ECARES

Outlier Detection in Nonparametric Frontier Models

Christopher Bruffaerts
ECARES-SBS-EM, Université Libre de Bruxelles

Bram De Rock
ECARES-SBS-EM, Université Libre de Bruxelles

Catherine Dehon
ECARES-SBS-EM, Université Libre de Bruxelles

ECARES working paper 2014-12

ECARES
ULB - CP 114/04
50, F.D. Roosevelt Ave., B-1050 Brussels BELGIUM
www.ecares.org

Outlier detection in nonparametric frontier models

Christopher Bruffaerts* Bram De Rock † Catherine Dehon‡

Abstract

In nonparametric frontier models, it is clear that outlying points can have a large influence on the efficiency of production units. The influence will however depend on the type of distance that is used to benchmark the production unit under analysis. The directional distance function measures the distance from a production unit to its full frontier along a linear path for a given direction while its robust counterpart considers its distance to the so-called order-α partial frontier. The first contribution is to study the robustness properties of this order-α directional efficiency estimator through the concept of influence function and breakdown point. To do so, the explicit link between the hyperbolic and the directional types of orientation is used. The second contribution is to propose a methodology using an adapted kernel density estimator to identify influential points with the directional efficiency estimator. This technique considers the density estimation of a transformed variable that depends on both the production unit being analysed as well as the direction used to benchmark its efficiency. To illustrate the results, some numerical examples are given as well as a real life example on the research efficiency of US universities.

Keywords: Nonparametric frontier models, Directional distance, Order-α frontiers, Outlier detection, Robust efficiency estimation.

1 Introduction

The field of production analysis is frequently used in economic applications that are characterized by multiple inputs and outputs (see for instance *Coelli et al. (2003)*, *Gattoufi et al. (2004)* and *Emrouzenjad et al. (2008)*). In this respect, it is of interest to know how efficiently a production unit is converting its inputs $x \in \mathbb{R}^p_+$ into outputs $y \in \mathbb{R}^q_+$. The benchmark used to determine the technical efficiency of a production unit (usually referred as a Decision Making Unit-DMU) is the production frontier which can be defined as the frontier of the following production set:

$$\Psi = \{(x, y) \in \mathbb{R}^{p+q}_+ \mid x \textit{ can produce } y\}.$$

*Université Libre de Bruxelles, ECARES (European Center for Advanced Research in Economics and Statistics). E-mail: cbruffae@ulb.ac.be.

†Université Libre de Bruxelles, ECARES. Bram De Rock gratefully acknowledges the European Research Council (ERC) for his starting grant.

‡Université Libre de Bruxelles, ECARES. Catherine Dehon acknowledges research support from the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy) and from the ARC contract of the Communauté Française de Belgique.

According to *Farrell (1957)*, technical efficiency is defined as a certain distance from the production unit that is being analysed to the production frontier. For a given direction, the directional distance function measures the distance from the production unit to the frontier along a linear path passing through the unit and pointing towards this direction.This generalized way of measuring efficiency was first introduced in *Chambers et al.(1996), Chambers et al.(1998)* and later in *Färe and Grosskopf (2000), Färe and Grosskopf (2004), Färe et al. (2008)*. For a unit performing at level (x, y) and for a given direction $d = (-d_x, d_y) \in \mathbb{R}_+^{p+q}$, it is formally defined as follows:

$$\delta(x, y \mid d_x, d_y) = \sup\{\delta \mid (x - \delta d_x, y + \delta d_y) \in \Psi\}.$$

This efficiency measurement provides the quantity by which, in the direction $(-d_x, d_y)$, the input levels x can be reduced and simultaneously the quantity by which the output levels y can be increased in order to attain the frontier. The production unit under analysis is efficient if $\delta(x, y \mid d_x, d_y) = 0$ and becomes more inefficient in the direction $d = (-d_x, d_y)$ as $\delta(x, y \mid d_x, d_y)$ increases. Both of the most popular distances in efficiency measurement which are the Farrell input and output distances are special cases of the directional distance function. It is therefore of great interest to study this generalized radial type of distance both from a theoretical and a practical point of view. The choice of the optimal direction $d = (-d_x, d_y)$ depends on the empirical application and corresponding data at hand. This falls beyond the scope of this paper but more discussions about it can be found in *Färe et al. (2008)*.

The production set is unknown and this is why it has to be estimated from the sample of n production units $\mathcal{S}_n = \left\{(X_i, Y_i) \in \mathbb{R}_+^{p+q},\ i = 1, \ldots, n\right\}$. In frontier models, the two most popular nonparametric estimation techniques are the Data Envelopment Analysis (DEA) introduced in *Charnes et al. (1984)* and the Free Disposal Hull (FDH) initiated in *Deprins et al. (1984)*. To estimate the production frontier, the former type of estimator which is based on linear programming techniques constructs the boundary of the convex hull formed by $\mathcal{S}_n$. The latter type of estimator is defined as the lowest monotone function covering all observations and can be seen as the non-convex version of DEA. Both of those methods are based on the idea of enveloping all data points. In particular, this means that a single anomalous production unit can change completely the estimation of the frontier and therefore impact all associated efficiency measurements.

To counter the drawback of non-robustness of both DEA and FDH estimation techniques, the concept of partial frontier was introduced. The first idea emerged in *Cazals et al. (2002)* in which the authors proposed the notion of order-m frontiers. Their main achievement was to redefine the production setting (production set, production frontier and related efficiency measurements) through a probabilistic formulation. The order-m type of frontier was later extended in *Aragon et al. (2005)* with the concept of order-α frontier. In a nutshell, a partial frontier, either of order-m or of order-α, is a frontier that is not as extreme as the production frontier of Ψ. In practice, the partial frontier leaves some points outside of the frontier and this is why it can be seen as a more robust version than both DEA and FDH estimation methods. As the parameter m and α go to ∞ and 1 respectively, the partial frontiers converge to the full production frontier. Even if those two types of frontiers are closely related (see *Daouia and Gijbels (2011)* for an explicit link), there is a major difference between them from a robustness point of view. The order-m type of frontier and therefore the associated order-m efficiency estimators

have unbounded influence functions as well as an asymptotic breakdown point of 0 (see *Daouia and Gijbels (2011)* for the input and output orientations). On the contrary, the order-α type of frontier and its related efficiencies exhibit interesting robustness properties. This motivates the use of the order-α type of estimators in the sequel. In particular, the order-α directional efficiency estimator that was introduced in *Simar and Vanhems (2012)* can be seen for a given choice of α as a robust version of the directional distance separating a DMU to the production frontier of Ψ.

It is in this context that the aim of the paper is twofold. First, along the lines of *Daouia and Gijbels (2011)* for both the input and output cases and *Bruffaerts et al. (2013a)* for the hyperbolic case, the robustness properties of the order-α directional efficiency estimator are studied via its influence function and breakdown point. Those results show the importance of the parameter α in both concepts of global and local robustness. The choice of the parameter α in practice remains a tedeous matter as underlined by many authors (see *Daouia et al. (2010)*, *Daouia and Gijbels (2011)* and *Daouia et al. (2012)*). The importance of this "trimming" parameter α motivates the second contribution of the paper which is the detection of outlying points in nonparametric frontier models. Assessing the efficiency of a given DMU when the sample is potentially contaminated with influential observations requires the practitioner to have an idea on the number of points that are influential for this DMU. As will be shown, the notion of influential point for a given DMU in frontier models depends crucially on the direction that is used to benchmark its efficiency. For instance, a production unit can have a large influence on the efficiency of a given DMU for a certain direction while it can have no influence at all for a different direction.

