterations until all candidate variables are either found irrelevant or included and no variables remain in the candidate set.

4. Results and analysis

For our Monte Carlo simulations we use MATLAB on an IBM p5-575 cluster, 64-bit, AIX 5.3 operating system. Table 3 gives the results. Pearson's correlation and mean square deviation are reported as a means of comparison for the four methods. Here, correlation coefficient and mean square deviation are calculated for each method relative to the true efficiency estimates and the average is taken over the 1000 trials. Table 4 shows the run times for the four methods with 100 observations for one trial.

Table 3 shows that while the four methods perform well under different experiments, no method dominates. In experiment 1, there is no correlation defined among inputs and ECM and RB perform better than PCA-DEA. A limitation of PCA-DEA is the impossibility of exactly recovering the efficiency levels, since the method defines only an 80% information retention level. Observe that the correlation between inputs has a definite impact on the performance of the four methods. For large sample sizes (300 observations as shown in experiment 10), the other methods are superior to PCA-DEA (see Section 4.3 for a discussion of sample size). Table 4 shows that PCA-DEA has the shortest run time due to its non-iterative characteristic.

BS necessitates the selection of a bandwidth parameter, which requires a considerable computational effort. Simar and Wilson (2000) suggest the normal referencing rule to reduce the computational burden. However, even when using the normal referencing rule, the BS algorithm results in a run time greater than 30 hours for experiments including 20 trials and 200 bootstrap replications. We perform an initial set of trial runs to determine the number of bootstrap replications, and expect to see a tradeoff between run time and performance. However, we find nothing substantial by studying the relationship between number of bootstrap replications and the correlation to true efficiency or mean square deviation (see Fig. 10). The graphs in the following sections provide further insights into the remaining three methods when BS is excluded not only to modest results in the initial analysis, but also because of its computational burden.

4.1. Impact of variations in correlations between inputs ($\rho_{x_1x_2}$ and $\rho_{x_1x_3}$)

Fig. 1 shows that when $\rho_{x_1x_2}$ and $\rho_{x_1x_3}$ are varied (Experiment 3), PCA-DEA significantly outperforms the other two methods. We

[4] Many of the experiments explored in this paper consider correlation; however, it is interesting to investigate Ruggiero's comment in quantitative terms to understand how his method is affected when correlation is present.

Table 3
Performance of variable selection methods.

Experiment	Correlation coefficient				Mean square deviation			
	ECM	PCA-DEA	RB	BS*	ECM	PCA-DEA	RB	BS*
1	0.9992	0.9842	0.9965	0.9252	0.0001	0.0008	0.0005	0.0101
2	0.9645	0.9875	0.9404	0.9465	0.0040	0.0006	0.0133	0.0073
3	0.9587	0.9853	0.8770	0.9166	0.0048	0.0007	0.0407	0.0175
4	0.9876	0.9846	0.9922	0.9008	0.0012	0.0008	0.0008	0.0134
5	0.9593	0.9876	0.8855	0.896	0.0044	0.0006	0.0363	0.0192
6	0.9747	0.9876	0.9726	0.9769	0.0025	0.0006	0.0034	0.0025
7	0.9995	0.9930	0.9994	0.9252	0.0000	0.0004	0.0001	0.0101
8	0.9839	0.9775	0.9779	0.8937	0.0011	0.0012	0.0052	0.0151
9	0.9072	0.9265	0.8429	0.7197	0.0076	0.0159	0.0518	0.0559
10	1.0000	0.9923	0.9993	1.0000	0.0000	0.0004	0.0000	0.0000

* Bootstrap for 200 replications and each experiment for 20 trials.

Table 4
Run time for one trial.

Method	Run time (minutes)
ECM	56
PCA-DEA	8
RB	19
BS *	> 30 h

* Bootstrap for 200 replications and each experiment for 20 trials.

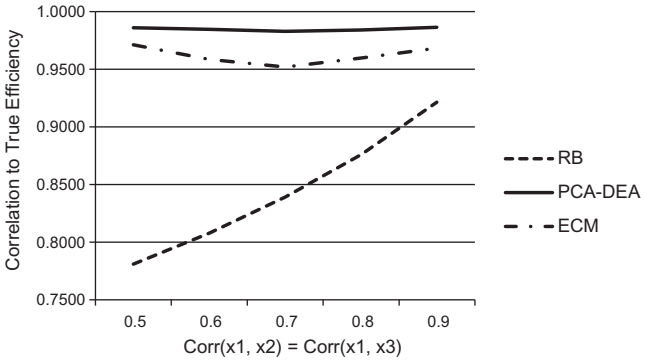

Fig. 1. Variation of correlation between inputs (experiment 3).

