European Journal of Operational Research 216 (2012) 200–207

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Forward search outlier detection in data envelopment analysis

Tiziano Bellini *

Università di Parma, Via Kennedy 6, I-43100 Parma, Italy

ARTICLE INFO

Article history:
Received 23 March 2010
Accepted 12 July 2011
Available online 21 July 2011

Keywords:
Data envelopment analysis (DEA)
Super-efficiency
Forward search
Outlier detection

ABSTRACT

In this paper we tackle the problem of outlier detection in data envelopment analysis (DEA). We propose a procedure where we merge the super-efficiency DEA and the forward search. Since DEA provides efficiency scores which are not parameters to fit the model to the data, we introduce a distance, to be monitored along the search. This distance is obtained through the integration of a regression model and the super-efficiency DEA. We simulate a Cobb–Douglas production function and we compare the super-efficiency DEA and the forward search analysis in both uncontaminated and contaminated settings. For inference about outliers, we exploit envelopes obtained through Monte Carlo simulations.

1. Introduction

DEA is a nonparametric linear programming model where n decision making units (*DMUs*) are evaluated according to their input consumption and output production (Charnes et al., 1978). The main idea of DEA is to identify the production frontier on which the *DMUs* are considered as efficient. *DMUs* that are not on this frontier are compared with their peers on the frontier to estimate their efficiency scores.

One of the main advantages of DEA is to allow the *DMUs* to have the full freedom to select linear programming weights. However, the efficient frontier can be influenced by outliers, therefore it is crucial to check for the presence of atypical *DMUs*.

In order to detect outliers, Banker and Chang (2006) exploit a typical backward approach, such as the super-efficiency procedure proposed by Banker and Gifford (1988), where each observation is excluded from its own reference set when the efficiency score is computed. This method identifies as outliers those observations whose super-efficiency exceeds a pre-specified level. In order to avoid this subjective evaluation and to avoid masking effects, we consider the forward search procedure originally introduced in linear and nonlinear regression by Atkinson and Riani (2000).

The key issue in applying the forward search to DEA is that a linear programming technique does not supply parameters to fit the model to the data. Thus, we do not have any measure of fitting error. One of the main contributions of our research is to define a measure of unit closeness integrating the Kuosmanen and Johnson (2010) regression model and the super-efficiency procedure proposed by Lovell and Rouse (2003). Another important issue that we face is the definition of boundaries through which to make inference on outliers.

We conduct our analysis on a Cobb–Douglas simulated function applying the BCC (Banker et al., 1984) model for variable returns to scale. We compare the super-efficiency DEA and our forward search procedure in both uncontaminated and contaminated settings.

The paper is organised as follows. In Section 2 we describe the mechanics of DEA and super-efficiency DEA. In Section 3 we show how the forward search can be extended to DEA. In Section 4 we consider simulated uncontaminated and contaminated data generated from a Cobb–Douglas production function, comparing the super-efficiency DEA and our forward search analysis. Section 5 contains concluding remarks and indicates directions for future research.

2. DEA mechanics

DEA was introduced by Charnes et al. (1978) by extending the idea of estimating technical efficiency with respect to a production frontier. Considering that DMU_k ($k = 1,\ldots,n$) has a consumption $X_{k,i}$ of input i ($i = 1,\ldots,r$) and produces $Y_{k,j}$ of output j ($j = 1,\ldots,p$), we wish to evaluate the efficiency score of DMU_k among $DMU_1,\ldots,DMU_n$. Under the variable returns to scale assumption,[1] we have to solve, for each DMU_k, the following linear program:

* Tel.: +39 3480701760.
E-mail address: tiziano.bellini@unipr.it

[1] We focus on this assumption, but our analysis can be extended to other scale returns assumptions.

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.07.023

$$\begin{aligned}
&\max\ \theta_k \\
&s.t. \\
&\theta_k Y_{k,j} \leqslant \sum_{l=1}^{n} \lambda_l Y_{l,j} \quad (j = 1,\ldots,p), \\
&X_{k,i} \geqslant \sum_{l=1}^{n} \lambda_l X_{l,i} \quad (i = 1,\ldots,r), \\
&\sum_{l=1}^{n} \lambda_l = 1, \\
&\lambda_l \geqslant 0 \quad (l = 1,\ldots,n),
\end{aligned} \tag{1}$$

where, following the notation exploited by Banker and Chang (2006), $\theta_k \geqslant 1$ is a scalar which represents the inefficiency of DMU_k while $\psi_k = \theta_k^{-1}$ is its efficiency[2] and λ_l are linear programming weights.

The above-described model classifies *DMUs* on the frontier as efficient. *DMUs* which are not on the frontier, given their current input consumption, should be able to increase their output production to the extent indicated by their inefficiency scores.

Starting from this model, Banker and Chang (2006) find out that the super-efficiency DEA is effective in outlier detection. In the next section we summarize the super-efficiency DEA procedure.

2.1. The super-efficiency DEA

Considering that conventional DEA models evaluate the efficiency for DMU_k relative to all observations included into the set, the super-efficiency DEA excludes each observation from its own reference set.

Originating from the BCC model described in the equation system (1), the output oriented super-efficiency can be obtained, as in Banker and Chang (2006), starting from the solution of the following linear program:

$$\begin{aligned}
&\max\ \theta_k^{SI} \\
&s.t. \\
&\theta_k^{SI} Y_{k,j} \leqslant \sum_{\substack{l=1 \\ l\neq k}}^{n} \lambda_l Y_{l,j} \quad (j = 1,\ldots,p), \\
&X_{k,i} \geqslant \sum_{\substack{l=1 \\ l\neq k}}^{n} \lambda_l X_{l,i} \quad (i = 1,\ldots,r), \\
&\sum_{\substack{l=1 \\ l\neq k}}^{n} \lambda_l = 1, \\
&\lambda_l \geqslant 0 \quad (l = 1,\ldots,n),
\end{aligned} \tag{2}$$

where the superscript *SI* stands for super-inefficiency and the super-efficiency score ψ_k^{SE} is the reciprocal of θ_k^{SI}.