Outlier detection in nonparametric frontier models has been studied in particular in *Wilson (1993, 1995)* and *Simar (2003)*. The former proposes to identify anomalous points by using a certain statistic for which the distribution is known because of the rather restrictive assumption that outputs in the production process are normally distributed. To detect anomalous points, the second paper makes uses of a leave-one-out approach by using the order-m type of estimator. Although simple and easy use, this methodology stuggles to identify outlying points that are hidden by other outliers (known as the masking effect). Those two methods enable the practitioner to find anomalous observations with respect to the sample, *i.e.* observations that can affect the estimated frontier in general. However, it can be that those anomalous points do not impact the efficiency of some of the DMUs of the sample. Removing those anomalous points from the sample might therefore lead to a loss of information in the computation of the efficiency of some of the production units (especially in a nonparametric context in which the curse of dimensionality is present). For this reason, anomalous observations are defined in this paper with respect to the production unit under analysis as the set of outlying points might differ from one DMU to another. We propose in this paper a methodology for identifying outliers which remains totally nonparametric and which takes into account the presence of the production frontier. The estimation of the efficiency of a given production unit and for a specific direction is tantamount to the estimation of a boundary point of a dimensionless variable which depends on both the DMU and the direction. To detect influential observations for a given DMU, the idea of the proposed method is to find approximately where the boundary point related to the production frontier is lying and flag as outliers points that are beyond this estimated boundary. To reach this ob-

jective, we first exploit the explicit link between the hyperbolic type of distance and the directional one (see *Simar and Vanhems (2012)*) and show that directional efficiencies can be computed from the quantile function of a univariate transformation of the inputs and outputs. Based on this dimensionless transformation, kernel density estimators are used to get an idea where the boundary is lying. Because of the boundary issue and the fact that the sample is potentially contaminated with outlying points, we make use of the skew-normal type of kernel to get to know whereabout the boundary is lying. This way of doing has the advantage of being totally nonparametric and does not require any distributional assumption. Moreover, it avoids the masking effect and is relatively simple and fast to use. As stressed in *Simar (2003)*, a methodology for detecting outlying points should be seen as a first step in the analysis of the data (exploratory analysis) and should not stop the researcher to do a further analysis on those potential anomalous observations.

In the next section, the probabilistic formulation of the directional efficiency estimator and its robust counterpart are introduced. In particular, it is shown how the hyperbolic type of distance enables to recover any directional type of distance from a given DMU to the production frontier. Section 3 treats the robustness properties of this order-α directional type of estimator through the concept of its influence function as well as its breakdown point. A methodology to identify influential points by using an adapted kernel density estimator of a transformed random variable is presented in Section 4. To illustrate the results, Section 5 covers in a first part some numerical examples under different contamination settings and in a second part presents an example on the research efficiency of US universities.

2 The directional distance function

A generalized and flexible measurement The directional distance function is a generalized way of measuring the efficiency of production units. First of all, for a specific choice of the direction, it encompasses the two most popular types of distance which are the Farrell input and output distances. For a production unit performing at level (x, y), the input and output efficiency measurements of this DMU are defined respectively as follows:

$$\theta(x,y) = \inf\{\theta > 0 \mid (\theta x, y) \in \Psi\},$$
$$\lambda(x,y) = \sup\{\lambda > 0 \mid (x, \lambda y) \in \Psi\}.$$

Those efficiency scores represent in the first case the contraction of inputs and in the second case the enlargment of outputs required for the DMU under analysis to be efficient. By specifying the direction $d = (d_x, d_y) = (x, 0) \in \mathbb{R}^{p+q}$, one recovers the input oriented efficiency: $\delta(x, y \mid d_x, d_y) = 1 - \theta(x, y)^{-1}$. Similarly, if $d = (d_x, d_y) = (0, y) \in \mathbb{R}^{p+q}$, we recover the output oriented efficiency: $\delta(x, y \mid d_x, d_y) = \lambda(x, y)^{-1} - 1$.

The directional distance function allows the decision-maker to have more flexibility in terms of its inputs and outputs. Contrarily to both input and output orientations, this type of measurement gives the opportunity to the decision-maker to play on both his inputs and outputs. This can be simply done by picking specific directions $d_x \in \mathbb{R}^p$

and $d_y \in \mathbb{R}^q$ which allows a certain flexibility on each input and each output according to the possibilities of the decision-maker. For instance, if a zero weight is allocated to a specific input in the analysis, this means that the decision-maker cannot change this specific input. It was shown in *Daraio and Simar (2014)* how the directional type of estimator can be computed when at least one of the components of $d_x \in \mathbb{R}^p$ and/or $d_y \in \mathbb{R}^q$ is zero.

Probabilistic formulation To introduce the concept of partial frontiers, the authors in *Cazals et al. (2002)* defined a probabilistic framework to the production setting in the context of efficiency analysis. For a production unit performing at level (x, y), the probability that this unit is being weakly dominated is given as follows:

$$H_{XY}(x, y) = P[X \leq x, Y \geq y].$$

This joint distribution allows to redefine completely the production set (see *Simar and Vanhems (2012)*):

$$\Psi = \{(x, y) \in \mathbb{R}_+^{p+q} \mid H_{XY}(x, y) > 0\}.$$

From this characterization of the production set Ψ, the directional distance function can be written in the following way:

$$\begin{aligned}\delta(x, y \mid d_x, d_y) &= \sup\{\delta > 0 \mid (x - \delta d_x, y + \delta d_y) \in \Psi\} \\ &= \sup\{\delta > 0 \mid H_{XY}(x - \delta d_x, y + \delta d_y) > 0\}.\end{aligned}$$

Beside providing a new formulation to the production process, the probabilistic formulation allows as well to introduce partial frontiers. The partial frontier is defined as the boundary of the following partial production set for $\alpha \in (0, 1)$:

$$\Psi_\alpha = \{(x, y) \in \mathbb{R}_+^{p+q} \mid H_{XY}(x, y) > 1 - \alpha\}.$$

The order-α type of frontier provides a less extreme benchmark than the frontier of the full production set Ψ. The directional order-α type of measurement is naturally defined as the distance to this partial frontier of order α. It is defined formally for $\alpha \in (0, 1)$ as follows:

$$\delta_\alpha(x, y \mid d_x, d_y) = \sup\{\delta > 0 \mid H_{XY}(x - \delta d_x, y + \delta d_y) > 1 - \alpha\}.$$

This partial efficiency score represents the amount by which a DMU should decrease its inputs and increase its outputs in the direction $(-d_x, d_y)$ such that it is being dominated with probability $1 - \alpha$. When $\alpha = 1$, one recovers the traditional directional distance function that was previously defined.