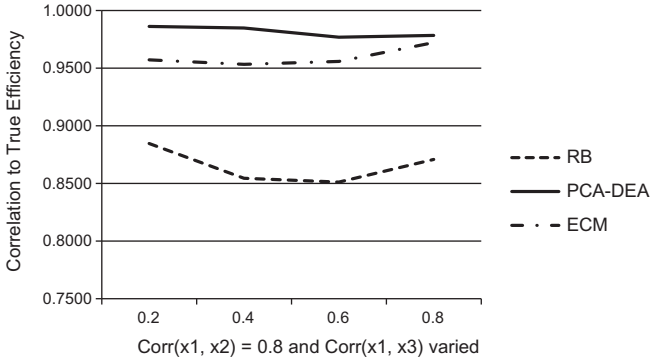

Fig. 2. Variation of input contribution and correlation between inputs (experiment 5).

conclude that PCA-DEA is robust to correlation between inputs, whereas ECM performs well for both low correlation (0.5) and high correlation (0.9). RB performance improves as correlation increases. Also, when the correlation between inputs varies along

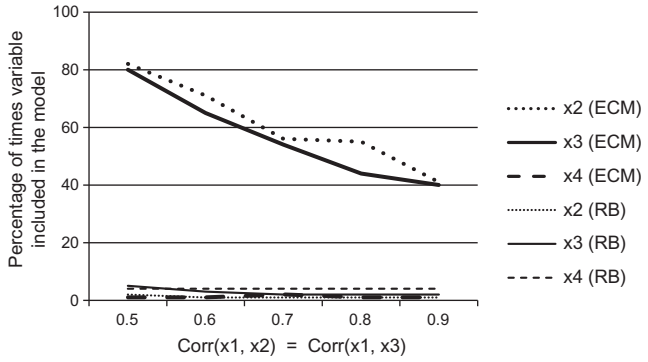

Fig. 3. Correlation between inputs' effect on percentage inclusion of x_2, x_3 and x_4 (experiment 3).

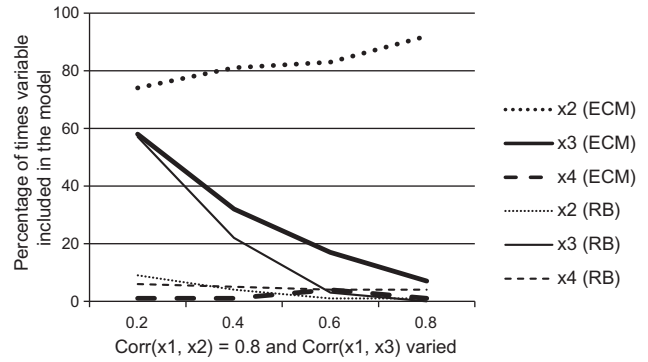

Fig. 4. Correlation between inputs' effect on percentage inclusion of x_2, x_3 and x_4 (experiment 5).

with the input contribution to output (Experiment 5), Fig. 2 shows that PCA-DEA performs better than the other two methods (RB performs worst).

Figs. 3 and 4 show the percentage inclusion (number of times the candidate variable is included out of 1000 trials) as a part of the production process via RB and ECM. Since PCA-DEA replaces the original variables with PCs it is not possible to obtain this information. Referring to Fig. 3, as the correlation between input increases, RB results in misspecification since and are not identified as part of the production process.

In experiment 5 (refer to Fig. 4), ECM chooses more x_2 and x_3 less as correlation increases. From the definition of experiment 5, the input contribution of x_2 is greater than that of x_3 and hence ECM selects the input with the larger contribution to output. However this is not true for RB where both x_2 and x_3 are included fewer times as correlation increases.

