

Robust Data Envelopment Analysis

Tiziano Bellini

Abstract Data envelopment analysis (DEA) is a nonstochastic and nonparametric linear programming technique where a set of units are evaluated according to their input consumption and output production (Charnes *et al.* (1978)). Given that DEA efficiency analysis can be influenced by the presence of outliers, Banker and Chang (2006) proposed a method to detect atypical units through the super-efficiency model (Banker and Gifford (1988)). When relying on super-efficiency, masking and swamping can affect the analysis. In order to overwhelm these drawbacks, we propose a DEA based on the forward search framework originally introduced by Atkinson and Riani (2000). We carry out the analysis on a simulated Cobb-Douglas production function and we contaminate the dataset checking the robustness of our statistical model. Stressing the importance to find theoretical boundaries for the inference on outliers, we build up envelopes through Monte Carlo simulations.

Key words: Data envelopment analysis (DEA), statistical modeling, forward search, outlier detection.

1 Introduction

The main idea of DEA is to identify the production frontier on which the decision making units (*DMUs*) are considered as efficient. This frontier can be influenced by atypical observations, therefore it is crucial to check for their presence. In order to detect outliers Banker and Chang (2006) exploit the backward super-efficiency model, proposed by Banker and Gifford (1988), where each observation is excluded from its own reference set when efficiency score is computed. This method identifies as outlier the observations whose super-efficiency score exceeds a pre-specified score level. We show that this approach can be affected by masking and swamping.

Universita' di Parma, Via Kennedy 6, I-43100, Parma Italy, e-mail: tiziano.bellini@unipr.it

In order to overwhelm these drawbacks we extend the forward search procedure monitoring the quantities of interest as a function of the sample size.

The first task to accomplish in the forward search analysis is the definition of an opportune measure of unit closeness. It would seem natural to use directly DEA scores, but scores depend on units belonging to the subset, then an exogenous rule would be required to build up subsets of increasing size. However, if we use an exogenous rule, the forward search loses its strength.

In linear regression, Atkinson and Riani (2000) base the search on *error* estimates. But in DEA we cannot directly apply this framework. Then, our main contribution is to interpret DEA as a linear regression issue, as Kuosmanen and Johnson (2010), and exploit the difference (*error*) between observed output values and their estimates in order to achieve the goal of detecting outliers. For each *DMU* we define a relative distance which depends on DEA efficiency scores (high scores imply high errors) and in order to emphasize the magnitude of errors, we exploit the super-efficiency procedure proposed by Lovell and Rouse (2003).

Another important facet of our analysis is to make inference on outliers. If Banker and Chang (2006) do not supply any specific rule to define the super-efficiency level crossing which *DMUs* can be considered as atypical units, we propose a procedure to build up envelopes as statistical inference boundaries.

We conduct our DEA analysis on a Cobb-Douglas simulated function applying the BCC model (Banker *et al.* (1984)) for variable return of scale. In order to verify whether our procedure is effective, we introduce some contaminations and we compare our findings to the results deriving from the application of the super-efficiency procedure.

The paper is organised as follows. In section 2 we describe DEA mechanics and we show our statistical modeling based on the forward search approach. In section 3 we apply our framework to a simulated Cobb-Douglas production function comparing super-efficiency and forward search DEA analysis. Section 4 contains concluding remarks and directions for future research.