From hyperbolic to directional If both input and output types of distance are nested cases of the directional distance function, it is not the case for the so-called hyperbolic type of distance. This specific efficiency measurement was originally proposed in *Färe et al. (1985)* and *Färe and Grosskopf (2000)* and is defined in the following way:

$$\gamma(x, y) = \sup\{\gamma > 0 \mid (\gamma^{-1}x, \gamma y) \in \Psi\}.$$

This quantity represents the simultaneous contraction of inputs and expansion of outputs required for the DMU to reach the frontier. As shown in *Simar and Vanhems (2012)*, for a given transformation of the sample $S_n = \{(X_i, Y_i) \in \mathbb{R}^{p+q}; i = 1, \ldots, n\}$, it is possible to recover the directional distance function from the hyperbolic one. By considering the following monotonic increasing transformation for the inputs $x^* = \exp(x./d_x)$ and $y^* = \exp(y./d_y)$ for the outputs (where ./ refers to the componentwise division of vectors with $d_x > 0$ and $d_y > 0$), the attainable set in this new system of coordinates becomes:

$$\Psi^* = \{(x^*, y^*) \in \mathbb{R}_+^{p+q} \mid x^* = \exp(x./d_x), y^* = \exp(y./d_y) \text{ for some } (x, y) \in \Psi\}.$$

This set can equivalently be written as:

$$\Psi^* = \{(x^*, y^*) \in \mathbb{R}_+^{p+q} \mid H_{X^*Y^*}(x^*, y^*) > 0\},$$

where $H_{X^*Y^*}(.,.)$ is the transformation of the probability distribution $H_{XY}(.,.)$ for the random variables $X^* = \exp(X./d_x)$ and $Y^* = \exp(Y./d_y)$[1]. It was shown by the authors that the following equality holds:

$$H_{XY}(x, y) = H_{X^*Y^*}(x^*, y^*).$$

Hence, the order-α directional type of distance can be defined as follows:

$$\begin{aligned}
\delta_\alpha(x, y \mid d_x, d_y) &= \sup\{\delta > 0 \mid H_{XY}(x - \delta d_x, y + \delta d_y) > 1 - \alpha\} \\
&= \sup\{\delta > 0 \mid H_{X^*Y^*}(x^* \exp(-\delta), y^* \exp(\delta)) > 1 - \alpha\} \\
&= \sup\{\gamma > 0 \mid H_{X^*Y^*}(\gamma^{-1} x^*, y^* \gamma) > 1 - \alpha\},
\end{aligned}$$

where $\gamma = \exp(\delta) > 0$.
In other words, the order-α directional distance function can be recovered from the order-α hyperbolic type of distance through the following relationship:

$$\delta_\alpha(x, y \mid d_x, d_y) = \log(\gamma_\alpha(x^*, y^*)),$$

where $\gamma_\alpha(x^*, y^*) = \sup\{\gamma > 0 \mid H_{X^*Y^*}(\gamma^{-1} x^*, \gamma y^*) > 1 - \alpha\}$ [2]. This is an important relationship that will allow us on the one hand to transpose the robustness properties of the hyperbolic type of estimator to the ones of the directional type of estimator and on the other hand to make the detection of anomalous observations easier.

The directional efficiency estimator Given that the production sets Ψ and Ψ_α are unknown, they have to be estimated from the sample $S_n = \{(X_i, Y_i) \in \mathbb{R}^{p+q}; i = 1, \ldots, n\}$ in order to benchmark the DMU that is being analysed. The joint distribution $H_{XY}(.,.)$ defined previously can be estimated by using its empirical counterpart:

$$\widehat{H}_{XY}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y).$$

[1] To recover both the input and output efficiency measurements, slightly different transformations have to be considered. For the former type of measurement, the transformations are: $X^* = \exp(X)$ and $Y^* = \exp(Y./d_y)$ while for the latter one the transformations are: $X^* = \exp(X./d_x)$ and $Y^* = \exp(Y)$.

[2] For the order-α input efficiency measurement, the relationship becomes $\delta_\alpha(x, y \mid d_x, 0) = \log(\gamma_\alpha(x^*, y^*))$ with $\gamma_\alpha(x^*, y^*) = \sup\{\gamma > 0 \mid H_{X^*Y^*}(\gamma^{-1} x^*, y^*) > 1 - \alpha\}$ while for the order-α output efficiency measurement it becomes $\delta_\alpha(x, y \mid 0, d_y) = \log(\gamma_\alpha(x^*, y^*))$ with $\gamma_\alpha(x^*, y^*) = \sup\{\gamma > 0 \mid H_{X^*Y^*}(x^*, \gamma y^*) > 1 - \alpha\}$.

The order-α directional efficiency estimator for the direction $d = (-d_x, d_y)$ and for a DMU performing at level (x, y) can therefore be estimated for $\alpha \in (0, 1)$ as follows:

$$\widehat{\delta}_\alpha(x, y \mid d_x, d_y) = \sup\{\delta > 0 \mid \widehat{H}_{XY}(x - \delta d_x, y + \delta d_y) > 1 - \alpha\}.$$

Equivalently, the efficiency estimator $\widehat{\delta}_\alpha(x, y \mid d_x, d_y)$ can be computed from the order-α hyperbolic efficiency estimator of the transformed sample in the following way (see *Simar and Vanhems (2012)*:

$$\widehat{\delta}_\alpha(x, y \mid d_x, d_y) = \log(\widehat{\gamma}_\alpha(x^*, y^*)),$$

where $\widehat{\gamma}_\alpha(x^*, y^*) = \sup\{\gamma > 0 \mid \widehat{H}_{X^*Y^*}(\gamma^{-1}x^*, \gamma y^*) > 1 - \alpha\}$ and for the same transformations mentioned previously. In the above estimation, if $\alpha = 1$, one recovers the non-robust FDH estimator of the directional type of distance (see *Deprins et al. (1984)*). When using order-α partial frontiers, it is often the case that a common value of α is chosen such that all DMUs are benchmarked with respect to this estimated partial frontier. The practitioner seeks through the concept of partial frontier to have a robust methodology to assess the efficiency of DMUs and at the same time to have the partial frontier to lie as closely as possible to the complete frontier of the production set. Because of the unique value of α, it can be that the partial frontier in some areas is very much inside the production set, particularly in areas where there are very small or very large production units. This can be explained by the fact that small or large production units might not be influenced by outlying points as much as the average production unit. Hence, picking an "optimal" value for α for the average production unit does not necessarily mean that it is optimal for small or large units. In this respect, it is important to note that the robust efficiency estimator $\widehat{\delta}_\alpha(x, y; d_x, d_y)$ is estimating the true directional efficiency $\delta(x, y; d_x, d_y)$ of the uncontaminated distribution only if $1 - \alpha$ is equal to the proportion of influential points in the direction $d = (-d_x, d_y)$. The choice of the order α is of course crucial and picking an optimal value for it is a tedeous matter. In the sequel, a methodology to identify influential points by means of an adapted kernel density based technique will be presented. The notion of influential point will therefore depend on both the DMU that is being analysed and on the direction that is used to assess its efficiency. If the sample is not contaminated with outlying points, the traditional DEA and FDH estimators yield an estimation of the distance separating the DMU to the full production frontier. If the sample is however contaminated, the partial order-α efficiency estimators provide an estimator of the partial frontier which can be seen for an adequate value of α as an estimator of the full frontier of the uncontaminated distribution. This is why being able to detect influential points for a DMU allows the researcher to have a robust estimation of the efficiency measurement related to the full production frontier.

3 Robustness of the order-α directional distance

Nowadays, it is important to provide robust estimators and do robust inference whenever performing any statistical analysis (*Hampel et al. (1986)*). Indeed, it could be that a certain percentage of outlying points are present in the data and therefore might influence the analysis made by the applied researcher. The robustness properties of

efficiency estimators was first studied in *Daouia and Ruiz-Gazen (2006)* and in *Daouia and Gijbels (2011)* for both the input and output orientations and in *Bruffaerts et al. (2013a)* for the hyperbolic orientation. In this section, the robustness properties of the directional type of estimator are studied through the concepts of influence function and breakdown point.

