Computers & Operations Research 37 (2010) 417 – 425

Contents lists available at ScienceDirect

Computers & Operations Research

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

A unified model for detecting efficient and inefficient outliers in data envelopment analysis

Wen-Chih Chen [a,*], Andrew L. Johnson [b]

[a] Department of Industrial Engineering and Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan
[b] Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA

ARTICLE INFO

Available online 21 June 2009

Keywords:
Data envelopment analysis
Outlier
Post analysis

ABSTRACT

Data envelopment analysis (DEA) uses extreme observations to identify superior performance, making it vulnerable to outliers. This paper develops a unified model to identify both efficient and inefficient outliers in DEA. Finding both types is important since many post analyses, after measuring efficiency, depend on the entire distribution of efficiency estimates. Thus, outliers that are distinguished by poor performance can significantly alter the results. Besides allowing the identification of outliers, the method described is consistent with a relaxed set of DEA axioms. Several examples demonstrate the need for identifying both efficient and inefficient outliers and the effectiveness of the proposed method. Applications of the model reveal that observations with low efficiency estimates are not necessarily outliers. In addition, a strategy to accelerate the computation is proposed that can apply to influential observation detection.

1. Introduction

Data Envelopment Analysis (DEA) introduced by Charnes et al. in [6] is a mathematical programming technique for evaluating the efficiency of an observation relative to a set of similar observations. Generally viewed as a success story for the operations research community [15], DEA's real-world relevance, diffusion, and global popularity are evident in literature such as Seiford et al. [27]. It has been applied variously to financial institutions (e.g., [10,14]), technology investment evaluation (e.g., [9]), among many other applications.

DEA efficiency estimates are quite sensitive to the presence of outliers since the method uses extreme observations to identify superior performance [28]. However, outliers can be difficult to identify, because each record describing an observation is typically a high-dimensional vector with multiple inputs and outputs. Some outliers are the results of measuring or recording errors, while others are the results of unusual characteristics, including factors related to the external environment, or uncontrollable factors. However, they can also be associated with low probabilities of occurrence. When the associated observations differ greatly from the remainder of the data set, the outliers can represent unexpected knowledge to be gained.

If the concept of outliers is intuitive, a rigorous definition is somewhat elusive. In the literature, outliers have been loosely defined as *an observation (or a set of observations) which appears to be inconsistent with the remainder of that set of data* [5]. In the efficiency estimation context, many authors term observations with significant influence on others' efficiency estimates as influential observations (e.g., [23,32]). An influential observation typically owes its influence to the fact that it is an outlier and supports part of the deterministic frontier. However, Pastor et al. [23] and Simar [30] observe that an outlier is not necessarily an influential observation, and that influential observation is not necessarily far away from the data cloud.

In the nonparametric efficiency analysis literature some studies with a particular interest in efficiency estimates attempt to detect influential observations (e.g., [23,25,32]), while others focus on detecting outliers removed from the data cloud (e.g., [13,31]). From the perspective of the theories utilized, still others (e.g., [23,32,25]) use estimates of DEA efficiency and the frontier concept. Wilson [31] and Fox et al. [13] examine the dissimilarity of an input–output record to other observations. Simar [30] notes that these types of approaches *do not take the frontier aspects of the problem into account*. This is a significant limitation, because researchers are mostly interested in detecting overly efficient outliers with the most influence upon the efficiency measures. As a result, another direction of the outlier literature considers statistical inferences in DEA's nonparametric context (e.g., [30]).

This paper is motivated by a Web-based DEA benchmarking tool for warehouse operations [19]. Although the Internet allows researchers to collect data quickly and securely, the data may be entered in error, or may not represent an actual facility. These drawbacks increase the need for effective data filtering techniques that

* Corresponding author.
E-mail addresses: wenchih@faculty.nctu.edu.tw (W.-C. Chen), ajohnson@tamu.edu (Andrew L. Johnson).

0305-0548/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2009.06.010

can identify inefficient outliers, a growing concern for researchers. Performance benchmarking is a set of processes and practices used to determine (i) reference values for selected performance indices, and (ii) factors for key processes affecting performance [21]. Both goals rely on quality data, and the latter is also affected by inefficient outliers. When examining commercial pharmacies, Johnson and McGinnis [18] demonstrate the effect of inefficient outliers on the results of a second stage regression to identify attributes that correlate positively with efficiency. They find that a particular attribute, population of the surrounding area, is unrelated to store efficiency, but that it correlates positively with efficiency when the inefficient outliers are removed.

Inefficiency outliers are also an issue in post analyses using DEA efficiency estimates, e.g., statistical testing (determining whether two populations are equally efficient), cross-validation (using the weights selected by each observation to define a common set of weights in a post analysis), distribution analysis (determining whether DEA efficiency estimates are consistent with economic theories of efficiency of markets), benchmarking (identifying best and worse practices), industry trends analysis (identifying the efficient observations used as benchmarks for a large set of observations), etc.

Only Johnson and McGinnis [18] employ the "inefficient frontier" concept to detect possible outliers that perform poorly. But the "inefficient frontier" concept is *ad hoc*, and is not consistent with the standard DEA axioms. Production theory assumes that observations are bounded by those with superior performance, and that interior points (with respect to the efficient frontier) are always feasible. Simply applying existing procedures, e.g., Pastor et al. [23] to the inefficient observations violates standard axioms of DEA (production theory) and thus is logically problematic.

This paper aims to identify outliers that influence both efficiency estimates and DEA post analysis. We approach the problem by identifying a set of axioms and developing an approach consistent with the axioms. Adopting a relaxed set of DEA axioms allows the detection and ranking of both efficient outliers that influence the efficiency estimates and inefficient outliers that may influence post analysis procedures. Identifying either efficient or inefficient outliers separately is also possible. Applications of the model to real-world case studies demonstrate that intuitively flagging the worst-performing observations as inefficient outliers is not necessarily correct. In addition, a strategy to accelerate the computation is proposed that can be applied to influential observation detection such as Pastor et al. [23].

The remainder of this paper is organized as follows. The next section proposes new outlier measures taking efficiency influence into account to measure the effect of an outlier or a set of outliers. Section 3 describes four case studies that demonstrate the proposed method. Section 4 offers a computational remark, and Section 5 states the conclusions.

2. Outlier measures

This section introduces the fundamentals of DEA and proposes new outlier measures considering efficiency influence.

2.1. Fundamental

Consider an input set I and an output set J. Denote $\mathbf{x} \in \Re_+^{|I|}$ as an input vector and $\mathbf{y} \in \Re_+^{|J|}$ as an output vector. The production possibility set (PPS) T, representing the feasibility of transforming inputs to outputs, is defined as

$$T \equiv \{(\mathbf{x},\mathbf{y}) : \mathbf{y} \text{ can be produced by } \mathbf{x}\}.$$

Shephard [29] defines the output distance function ($D_o(\mathbf{x}',\mathbf{y}')$) and input distance function ($D_i(\mathbf{x}',\mathbf{y}')$) between any specific input–output bundle $(\mathbf{x}',\mathbf{y}')$ and boundary of T as follows:

$$D_o(\mathbf{x}',\mathbf{y}') \equiv \inf\{\alpha : (\mathbf{x}',\mathbf{y}'/\alpha) \in T\},$$

$$D_i(\mathbf{x}',\mathbf{y}') \equiv \sup\{\alpha : (\mathbf{x}'/\alpha,\mathbf{y}') \in T\}.$$

Distance functions measure how far to locate $(\mathbf{x}',\mathbf{y}')$ on the boundary of T by changing either its outputs or inputs proportionally. Any $(\mathbf{x}',\mathbf{y}')$ with a distance function of one is referred to as being on the boundary (frontier) of T.