The difference between the above equation system and that described in equation system (1) is that in this latter model the observation k under evaluation is not included in the reference set for the constraints.

In our analysis we exploit the BCC model implemented in the DEA *R package*.[3] Starting from this software we implement the equivalent super-efficiency DEA model proposed by Lovell and Rouse (2003). This model generates the same super-efficiency scores as conventional super-efficiency models for all units having a feasible solution and generates a feasible solution for all units not having a feasible solution under traditional DEA super-efficiency models. The Lovell and Rouse (2003) model can be represented as follows:

$$\begin{aligned}
&\max\ \theta_k^{SI,LR} \\
&s.t. \\
&\theta_k^{SI,LR} Y_{k,j} \leqslant \sum_{\substack{l=1 \\ l\neq k}}^{n} \lambda_l Y_{l,j} + \gamma\lambda_k Y_{k,j} \quad (j = 1,\ldots,p), \\
&X_{k,i} \geqslant \sum_{\substack{l=1 \\ l\neq k}}^{n} \lambda_l X_{l,i} + \lambda_k X_{k,i} \quad (i = 1,\ldots,r), \\
&\sum_{\substack{l=1 \\ l\neq k}}^{n} \lambda_l + \lambda_k = 1, \\
&\lambda_l, \lambda_k \geqslant 0 \quad (l,k = 1,\ldots,n),
\end{aligned} \tag{3}$$

where the superscript *SI*, *LR* stands for super-inefficiency according to Lovell and Rouse. For each output $j \in \{1,\ldots,p\}$, it is required to select $\min Y_{k,j} > 0$ to remove any zero values, calculate $\gamma_j = \frac{\max Y_{k,j}}{\min Y_{k,j}} + 1$ and set $\gamma = [\max(\gamma_1,\ldots,\gamma_p)]^{-1}$. We compute $\theta_k^{SI,LR}$ and then we estimate $\psi_k^{SE,LR} = (\theta_k^{SI,LR}\gamma)^{-1}$. According to this approach, the upper super-efficiency score level is γ^{-1}.

In order to carry out the search, the problem of ranking *DMUs* arises. At first, we could think of exploiting efficiency scores, but these scores are a function of all *DMUs*. Stressing that we cannot directly use DEA scores, we need to define a framework where such scores can be exploited to obtain a measure of unit closeness. In what follows we describe how DEA can be seen from a regression point of view which allows us to find a distance that is crucial for our purposes.

2.2. DEA as a nonparametric regression

In the case where we consider only one output ($p = 1$) and multiple inputs ($r > 1$), we can represent the production function as follows:

$$Y_k = f(\mathbf{X}_k) + \epsilon_k, \tag{4}$$

where $\mathbf{X}_k$ denotes the input vector of DMU_k and ϵ_k represents the deviation of DMU_k from the efficient frontier. According to Kuosmanen and Johnson (2010), relying on Eq. (4) and considering $f^{DEA}(\mathbf{X}_k) = \psi_k Y_k$, we notice that the deviation from the efficient frontier can be considered as a linear regression error and it is obtained, for each DMU_k, as the optimal solution to the following linear program:

$$\begin{aligned}
&\min\ \epsilon_k \\
&s.t. \\
&Y_k = \sum_{l=1}^{n} \lambda_l Y_l + \epsilon_k, \\
&X_{k,i} \geqslant \sum_{l=1}^{n} \lambda_l X_{l,i} \quad (i = 1,\ldots,r), \\
&\sum_{l=1}^{n} \lambda_l = 1, \\
&\lambda_l \geqslant 0 \quad (l = 1,\ldots,n).
\end{aligned} \tag{5}$$

The above equation system represents the general framework to link DEA to regression analysis. For a more detailed representation of this functional relationship and constraints, see Kuosmanen and Johnson (2010). We can therefore stress that, in the single-output setting, the DEA approach described in the equation system (5) is equivalent to the standard output oriented variable return to scale DEA in the sense that for each DMU_k we have:

$$\epsilon_k^{DEA} = Y_k - f^{DEA}(\mathbf{X}_k) = Y_k - \psi_k Y_k, \tag{6}$$

[2] In the case where DMU_k is efficient $\theta_k = \psi_k = 1$.

[3] The software is available on the website <http://cran.bic.nus.edu.sg/src/contrib/Archive/DEA>. The website was last accessed in July 2011.

where ϵ_k^{DEA} is the optimal solution of equation system (5) for DMU_k, ψ_k is the efficient solution of equation system (1) and $\mathbf{X}_k$ is the matrix of r inputs for DMU_k. Then, for $Y_k > 0$, we can define the following distance:

$$d_k^{DEA} = \frac{|Y_k - \psi_k^{SE,LR} Y_k|}{Y_k}. \quad (7)$$

In order to compute d_k^{DEA}, we do not use the equation system (1), but the super-efficiency $\psi_k^{SE,LR}$ obtained from equation system (3). This choice is due to the fact that super-efficiency DEA highlights *DMUs* which are very different from the others, while the traditional approach does not emphasize these differences. In particular, because of its mechanics, this approach highlights super-efficient units. Therefore, as will be evident in the next sections, non-efficient *DMUs* are not detected as atypical in our forward search analysis which relies on the distance in Eq. (7).