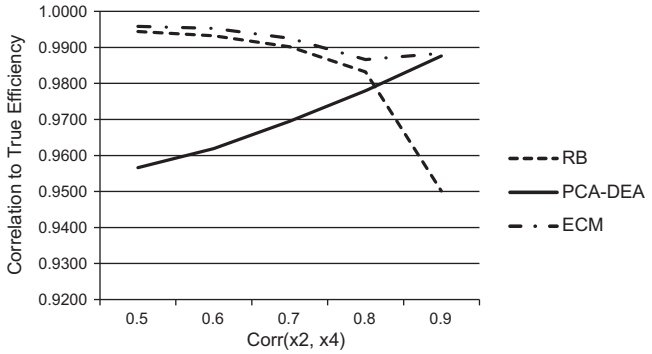

Fig. 5. Variation of correlation between relevant input and random variable (experiment 8).

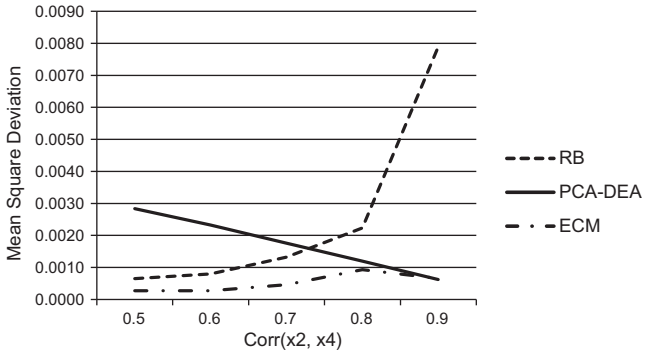

Fig. 6. Variation of correlation between relevant input and random variable (experiment 8).

4.2. Impact of variation in correlation between relevant input (x_2) and random variable (x_4)

Experiment 8 varies the correlation between candidate variable x_2 and random variable x_4 to understand the effects of correlation on the three methods. We find that RB and ECM perform better than PCA-DEA (Figs. 5 and 6). The efficiency estimates deteriorate significantly in PCA-DEA when x_4 is included.

However, the effect becomes less severe when the correlation between x_2 and x_4 increases, since x_4 contains the same information as x_2. On the other hand, as correlation increases, both ECM and RB tend to include x_4 and hence performance deteriorates. The percentage inclusion information for RB and ECM appears in Fig. 7.

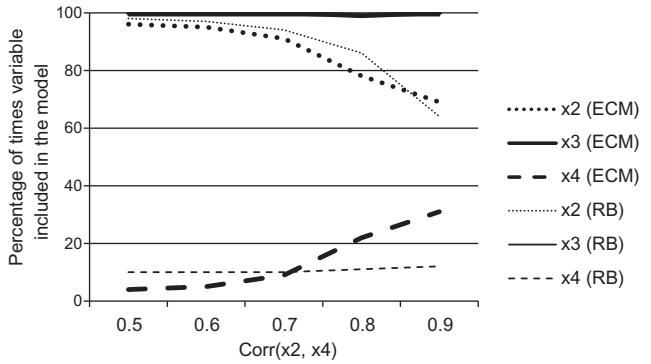

Fig. 7. Correlation between relevant input and random variable effect on percentage inclusion of x_2, x_3 andx_4 (experiment 8); Note: x3 (ECM) and x4 (RB) overlap at nearly 100%.

4.3. Impact of variation in sample size (n)

Fig. 8 shows the impact of sample size and performance for the three methods. Note that RB is the method most affected by a small sample size, but all three methods perform better as the sample size increases. This provides further support that PCA-DEA works well for small samples, but both ECM and RB outperform PCA-DEA as the sample size increases.

Fig. 9 shows that RB fails to identify the correct set of variables in the production process for smaller sample sizes and gives poor estimates compared to ECM and PCA-DEA. In general, PCA-DEA can be employed for better efficiency estimation in the case of smaller sample sizes.

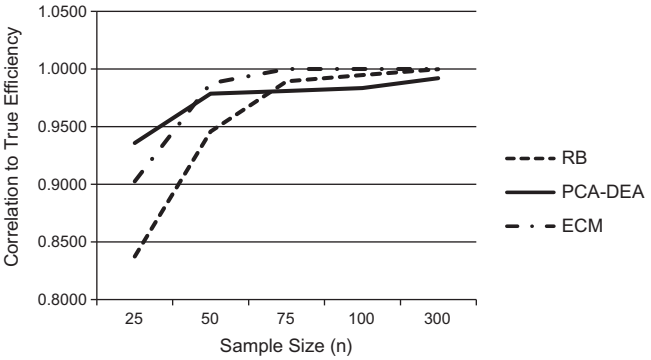

Fig. 8. Variation of Sample Size (Experiment 9).