3.1 Influence function

One of the goals of robustness is to quantify the degree to which an estimator is robust with respect to outliers. A local quantitative concept of robustness is given by the influence function of an estimator. We define the functional $W^{\alpha}_{xy;d_x,d_y}$ that corresponds to the directional type of estimator (in the direction $d = (-d_x, d_y)$) for a unit operating at (x, y) in the following way:

$$\delta_\alpha(x, y \mid d_x, d_y) = W^{\alpha}_{xy;d_x,d_y}(H_{XY}) = \sup\{\delta > 0 \mid H_{XY}(x - \delta d_x, y + \delta d_y) > 1 - \alpha\}.$$

Proposition 3.1. *For $\alpha \in (0, 1)$ and for any $(x, y) \in \mathbb{R}^{p+q}_+$ in the interior of Ψ such that $\gamma_\alpha(x^*, y^*) > 0$, let the function $\phi(w) = H_{X^*Y^*}(x^*w^{-1}, wy^*)$. Assuming that $\phi(w)$ is differentiable at $\gamma_\alpha(x^*, y^*)$ with strictly negative derivative, then for any $(x_0, y_0) \in \mathbb{R}^{p+q}_+$:*

$$IF\left((x_0, y_0), W^{\alpha}_{xy;d_x,d_y}, H_{XY}\right) = \frac{\alpha - \mathbb{1}\left(Z_{xy}(x_0, y_0) \leq \gamma_\alpha(x, y)\right)}{\gamma_\alpha(x^*, y^*)\phi'(\gamma_\alpha(x^*, y^*))},$$

where $Z_{xy}(x_0, y_0) = \min\left\{\min_{1\leq k\leq p} \frac{x^{(k)}}{x_0^{(k)}}, \min_{1\leq l\leq q} \frac{y_0^{(l)}}{y^{(l)}}\right\}$ *and* $\mathbb{1}(.)$ *represents the indicator function.*

When integrating the square of the influence function of the order-α directional distance, we recover the asymptotic variance given in Theorem 4.2 in *Simar and Vanhems (2012)*. By considering the worst case scenario for the influence function by taking its supremum on the whole input/output space, the following gross error sensitivity measure (GES) is obtained.

Corollary 3.1. *For $\alpha \in (0, 1)$ and for any $(x, y) \in \mathbb{R}^{p+q}_+$ in the interior of Ψ such that $\gamma_\alpha(x^*, y^*) > 0$, let the function $\phi(w) = H_{X^*Y^*}(x^*w^{-1}, wy^*)$. Assuming that $\phi(w)$ is differentiable at $\gamma_\alpha(x^*, y^*)$ with strictly negative derivative, then for any $(x_0, y_0) \in \mathbb{R}^{p+q}_+$:*

$$GES\left(W^{\alpha}_{xy;d_x,d_y}, H_{XY}\right) = \frac{\max(\alpha, 1 - \alpha)}{\gamma_\alpha(x^*, y^*)\ \phi'(\gamma_\alpha(x^*, y^*))}.$$

As noticed, the gross-error sensitivity of the order-α directional type of estimator is finite, which is a desirable property an estimator can have.

3.2 Breakdown point

According to *Donoho and Huber (1983)*, a global measure of robustness of an estimator is given by its so-called breakdown point, which gives the minimum number of data points that can be contaminated before the estimator can arbitrarily be sent to the bound of the possible set of values (usually $\pm\infty$).

Proposition 3.2. *Let* $(x, y) \in \mathbb{R}_+^{p+q}$ *and* $(d_x, d_y) \in \mathbb{R}_+^{p+q}$. *Then, for* $\alpha \in (\frac{1}{2}, 1)$*:*

$$BP\left(\widehat{\delta}_{\alpha,n}(x, y \mid d_x, d_y)\right) = \begin{cases} \dfrac{n(1-\alpha)+1}{n} & \text{if } \alpha n \in \mathbb{N}, \\ \dfrac{n-[\alpha n]}{n} & \text{otherwise,} \end{cases}$$

where [.] *denotes the integer part of a number.*

Proof of Proposition 3.2. Given the existing link between the hyperbolic and directional types of distance, it follows that the empirical breakdown point of the order-α directional type of estimator is the same as the breakdown point of the order-α directional estimator (see *Bruffaerts et al. (2013a)*). □

Because of the close relationship between the breakdown point of the estimator and the parameter α, it is crucial to choose an adequate value for α. On the one hand, a too large value might lead the efficiency estimator to break down when estimating the production frontier. On the other hand, a too small value of α implies that the efficiency estimator is farther away from the true production frontier. Given that in any practical situation the number of influential points is unknown, one needs to have a method to detect them such that an appropriate value for α can be used.

4 Detection of influential points

To obtain a robust efficiency measurement for a specific production unit, the practitioner needs to have an idea of the DMUs that can have a large influence on its efficiency. Given that there is no clear definition of what an outlier is and that there are many possible reasons for which it can be part of the sample, detecting outliers is a difficult problem. Moreover, given that there is no underlying model nor by any assumption on the distribution of inputs and outputs, outliers can only be defined as observations that are behaving differently with respect to other observations. As pointed out in *Simar (2003)*, the detection of outlying points in the context of frontier models is even more specific because of the boundary specification. To detect the presence of outlying points in frontier models, *Simar (2003)* proposed a leave-one-out approach using an order-m type of estimator. To know whether some points are potential outliers, the idea of this method is to compute for each DMU their order-m efficiency score from a sample that does not contain the DMU itself (leave-one-out approach). This operation is repeated for different values of m and the proportion of DMUs that are left outside the order-m frontier is plotted as a function of m. From this plot (non-increasing function of m), one can choose a threshold value for m and flag points that are beyond this order-m frontier as outlying points; the threshold value can for instance be chosen where an elbow effect occurs on the plot. This methodology however struggles to detect the presence of outliers that are close to each other in the data, *i.e.* a cluster of anomalous points. To counter this drawback that is known as the masking effect, the author proposes to delete points that are flagged as outliers and redo the procedure until no outlying point is found. Beside being a repetitive procedure, this "sequential deletion" of points might lead to the removal of most of the points in the sample and therefore lead to a loss of information

for some of the production units in the sample. It could be that a swamping effect takes place (*i.e. removing observations that are actually not outlying.*

In this paper, the focus is rather on knowing whereabout the true production frontier is lying and flag points that are beyond this threshold as outliers. The estimation of a boundary point (or even a discontinuity point) has been well studied in the field of deconvolution problems (see *Hall and Simar (2002)* and *Delaigle and Gijbels (2006)*). In the latter paper, the authors use a "diagnostic function" (the derivative of a kernel density estimator) and let the bandwidth of the kernel change in order to know whereabout the boundary is lying. Similar ideas are used in the following methodology. In the sequel, we consider the efficiency of a DMU that is performing at level (x, y) and look closer at production units that can have a large influence on its directional type of efficiency. To do so, we make use of the explicit relationship between both the hyperbolic and directional distance functions.