In our simulation analysis, in the next sections, we focus on the assumption of a production function with only one output ($p = 1$) and multiple inputs ($r > 1$). However, it is useful to notice that the above described approach can be extended to multiple-output and multiple-input settings. Stressing that the key issue is to find a useful distance to monitor during the forward search and pursuing the goal to emphasize atypical *DMUs*, we rely on the fact that the efficiency score is the same across all outputs. Thus, in the multiple output environment, we consider separately each output and we concentrate on what follows:

$$d_k^{DEA,mo,oo} = \max_{1 \leqslant j \leqslant p} d_{k,j}^{DEA}, \quad (8)$$

where *mo*, *oo* stands for multiple outputs, output oriented.

A similar approach can be followed in the input oriented setting where inputs have to be minimized subject to constraints on outputs. Even in this case we concentrate on the maximum of the estimation error which, in this case, is computed over all r inputs.

In the next section we describe how we extend the forward search to DEA analysis.

3. The forward search in DEA analysis

The forward search is a statistical method originally introduced in linear and nonlinear regression by Atkinson and Riani (2000) and this is the first attempt to extend the approach to a linear programming technique.

The forward search is made up of three steps: initializing, progressing and monitoring. The first task is to find the appropriate starting subset, then it is required to specify the way to progress and highlight some statistics to monitor step by step along the search.

As we outlined above, one of the purposes of the forward search is to identify observations which are different from the majority of the data and to compute the effect of these observations on inferences made about the model. There may be a few outliers or it may be that observations can be divided into groups. Although it is convenient to refer to such observations as outliers, they may well form a large part of the data and they can have an independent unsuspected structure. It is often impossible to detect these structures from a model fitted to all the data because the effect of outliers is masked. Backward methods based on a deletion approach fail to show any important feature. The difficulty arises because outliers are included into the dataset employed for fitting the model.

Many methods for outlier detection seek to divide the dataset into two parts: a clean subset and an outlier subset. The clean data are used to estimate model parameters. By contrast, the forward search is based on the following idea: we build up an initial subset of a few units and, step by step, we add one additional unit. The increasing subset is made up by those units that are the closest according to a predefined measure.

The crucial issue in applying the forward search to DEA can be summarized as follows:

- When we consider a statistical model, for example a regression model, we can estimate its parameters on a subset of data and compute fitting errors. Then, we can rank units according to these errors as in Atkinson and Riani (2000).
- When we focus on DEA, we compute efficiency scores which are not parameters to fit the model to data. Thus we cannot compute fitting errors.

In order to face these issue, as we anticipated, we rely on the Kuosmanen and Johnson (2010) model where DEA is expressed as a regression model and we can exploit d_k^{DEA} of Eq. (7) to rank *DMUs*. In what follows we describe how we propose to carry out the above mentioned three steps in the linear programming DEA environment.

3.1. Initializing, progressing and monitoring the search

The first issue in the forward search is to identify the initial subset. In other words, we need to find a subset S_b of size$b \ll n$ that is outlier free. According to Atkinson et al. (2004), we can obtain this initial subset by applying the following alternative procedures:

- Bivariate boxplots from peeling. A natural nonparametric way of finding a central region in two-dimensions is to use convex hull peeling. The output of peeling is a series of nested convex polygons (hulls) which might be fitted through *B*-spline curves to obtain smooth contours. In order to find a central part of the data, as described in Zani et al. (1998), a robust bivariate centroid is found based on the observations inside the inner region defined by the fitted splines. In this way both the efficiency property of the arithmetic mean and the natural trimming offered by the hulls are exploited. The initial subset S_b consists of the closest b units in the central part of the data.
- Bivariate boxplots from ellipses. The bivariate boxplots calculated from *B*-splines are over elaborate to find a central part of the data. Therefore it can be useful to exploit a simpler method in which ellipses with a robust centroid are fitted to the data as in Riani and Zani (1997). The robust centroid of the ellipse is found as the component wise median of the two variables in the scatterplot. The shape of the contours are based on a covariance matrix in which the univariate medians are used, but which is otherwise calculated in the usual way. In S_b there are the closest units to the centroid.

In both cases we obtain a subset S_b of size b which is outlier free. We need to stress that the choice of the initial subset does not affect the final results of the forward search. In other words, it is useful to provide an effective rule to start the search, but the last steps of the procedure, on which the most important findings are concentrated, are not affected by the choice of S_b.

In order to progress in the search, at each step m, Atkinson and Riani (2000) propose to use regression parameters estimated from the model fitted to the subset of size $(m-1)$ and compute regression errors. The subset S_m is made up by the m units with the lowest regression errors. By contrast, DEA does not supply parameters, but efficiency scores which derive from the comparison of all *DMUs* belonging to the monitored set. For this reason, in order to compute d_k^{DEA}, at step m, we need to consider not all n units, but a subset whose size is as close as possible to $(m-1)$. In what follows we show how to progress in DEA analysis, distinguishing between the subset of size $(b+1)$ and the generic subset of size $m > (b+1)$. The

way to progress is the same in both cases, but it is useful to point out how to build up the first subset after the initialization and, then, highlight how to obtain the generic S_m.