4.4. Kolmogorov–Smirnov tests for comparison of test results

The non-parametric Kolmogorov–Smirnov (KS) test is commonly used to compare two samples of data to test the null hypothesis that the data are from the same distribution (see Gibbons (1985) for details of the KS test). We conduct a KS test on the correlation coefficients vector and mean square deviations vector for 1000 trials performed for each of the 10 experiments. The values in the table are the indicator variable values for the correlation coefficients. When the indicator variables agree for the correlation coefficients and mean square deviations, the common indicator variable is reported. The two measures agree except for experiment 9 (comparing ECM and RB).[5] Observe that for the majority of experiments the distributions of data significantly differ for both correlation coefficients and mean square deviations even though the values are similar or close in Table 3. Note that column 1 in Table 5 has several zero values for the indicator variable which implies that the differences in the performance of ECM and RB are statistically insignificant. The large number of ones in Table 5 indicates that the four methods differ in performance behavior.

It is clear from Table 3 that the methods outperform BS and hence it is excluded in experiments 11 to 14. These additional experiments augment the base case by considering alternatives such as VRS, changing dimensionality of the data set and distribution for the inefficiency distribution. The results are shown in Table 6.

Experiment 11 shows that RB is robust to the returns to scale effect. Comparing experiment 1 and experiment 11 which differ only by CRS and VRS technology, we observe that PCA-DEA's performance considerably decreases. We conclude that PCA-DEA is vulnerable to the choice of technology. Experiment 12 increases

[5] For this case, the unbracketed value is the indicator variable for the correlation coefficient and the bracketed value is the indicator variable value for the mean square deviation.

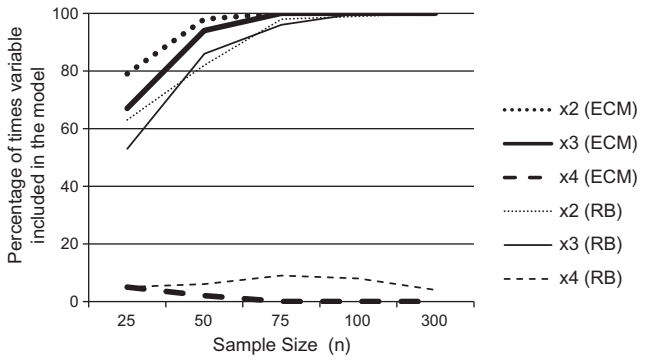

Fig. 9. Variation of Sample Size Effect on Percentage Inclusion of x_2, x_3 and x_4 (Experiment 9).

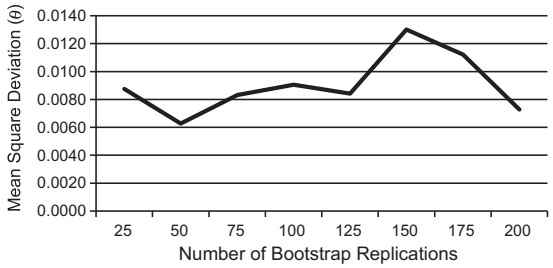

Fig. 10. Mean Square Deviation vs. Bootstrap Replications (Experiment 2).

Table 5
Results of Kolmogorov–Smirnov tests (1 indicates a statistically significant difference).

Experiment	ECM vs. RB	ECM vs. PCA-DEA	RB vs. PCA-DEA	ECM vs. BS *	RB vs. BS *	PCA-DEA vs. BS *
1	1	1	1	1	1	1
2	1	1	1	0	0	1
3	1	1	1	1	1	1
4	1	1	1	1	1	1
5	1	1	1	1	0	1
6	0	1	1	0	0	1
7	0	1	1	1	1	1
8	1	1	1	1	1	1
9	1 [0]	1	1	1	1	1
10	1	1	1	0	0	1

* Bootstrap for 200 replications and each test for 20 trials.

Table 6
Performance of variable selection methods (additional experiments).