2 DEA forward search

DEA was introduced by Charnes *et al.* (1978) extending the idea of estimating technical efficiency with respect to a production frontier. Considering that DMU_k ($k = 1, \dots, n$) has a consumption $X_{k,i}$ of input i ($i = 1, \dots, m$) and produces $Y_{k,j}$ of output j ($j = 1, \dots, p$), we wish to evaluate the efficiency score of DMU_k among $DMU_1, \dots, DMU_n$. Under a variable returns to scale assumption¹, we have to solve, for each DMU_k and considering $i \in 1, \dots, m$ and $j \in 1, \dots, p$, the following linear program

¹ We focus on the variable return of scale assumption, but our analysis can be extended to other scale return assumptions.

$$\max \{ \theta_k \mid \theta_k Y_{k,j} \leq \sum_{l=1}^n \lambda_l Y_{l,j}; \ X_{k,i} \geq \sum_{l=1}^n \lambda_l X_{l,i}; \ \sum_{l=1}^n \lambda_l = 1; \ \lambda_l \geq 0 \}, \quad (1)$$

where θ_k is a scalar which represents the efficiency of DMU_k and λ_l are linear programming weights.

Starting from this model, Banker and Chang (2006) proposed the super-efficiency DEA as an effective procedure for outlier detection. In super-efficiency DEA each observation is excluded from its own reference set when its efficiency is estimated.

Once introduced the traditional and the super-efficiency DEA analysis, it is now interesting to underline the theoretical link between DEA and linear regression. In the case where we consider only one output ($p = 1$), under assumption of monotonicity and concavity, we can consider the following linear program

$$f^{DEA}(\mathbf{X}) = \max \{ Y \mid Y = \sum_{l=1}^n \lambda_l Y_l; \ X_i \geq \sum_{l=1}^n \lambda_l X_{l,i}; \ \sum_{l=1}^n \lambda_l = 1; \ \lambda_l \geq 0 \}, \quad (2)$$

where Y is the production function to be maximized. According to Kuosmanen and Johnson (2010), in the case where we substitute in equation system (2) the estimator of equation system (1), we notice that the DEA efficiency estimate ϵ_k^{DEA} can be considered as linear regression *error* and is obtained, for each DMU_k , as the optimal solution to the following linear program

$$\epsilon_k^{DEA} = \min \{ \epsilon_k \mid Y_k = \sum_{l=1}^n \lambda_l Y_l + \epsilon_k; \ X_{k,i} \geq \sum_{l=1}^n \lambda_l X_{l,i}; \ \sum_{l=1}^n \lambda_l = 1; \ \lambda_l \geq 0 \}. \quad (3)$$

We can therefore underline that, in the single-output setting, the DEA approach described on the equation system (3) is equivalent to the standard output oriented variable return of scale DEA in the sense that for each DMU_k we have

$$\begin{aligned} f_k^{DEA}(\mathbf{X}) &= \theta_k Y_k, \\ \epsilon_k^{DEA} &= Y_k - f_k^{DEA}(\mathbf{X}) = Y_k - \theta_k Y_k, \end{aligned} \quad (4)$$

where ϵ_k^{DEA} is the optimal solution of equation system (3) for DMU_k , θ_k is the optimal solution of equation system (1) and $\mathbf{X}$ is the matrix of m inputs. Then, we can define the following *relative* distance

$$d_k^{DEA} = \frac{|Y_k - \theta_k^{SE,LR} Y_k|}{Y_k}, \quad (5)$$

where $\theta_k^{SE,LR}$ is the optimal solution of the super-efficiency procedure proposed by Lovell and Rouse (2003). Equation (5) is crucial as measure of unit closeness in our forward search analysis and it allows to define an *endogenous* rule to progress the search ordering $DMUs$ at each subset size. It is interesting to notice that we do not exploit the traditional DEA model, but we use the super-efficiency framework in order to emphasize extreme scores.

Once defined an effective measure through which to carry out the search, it is now interesting to show in detail how to initialize, progress and monitor the search.

The first issue in our statistical modeling is to identify the so called initial subset. In other words, we need to find a subset S_b of size $b \ll n$ which contains units which can be considered as outlier free. In DEA analysis, the efficient frontier is defined considering the comparison among all units belonging to the set. We can concentrate on the following alternative procedures: bivariate boxplots from peeling and bivariate boxplots from ellipses. In both cases we find out a subset S_b of size b which can be considered as outlier free.