- Subset $S_b + 1$. We apply DEA mechanics to the subset S_b obtaining, for each $DMU_k \in S_b$, the super-efficiency score $\psi_k^{SE,LR}$. Applying Eq. (7) we estimate d_k^{DEA} for all $k \in S_b$. For units not belonging to the subset S_b, we need to consider subsets whose size is as close as possible to b, then we focus on subsets of size $(b+1)$. In order to compute d_k^{DEA} for $k \notin S_b$ we consider a subset obtained through the union of S_b and DMU_k. Considering this subset, we estimate $\psi_k^{SE,LR}$ which is exploited in Eq. (7) to obtain d_k^{DEA}. We repeat the same analysis for all $(n-b)$ units not belonging to S_b. After this computation, we have n distances $d_k^{DEA}(k = 1, \ldots, n)$ and the subset S_{b+1} is made up by the $(b+1)$ *DMUs* with the lowest d_k^{DEA}.
- Subset of size $m > (b+1)$. At each step m of the search we perform the procedure described above, but in this case we consider *DMUs* belonging to the subset S_{m-1} to compute d_k^{DEA} for all $k \in S_{m-1}$. For *DMUs* not belonging to S_{m-1} we add the unit DMU_k to S_{m-1} and we compute d_k^{DEA}. We carry out the same for all $k \notin S_{m-1}$. The subset S_m is made up by the m *DMUs* with lowest d_k^{DEA}. We continue in this way for all steps until all *DMUs* are included in the subset.

As described above, the key contribution of our research is to specify a measure, d_k^{DEA}, which allows us to perform the search and monitor distances among *DMUs* detecting the presence of data structures and outliers.

A major advantage of the forward search is to provide the user with a number of informative pictures displaying all the diagnostics computed along the search. Stressing that the subset S_m is made up by the m units with the lowest d_k^{DEA}, at each step of the search, it is useful to monitor $d^{DEA,\max}(m)$, the maximum distance for units included in the subset, and $d^{DEA,\min}(m)$, the minimum distance of units out of the subset. We use this notation to emphasize that the distance is a function of the subset size m and we focus, in the first case, on the maximum of this distance for units belonging to S_m and, in the second case, on the minimum for units not belonging to S_m. Drawing a point-wise curve of both these distances along the search, we expect that the inclusion of an atypical unit into the subset causes a jump in $d^{DEA,\min}$ at step $(m-1)$ while $d^{DEA,\max}$ is expected to jump at step m (Bellini, 2010). In order to make inference about outliers, we need to compare these curves to statistical boundaries which, in the forward search literature, are called envelopes.

Envelopes are lower and upper bounds within which, at a certain confidence level, $d^{DEA,\max}(m)$ and $d^{DEA,\min}(m)$ are likely to stay. The crossing of envelopes highlights the presence of atypical units. As detailed in Section 4.2, we build up envelopes through Monte Carlo simulations. We apply the forward search to each of the N simulated realizations of $(Y, \mathbf{X})$, obtained from a stochastic process which imitates the variables of the observed dataset, and we compute, for each step m, lower and upper percentiles of the maximum distance into the subset and the minimum distance outside the subset. For each step of the search, envelopes are the collection of lower and upper percentiles of $d^{DEA,\max}(m)$ and $d^{DEA,\min}(m)$ computed on the N simulated realizations of $(Y, \mathbf{X})$. From a graphical point of view, we invite the reader to have an idea of these distances compared to their envelopes in Figs. 5 and 6, where we apply the forward search to a Cobb–Douglass production function.

Once the main steps of the forward search are defined, in the next section, we compare the super-efficiency procedure, developed according to Lovell and Rouse (2003), and our forward search procedure applied on a simulated Cobb–Douglas production function.

4. Super-efficiency and forward search analysis on a simulated production function

We apply our analysis to the simulation environment exploited by Banker and Chang (2006). In particular, we consider a production technology with a single output and two inputs that can be represented through the following Cobb–Douglas production function:

$$Z_k = (X_{k,1} - \alpha_{k,1})^{\beta_{k,1}} (X_{k,2} - \alpha_{k,2})^{\beta_{k,2}}, \tag{9}$$

where, as in Banker and Chang (2006), $\alpha_{k,1} = \alpha_{k,2} = 5$ for all k. X_1, X_2 are generated randomly from independent uniform distributions in the interval $[10, 20]$ and β_1, β_2 are simulated from independent uniform distributions in the interval $[0.4, 0.5]$. Since the sum of β_1 and β_2 is less than one, the production function in (9) satisfies the BCC assumption of concavity, while the shifts α_1, $\alpha_2 > 0$ allow both increasing and decreasing returns to scale.

Once Z_k has been generated as the efficient output for DMU_k, we compute the real output Y_k as follows:

$$Y_k = \frac{Z_k}{exp(u_k)}, \tag{10}$$

where $u_k = \ln \theta_k$. We generate randomly u_k from a half normal distribution $|N(0, \sigma^2)|$ allowing $\theta_k \geqslant 1$. In order to simplify the generation of Y_k, without loss of generality, we assume $\sigma^2 = 1$. Banker and Chang (2006) concentrate on detailed simulation experiments aiming to evaluate the performance of super-efficiency DEA in both ranking efficient units and identifying outliers, while we focus only on this latter goal. Furthermore, it is useful to stress that they find that the ranking procedure does not perform in a satisfactory way. In fact, the correlations between the true efficiency and the estimated super-efficiency are negative for the subset of efficient observations. We carry out the same analysis simulating $N = 1000$ times the above described model with $n = 50$. For observations which have BCC efficiency = 1, we obtain the results of Table 1.

When we consider all observations, as it is evident from Table 2, we obtain high correlations as do Banker and Chang (2006). This is particularly true in the case where we focus on the BCC approach, while the mean and median of the Pearson correlations between the true and the super-efficiency scores according to Lovell and Rouse (2003) are much smaller than the corresponding Spearman correlations. The Pearson (product moment) correlation coefficient is a measure of the linear relationship between two quantities (in this case, true and super-efficiency scores) while Spearman rank correlation coefficient is a measure of the *monotone* relationship

Table 1
Means and medians of correlation coefficients between the true and the super-efficiency scores for observations that have BCC efficiency = 1.