Experiment	Correlation coefficient			Mean square deviation		
	ECM	PCA-DEA	RB	ECM	PCA-DEA	RB
11	0.9705	0.9313	0.9918	0.0025	0.0028	0.0004
12	0.9510	0.9279	0.9831	0.0044	0.0251	0.006
13	1.0000	0.9086	0.9985	0.0000	0.0525	0.0001
14	1.0000	0.9934	0.9878	0.0000	0.0007	0.0094

the dimensionality of the production function, and again RB is robust because the variables are selected through regression, whereas ECM heavily depends on DEA. Even though the variables are generated independently the sample correlations are positive and with more variables there are more pairs of inputs with positive sample correlations. Thus the principle components that are uncorrelated less resemble the original input vectors and thus PCA-DEA performance is diminished. Decreasing the dimension (experiment 13) results in improved accuracy for all methods except PCA-DEA. PCA-DEA now suffers because the information retention level remains at 80%, but the decrease in relevant inputs increases the relative contribution of x_4 to the estimated components. An exponential inefficiency distribution is considered in experiment 14 and all of the methods perform similarly to experiment 1; hence we conclude that the four methods are robust to inefficiency distribution.

5. Conclusion

One objective of this paper is to provide insights into the performance of the four methods. ECM has relatively long run times compared to PCA-DEA and RB, but performs well under most scenarios, and provided the correlation is low, identifies input variable contribution to outputs. We conclude that PCA-DEA is a robust technique in which some amount of information is retained from each of the original variables, unlike the other three methods which select or discard one entire variable. PCA-DEA also has the smallest run time, works best with smaller sample sizes, and is robust to the high correlations between inputs and irrelevant variables. Ruggiero's regression-based method is easily implemented, performs better than the bootstrap approach, and takes less computational time. In the case of highly correlated inputs and smaller sample sizes, RB may not perform as well as the other three methods. RB or ECM is preferred to PCA-DEA for large sample sizes. For larger sample sizes with low correlations among candidate variables, RB performs very well and accurately identifies the variables involved in the production process. However, we find that the bootstrap requires a long run time and has either similar or slightly worse performance. Our conclusions are summarized below.

- PCA-DEA
  - ◦Smallest run time
  - ◦Works well with smaller sample sizes ($n \sim 25$)
  - ◦Robust to high correlations (>0.80) between relevant and irrelevant variables
  - ◦Vulnerable to choice of technology (CRS or VRS)
  - ◦Robust to inefficiency distribution
  - ◦May not work well with higher dimension datasets
  - ◦Not clear how many PCs are needed
  - ◦Can never obtain true efficiency level
- RB
  - ◦Works well with low correlation (< 0.2) among inputs and a large sample size (n > 100)
  - ◦Less vulnerable to the curse of dimensionality
  - ◦Robust to inefficiency distribution
  - ◦Robust to choice of technology (CRS or VRS)
  - ◦May not work well with high correlation between variables (>0.8)
  - ◦Easy implementation
- ECM
  - ◦Performs moderately well under most scenarios
  - ◦Works well with low correlation (<0.2) among inputs and a large sample size (n > 100)
  - ◦Performs better than RB given correlation and sample size conditions above
  - ◦Can also identify input contribution to output
  - ◦Slightly effected by choice of technology (CRS or VRS)
  - ◦May not work well with high correlation between variables (>0.8)
  - ◦Vulnerable to the curse of dimensionality
  - ◦Robust to inefficiency distribution
- BS
  - ◦Heavy computational burden

- Number of bootstrap replications needed unclear
- Poor performance

Comparing the efficiency estimates resulting from the production models specified by the four methods did reveal significant differences, which indicates that they will behave differently even under similar conditions. It reinforces our conclusion that the user must take care when selecting the best-fit method for identifying relevant and irrelevant variables in the production process. We suggest that highly correlated inputs could adopt PCA-DEA to overcome the curse of dimensionality, whereas for large sample sizes RB or ECM are best suited provided there is a low correlation among the variables.

Suggestions for future research include evaluation of the performance of variable selection methods under different production functions and different distributions to generate the inputs. This would provide further understanding of the robustness of the results presented in this paper. We also observe that techniques to determine the values of the parameters based on the characteristics of the data set for any of the methods are not currently available in the literature; this is a useful area for research.

Acknowledgments

We acknowledge the Texas A& M Supercomputing Facility (http://sc.tamu.edu/) for providing the computing resources (see complete configuration of the supercomputer at http://sc.tamu.edu/systems) essential in the research reported in this paper.

We also thank N. Adler and E. Yazhemsky for generously sharing the PCA-DEA code.