4.1 From multivariate to univariate

The problem of efficiency measurement in frontier models is typically a univariate problem as the efficiency scores defined previously are scalar numbers. In particular, the hyperbolic type of measurement for a unit performing at level $(x, y) \in \mathbb{R}_+^{p+q}$ can be recovered from a dimensionless transformation of the random vector $(X, Y) \in \mathbb{R}_+^{p+q}$. This transformation which is denoted as $Z_{xy}(X, Y)$ is expressed as follows: (see *Bruffaerts et al. (2013a)*):

$$Z_{xy}(X,Y) = \min\left\{ \min_{1\leq k\leq p} \frac{x^{(k)}}{X^{(k)}}, \min_{1\leq l\leq q} \frac{Y^{(l)}}{y^{(l)}} \right\}.$$

The order-α hyperbolic efficiency measurement can be determined from the quantile function of the distribution of the above transformation:

$$\gamma_\alpha(x,y) = Q^\alpha\big(F_{Z_{xy}(X,Y)}\big),$$

where $F_{Z_{xy}(X,Y)}$ is the cumulative distribution function of the random variable $Z_{xy}(X, Y)$ and $Q^\alpha(.)$ represents the quantile function of order α. In particular, for a value $\alpha = 1$, one recovers the traditional hyperbolic efficiency measurement.

Given the close relationship between the hyperbolic and directional orientations, the former can be used to identify influential points. Identifying influential points in a direction $d = (-d_x, d_y)$ for a DMU performing at level (x, y) with the sample $S_n = \{(X_i, Y_i) \in \mathbb{R}^{p+q}; i = 1, \ldots, n\}$ is equivalent to identify influential points for the unit perfoming at level (x^*, y^*) with the hyperbolic type of estimator with the transformed sample $S_n^* = \{(X_i^*, Y_i^*) \in \mathbb{R}^{p+q}; i = 1, \ldots, n\}$. Because the hyperbolic efficiency estimator for the unit performing at level (x, y) can be determined from the dimensionless transformation $Z_{xy}(X, Y)$, the identification of outliers is based on this dimensionless variable. The boundary point of this transformed variable $Z_{xy}(X, Y)$ yields the hyperbolic efficiency measurement for the DMU performing at level (x, y). Hence, the estimation of the boundary point related to a point (x, y) in a direction $d = (-d_x, d_y)$ boils down to the estimation of the boundary point of the variable $Z_{x^*y^*}(X^*, Y^*)$. This is why the methodology for detecting outliers presented here below concerns a unit performing at level $(x, y) \in \mathbb{R}^{p+q}$ and is for the hyperbolic type of distance.

4.2 Outlier detection methodology

General idea As shown previously, the estimation of the true hyperbolic efficiency measurement for the unit performing at level $(x, y) \in \mathbb{R}^{p+q}$ is tantamount to find the upper end point of the dimensionless transformed variable $Z_{xy}(X, Y)$. Although univariate, the detection of outliers is not at all straightforward mainly due to the boundary issue and to the fact that characteristics of the density function of the variable $Z_{xy}(X, Y)$ are totally unknown. This is why the technique considered here below aims at recovering this density function so that the boundary points can be located. An upper end point is itself characterized by the fact that the density function of the variable beyond this point is zero. A natural but naive idea to know where is it lying is to estimate the density function and consider the upper end point as the point where this estimated density is zero. Based on the sample $\{Z_{xy}(X_1, Y_1), \ldots, Z_{xy}(X_n, Y_n)\}$, it is possible to estimate via standard kernel functions the density of the variable $Z_{xy}(X, Y)$. However, there are three problems related to this approach.

First and most importantly, standard kernel density estimation techniques are generally based on symmetric types of kernels (such as Normal, Uniform and Epanechnikov) and are not appropriate for variables with upper and/or lower end points. Standard methods indeed allocate weight outside the density support. In other words, the bias of the estimated density for points that are close to boundaries is very large and does not vanish as the sample size grows (usually referred as the boundary bias). To get over this problem, different types of kernels have been proposed such as boundary kernels, cut and normalized kernels, beta kernels (see *Chen (1990)*), gamma kernels (see *Chen (2000)*) for instance (see *Karunamuni and Alberts (2005)* for an extensive review). All of those methods correct the boundary bias for points that are close to the boundary but preserve the properties of the standard kernel density estimator for points that are not close to the boundary (*i.e.* interior points). Those specific kernels work well when the boundary points are known. In our case, they are unknown which makes the problem more challenging. This is why, to our purpose, an adapted skew-normal kernel function is used to account for the fact that the variable $Z_{xy}(X, Y)$ has an upper end point which is unknown.

A second issue in kernel density estimation is the bandwidth choice. Different rules and techniques exist to find the optimal value of the bandwidth (in the sense that it minimizes the asympotic mean squared error (AMSE) or its integrated version (AMISE)). In practice, it is unusual that the practitioner comes up with a satisfactory rule for the many different settings that may arise. This is the reason why in our context, we let the bandwidth range over a grid of values to allow more flexibility when detecting outlying points. This way of doing is in line with the ideas presented in *Delaigle and Gijbels (2006)* in the case of deconvolution problems.

The third problem is that the sample is potentially contaminated with outlying points. In other words, the sample might contain points that are beyond the true upper end point of the variable $Z_{xy}(X, Y)$. Because of the presence of those outlying points, it might be that the upper end point will be wrongly estimated. However, this problem is no longer an issue because we use on the one hand an adapted kernel function to deal with the boundary issue and on the other hand a variable bandwidth. The reason for which outliers may not be such a big issue is that kernel functions consider the local neighbourhood of the point at which the density is estimated. As will be shown in the

simulations, using an adapted kernel function such as the skew-normal one with a varying bandwidth inherently solves the problem of having a potentially contaminated sample.

The skew normal kernel function Suppose the DMU being analysed is performing at level $(x, y) \in \mathbb{R}_+^{p+q}$ and that the hyperbolic efficiency measurement is considered. The standard kernel density estimator (KDE) of the univariate transformed sample $\{Z_{xy}(X_1, Y_1), \ldots, Z_{xy}(X_n, Y_n)\}$ (which will be simply denoted as $\{Z_1, \ldots, Z_n\}$ in the sequel) is defined as:

$$\widehat{f}_{KDE}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\Big(\frac{z - Z_i}{h}\Big),$$

with $k(.)$ being a univariate kernel function and $h > 0$ being the bandwidth. It is common to use a kernel which is unimodal and symmetric about zero such as the Normal type of kernel which is defined as follows:

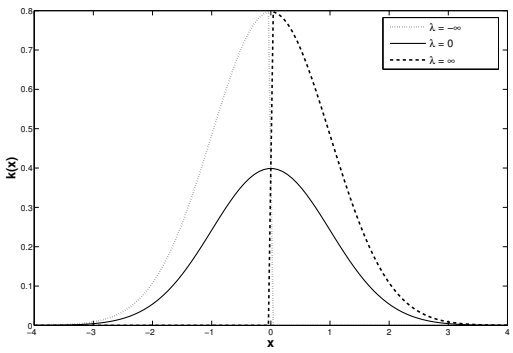
$$k(z) = \phi(z) = \frac{1}{\sqrt{2\pi}}\exp\Big(\frac{-z^2}{2}\Big).$$

Azzalini (1985, 1986) proposed a more general version of the Normal type of kernel which is the skew-normal type of kernel that is given as follows:

$$g(z; \lambda) = 2\phi(z)\Phi(\lambda z), \quad \lambda \in \mathbb{R}$$

where ϕ and Φ denote the pdf and cdf of the standard normal distribution respectively. The additional parameter λ controls the skewness of the kernel. Figure 1 depicts the skew normal kernel for specific values of λ and shows how the shape of the kernel changes according to the value of λ.

Figure 1: Plot of the skew normal kernel function for different values of λ.