	Mean	Median
Pearson correlation	−0.524	−0.552
Spearman correlation	−0.316	−0.357

Table 2
Means and medians of correlation coefficients between the true and the super-efficiency scores for all observations considering both the standard BCC and the Lovell and Rouse (2003) super-efficiency models.

	BCC		Super-Eff^{LR}	
	Mean	Median	Mean	Median
Pearson correlation	0.781	0.785	0.248	0.261
Spearman correlation	0.723	0.724	0.701	0.702

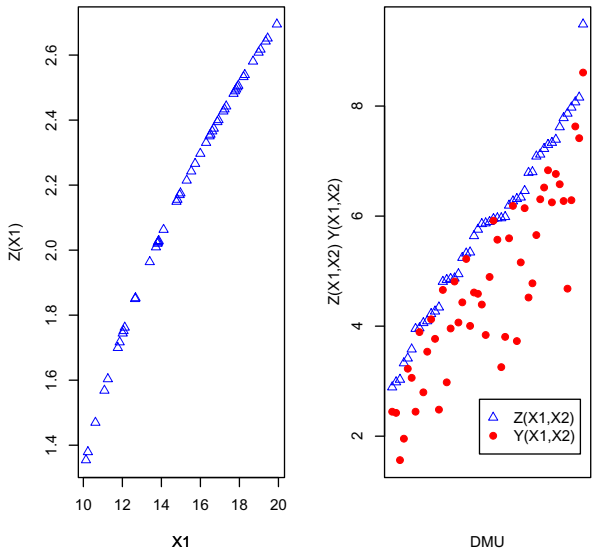

Fig. 1. Cobb–Douglas production function. The left panel shows the shape of Z_k as a function of $X_{k,1}$ when we consider as fixed $\beta_{k,1} = 0.4$, for all k, and $(X_{k,2} - \alpha_{k,2})^{\beta_{k,2}} = 1$, for all k. In the right panel the simulated efficient outputs Z_k of Eq. (9) are compared with their corresponding real outputs Y_k of Eq. (10). In this latter panel, we show Z_k and Y_k according to Z_k in ascending order.

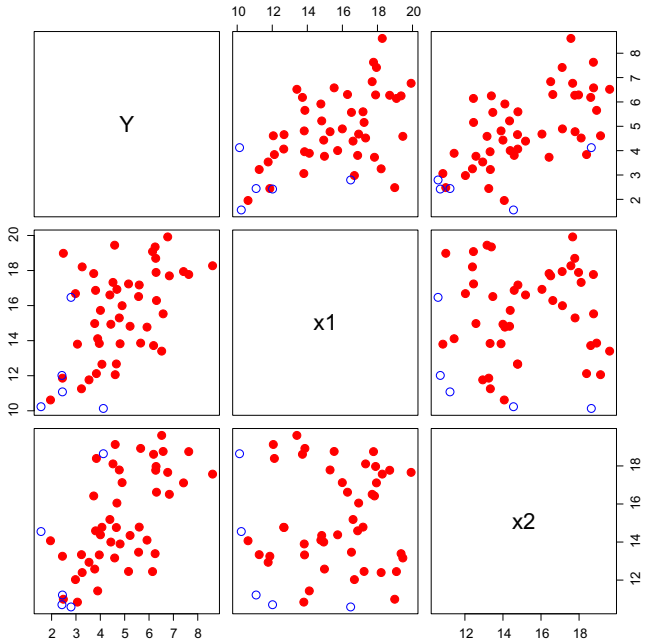

Fig. 2. Super-efficiency analysis: scatterplot matrix for the uncontaminated setting. *DMUs* with super-efficiency score corresponding to the upper level $\psi_k^{SE,LR} = 2.96$ are depicted through open circles, while non super-efficient *DMUs* are depicted through solid points.

between two quantities. As described below Eq. (3), the score $\psi_k^{SE,LR} = \left(\theta_k^{SI,LR}\gamma\right)^{-1}$ assumes high values for super-efficient units. This mechanics affects the product moment correlation coefficient, while the Spearman ones is not influenced by high super-efficiency scores.

In the next sections we concentrate on a single generation of Y_k, $k = 1,\ldots,50$, to which we apply the super-efficiency DEA and the forward search. In particular, in order to check how these procedures work, we obtain, for each DMU_k, the efficient output Z_k of Eq. (9) and the real output Y_k of Eq. (10) as depicted in Fig. 1.

4.1. The super-efficiency procedure for outlier detection applied to a simulated Cobb–Douglas production function

We start our empirical analysis focusing on the real output Y_k depicted on the right panel of Fig. 1. We consider this setting as uncontaminated because it is obtained through the mechanics described in the previous section without introducing exogenous contaminations. Fig. 2 shows that, when applying the Lovell and Rouse (2003) algorithm to this uncontaminated setting, five *DMUs* reach the upper super-efficiency level $\left(\psi_k^{SE,LR} = 2.96\right)$.

According to Banker and Chang (2006), it is necessary to define a threshold to verify whether these *DMUs* can be considered as outliers. It is interesting to notice that these five *DMUs* are generated through the same parameter set as all n *DMUs*; they cannot be considered as atypical *DMUs*.

In order to further check the robustness of the super-efficiency procedure in outlier detection, we introduce extreme contamination. In particular, we exogenously define the output and inputs values of DMU_{20}, DMU_{21} and DMU_{22} as follows: $Y_{20} = 0.5$, $Y_{21} = 0.7$, $Y_{22} = 0.55$ and $X1_{20} = X2_{20} = 4$, $X1_{21} = X2_{21} = 5$ and $X1_{22} = X2_{22} = 6$. Fig. 3 shows that only DMU_{20} lies on the super-efficiency frontier showing an evident masking effect. In other words, DMU_{20}, DMU_{21} and DMU_{22} are outliers, but the super-efficiency procedure detects the most extreme DMU_{20} with $\psi_{20}^{SE,LR} = 5.98$, while DMU_{21} and DMU_{22} with $\psi_{21}^{SE,LR} = 0.65$ and $\psi_{22}^{SE,LR} = 0.33$ are not super-efficient and are not identified as atypical *DMUs*.