In practice, T is unknown. Given a set of observations S with input–output vector $(\mathbf{x}_r,\mathbf{y}_r)$ $r \in S$, the empirical production possibility set (EPPS) can be approximated using the following axioms [4]:

1. (Convexity) If $(\mathbf{x},\mathbf{y}) \in T$ and $(\mathbf{x}',\mathbf{y}') \in T$, then $\lambda(\mathbf{x},\mathbf{y})+(1-\lambda)(\mathbf{x}',\mathbf{y}') \in T$, for $\lambda \in [0,1]$.
2. (Free disposability) $(\mathbf{x}',\mathbf{y}') \in T$, if $(\mathbf{x},\mathbf{y}) \in T$, $\mathbf{x}' \geq \mathbf{x}$ and $\mathbf{y}' \leq \mathbf{y}$.

The EPPS can then be expressed as a set of linear inequalities in $|S|$ nonnegative variables and denoted as

$$\hat{T} \equiv \left\{(\mathbf{x},\mathbf{y}) : \sum_{r\in S}\mathbf{x}_r\lambda_r \leq \mathbf{x};\ \sum_{r\in S}\mathbf{y}_r\lambda_r \geq \mathbf{y};\ \sum_{r\in S}\lambda_r = 1;\ \lambda_r \geq 0,\ r \in S\right\}.$$

Then the practicable estimations for distance functions are as follows:

$$[\hat{D}_i(\mathbf{x}',\mathbf{y}')]^{-1} = \min\left\{\alpha : \sum_{r\in S}\mathbf{x}_r\lambda_r \leq \alpha\mathbf{x}';\ \sum_{r\in S}\mathbf{y}_r\lambda_r \geq \mathbf{y}';\ \sum_{r\in S}\lambda_r = 1;\ \lambda_r \geq 0,\ r \in S\right\},$$

$$[\hat{D}_o(\mathbf{x}',\mathbf{y}')]^{-1} = \max\left\{\beta : \sum_{r\in S}\mathbf{x}_r\lambda_r \leq \mathbf{x}';\ \sum_{r\in S}\mathbf{y}_r\lambda_r \geq \beta\mathbf{y}';\ \sum_{r\in S}\lambda_r = 1;\ \lambda_r \geq 0,\ r \in S\right\}.$$

If $(\mathbf{x}',\mathbf{y}')$ is observed and thus is feasible, the distance function measures can be interpreted as the technical efficiency [12][1] which estimates the relative efficiency for a particular record $k \in S$ comparing against all observations in S. This leads to one of the well-known DEA models proposed by Banker et al. in [4]:

$$\alpha_k^S = \min_{\alpha,\lambda}\left\{\alpha :\ \sum_{r\in S}\mathbf{x}_r\lambda_r \leq \alpha\mathbf{x}_k;\ \sum_{r\in S}\mathbf{y}_r\lambda_r \geq \mathbf{y}_k;\ \sum_{r\in S}\lambda_r{=}1;\ \lambda_r \geq 0, r \in S\right\}, \quad \text{(BCC.I)}$$

$$\beta_k^S = \max_{\beta,\lambda}\left\{\beta :\ \sum_{r\in S}\mathbf{x}_r\lambda_r \leq \mathbf{x}_k;\ \sum_{r\in S}\mathbf{y}_r\lambda_r \geq \beta\mathbf{y}_k;\ \sum_{r\in S}\lambda_r{=}1;\ \lambda_r \geq 0, r \in S\right\}. \quad \text{(BCC.O)}$$

In summary, (BCC.I) and (BCC.O) provide the radial efficiency estimates using $\hat{T}$, which is constructed by observations according to axioms of convexity and free disposability. Observation k is said to be efficient and on the efficient frontier when its efficiency estimate is one.

It should be noted that $\hat{T}$ assumes free disposability, which implies that using more inputs and producing less outputs is always feasible. Influential measures based on (BCC.I) and (BCC.O) (e.g., [23]) will not flag inefficient observations as outliers, regardless of how poorly they perform. Free disposability adopted in standard DEA models actually assumes that all inefficient observations are part of the

[1] In fact, it is the reciprocal of the distance function.

PPS and does not allow for the concept of inefficient outliers; however, this paper finds that inefficient outliers matter. Johnson and McGinnis [18] develop the idea of the "inefficient frontier" to flag overly inefficient observations; however, they do not approach the problem with consideration of DEA axioms and their model is thus not consistent with the free disposability axiom. This paper relaxes the assumption of free disposability and simply uses part of $\hat{T}$ based on convexity, allowing for potential outliers to be ranked based on their influence.

2.2. New measures of outliers

This section proposes an outlier measure that can identify both efficient and inefficient outliers. These outliers are measured relative to a set constructed consistent with a subset of DEA axioms, and the individual outliers are ranked based on their influence on the measures. For a data set S, adopting the convexity assumption, the convex hull of S is as follows:

$$\hat{T}^S_{conv} \equiv \left\{(\mathbf{x},\mathbf{y}) : \sum_{r\in S}\mathbf{x}_r\lambda_r = \mathbf{x};\ \sum_{r\in S}\mathbf{y}_r\lambda_r = \mathbf{y};\ \sum_{r\in S}\lambda_r = 1;\ \lambda_r \geq 0,\ r\in S\right\}.$$

$\hat{T}^S_{conv}$ is a part of $\hat{T}$ ($\hat{T}^S_{conv} \subset \hat{T}$). Extending the definition of free disposability, the free disposal hull of a set $A \subset \Re^{|I|}_+ \times \Re^{|J|}_+$ is [16]:

$$FDH(A) \equiv \{(\mathbf{x}',\mathbf{y}') : \mathbf{x}' \geq \mathbf{x}, \mathbf{y}' \leq \mathbf{y} \quad \text{for some } (\mathbf{x},\mathbf{y}) \in A\}.$$

It is clear that $\hat{T} = FDH(\hat{T}^S_{conv})$ [16], namely $\hat{T}^S_{conv}$ is an essential component of EPPS without applying free disposability that makes identifying inefficient outliers impossible. Therefore, $\hat{T}^S_{conv}$ characterizes most important properties of $\hat{T}$, and can be used to identify both efficient and inefficient outliers.

To identify outliers that influence both the efficiency estimates and DEA post analysis, radial measures with respect to $\hat{T}^S_{conv}$ are proposed. For output-oriented analyses, a measure η^S_k is defined as

$$\eta^S_k \equiv \max_{\eta}\{\eta : (\mathbf{x}_k, \eta\mathbf{y}_k) \in \hat{T}^S_{conv}\} = \max_{\eta,\lambda}\left\{\eta : \sum_{r\in S}\mathbf{x}_r\lambda_r = \mathbf{x}_k;\ \sum_{r\in S}\mathbf{y}_r\lambda_r = \eta\mathbf{y}_k;\ \sum_{r\in S}\lambda_r = 1;\ \lambda_r \geq 0,\ r\in S\right\}. \tag{1}$$

The value of η^S_k measures how much the outputs of observation k can be scaled up proportionately while remaining in $\hat{T}^S_{conv}$. $\eta^S_k \geq 1$ and $(\mathbf{x}_k, \eta\mathbf{y}_k) \notin \hat{T}^S_{conv}$ if $\eta > \eta^S_k$; the projected point $(\mathbf{x}_k, \eta^S_k\mathbf{y}_k)$ refers to as on the outer boundary. Further, η^S_k can be interpreted as the "distance" between k and the outer boundary; it is said that the "distance" to the outer boundary is $(100\times\eta^S_k)\%$ of $\mathbf{y}_k$. $\eta^S_k = 1$ suggests that k is on the outer boundary since it cannot be scaled up while maintaining in $\hat{T}^S_{conv}$. This radial measure is identical to the output-oriented efficiency estimate, but with respect to $\hat{T}^S_{conv}$, not $\hat{T}$. As a result, (1) has a structure similar to (BCC.O), and thus ties directly to efficiency estimation. Observations with output efficiency estimates equal to one will have $\eta^S_k = 1$; this is formally stated as follows:

Proposition 1. *For $k \in S$, $\eta^S_k = 1$ if $\beta^S_k = 1$.*

Proof. See the appendix. □

Another measure related to k, γ^S_k, is defined as

$$\gamma^S_k \equiv \min_{\gamma}\{\gamma : (\mathbf{x}_k, \gamma\mathbf{y}_k) \in \hat{T}^S_{conv}\} = \min_{\gamma,\lambda}\left\{\gamma : \sum_{r\in S}\mathbf{x}_r\lambda_r = \mathbf{x}_k;\ \sum_{r\in S}\mathbf{y}_r\lambda_r = \gamma\mathbf{y}_k;\ \sum_{r\in S}\lambda_r = 1;\ \lambda_r \geq 0,\ r\in S\right\}. \tag{2}$$

Eq. (2) has the same interpretation as (1) but scaling k in the opposite direction. $0 \leq \gamma^S_k \leq 1$ and $(\mathbf{x}_k, \gamma\mathbf{y}_k) \notin \hat{T}^S_{conv}$ if $\gamma < \gamma^S_k$. The projected point $(\mathbf{x}_k, \gamma^S_k\mathbf{y}_k)$ refers to is located on the inner boundary of $\hat{T}^S_{conv}$. Similarly, γ^S_k represents the distance between k and the inner boundary; it suggests the distance is $(100\times\gamma^S_k)\%$ of $\mathbf{y}_k$. $\gamma^S_k = 1$ suggests that k is on the inner boundary, and the movement passes through the output origin $(\mathbf{x}_k,\mathbf{0})$ when $\gamma^S_k = 0$.

The "difference" between projected points $(\mathbf{x}_k, \eta^S_k\mathbf{y}_k)$ and $(\mathbf{x}_k, \gamma^S_k\mathbf{y}_k)$ is the width of segment constructed by identifying a ray within $\hat{T}^S_{conv}$ from the (output) origin $(\mathbf{x}_k,\mathbf{0})$ through observation $k \in S$. To be precise, the "width" is defined as $(\eta^S_k\mathbf{y}_k - \gamma^S_k\mathbf{y}_k)/\mathbf{y}_k = \eta^S_k - \gamma^S_k$, which specifies the width as a percentage of $\mathbf{y}_k$. In a single-output example, suppose $y_k = 100$ and the projected points are 120 and 70. Therefore, $\eta^S_k = 1.2$ and $\gamma^S_k = 0.7$ results in $\eta^S_k - \gamma^S_k = 0.5$; it is consistent with the original units of measure.

When the observation set R is removed from S ($R \subset S$, $k \notin R$), the corresponding convex hull $\hat{T}^{S\backslash R}_{conv}$ may change. The associated measures are denoted as $\eta^{S\backslash R}_k$ and $\gamma^{S\backslash R}_k$, and are computed as follows:

$$\eta^{S\backslash R}_k \equiv \max_{\eta,\lambda}\left\{\eta : \sum_{r\in S\backslash R}\mathbf{x}_r\lambda_r = \mathbf{x}_k;\ \sum_{r\in S\backslash R}\mathbf{y}_r\lambda_r = \eta\mathbf{y}_k;\ \sum_{r\in S\backslash R}\lambda_r = 1;\ \lambda_r \geq 0,\ r\in S\backslash R\right\}, \tag{3}$$

$$\gamma^{S\backslash R}_k \equiv \min_{\gamma,\lambda}\left\{\gamma : \sum_{r\in S\backslash R}\mathbf{x}_r\lambda_r = \mathbf{x}_k;\ \sum_{r\in S\backslash R}\mathbf{y}_r\lambda_r = \gamma\mathbf{y}_k;\ \sum_{r\in S\backslash R}\lambda_r = 1;\ \lambda_r \geq 0,\ r\in S\backslash R\right\}. \tag{4}$$

Since $k \notin R$, (3) and (4) are always feasible. It is always possible to set $\lambda_k = 1$ and achieve a feasible solution. Without R, the width becomes $\eta^{S\backslash R}_k - \gamma^{S\backslash R}_k$. Accordingly, based on the above metrics, the width related to k changes from $\eta^S_k - \gamma^S_k$ to $\eta^{S\backslash R}_k - \gamma^{S\backslash R}_k$ after R is removed, and the influence on observation k due to R is measured as

$$\delta^{o+i}_k(R) \equiv (\eta^S_k - \gamma^S_k) - (\eta^{S\backslash R}_k - \gamma^{S\backslash R}_k). \tag{5}$$

The value of $\delta^{o+i}_k(R)$ gives the change in the width of the convex hull with respect to k and R. Clearly, $\eta^S_k \geq \eta^{S\backslash R}_k \geq 1$ and $\gamma^S_k \leq \gamma^{S\backslash R}_k \leq 1$. Hence, $\delta^{o+i}_k(R) \geq 0$, and larger values indicate more significant changes in the width of the convex hull with respect to k. $\delta^{o+i}_k(R) = 0$ suggests that removing the set R does not affect radial measures through k. R has a significant effect on k if $\delta^{o+i}_k(R)$ is significantly large.

Notably, $\eta^S_k = \gamma^S_k = 1$ is possible and implies k is on both the inner and outer boundaries. Observations for which this condition holds are typically extreme element of the EPPS, such as maximum or minimum scale. In this case, $\eta^{S\backslash R}_k = \gamma^{S\backslash R}_k = 1$ ($k \notin R$), which indicates that k is unaffected by the removal of any observation set absent k. However, k may affect others and be flagged as an outlier. This additional information regarding observation k allows us to characterize and classify the possible source of k's dissimilarity, e.g., extreme in scale. However, the interpretation and use of this additional information are case dependent and subject to the user's judgment.

Other measures that consider only the change in width associated with either the inner or the outer boundary can be similarly defined. $\delta^o_k(R)$ is the change caused by the outer boundary shift that is associated with observation k and $\delta^i_k(R)$ is the result of the inner

boundary shift as follows:

$$\delta_k^o(R) \equiv \eta_k^S - \eta_k^{S\backslash R}, \tag{6}$$

$$\delta_k^i(R) \equiv \gamma_k^S - \gamma_k^{S\backslash R}. \tag{7}$$

Note that $\delta_k^o(R) \geq 0$ and $0 \geq \delta_k^i(R) \geq -1$. Further, $\delta_k^{o+i}(R)$ can be expressed as a combination of $\delta_k^o(R)$ and $\delta_k^i(R)$:

$$\begin{aligned}\delta_k^{o+i}(R) &\equiv (\eta_k^S - \gamma_k^S) - (\eta_k^{S\backslash R} - \gamma_k^{S\backslash R}) \\ &= (\eta_k^S - \eta_k^{S\backslash R}) - (\gamma_k^S - \gamma_k^{S\backslash R}) \\ &= \delta_k^o(R) - \delta_k^i(R) \\ &= |\delta_k^o(R)| + |\delta_k^i(R)|. \end{aligned} \tag{8}$$

Eq. (8) states that the total difference between the widths is the sum of the inner and outer parts. $\delta_k^o(R)$ and $\delta_k^i(R)$ can be considered separately to classify R as either an efficient or an inefficient outlier. Eq. (8) assumes equal importance for both efficient and inefficient outliers. However, this assumption is unnecessary. Weights can be assigned for $|\delta_k^o(R)|$ and $|\delta_k^i(R)|$ in (8) to represent differences in the importance of the two types of outliers, and are typically determined based on subjective judgment.

We measure the influence of R for k based on the absolute difference shown in (5)–(7) while Pastor et al. [23] use ratios to represent the influence level. Applying Pastor et al.'s radial measure gives $(\eta_k^S - \gamma_k^S)/(\eta_k^{S\backslash R} - \gamma_k^{S\backslash R})$, $(\eta_k^S/\eta_k^{S\backslash R})$ and $(\gamma_k^S/\gamma_k^{S\backslash R})$ associated with (5)–(7), respectively. Ratio measures have particular drawbacks in this context. First, the ratio measures are the percentage change of width as a percentage of $\mathbf{y}_k$, and losses are the original geometric interpretation (e.g., the length in a one-output case). Second, (5)–(7) allow the effect of the inner boundary shift and the outer boundary shift to be quantified and to aggregate (and decompose) information as shown in (8). However, methods for aggregating the ratio measures are not obvious.