In order to progress the search we perform the following steps:

- Subset S_{b+1} . We apply DEA mechanics to the subset S_b obtaining, for each $DMU_k \in S_b$, the efficiency score $\theta_k^{SE,LR}$. Applying equation (5) we estimate $d_k^{DEA} \forall k \in S_b$. For units not belonging to the subset, we perform DEA analysis considering all units belonging to S_b and one unit not belonging to S_b . In other words, for all $(n - b)$ units not belonging to the initial subset, we consider subsets of size $(b + 1)$ made up by the subset S_b and the additional DMU_k for which we need to compute d_k^{DEA} . After this computation, we have n distances d_k^{DEA} ($k = 1, \dots, n$) and the subset S_{b+1} is made up by the $(b + 1)$ $DMUs$ with lowest d_k^{DEA} .
- Subset S_m . At each step m of the search we consider the $DMUs$ belonging to the subset S_{m-1} to compute $d_k^{DEA} \forall k \in S_{m-1}$. For $DMUs$ not belonging to S_{m-1} we add one unit DMU_k and we compute $d_k^{DEA} \forall k \notin S_{m-1}$. The subset S_m is made up by the m $DMUs$ with lowest d_k^{DEA} . We continue in such a way until all $DMUs$ are included into the subset.

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

Considering the need to find theoretical boundaries for the inference on outliers, we build up envelopes (Riani and Atkinson (2007)) through Monte Carlo simulations and we compare, at each subset size m , both the maximum distance within the subset $d_m^{DEA,max}$ and the minimum distance out of the subset $d_m^{DEA,min}$ with their corresponding envelopes. When envelopes are crossed it is likely, at a certain confidence level, that outliers have been included.

Once defined the main steps of the forward search, in the next section, we compare backward and forward analysis applied to a simulated Cobb-Douglas production function.

3 Super-Efficiency and Forward Search DEA Analysis on a Simulated Production Function

We apply our analysis to the same simulation environment exploited by Banker and Chang (2006). In particular, we consider a single output and two inputs production

technology which can be represented through a Cobb-Douglas production function specified as follows

$$Z = (X_1 - \alpha_1)^{\beta_1} (X_2 - \alpha_2)^{\beta_2}, \quad (6)$$

where, as in Banker and Chang (2006), $\alpha_1 = \alpha_2 = 0.5$, X_1, X_2 are generated randomly from independent uniform distributions into the interval $[10, 20]$ and β_1, β_2 are simulated from independent uniform distributions into the interval $[0.4, 0.5]$.

Once defined this function as the efficient frontier, we consider as DMU_k output

$$Y_k = \frac{Z(X_{k,1}, X_{k,2})}{\exp(u_k)}, \quad (7)$$

where $\frac{1}{\exp(u_k)}$ is the reciprocal of the efficiency score. In order to obtain this score, we generate randomly u_k from a half normal distribution $|N(0, 1)|$.

We can summarize our simulation procedure by stating what follows. First we simulate Z_k of equation (6) obtaining the efficient frontier, then we simulate u_k which is required to compute $\frac{1}{\exp(u_k)}$ and, at the end, we obtain Y_k as described in equation (7).

Once generated Y_k ($k \in 1, \dots, n$), we apply the super-efficiency as well as the forward search DEA described in the previous sections. In particular, in order to check how these procedures work, using $n=50$ we obtain the efficient frontier $Z(X_1, X_2)$ and the *real* frontier $Y(X_1, X_2)$ of equation (7). We define this random generation as *true* simulation.

We start our empirical analysis by focusing on the uncontaminated true simulation. As we can see from Figure 1, carrying out the super-efficiency DEA Lovell and Rouse (2003) algorithm in the uncontaminated true simulation, five *DMUs* reach the upper super-efficiency limit ($\theta_k^{SE,LR} = 2.96$). Following Banker and Chang (2006), it is necessary to define a score level in order to verify whether these *DMUs* can be considered as outliers. We notice that these *DMUs* are generated through the same parameter set as others and cannot be considered as atypical.