A more flexible way of estimating a density function than with the standard Normal kernel is to use the skew-normal kernel which yields the following density estimation:

$$\widehat{f}_{SKDE}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}g\Big(\frac{z - Z_i}{h}; \lambda\Big),$$

where $\lambda = \lambda(h, z)$. The advantage of doing kernel density estimation with such a kernel function is that the shape of the kernel depends on the point at which the density is estimated. In many applications where boundary problems arise, the upper or lower end points are usually known. If a variable has for instance a lower end point at 0, one can use $\lambda = \lambda(h, z) = h/z$ as a way of controlling the skewness of the kernel for points that are close to 0. In our case, the upper end point is unknown and therefore the λ parameter can not only depend on h and z. An alternative way of doing is to let the parameter λ depend as well on the sample $\{Z_1, \ldots, Z_n\}$.

At the point where the density is to be estimated, we capture the "local" skewness by looking at the number of sample points that are to the right and to the left of the point and at the same time within a perimeter of size h. Hence, a possible way to adapt the way the kernel should look like is by choosing the following choice:

$$\lambda(z, h, \mathcal{Z}_n) = h \, \frac{\sum_{i=1}^{n} \mathbb{1}_{[z-h,z]}(Z_i) - \sum_{i=1}^{n} \mathbb{1}_{[z,z+h]}(Z_i)}{\left(\sum_{i=1}^{n} \mathbb{1}_{[z-h,z]}(Z_i)\right)\left(\sum_{i=1}^{n} \mathbb{1}_{[z,z+h]}(Z_i)\right)},$$

where $\mathcal{Z}_n = \{Z_i \, ; i = 1, \ldots, n\}$. If the density is estimated at z and that there are points to its left but none to its right (in a window controlled by h), λ takes the value $-\infty$ and therefore the kernel takes the form of the cut and normalized kernel. The same happens for the lower end point where λ takes the value ∞. For interior points, the λ parameter should be relatively close to 0 and the kernel used should therefore be similar to the standard Normal one.

A range of bandwidths The idea of the method is to let the bandwidth range over a grid of values and each time look where the candidate for the upper end point is. By doing so, we allows ourselves more flexibility than by choosing a single value for h. For this range of values, the kernel density estimation with the skew-normal type of kernel is applied. As in *Delaigle and Gijbels (2006)* for deconvolution problems, we pick a geometric decreasing grid for the bandwidth such that as h gets smaller, the grid becomes thinner. This allows to be more careful about a possible candidate for the boundary point as the bandwidth converges towards 0. The candidate necessarily needs to be above the median of all observations otherwise the sample would include more outliers than normal observations. Among all possible candidates, we choose the one which appeared the most often and which is the most likely to be the boundary point. As a starting value for h, one can use the robust rule $h = 2.34 \min(\widehat{\sigma}_Z, Qn(Z))n^{-1/5}$, where $\widehat{\sigma}_Z$ and $Qn(Z)$ are the standard deviation and the robust scale estimator (see *Rousseeuw and Croux (1993)*) of the variable Z. This starting value is a robust rule for the case of kernel density estimation for variables with unbounded support. This choice for the starting value of h guarantees "oversmoothing" for the case where the variable has a bounded support.

4.3 Algorithm

The proposed method to identify influential points for a DMU performing at level (x, y) in the direction $d = (-d_x, d_y)$ can be summarized as follows:

1. Given the sample $S_n = \{(X_i, Y_i) \in \mathbb{R}^{p+q}; i = 1, \ldots, n\}$, apply the following transformations: $X_i^* = \exp(X_i./d_x)$ and $Y_i^* = \exp(Y_i./d_y)$. Apply the same type of

transformation to the analysed DMU: $x^* = \exp(x./d_x)$ and $y^* = \exp(y./d_y)$. [3]

2. Transform the data in the following way:

$$Z_{x^*y^*}(X_i^*, Y_i^*) = \min\left\{ \min_{1\leq k\leq p} \frac{x^{*(k)}}{X_i^{*(k)}}, \min_{1\leq l\leq q} \frac{Y_i^{*(l)}}{y^{*(l)}} \right\}, \qquad i = 1, \ldots, n.$$

3. Pick a grid of values for the bandwidths. For instance, generate a sequence $h_{i+1} = 0.85\ h_i$ with as initial value $h = 2.34 \min(\widehat{\sigma}_{Z^*}, Qn(Z^*))n^{-1/5}$ and until h is very close to 0.

4. Based on the sample $\{Z_{x^*y^*}(X_1^*, Y_1^*), \ldots, Z_{x^*y^*}(X_n^*, Y_n^*)\}$, estimate the density for each bandwidth using the skew-normal kernel with the parameter λ as defined previously. For each bandwidth, select the candidate upper end point (*i.e.* the point for which the density becomes zero and which is greater than the median value of all observations).

5. Given this set of candidates, the point that is most likely to be the closest to the boundary point is the most frequent value (*i.e.* the mode) among all those candidates.

6. Influential points for the DMU performing at level (x, y) in the direction $d = (-d_x, d_y)$ are points that are beyond this threshold.

5 Numerical applications

The goal of this section is threefold. First, the outlier detection procedure is demonstrated via some simulations. In those examples, we only consider the hyperbolic type of orientation as the purpose of this first experiment is to see whether the methodology to identify influential points can discriminate efficiently influential points from non-influential ones. Second, still using simulations, we show that when identifying anomalous observations, the direction that is used to measure the efficiency of a DMU matters. Thirdly, we show how the proposed methodology can be applied on a real life example concerning the research efficiency of US universities.

5.1 Detecting influential points

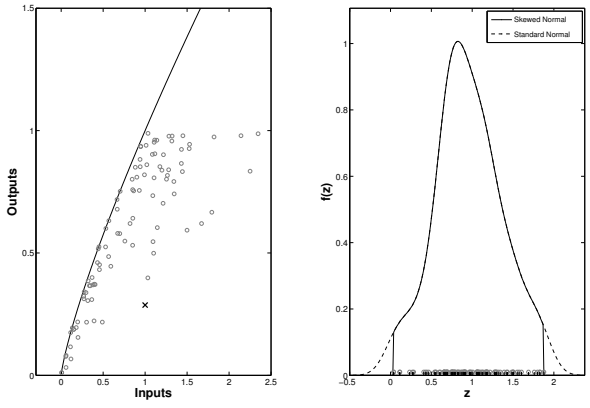
Bivariate The purpose of this section is to show on simple examples with only one input and output (*i.e.* $p = q = 1$) how the above methodology can be used to identify influential points. By doing simulations and thus controlling the underlying DGP, it is clear which observations are outlying as they are generated from a totally different DGP. The advantage of considering this bivariate case is that it can be clearly seen which points are influencing the efficiency of DMUs for different directions. The following standard Data Generating Process (DGP) is considered: $\tilde{X} \sim U[0, 1]$, $Y = \tilde{X}^{0.8}$, $X = \tilde{X} \times e^{-U}$ with $U \sim \exp(1/3)$[4]. Outlying points are generated as $X_{out} \sim \mathcal{N}_2(|\ \mu\ |, \Sigma)$, where

[3] For the input and output orientations for which $d_y = 0$ and $d_x = 0$ respectively, see the transformations given earlier in footnote 1.

[4] The way the data is generated is not important here as the goal of the experiment is only to illustrate the methodology to detect anomalous observations.