Rather than judge whether R is outlying, this paper intends to rank the importance of potential outliers based on their influence. $\delta_k^*(R)$ only quantifies the effect on an individual observation k. A summary statistic of the overall influence of R on the data set provides information to characterize R. This summary statistic is used to prioritize and identify the observations that justify the costly activity of further examination to confirm the validity of suspicious data. Several summary statistics are available; for example, Wilson [32] uses total value and the average number of individual influences, where the number of observations that are affected is also of interest. Pastor et al. [23] summarize individual influences by a statistical model so that statistical inference is possible. Metrics such as total influence $\sum_{k\in S}\delta_k^*(R)$ and average influence are reported in this work so that the graphical meaning of data can be understood easily. To make a judgment, a threshold must be defined. The selection on a threshold level is case dependent, and represents the tradeoffs between the cost of confirmation and the expense of including questionable data in the analysis. A "loose" criterion increases the risk of the outlier's existence and a "strict cut" costs more in confirmation.

The applications of the suggested model may be problematic if the data set is ill-conditioned, i.e., the number of observations is small and the variables do not vary over a sufficiently wide range [22]. Outlier detection assumes the input–output space is stable; otherwise it is difficult to distinguish whether the influence owes to dissimilarity of the observation or the ill-conditioned data set. To aggregate the input–output space to decrease the number of variables relative to the number of observations, methods proposed by Olesen and Petersen [22] or Pastor et al. [24] may be appropriate.

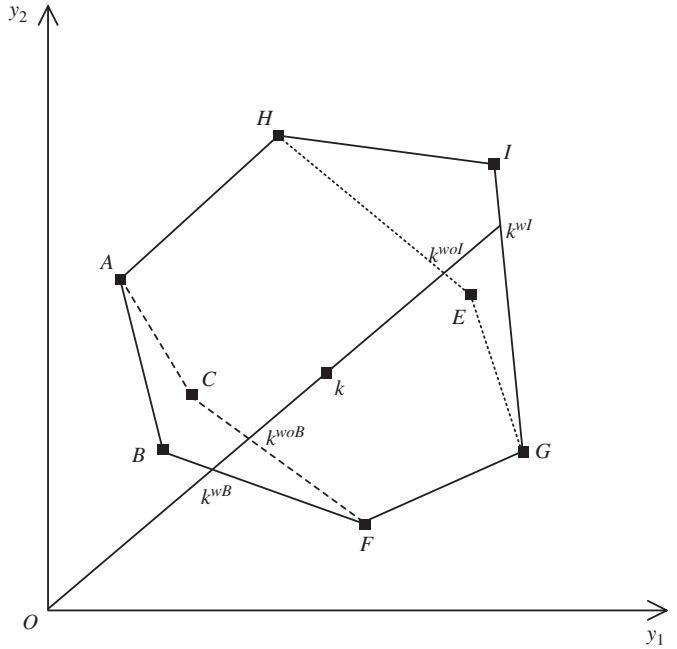

Fig. 1. A two-output equal-input illustration for convex hall approximation and influential measurements.

We note that if few outliers are close to each other such that one outlier does not differ significantly from the rest with respect to any characteristic of interest, the approaches measuring the influence of an observation's presence will have difficulty identifying this type of outlying data. This is termed the masking effect, and is stated and evident in many cases [30]. To eliminate masking, a combination of different observations should be removed in each stage with $|R| \geq 2$ yielding corresponding influential measures. Considering subsets allows heterogeneous subgroups within the analysis to distinguish themselves and they can be removed for a separate analysis.

2.3. Example

Fig. 1 presents a two-output equal-input example. Consider an observation set $S = \{A, B, C, E, F, G, H, I, k\}$; the convex hull is *ABFGIH*. Point k can be scaled up to k^{wI} ($\eta_k^S = (Ok^{wI}/Ok)$) on the outer boundary (*HIG*), and/or scaled down to k^{wB} ($\gamma_k^S = (Ok^{wB}/Ok)$) on the inner boundary (*ABF*). The width of ray *Ok* in the convex hull ($k^{wI}k^{wB}$) can be measured as $\eta_k^S - \gamma_k^S = (Ok^{wI} - Ok^{wB})/Ok$. If B is dropped from S ($R = \{B\}$), then the distance to the outer boundary remains unchanged ($\eta_k^{S\backslash\{B\}} = (Ok^{wI}/Ok)$), while the inner boundary is shifted to *ACF* such that $\gamma_k^{S\backslash\{B\}} = Ok^{woB}/Ok$, and then the width is $\eta_k^{S\backslash\{B\}} - \gamma_k^{S\backslash\{B\}} = (Ok^{wI} - Ok^{woB})/Ok$.

For k, the difference between the widths due to the existence of B can be measured by using $\delta_k^{o+i}(\{B\}) = (Ok^{wI} - Ok^{wB})/Ok - (Ok^{wI} - Ok^{woB})/Ok = (Ok^{woB} - Ok^{wB})/Ok$. The inner boundary shift, $\delta_k^i(\{B\}) = \gamma_k^S - \gamma_k^{S\backslash\{B\}} = (Ok^{wB} - Ok^{woB})/Ok$, is measured. The outer boundary shift is $\delta_k^o(\{B\}) = (Ok^{wI} - Ok^{wI})/Ok = 0$ since B does not affect the outer boundary. Similarly, when only observation I ($R = \{I\}$) is dropped, the inner boundary is the same, but the outer boundary changes to *HEG*. The new width of the convex hull that is associated with k is $\eta_k^{S\backslash\{I\}} - \gamma_k^{S\backslash\{I\}} = (Ok^{woI} - Ok^{wB})/Ok$. The influence of I, $\delta_k^{o+i}(\{I\})$, $\delta_k^o(\{I\})$ and $\delta_k^i(\{I\})$, can be obtained.

2.4. Input-oriented cases

Analogously, for any observation k the following apply in the input-oriented cases:

$$\phi_k^S \equiv \min_{\phi,\lambda} \left\{ \phi : \sum_{r\in S} \mathbf{x}_r\lambda_r = \phi\mathbf{x}_k;\ \sum_{r\in S} \mathbf{y}_r\lambda_r = \mathbf{y}_k;\ \sum_{r\in S} \lambda_r = 1;\ \lambda_r \geq 0,\ r \in S \right\}, \tag{9}$$

$$\phi_k^{S\backslash R} \equiv \min_{\phi,\lambda} \left\{ \phi : \sum_{r\in S\backslash R} \mathbf{x}_r\lambda_r = \phi\mathbf{x}_k;\ \sum_{r\in S\backslash R} \mathbf{y}_r\lambda_r = \mathbf{y}_k;\ \sum_{r\in S\backslash R} \lambda_r = 1;\ \lambda_r \geq 0,\ r \in S\backslash R \right\}, \tag{10}$$

$$\pi_k^S \equiv \max_{\pi,\lambda} \left\{ \pi : \sum_{r\in S} \mathbf{x}_r\lambda_r = \pi\mathbf{x}_k;\ \sum_{r\in S} \mathbf{y}_r\lambda_r = \mathbf{y}_k;\ \sum_{r\in S} \lambda_r = 1;\ \lambda_r \geq 0,\ r \in S \right\}, \tag{11}$$

$$\pi_k^{S\backslash R} \equiv \max_{\pi,\lambda} \left\{ \pi : \sum_{r\in S\backslash R} \mathbf{x}_r\lambda_r = \pi\mathbf{x}_k;\ \sum_{r\in S\backslash R} \mathbf{y}_r\lambda_r = \mathbf{y}_k;\ \sum_{r\in S\backslash R} \lambda_r = 1;\ \lambda_r \geq 0,\ r \in S\backslash R \right\}, \tag{12}$$

where $k \notin R$. These equivalencies specify the relationship between k and the corresponding convex hull boundaries. Applying the same argument to output-oriented cases, (10) and (12) are always feasible. Based on similar arguments addressed in Section 2.2, the measure of the effect on observation k due to R becomes

$$\delta_k^o(R) \equiv \pi_k^S - \pi_k^{S\backslash R}, \tag{13}$$

$$\delta_k^i(R) \equiv \phi_k^S - \phi_k^{S\backslash R}, \tag{14}$$

$$\delta_k^{o+i}(R) \equiv (\pi_k^S - \phi_k^S) - (\pi_k^{S\backslash R} - \phi_k^{S\backslash R}). \tag{15}$$