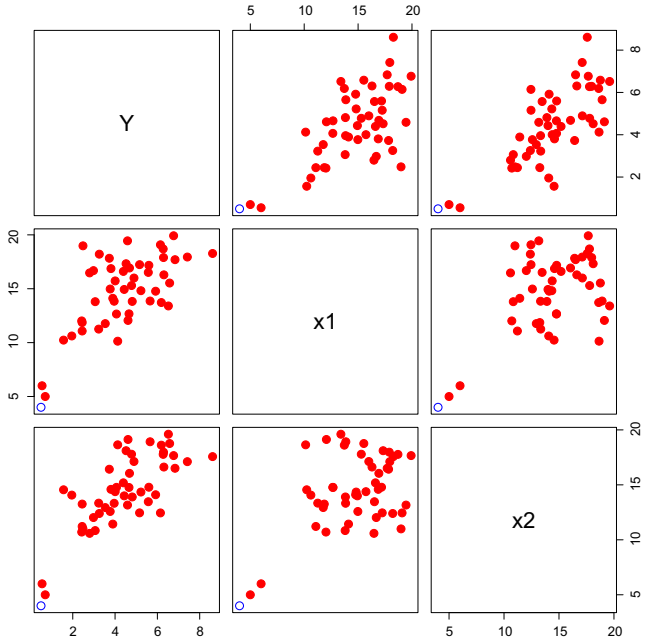

Fig. 3. Super-efficiency analysis: scatterplot matrix for the contaminated setting. DMU_{20}, DMU_{21} and DMU_{22} are exogenously contaminated on both the output and inputs sides ($Y_{20} = 0.5$, $Y_{21} = 0.7$, $Y_{22} = 0.55$ and $X1_{20} = X2_{20} = 4$, $X1_{21} = X2_{21} = 5$, $X1_{22} = X2_{22} = 6$). They constitute a separated cluster, but the super-efficiency procedure highlights only DMU_{20} with $\psi_{20}^{SE,LR} = 5.98$ (circle with an open white interior), while $\psi_{21}^{SE,LR} = 0.65, \psi_{22}^{SE,LR} = 0.33$ are not super-efficient (solid points). There is a masking phenomenon.

4.2. The forward search applied to a simulated Cobb–Douglas production function

We apply the steps of the procedure outlined in Section 3. The first task is to find an outlier free initial subset. We carry out both

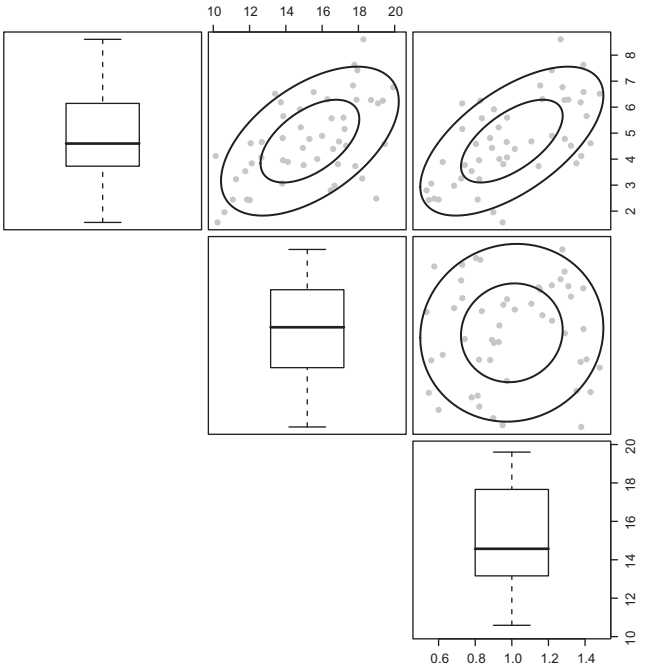

Fig. 4. Bivariate boxplots from ellipses: uncontaminated setting. Pairs plot with sumperimposed contours of bivariate boxplots calculated from *B*-splines.

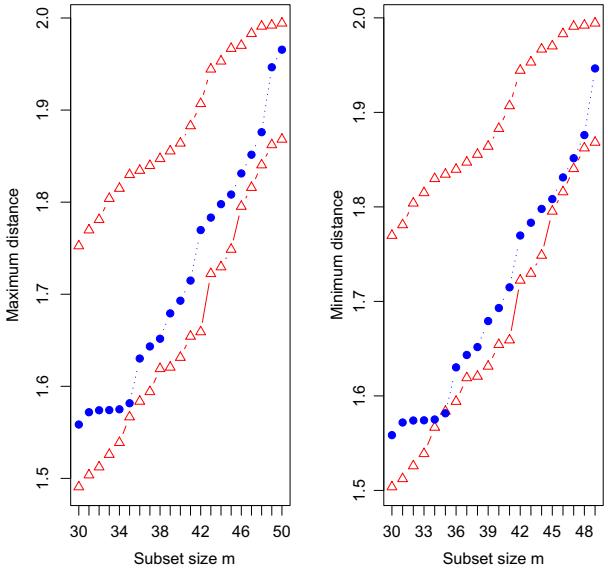

Fig. 5. Forward search $d^{DEA,\max}(m)$,$d^{DEA,\min}(m)$ and envelopes: uncontaminated setting. Maximum distance within the subset $d^{DEA,\max}(m)$ (solid points on the left panel) and minimum distance out of the subset $d^{DEA,\min}(m)$ (solid points on the right panel) compared to their 1% and 99% envelopes (triangles). It is evident that, in this uncontaminated setting, both distances stay within envelopes. As we expected, no outliers are detected.

initialization procedures described in Section 3.1 and, because of its simplicity, we decide to exploit the bivariate boxplots from ellipses as shown in Fig. 4. As we stressed above, the initialization procedure does not affect the final steps of the search where the main results are concentrated, but an effective start quickens the search. Other random procedures have been exploited and we obtained the same final forward search results.