$\mu = (0,1)'$ and $\Sigma = [0.01, 0; 0, 0.01]$. A sample of $n = 100$ points is generated according to this DGP and the point being analysed performs at level $(x_0, y_0) = (1, 0.287)$ in such a way that its hyperbolic efficiency $\gamma(x_0, y_0) = 2$. The first situation that is considered is the one in which no outlying points are added to the data and is depicted on the left plot of Figure 2. The right plot of this Figure shows both density estimates (the skew-normal and the standard normal kernels) with the "optimal" value of h. As can be seen, we learn from both densities that there are no outliers. In Figure 3, one outlier has been added to the above setting in order to see whether the methodology is able to detect this outlying point. The right plot of this figure indicates clearly the presence of one outlier. Figure 4 shows the results for $\varepsilon = 5\%$ of outliers in the data and as can be seen all of the outlying points are detected from the adapted kernel density estimator. As can be noticed, both kernel density estimators are similar for interior points whereas they behave clearly differently near isolated data points and in particular near boundaries. Let us notice that the outliers that have been added are relatively close to the true production frontier. The same simulations as above were performed with a sample size $n = 1000$ and the kernel density estimators are shown in Figure 5 for the three contamination levels.

Figure 2: Simulated dataset with $n = 100$ with no contamination and plot of the kernel density estimators (both standard Normal and skew-normal kernels) of the transformed variable relative to the analysed DMU (cross).



Sensitivity and specificity The above simulations are based for different settings. To convince ourself of the discriminating power of this outlier detection technique, the same exercice as above was performed a large number of times (B=1000). The same DGP as above was used and outliers were generated according to the same process (outlying points which are not far from the true production frontier). To assess how good or bad the detection of outliers is, one can use sensitivity and specificity measures. The former measure yields the probability of detecting non-outlying observations among non-outlying observations while the other gives the probability of identifying outliers among outlying observations. Those measures naturally depend on the contamination

Figure 3: Simulated dataset with $n = 100$ with $\varepsilon = 1\%$ and plot of the kernel density estimators (both standard Normal and skew-normal kernels) of the transformed variable relative to the analysed DMU (cross).

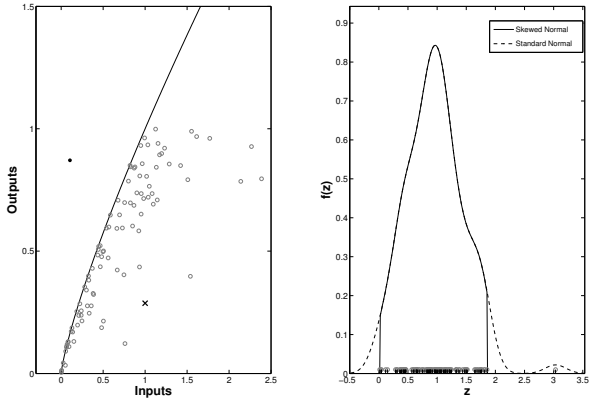
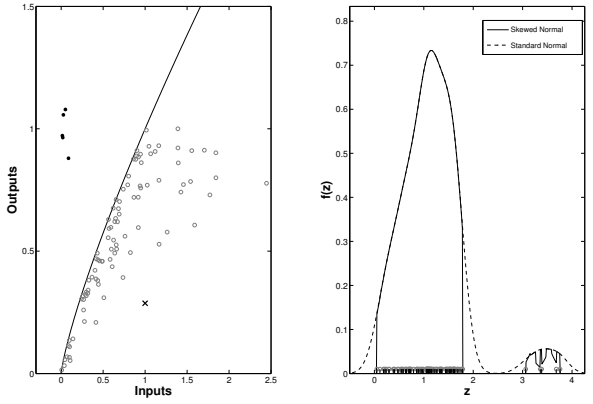
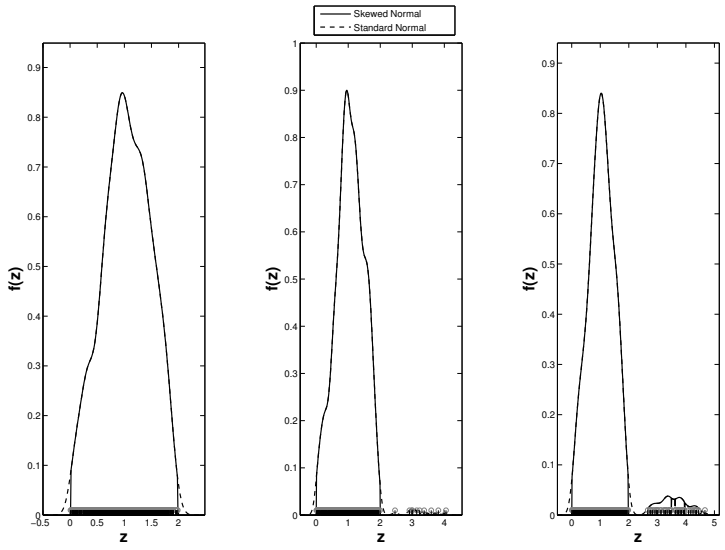


Figure 4: Simulated dataset with $n = 100$ with $\varepsilon = 5\%$ and plot of the kernel density estimators (both standard Normal and skew-normal kernels) of the transformed variable relative to the analysed DMU (cross).



that is added to the sample. However, it can give an idea of the discriminating power of the technique for "moderate" outliers (the same type as in the previous simulations). Table 1 presents the results for the two sample sizes and for different contamination levels. This table shows that the discriminating power of the method is relatively high as all measures of sensitivity and specificity are close to 1.

Figure 5: Kernel density estimators of the transformed variable relative to the analysed DMU for three simulations with $n = 1000$ with $\varepsilon = 0\%$, $\varepsilon = 1\%$, and $\varepsilon = 5\%$ respectively.



	n=100		**n=1000**	
	sensitivity	*specificity*	*sensitivity*	*specificity*
$\varepsilon = 0\%$	0.994	-	1	-
$\varepsilon = 1\%$	1	0.996	1	0.997
$\varepsilon = 5\%$	1	0.996	1	0.997
$\varepsilon = 10\%$	0.999	0.994	1	0.98

Table 1: Sensitivity and specificity of the outlier detection methodology (computed from 1000 simulations) for different sample sizes (n=100, 1000) and for different contamination levels ($\varepsilon = 0\%, 1\%, 5\%, 10\%$).