In the input-oriented cases, the outer boundary is associated with the inefficient observations. The corresponding measures are given by (11) and (12), and the related difference is defined by (13). Since $\pi_k^S \geq \pi_k^{S\backslash R} \geq 1, \delta_k^o(R) \geq 0$. Similarly, (14) is related to the change of the boundary that is closer to the origin, which is related to efficient observations in input-oriented analyses. $0 \geq \delta_k^i(R) \geq -1$, because $\phi_k^S \leq \phi_k^{S\backslash R} \leq 1$. Based on the arguments used in Section 2.1, $\delta_k^{o+i}(R)$, defined by (15), is the total change in the width associated with outlier candidate R, and combines the inner and the outer parts, such that $\delta_k^{o+i}(R) = |\delta_k^o(R)| + |\delta_k^i(R)|$.

Depending on the purpose of the analysis, either input- or output-oriented approaches should be adopted. If an input-oriented DEA model is selected to measure efficiency, an input-oriented influential measure should be used to avoid biased conclusions, and output-oriented influential measures should be selected when an output-oriented analysis is used to quantify efficiency. However, if the orientation of the analysis has not been determined, both metrics are recommended to fully explore the data set to discover unexpected knowledge.

3. Case studies

This section applies the proposed model using four DEA cases. The first two are simulated cases and illustrate the effectiveness through scatters of the data. The third case compares the model to earlier works via a common testbed, and the fourth identifies possible outlier suspects in a warehouse data set and shows their impact in a post analysis.

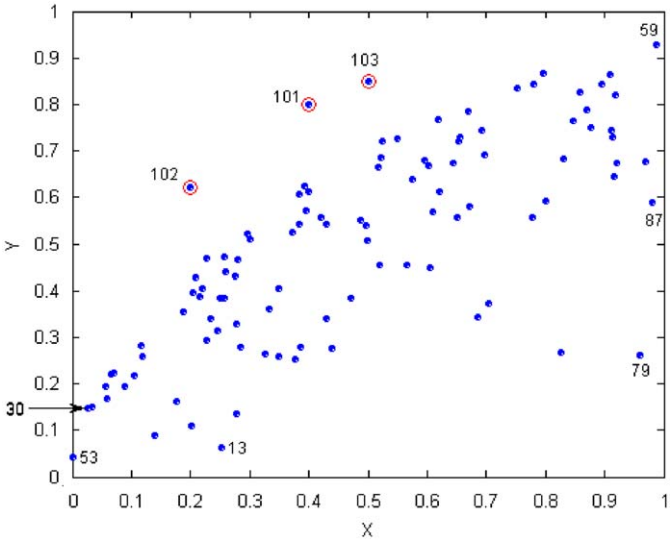

Fig. 2. The scatter plot of case A.

Table 1
Ranking of outliers (case A, output-oriented).

Rank	$\lvert\delta^o\rvert$			$\lvert\delta^i\rvert$			Observation ranked by	
	Observation	Tol.	Avg.	Observation	Tol.	Avg.	Tol. δ^{i+o}	Avg. δ^{i+o}
1	102	15.16	0.344	13	5.48	0.057	102	102
2	59	3.51	0.080	53	5.01	0.209	13	53
3	101	1.39	0.034	79	4.88	0.066	53	87
4	103	1.35	0.025	87	0.097	0.097	79	59
5	30	0.53	0.045				59	79
6							101	13
7							103	30
8							30	101
9							87	103

3.1. Case A – simulated bivariate case

Case A simulates a single-input single-output data set in which 100 observations are generated according to the function [30]

$$Y = X^{0.5} \cdot \exp(-U)$$

where $X \sim uniform(0, 1)$ and U is exponentially distributed with mean 1/3. Three extremely efficient outliers, 101, 102 and 103, are also added. Fig. 2 plots all 103 data points.

Table 1 summarizes the outlier ranking (by total influence, $\sum_{k\in S}\delta_k^*(R)$) for an output-oriented analysis using the proposed method[2] ; the associated points are also indicated in Fig. 2. The first panel of Table 1 corresponds to the outer boundary, and only five observations affect this boundary, including the extremely efficient outliers 101, 102 and 103. The second and third columns present the total influence and the average influence (the total influence divided by the number of observations affected, respectively).

[2] The input-oriented analysis was calculated and similar results were developed. The analysis is available upon request.

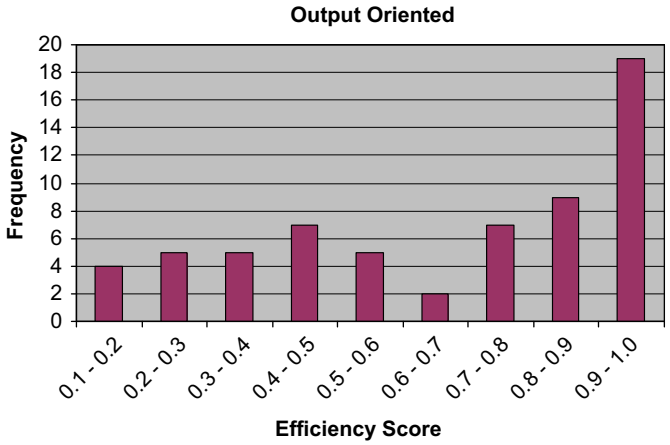

Fig. 3. Histogram of BCC estimates from Scheel [26].

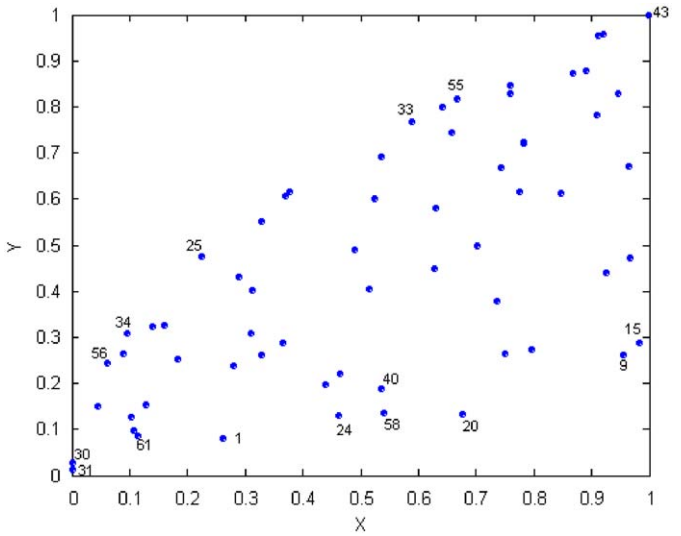

Fig. 4. The scatter plot of case B.

Similar to the first group, the second group ranks the total influence associated with the inner boundary. Only four outliers can affect the inner boundary, verified in Fig. 2. The third group ranks outliers by the changes in the total and average convex hull widths, which are the sums of inner and outer parts as specified by (8).