References

Adler, N., Golany, B., 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), 260–273.

Adler, N., Golany, B., 2002. Including principal component weights to improve discrimination in data envelopment analysis. The Journal of the Operational Research Society 53 (9), 985–991.

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

Banker, R.D., 1996. Hypothesis tests using data envelopment analysis. Journal of Productivity Analysis 7 (2), 139–159.

Banker, R.D., Charnes, A., Copper, W.W., 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30 (9), 1078–1092.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2 (6), 429–444.

Chen, W.-C., Johnson, A.L., 2010. The dynamics of performance space of major league baseball pitchers 1871-2006. Annals of Operational Research 181 (1), 287–302.

Dario, C., Simar, L., 2007. Advanced Robust and Nonparametric Methods in Efficiency Analysis. Springer, XXII: 248.

Dyson, R.G., Allen, R., Camanho, A.S., Podinovski, V.V., Sarrico, C.S., Shale, E.A., 2001. Pitfalls and protocols in DEA. European Journal of Operational Research 132 (2), 245–259.

Fanchon, P., 2003. Variable selection for dynamic measures efficiency in the computer industry. International Advances in Economic Research 9 (3), 175–188.

Fried, H.O., Lovell, C.A.K., Schmidt, S.S., 2008. The Measurement of Productive Efficiency and Productivity Growth. Oxford University Press.

Gibbons, D.J., 1985. Nonparametric Statistical Inference, 2nd ed. McGraw-Hill Inc.

Golany, B., Roll, Y., 1989. An application procedure for DEA. Omega 17 (3), 237–250.

Jenkins, L., Anderson, M., 2003. A multivariate statistical approach to reducing the number of variables in data envelopment analysis. European Journal of Operational Research 147 (1), 51–61.

Lewin, A.Y., Morey, R.C., Cook, T.J., 1982. Evaluating the administrative efficiency of courts. Omega 10 (4), 401–411.

Norman, M., Stoker, B., 1991. Data Envelopment Analysis: The Assessment of Performance. John Wiley and Sons Chichester, England.

Olson, J.A., Schmidt, P., Waldman, D.M., 1980. A Monte Carlo study of estimators of stochastic frontier production functions. Journal of Econometrics 13 (1), 67–82.

Pastor, J., 1996. Translation invariance in DEA: a generalization. Annals of Operations Research 66, 93–102.

Pastor, J.T., Ruiz, J.L., Sirvent, I., 2002. A statistical test for nested radial dea models. Operations Research 50 (4), 728–735.

Ruggiero, J., 2005. Impact assessment of input omission on DEA. International Journal of Information Technology & Decision Making 04 (03), 359–368.

Sexton, T.R., Silkman, R.H., Hogan, A.J., 1986. Data envelopment analysis: critique and extensions. New Directions for Program Evaluation 1986 (32), 73–105.

Sigala, M., Airey, D., Jones, P., Lockwood, A., 2004. ICT paradox lost? a stepwise dea methodology to evaluate technology investments in tourism settings. Journal of Travel Research 43 (2), 180–192.

Simar, L., Wilson, P.W., 2000. A general methodology for bootstrapping in non-parametric frontier models. Journal of Applied Statistics 27 (6), 779–802.

Simar, L., Wilson, P.W., 2001. Testing restrictions in nonparametric efficiency models. Communications in Statistics 30 (1), 159–184.

Sirvent, I., Ruiz, J.L., Borras, F., Pastor, J.T., 2005. A Monte Carlo evaluation of several tests for the selection of variables in dea models. International Journal of Information Technology & Decision Making 4 (3), 325–343.

Smith, P., 1997. Model misspecification in data envelopment analysis. Annals of Operations Research 73 (0), 233–252.

Ueda, T., Hoshiai, Y., 1997. Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. Journal of the Operations Research Society of Japan 40 (4), 466–478.

Valdmanis, V., 1992. Sensitivity analysis for DEA models: an empirical example using public vs. NFP hospitals. Journal of Public Economics 48 (2), 185–205.

Wagner, J.M., Shimshak, D.G., 2007. Stepwise selection of variables in data envelopment analysis: procedures and managerial perspectives. European Journal of Operational Research 180 (1), 57–67.

Wang, H.J., Schmidt, P., 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. Journal of Productivity Analysis 18 (2), 129–144.