Considering the forward search statistical modeling described above, the first task is to find an outlier free initial subset. We carry out both initialization procedures described on section 2 and we decide to exploit the bivariate boxplots from ellipses because of its simplicity.

Once outlined how to carry out the initialization, it is now interesting to highlight that at each subset size m , we concentrate on monitoring the maximum distance within the subset $d_m^{DEA,max}$ and the minimum distance out of the subset $d_m^{DEA,min}$.

When we carry out the forward search, we need to find theoretical boundaries for the inference on outliers, then 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 out of the subset of size m .

We plot both the maximum distance within the subset and the minimum distance out of the subset with their envelopes in Figure 2. We consider percentiles: 1% and 99%.

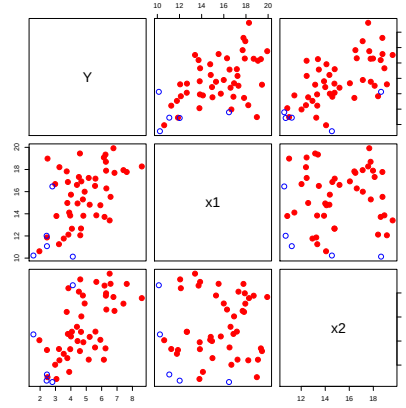


Fig. 1 Super-efficiency analysis: scatterplot matrix uncontaminated setting. *DMUs* with super-efficiency score corresponding to the upper limit $\theta_k^{SE,LR} = 2.96$ are depicted through filled circles while non super-efficient *DMUs* are depicted through solid points.

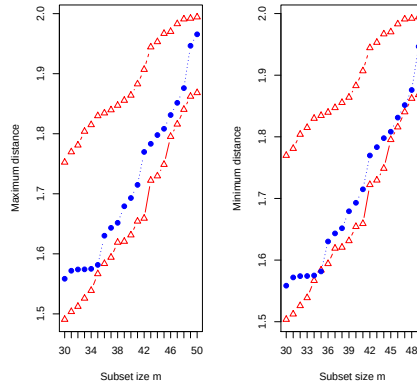


Fig. 2 True distances and envelopes: uncontaminated setting. Maximum distance within the subset $d_m^{DEA,max}$ (solid points on the left panel) and minimum distance out of the subset $d_m^{DEA,min}$ (solid points on the right panel) compared with their 1% and 99% envelopes (filled triangles). It is evident that, in this uncontaminated setting, both distances computed on the simulated function of equation (7) stay within envelopes. It means that no outliers are detected.

We notice from Figure 2 that all *DMUs* stay within envelopes. We can therefore state that our procedure does not detect any outlier.

In order to check the robustness of the super-efficiency procedure in outlier detection, we introduce an extreme contamination. In particular, we exogenously define the output and inputs values of DMU_{20} and DMU_{21} as follows: $Y_{20} = 0.5$, $Y_{21} = 0.7$ and $X1_{20} = X2_{20} = 4$, $X1_{21} = X2_{21} = 5$. Carrying out the same analysis as above,

we obtain that only DMU_{20} lies on the super-efficiency frontier showing an evident masking effect. In other words, both DMU_{20} and DMU_{21} are outliers, but the super-efficiency procedure detects the most extreme DMU_{20} with $\theta_{20}^{SE,LR} = 5.98$, while DMU_{21} has a super-efficiency score $\theta_{21}^{SE,LR} = 0.65$. DMU_{21} is a masked outlier.

We are now interested in carrying out the forward search on this contaminated setting. We execute the same steps as described in the uncontaminated setting obtaining the distance $d_m^{DEA,max}$ and the $d_m^{DEA,min}$ for the contaminated environment. Therefore, we exploit Monte Carlo simulations in order to obtain lower and upper envelopes for both $d_m^{DEA,max}$ and $d_m^{DEA,min}$. Figure 3 shows that all contaminated units (Y_{20}, Y_{21}) stay out of the envelopes showing that our procedure is effective in detecting atypical $DMUs$.