101, 102 and 103 are flagged as outliers; this result is consistent with the data generating process, in which 101, 102 and 103 are purposely added as efficient outliers. Further, the average influence of all listed observations, except for observations 102 and 53, does not exceed 0.1 (10%), and shows that the threshold is case-dependent if the gauging process is used to filter each new observation added to the data set. This observation is also pointed out by Simar [30]. Finally, the masking effect does exist as seen in Fig. 2. Observations 59 and 53 are flagged as outliers, although they are extreme in scale and differ somewhat from 101, 102 and 103. As discussed in Section 2.4, both 59 and 53 have $\eta_k^S = \gamma_k^S = 1$ (unlike101, 102 and 103) and can be easily classified as extreme in scale, but belonging in the analysis.

3.2. Case B – simulated bivariate case with empirical efficiency estimate distribution

Although it is necessary to identify the inefficient outliers, one can argue that it is easier and sufficient to check the empirical distribution of the efficient estimates, flagging any observations below a defined efficiency threshold as outliers (referred to in this paper as the trimming method). The following example demonstrates that this simple idea cannot be applied effectively in some circumstances.

Simulated efficiency estimates are commonly assumed to have exponential or half-normal distributions with significant tail [20]; a percentage of observations in the tail thus can be specified and flagged as outliers. However, DEA efficiency estimates rarely follow this pattern in many observed applications. Fig. 3 plots the distribution of the output-oriented BCC efficiency estimates based on the data collected by Scheel [26], in which 63 observations each had four inputs and two outputs.[3] The estimate distribution does not fit either the exponential or half-normal distributions, and it is difficult to identify extremely inefficient observations. Fig. 3 shows a clear gap in the distribution between 0.6 and 0.7. If a trimming approach is used and 0.6 could be selected as the efficiency level below which data is removed, then 26 observations will be removed. However, these observations are not necessarily distant from other observations when mapped in input–output space as Fig. 4 illustrates.

Table 2
Ranking of outliers (case B, output-oriented).

	$\lvert\delta^o\rvert$			$\lvert\delta^i\rvert$			Observation ranked by	
Rank	Observation	Tol.	Avg.	Observation	Tol.	Avg.	Tol. δ^{i+o}	Avg. δ^{i+o}
1	43	4.77	0.795	20	2.91	0.052	43	43
2	34	0.715	0.089	31	1.77	0.047	20	15
3	25	0.648	0.043	15	0.62	0.310	31	56
4	56	0.193	0.096	9	0.21	0.011	34	34
5	33	0.080	0.007				25	20
6	55	0.027	0.002				15	31
7	30	0.022	0.022				9	25
8							56	30
9							33	9
10							55	33

Ranking of efficiency estimates from the bottom: 1, 20, 58, 24, 61, 40, 9, 15.

Case B is a bivariate output-oriented case. Rather than following the exponential distribution as in Case A, efficiency estimates follow the empirical distribution obtained from Scheel [26] (Fig. 3). Sixty-three points are generated according to $Y = X^{0.5} \cdot E$ where $X \sim uniform(0,\ 1)$, and output efficiency estimates E from Scheel's data are randomly assigned. Fig. 4 displays the scatter plot of 63 points. Table 2 ranks the outliers.

Seven observations influence the outer boundaries, but only observation 43 has a strong impact on the other observations with an average change of more than 0.795. However, Fig. 4 reveals that observation 43 has an extreme scale and can be identified, since $\eta_{43}^S = \eta_{43}^S = 1$, which is consistent with the data generating process. For the inner boundary, four outliers have inefficient output estimates. The bottom of Table 2 presents the observations ranked by lowest efficiency estimates. These observations are not necessarily the outliers through visual observation or through the proposed outlier detection scheme.

This example shows that the proposed approach can detect efficient and inefficient outliers. In particular, it demonstrates that in cases of an empirical DEA efficiency estimate distribution, simply flagging the worst-performing observations as inefficient outliers can yield misleading results.

[3] Data set is available at http://www.wiso.uni-dortmund.de/lsfg/or/scheel/doordea.htm.

Table 3
Ranking of outliers (case C, input-oriented).

	Tol			Avg			Fox 04				
Rank	δ^i	δ^o	δ^{o*i}	δ^i	δ^o	δ^{o*i}	Mix	Scale	AD	W93	W95
1	47	15	47	47	31	10	66	59	59	59	59
2	44	31	10	10	10	47	48	32	32	44	44
3	10	10	15	57	15	31	15	69	69	33	52
4	57	43	44	59	54	54	56	5	5	66	69
5	49	54	31	20	7	15	69	62	62	35	62
6	59	8	57	44	43	59	49	44	44	54	56
7	52	7	54	54	58	57	68	29	29	68	15
8	66	51	49	68	8	20	5	61	61	67	58
9	20	58	66	48	24	44	61	38	48	8	45
10	54	38	8	45	51	43	67	48	38	50	17

3.3. Case C – empirical multi-input and multi-output case

Data collected in Charnes et al. [7] are used as a multi-input multi-output example. These data, containing 70 observations each with five inputs and three outputs, constitute a common testbed for outlier detection studies. Input-oriented analysis is applied to compare the resulting ranks (for both total and average measures) against Wilson [31,32] (W93 and W95, respectively) and Fox et al. [13] (Fox04) (Table 3).

Wilson [32] measures the influence based on the change in the super efficiency estimates as defined by Andersen and Petersen [1]. Wilson [31] extends the outlier measure suggested by Andrews and Pregibon [2], which identifies *as outliers those observations which contribute the largest proportion of the volume of the full data set*, to the case of multiple outputs to examine the geometric properties of input–output data directly. Fox et al. [13] propose metrics measuring dissimilarity between any two input–output vectors in scale and mix aspects (and also the composition of scale and mix). Observations with highest summary dissimilarity are considered as outliers. W93 and Fox04 are listed in Fox et al. [13] and W95 uses total influence.

In [32], observation 59 is undefined using super efficiency, because the scale of this observation is extremely large in input-oriented analyses or extremely small in output-oriented analyses. Fox et al. [13] also present evidence of this finding. However, unlike in other investigations, observation 59 is not ranked as a top outlier using the proposed method, because points such as 1, 21, 44 and 54 are now on the boundaries of the convex hull, such that 59 do not affect them. Hence, with fewer points affected, observation 59 has less overall effect on the entire data set.

The results obtained using various methods reveal some discrepancies. As Fox et al. [13] note, the outlier detecting schemes are related to different aspects and produce different conclusions. Wilson [32] also suggests that more than one approach should be applied to detect outliers. The consistency among the conclusions based on different approaches is a useful index for prioritizing the data to be investigated: a data point is more likely to be an outlier when flagged by several methods. Further, the inconsistency among the different methods can suggest a direction for more study to better understand the data.

3.4. Case D – warehouse performance

In this case, warehouse performance data collected by Hackman et al. [17] are used to demonstrate the effectiveness of the proposed outlier detection scheme, and especially the influence of suspect records. There are 57 warehouse records, each having three inputs and five outputs; some warehouses have union while others do not. Eight records (1, 3, 6, 28, 35, 38, 46 and 50) are flagged as potential outliers based on (8) and a threshold 0.05. Three of the eight flagged observations are located on the inefficient frontier, but do not have the lowest efficiency estimates; they rank third, 18th and 27th least efficient of the total records. This supports the insight that the outlier detection method does not simply identify the records with the lowest efficiency estimates; rather it identifies the observations that most significantly distort the production possibility set.

Table 4
Summary of warehouse performance comparisons.

	Full data (57 records)		Without outliers (49 records)	
Investment	$>$ \$1 M	$\leq$ \$1 M	$>$ \$1 M	$\leq$ \$1 M
Observation no.	36	21	29	20
Sample standard deviation	0.293	0.115	0.282	0.109
p-Value	0.147		0.0989	

To investigate the relationship between warehouse performance and capital investment, particularly the performance of those with equipment investment of more than \$1 million, an output-oriented BCC analysis (BCC.O) is conducted. All warehouses are pooled to obtain their efficiency estimates.