For each subset of size m, we concentrate on monitoring the maximum distance within the subset $d^{DEA,\max}(m)$ and the minimum distance out of the subset $d^{DEA,\min}(m)$.

When we carry out the forward search, we need to find theoretical boundaries for inferences about outliers, for which we generate envelopes. As we mentioned above, we apply Monte Carlo simulations in order to obtain lower and upper envelopes for both the maximum distance within the subset and the minimum distance outside the subset of size m. The procedure can be summarized as follows:

- We simulate N times inputs $X_{k,1}$, $X_{k,2}$ and parameters $\beta_{k,1}$, $\beta_{k,2}$ as described in previous sections.
- We simulate N times the production output of Eq. (10).
- We run the forward search on each of these N simulations obtaining N distributions of the maximum distance within the subset and minimum distance out of the subset.
- At each subset size, we order these N distances $d^{DEA,\max}(m)$, $d^{DEA,\min}(m)$ and we consider the lower and upper percentiles obtaining point-wise curves which represent lower and upper envelopes within which the distance observed on the real simulation is likely to stay at a certain confidence level (Bellini and Riani, 2011).

In Fig. 5, we plot both the maximum distance within the subset and the minimum distance out of the subset with their envelopes. We consider the 1st and the 99th percentiles. As we expected, Fig. 5 shows that all *DMUs* stay within envelopes.

Unlike Banker and Chang (2006) who need to define a subjective threshold to state whether there are atypical observations, our procedure highlights that at the specified confidence level there are no outliers. In order to check the robustness of our procedure and compare our approach to the super-efficiency examined in Section 4.1, which substantially corresponds to the approach of Banker and Chang (2006), we introduce the same contaminations as above on DMU_{20}, DMU_{21} and DMU_{22}. We carry out the forward search on this contaminated setting. We execute the same steps as described in the uncontaminated setting obtaining the distances $d^{DEA,\max}(m)$ and $d^{DEA,\min}(m)$ for the contaminated environment. Therefore, we exploit Monte Carlo simulations in order to obtain lower and upper envelopes for both $d^{DEA,\max}(m)$ and $d^{DEA,\min}(m)$. Fig. 6 shows that all contaminated units stay out of the envelopes highlighting that our procedure is effective in detecting atypical *DMUs*. We further superimpose envelopes and we confirm that the outliers are correctly identified.

According to what we showed above, we can summarize that, by comparing the Banker and Chang (2006) approach to our procedure, the forward search detects outliers without introducing subjective evaluations. In fact when we carry out Monte Carlo simulations to build up envelopes, which are thresholds to make inference on outliers, we do not introduce any subjectivity. At the same time, as we highlighted in the contaminated environment, the Banker and Chang (2006) super-efficiency method is subject to masking. In fact when we contaminate three *DMUs*, only one of them is identified as an outlier, while our forward search correctly detects all atypical units.

In order to further verify our results, in the next section we carry out a correlation analysis between the true efficiency and the estimated ones in both the uncontaminated and the contaminated settings.

4.3. Correlation analysis on the simulated Cobb–Douglas production function

In order to complete the correlation study carried out in Section 4, we concentrate on the simulated setting in which we developed both the super-efficiency and the forward search analysis. In the uncontaminated setting, as is shown in Table 3, we obtain results

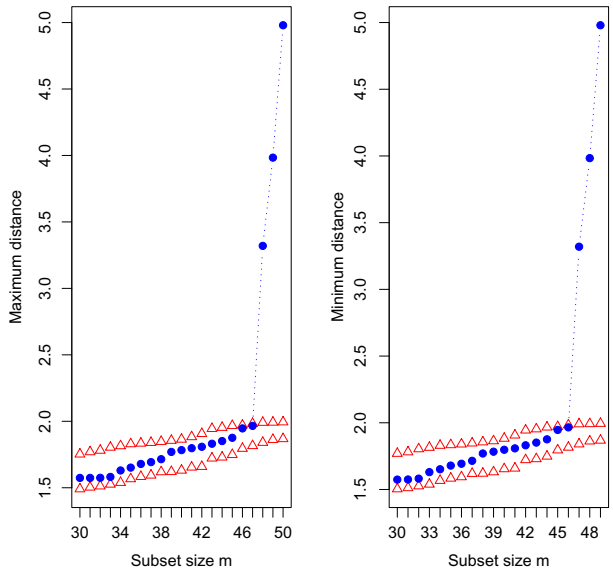

Fig. 6. Forward search $d^{DEA,\max}(m)$,$d^{DEA,\min}(m)$ and envelopes: contaminated setting. Maximum distance within the subset $d^{DEA,\max}(m)$ (solid points on the left panel) and minimum distance out of the subset $d^{DEA,\min}(m)$ (solid points on the right panel) compared to their 1% and 99% envelopes (triangles). It is evident that, in this contaminated setting, envelopes are crossed on the last steps of the forward search. Contaminated DMU_{20}, DMU_{21} and DMU_{22} are correctly detected.

Table 3
Correlation between the true and the estimated efficiency in the uncontaminated setting. In the first and second columns all *DMUs* are considered while in the third column the super-efficiency is computed considering only units with BCC efficiency = 1.