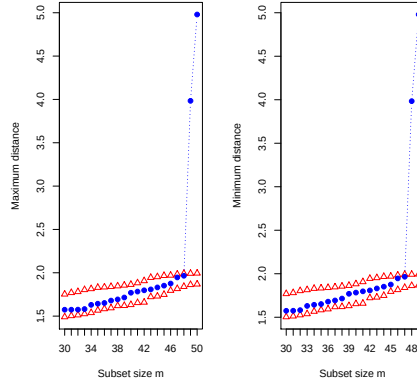


Fig. 3 True distances and envelopes: contaminated setting. Maximum distance within the subset $d_m^{DEA,max}$ (solid points on the left panel) and minimum distance out of the subset $d_m^{DEA,min}$ (solid points on the right panel) compared with their 1% and 99% envelopes (filled triangles). It is evident that, in this contaminated setting, both distances cross envelopes on the last steps of the forward search. Contaminated DMU_{20} and DMU_{21} are correctly detected.

4 Concluding Remarks

We are pioneer in extending the forward search originally proposed in linear and nonlinear regression by Atkinson and Riani (2000) to DEA analysis. Our main contribution, in order to start, progress and monitor the search, is to propose a distance obtained exploiting both the linear regression framework proposed by Kuosmanen and Johnson (2010) as well as the Lovell and Rouse (2003) super-efficiency DEA procedure. In addition, in order to make inference on outliers, we propose a statistical procedure to build up envelopes through Monte Carlo simulations.

We carry out the analysis on simulated data. We consider a Cobb-Douglas production function with one output and two inputs. We generate sub-optimal Cobb-Douglas frontier on which we apply the BCC model. We carry out both the super-efficiency and the forward search analysis on an uncontaminated setting as well as on an contaminated environment. We show that the super-efficiency procedure is affected by drawbacks such as the subjective score level to classify outlying *DMUs* as well as the masking. We perform the forward search and considering envelopes to make inference on atypical units. We show that our statistical model is not affected by subjective estimates and is effective in detecting outliers and data structures avoiding masking issues.

This work is the first step in forward search DEA analysis. We stress the importance to use effective measures of unit closeness and, we point out the need to base the analysis on robust envelopes. Our analysis is conducted into a single output setting and we follow a DEA output oriented approach. There is the need to extend the analysis to multi-output settings and input-oriented environment with multiple inputs and outputs. Further studies need to be devoted to real data where the functional relationship between outputs and inputs is not exogenously specified as in our research and envelopes generation cannot be obtained as in our example.

References

1. Atkinson, A.C., Riani, M.: Robust diagnostic regression analysis. Springer - Verlag, New York (2000)
2. Banker, R.D., Gifford, J.L.: A relative efficiency model for the evaluation of public health nurse productivity. Mellon University Mimeo, Carnegie (1988)
3. Banker, R.D., Chang, H.: The super-efficiency procedure for outlier identification, not for ranking efficient units. *European Journal of Operational Research*. **175**, 1311-1320 (2006)
4. Banker, R.D., Charnes, A., Cooper, W.W.: Some models for estimating technical and scale efficiencies in data envelopment analysis. *Management Science*. **30**, 1078-1092 (1984)
5. Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. *European Journal of Operational Research*. **2**, 429-444 (1978)
6. Kuosmanen, T., Johnson, A.L.: Data envelopment analysis as nonparametric least squares regression. *Operations Research*. Forthcoming (2010)
7. Lovell, C.A.K., Rouse, A.P.B.: Equivalent standard DEA models to provide super-efficiency scores. *Journal of the Operational Research Society*. **54**, 101-108 (2003)
8. Riani, M., Atkinson, A.C.: Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in Data Analysis and Classification*. **1**, 123-141 (2007)