The hypothesis test, originally applied in DEA by Banker [3], is used to test whether the two groups (greater than/less than \$1 million of equipment investment) perform identically. The hypothesis test assumes that two groups with identical performance should have the same parameters of the efficiency estimate distributions. Banker suggests the use of a half-normal distribution, and the standard deviation of the half-normal distribution completely characterizes the distribution, because the mean is zero by definition. Therefore, the hypothesis tested is

$$H_0 : \sigma_H = \sigma_L \text{ against } H_1 : \sigma_H \neq \sigma_L$$

where σ_H and σ_L are the population standard deviations for the high and low equipment investment warehouses, respectively.

Table 4 summarizes the efficiency estimates using the total records and removing eight potential outliers. Using all 57 records as peers, the standard deviation of efficiency estimates[4] for warehouses with $>$ \$1 million investment is larger than those with $\leq$ \$1 million capital investment (0.293 vs. 0.115); the difference is statistically insignificant. The p-value is 0.147 using the full data set as the sample. Thus we fail to reject the null hypothesis at significant level 0.1 and can conclude that a warehouse's equipment level does not affect warehouse performance. After identifying and removing eight possible outliers, the difference in standard deviation of the two populations is 0.183 (0.282 vs. 0.109). The p-value is 0.0989; we reject H_0 and can conclude that equipment level does affect warehouse performance.

The results show that this paper's proposed outlier detecting scheme identifies both efficient and inefficient outliers that can affect the analysis results and produce different conclusions. However, it is important to note that the finding flags the observations that are most dissimilar to the other observations in the data set as measured by their influence on the EPPS, which suggests further investigation of this set of observations. Should added confirmation result in removing all observations, our results indicate an impact on the results of post analysis.

4. Computational remark

The computation procedure of $\delta_k^*(R)$ is based on removing an observation (or a set of observations) and calculating the influence on the remaining observations. This type of method requires a massive computational effort, particularly when it is necessary to eliminate the masking effect. We suggest a computation strategy that

[4] In fact, it is the reciprocal of the optimal value of (BCC-O).

will greatly reduce the computation time and that can be used to investigate the masking effect.

When $|R|=1$ and all observations are tested as potential outliers, there are $2\times|S|\times|S|$ linear programming (LP) problems to be solved, because for every observation we measure the influence on every other observation for both boundaries. In reality, we only have to calculate the effect that removing observations on a boundary has on observations that are not on the boundary. Without loss of generality, for output-oriented analyses the optimal solution of the corresponding LP problems (1) and (2) has at most $|I|+|J|$ λ's in the basis that are non-zero. Thus at least $|S|-|I|-|J|$ λ's are zero, and (3) and (4) will result in the same optimal values obtained by (1) and (2) when each of these observations is a removal candidate. This can be stated formally in the following propositions:

Proposition 2. *For $k\in S$ and $p\in S$, $\eta_k^S=\eta_k^{S\setminus\{p\}}$ if $\lambda_p^*=0$ is the optimal solution of* (1) *providing η_k^S.*

Proof. See the appendix. □

Proposition 3. *For $k\in S$ and $p\in S$, $\gamma_k^S=\gamma_k^{S\setminus\{p\}}$* if *$\lambda_p^*=0$ is the optimal solution of* (2) *providing γ_k^S.*

Proof. See the appendix. □

That is, only $|I|+|J|$ observations affect a given observation k's reference point on the boundary, so the removal of at least $|S|-|I|-|J|$ observations will not affect the outer boundary corresponding to $k\in S$. Therefore, computing δ_k^o in (6) does not require examining the removal of all $|S|-1$ observations, but only the points with zero λ's in the optimal solution of (1). Identical arguments are made for δ_k^i in (7). The observations result in a simplified procedure that solves at most $2\times|S|\times(|I|+|J|+1)$ LP problems, and greatly reduces the number of LP problems to be solved (especially, $|S|\gg|I|+|J|$ which is typical).

Further, we observe that observations on the outer (inner) boundaries will not be affected by removal of any other observations when measuring the distance to the outer (inner) boundary. Namely, there is no influence on the outer (inner) boundary as measured through k when $\eta_k^S=1$ ($\gamma_k^S=1$). This can be stated formally by the following:

Proposition 4. *For $k\in S$, $\eta_k^S=1$ implies $\eta_k^{S\setminus R}=1$ where $R\subset S$ and $k\notin R$.*

Proof. See the appendix. □

Proposition 5. *For $k\in S$, $\gamma_k^S=1$ implies $\gamma_k^{S\setminus R}=1$ where $R\subset S$ and $k\notin R$.*

Proof. See the appendix. □

Thus, it is not necessary to solve (3) and (4) regarding k, which satisfies the sufficient condition of Propositions 4 and 5, respectively. This observation further reduces the number of LP problems needed to be solved depending on data distribution. For example, if 20% of the data are on the outer boundary ($\eta_k^S=1$) and 15% of the data on the inner boundary ($\gamma_k^S=1$), Propositions 4 and 5 can be applied to indicate that $0.2\times|S|\times(|I|+|J|)+0.15\times|S|\times(|I|+|J|)$ problem solving can be saved from $0.2\times|S|\times(|I|+|J|+1)$.

The procedure can be extended to cases with $|R|\geq 2$. For example, there are $(|S|-1)\times(|S|-2)$ possible combinations of R ($|R|=2$) for each k, but Proposition 2 suggests that only $(|I|+|J|)\times(|S|-2)$ of them are needed for solving $\eta_k^{S\setminus R}$. Moreover, by integrating other methods, such as Chen and Cho [8] and Dula [11], which accelerate solving the single DEA problem, computational time can be further reduced. The proposed idea can also be applied to radial influence measures such as Pastor et al. [23]. Indeed, outlier detection increases the need to accelerate DEA computations and provides an application for a variety of acceleration methods.

5. Conclusion

This paper presents an outlier detection method that ranks the importance of the outliers to be investigated based upon their influence. Unlike previous outlier detection schemes, this method also identifies inefficient outliers that could impact post-efficiency estimation analysis. Where previous literature does not reconcile their approaches with the axioms of DEA, the method presented in this paper use the convex hull of the data by relaxing the free disposability axiom and allows the detection of inefficient outliers. In the case studies presented, the proposed method effectively ranks outliers and provides added information about their locations in the input–output space. The case studies demonstrate counter-examples to the intuitive misunderstanding that observations with poor efficiency estimates are more likely to be outliers. A real-world case also shows that outliers detected may lead to improper conclusions in post analysis based on DEA efficiency, such as testing the difference in efficiency of two populations using the Kolmogorov–Smirnov test. Moreover, we propose a strategy to reduce the computation time of outlier detection, and suggest that the strategy can be applied to other computational intensive influence measures such as suggested in Pastor et al. [23].

Acknowledgements

This research was supported in part by the National Science Council, Taiwan under Grant NSC 94-2213-E-009-078 and NSC 97-2221-E-009-113. The authors gratefully acknowledge the computing assistance of Mr. Chin-Chia Kuo.

Appendix.