	BCC_{All}	Super-Eff^{LR}_{All}	Super-$Eff^{LR}_{Eff=1}$
Pearson correlation	0.674	0.076	−0.572
Spearman correlation	0.661	0.604	−0.496

Table 4
Correlation between the true and the estimated BCC efficiency in the contaminated setting. In the first column the analysis is carried out considering all *DMUs*. In the second column, according to the super-efficiency approach, only DMU_{20} is excluded. In the third column, according to the forward search, DMU_{20}, DMU_{21} and DMU_{22} are excluded from the estimation of correlation highlighting the increase of both Pearson and Spearman coefficients.

	DMU_{All}	$DMU_{All\neq 20}$	$DMU_{All\neq 20,21,22}$
Pearson correlation	0.495	0.534	0.668
Spearman correlation	0.682	0.714	0.723

aligned with those achieved on the 1000 simulations described above. As in the super-efficiency analysis of Table 2, Pearson correlation coefficient is affected by high super-efficiency scores while Spearman, measuring the monotone relationship between two quantities, is not affected by these large values.

In the contaminated setting, considering all units, when we apply the Lovell and Rouse (2003) super-efficiency approach, we obtain negative Pearson correlation coefficients while the Spearman ones are positive, but very small in absolute value. For this reason, hereafter, we will focus on BCC scores. Table 4 highlights that excluding DMU_{20}, the outlier detected through the Banker and Chang (2006) approach, the correlation increases when compared to that computed on all *DMUs*. However, in the case where we exclude all contaminated units (DMU_{20}, DMU_{21} and DMU_{22}) correctly detected through the forward search, there is an additional increase of the correlation between the true and the estimated efficiencies.

In order to detect atypical units, DEA users can exploit the super-efficiency analysis proposed by Banker and Chang (2006), eventually using the Lovell and Rouse (2003) approach. In addition, as we have emphasized, it can be very useful to carry out the forward search. Backward and forward analyses are not opposite one to the other. Indeed, the forward search complements the super-efficiency approach highlighting masked atypical *DMUs* when the underlying production function is known.

5. Concluding remarks

We are the first to extend to DEA the forward search originally proposed by Atkinson and Riani (2000). The forward search is based on the monitoring of unit closeness as a function of the sample size. DEA, however, does not provide parameters to fit the model to data. Then, the key issue of our research is to define a framework through which to bring a nonparametric linear programming analysis to a statistical fitting model. We integrate the linear regression DEA framework proposed by Kuosmanen and Johnson (2010) and the super-efficiency procedure of Lovell and Rouse (2003). We obtain a distance to monitor along the search in both the output and input oriented environments also considering multiple outputs and inputs. In addition, to aid inference on outliers, we propose a Monte Carlo procedure to obtain envelopes.

We conduct our research on data simulated from a Cobb–Douglas production function with one output and two inputs. We start by considering 1000 simulations on which we study the correlation between the true efficiency and the estimated ones. As do Banker and Chang (2006), we can state that the super-efficiency algorithm is not very useful for ranking efficient units. Thus, we concentrate on an individual simulated scenario. We carry out both the super-efficiency DEA and the forward search analysis in uncontaminated and contaminated settings.

The super-efficiency procedure proposed by Banker and Chang (2006) suggests a need to identify a subjective threshold to classify *DMUs* as outliers and to consider the potential effect of masking. By concentrating on the comparison of the maximum distance within the subset and the minimum distance outside the subset, our forward approach allows us to avoid subjectivity in outlier identification and is not affected by masking.

This work is the first step in forward search DEA analysis. Further studies need to be devoted to real data analysis where the functional relationship between outputs and inputs is not exogenously specified as in our research and envelopes cannot be generated from a given function as in our analysis.

Acknowledgments

We are grateful to an Editor for very constructive suggestions, to two anonymous referees and to Professor Anthony C. Atkinson for valuable comments on earlier drafts.

References

Atkinson, A.C., Riani, M., 2000. Robust Diagnostic Regression Analysis. Springer-Verlag, New York.
Atkinson, A.C., Riani, M., Cerioli, A., 2004. Exploring Multivariate Data with the Forward Search. Springer-Verlag, New York.
Banker, R.D., Gifford, J.L., 1988. A Relative Efficiency Model for the Evaluation of Public Health Nurse Productivity. Carnegie Mellon University Mimeo, Pittsburg, PA, USA.
Banker, R.D., Chang, H., 2006. The super-efficiency procedure for outlier identification, not for ranking efficient units. European Journal of Operational Research 175, 1311–1320.
Banker, R.D., Charnes, A., Cooper, W.W., 1984. Some models for estimating technical and scale efficiencies in data envelopment analysis. Management Science 30, 1078–1092.

Bellini, T., 2010. Detecting atypical observations in financial data: the forward search for elliptical copulas. Advances in Data Analysis and Classification 4, 287–299.
Bellini, T., Riani, M., 2011. Robust analysis of default intensity. Computational Statistics and Data Analysis, in press. doi:10.1016/j.csda.2011.03.007.
Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429–444.
Kuosmanen, T., Johnson, A.L., 2010. Data envelopment analysis as nonparametric least squares regression. Operations Research 58 (1), 149–160.
Lovell, C.A.K., Rouse, A.P.B., 2003. Equivalent standard DEA models to provide super-efficiency scores. Journal of the Operational Research Society 54, 101–108.
Riani, M., Zani, S., 1997. An iterative method for the detection of multivariate outliers. Metron 55, 101–117.
Zani, S., Riani, M., Corbellini, A., 1998. Robust bivariate boxplots and multiple outlier detection. Computational Statistics and Data Analysis 24, 257–270.