Proposition 1. *For $k\in S$, $\eta_k^S=1$ if $\beta_k^S=1$.*

Proof. (1) is identical to (BCC.O) but with equalities for all constraints. The feasible region of (1) is smaller than that of (BCC.O), and thus $\beta_k^S\geq\eta_k^S$. $\beta_k^S=1$ leads to $\eta_k^S=1$ because $\eta_k^S\geq 1$. □

Proposition 2. *For $k\in S$ and $p\in S$, $\eta_k^S=\eta_k^{S\setminus\{p\}}$ if $\lambda_q^*=0$ is the optimal solution of* (1) *providing η_k^S.*

Proof. The dual of (1) is

$$\begin{aligned}&\min_{u_i,v_j,u_0}\sum_{i\in I}u_ix_{ik}+u_0\\&\text{s.t. }\sum_{i\in I}u_ix_{ir}+u_0\geq\sum_{j\in J}v_iy_{jr}\ r\in S,\\&\sum_{j\in J}v_iy_{jk}=1.\end{aligned}\tag{D1}$$

To satisfy $\sum_{r\in S}\lambda_r=1$ in (1), there must exist $q\in S$ such that $\lambda_q^*>0$. According to complementary slackness, $\left(\sum_{i\in I}u_i^*x_{iq}+u_0^*-\sum_{j\in J}v_j^*y_{jq}\right)\lambda_q^*=0$ where u_i^*, u_0^* and v_j^* are optimal solutions for (D1), and thus $\sum_{i\in I}u_i^*x_{iq}+u_0^*-\sum_{j\in J}v_j^*y_{jq}=0$. The type 1 constraint associated with q in (D1) is binding. It is clear that $q\neq p$, and the optimal value remains the same when removing constraint $\sum_{i\in I}u_ix_{iq}+u_0^*\geq\sum_{j\in J}v_j^*y_{jq}$. Therefore, $\eta_k^S=\eta_k^{S\setminus\{p\}}$. □

Proposition 3. *For $k\in S$ and $p\in S$, $\gamma_k^S=\gamma_k^{S\setminus\{p\}}$* if *$\lambda_p^*=0$ is the optimal solution of* (2) *providing γ_k^S.*

Proof. Apply the same arguments for Proposition 2. □

Proposition 4. *For $k \in S$, $\eta_k^S = 1$ implies $\eta_k^{S\setminus R} = 1$ where $R \subset S$ and $k \notin R$.*

Proof. It is clear that $\eta_k^{S\setminus R} \geq 1$. When $\eta_k^S = 1$ is the optimal value in (1), $\lambda_k^* = 1$ and $\lambda_r^* = 0$ for $r \in S\setminus\{k\}$ is the optimal solution (or one of the optimal solutions) for (1). According to complementary slackness $(\sum_{i\in I} u_i^* x_{ik} + u_0^* - \sum_{j\in J} v_j^* y_{jk})\lambda_k^* = 0$ where u_i^*, u_0^* and v_j^* are optimal solution for the dual (D1), and thus $\sum_{i\in I} u_i^* x_{ik} + u_0^* - \sum_{j\in J} v_j^* y_{jk} = 0$. The type 1 constraint associated with k in (D1) is binding, and removing type 1 constraints associated with R in (D1) will remain the same objective value and cannot be better. That is (D1) without type 1 constraints associated with R will remain the same optimal value, and it is the dual of (3). Therefore, $\eta_k^{S\setminus R} = 1$. □

Proposition 5. *For $k \in S$, $\gamma_k^S = 1$ implies $\gamma_k^{S\setminus R} = 1$ where $R \subset S$ and $k \notin R$.*

Proof. Apply the same arguments for Proposition 4. □

References

[1] Andersen P, Petersen NC. A procedure for ranking efficient units in data envelopment analysis. Management Science 1993;39(10):1261–4.
[2] Andrews DF, Pregibon D. Finding the outliers that matter. Journal of the Royal Statistical Society, Series B 1978;40:85–93.
[3] Banker RD. Maximum-likelihood, consistency and data envelopment analysis – a statistical foundation. Management Science 1993;39(10):1265–73.
[4] Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiency in data envelopment analysis. Management Science 1984;30(9):1078–92.
[5] Barnett V, Lewis T. Outliers in statistical data. New York: Wiley; 1995.
[6] Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. European Journal of Operational Research 1978;2:429–44.
[7] Charnes A, Cooper WW, Rhodes E. Evaluating program and managerial efficiency: an application of data envelopment analysis to program flow through. Management Science 1981;27(6):668–97.
[8] Chen W-C, Cho W-J. A procedure for large-scale DEA computations. Computers and Operations Research 2009;36(6):1813–24.
[9] Chen Y, Liang L, Yang F, Zhu J. Evaluation of information technology investment: a data envelopment analysis approach. Computers and Operations Research 2006;33(5):1368–79.
[10] Chen Y, Gregoriou GN, Rouah FD. Efficiency persistence of bank and thrift CEOs using data envelopment analysis. Computers and Operations Research 2009;36(5):1554–61.
[11] Dula JH. A computational study of DEA with massive data sets. Computers and Operations Research 2008;35:1191–203.
[12] Farrell MJ. The measurement of productivity efficiency. Journal of the Royal Statistical Society 1957;120:377–91.
[13] Fox KJ, Hill RJ, Diewert WE. Identifying outliers in multi-output models. Journal of Productivity Analysis 2004;22:73–94.
[14] Fried HO, Lovell CAK, Turner JA. An analysis of the performance of university-affiliated credit unions. Computers and Operations Research 1996;23(4):375–84.
[15] Gattoufi S, Oral M, Kumar A, Reisman A. Content analysis of data envelopment analysis literature and its comparison with that of other OR/MS fields. Journal of the Operational Research Society 2004;55(9):911–35.
[16] Hackman ST. Production economics: integrating the microeconomic and engineering perspectives. Berlin, Heidelberg: Springer; 2008.
[17] Hackman ST, Frazelle EH, Griffin P, Griffin SO, Vlatsa DA. Benchmarking warehousing and distribution operations: an input–output approach. Journal of Productivity Analysis 2001;16:79–100.
[18] Johnson AL, McGinnis LF. Outlier detection in two-stage semiparameteric DEA models. European Journal of Operational Research 2008;187(2):629–35.
[19] Johnson AL, Chen W-C, McGinnis LF. Internet-based benchmarking for warehouse operations. Working Paper, 2008.
[20] Kumbhakar SC, Lovell CAK. Stochastic frontier analysis. Cambridge, UK: Cambridge University Press; 2000.
[21] Muñiz M, Paradi J, Ruggiero J, Yang Z. Evaluating alternative DEA models used to control for non-discretionary inputs. Computers and Operations Research 2006;33(5):1173–83.
[22] Olesen OB, Petersen NC. Indicators of ill-conditioned data sets and model misspecification in data envelopment analysis: an extended facet approach. Management Science 1996;42(2):205–19.
[23] Pastor JT, Ruiz JL, Sirvent I. A statistical test for detecting influential observations in DEA. European Journal of Operational Research 1999;115(3):542–54.
[24] Pastor JT, Ruiz JL, Sirvent I. A statistical test for nested radial DEA models. Operations Research 2002;50(4):728–35.
[25] de Sousa MDCS, Stosic B. Technical efficiency of the Brazilian municipalities: Correcting nonparametric frontier measurements for outliers. Journal of Productivity Analysis 2005;24:157–81.
[26] Scheel H. Continuity of the BCC efficiency measure. In: Westermann G, editor. Data envelopment analysis in the service sector. Wiesbaden, Germany: Gabler; 1999.
[27] Seiford L, Fare R, Lovell CAK, Banker RD, Simar L, Forsund F. et al. Summary of some of the discussion at the Advanced Research Workshop on Efficiency Measurement, held at Odense University, May 22–24, 1995. Journal of Productivity Analysis 1996;7(2–3):341–5.
[28] Sexton TR, Silkman RH, Hogan AJ. Data envelopment analysis: critique and extensions. In: Silkman RH, editor. Measuring efficiency: an assessment of data envelopment analysis. San Francisco, CA: Jossey-Bass; 1986.
[29] Shephard RW. Theory of cost and production functions. Princeton, NJ: Princeton University Press; 1970.
[30] Simar L. Detecting outliers in frontier models: a simple approach. Journal of Productivity Analysis 2003;20:391–424.
[31] Wilson PW. Detecting outliers in deterministic nonparametric frontier models with multiple outputs. Journal of Business and Economic Statistics 1993;77(6):779–802.
[32] Wilson PW. Detecting influential observations in data envelopment analysis. Journal of Productivity Analysis 1995;6:27–45.