Multivariate In the situation where more than one input and more than one output are considered, it is difficult to visualize the points that are used in the construction of the estimated frontier and which can potentially be outlying. In the same lines as in *Park et al. (2000)*, *Simar (2003)* and *Daouia and Gijbels (2011)*, we illustrate the methodology in a multiple input and multiple output case ($p = 2, q = 2$). The goal of this experiment is to see whether the same outliers can be detected as in the above references. The following DGP generates efficiencies along output rays. The efficient

frontier is given by the following function:

$$y^{(2)} = 1.0845(x^{(1)})^{0.3}(x^{(2)})^{0.4} - y^{(1)},$$

where $y^{(j)}$ and $x^{(j)}$ represent the j^{th} component of y and x respectively ($j \in \{1, 2\}$). Let $X_i^{(j)} \sim U[1, 2]$ and $\tilde{Y}_i^{(j)} \sim U[0.2, 5]$. In the output space, the random rays are given by the slopes $K_i = \tilde{Y}_i^{(2)}/\tilde{Y}_i^{(1)}$. The generated random points on the frontier are given as follows:

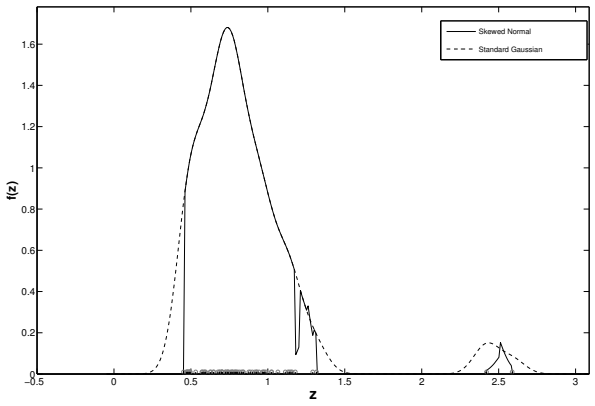
$$Y_{i,eff}^1 = \frac{1.0845(X_i^{(1)})^{0.3}(X_i^{(2)})^{0.4}}{K_i + 1},$$
$$Y_{i,eff}^2 = 1.0845(X_i^{(1)})^{0.3}(X_i^{(2)})^{0.4} - Y_{i,eff}^1.$$

The efficiencies are generated as e^{-U_i} where $U_i \sim \exp(1/3)$, so that the final outputs are $Y_i = Y_{i,eff} \times e^{-U_i}$. As in *Simar (2003)* and *Daouia and Gijbels (2011)*, three outliers are added to the above DGP:

- $X_{out1} = (1.5, 1.5)$ and $K_{out1} = 1$,
- $X_{out2} = (1.25, 1.75)$ and $K_{out2} = 1/2$,
- $X_{out3} = (1.75, 1.25)$ and $K_{out3} = 2$.

The DMU under analysis is a DMU performing at level $(x_0, y_0) = (1.8, 1.8, 0.5, 0.5)$ and the orientation used to assess its efficiency is the output one. The methodology for identifying outliers is applied to this DMU for the output direction. Figure 5.1 shows clearly that the three points have been identified.

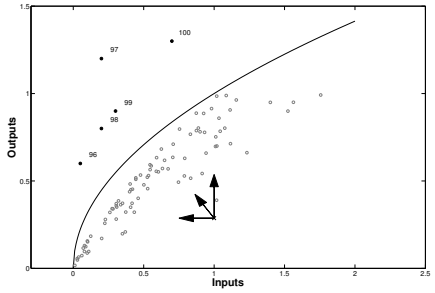
Figure 6: Kernel density estimators of the transformed variable relative to the analysed DMU for the output orientation with three outlying points.



5.2 Direction matters

Bivariate The second purpose of this experiment is to show that the concept of outlying point in frontier models will depend mainly on the type of distance that is being used to assess efficiency. Back to the simple bivariate case described above, three different directions are considered to benchmark the DMU performing at level (x_0, y_0): the input and output directions as well as a "median" direction for which $d = (med(X_i), med(Y_i))$ (*i.e.* the median of the inputs and outputs of all DMUs respectively). This median direction is the robust equivalent of the choice that was considered in *Simar and Vanhems (2012)* where the authors use the empirical means as directions. Data points ($n = 100$) are generated according to the same DGP as before and five outlying points are deliberately added to the sample. The left plot of Figure 7 depicts the situation considered and shows the three directions that are used. By using the outlier detection methodology, different outlying points are found for each direction considered. For instance, in the input direction, the DMU being analysed is not much influenced by other production units whereas it is influenced by other units for the output and median directions. This small example shows that when identifying DMUs that are influential for a given production unit, it is important to take into account the direction that is used to benchmark its efficiency. Given that we are in a nonparametric context and therefore are exposed to the curse of dimensionality, it is important not to remove from the sample some production units that are influential for some of the DMUs but that are as well informative on the efficiency of other DMUs. Hence, an individual analysis on each DMU is preconized when the researcher's goal is to infer on individual efficiency scores. More importantly, the direction used in the analysis has clearly an impact on the number of influential points. Contrarily to other settings in statistics, the notion of influential point depends totally on the DMU being analysed as well as on the direction used to assess its efficiency.

Figure 7: Simulated dataset with $n = 100$ points among which five outliers and corresponding influential points of (x_0, y_0) for different directions.



Orientation	**Outliers**
Input $(d_x, d_y) = (0, 1)$	-
Output $(d_x, d_y) = (1, 0)$	97, 100
Directional $(d_x, d_y) = (med(X_i), med(Y_i))$	97, 98, 99

Multivariate The same exercise as above is performed in the multivariate context with the same DGP as in the previous section but for other directions than just the output one. The DMU analysed is simply the DMU performing at level $(x_0, y_0) = (\bar{X}, \bar{Y})$, where $\bar{X}$ and $\bar{Y}$ represent the means of the input and output vectors respectively of the underlying DGP (without contamination). This DMU represents an "average" DMU

from the sample. Table 2 gathers results regarding this DMU for the different directions considered as well as the numberof production units in the dominating set for each of those directions. As can be observed among the three added outliers, only one has been detected as outlier (in the output direction). This is simply due to the fact that the three outliers are not always in the dominating set of the DMU which indicates that by construction they are not influential for measuring efficiency.

Orientation	**Size of reference set**	**Outliers**
Input	24	-
Output	26	out1
Directional	103	-

Table 2: Influential points in the multivariate case ($p = q = 2$) for the DMU under analysis for different directions.

Conclusion Once the influential points of a DMU for a specific direction have been identified, the practitioner can use in his analysis as trimming parameter the value $\alpha = 1 - k/n$, where k represents the number of influential points that was detected. By identifying outlying points for each DMU in the sample, a specific value of α can be chosen for each of them. Compared to the usual applications of partial frontiers in which a single value of α for all DMUs is used, this individual analysis of DMUs allows not to underestimate the efficiency of some of the production units. By picking a value for α based on the number of outlying points, the researcher knows that for the DMU being analysed the partial frontier is close to the frontier of the production set and therefore its associated order-α efficiency score is close to the true efficiency measurement.

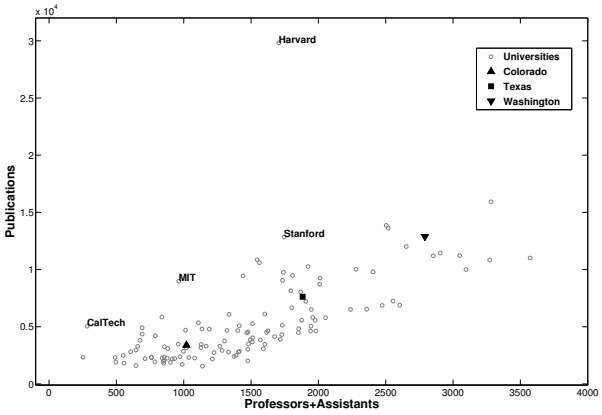
5.3 Real life dataset

Results from the previous sections are illustrated on a dataset of 124 US universities that was analysed in *Bruffaerts et al. (2013b)*. This dataset was used to measure the research efficiency of American universities. In the following experiments, focus is devoted to three particular universities: the Colorado State University, the University of Texas at Austin and the University of Washington. Relatively to the sample at hand, these universities are respectively small scale, medium scale and large scale universities. The goal is on the one hand to identify universities that are influential for the three analysed institutions and on the other hand to provide a robust efficiency measurement for those three institutions and for different directions.

Bivariate As a first experiment, we consider as input the aggregated number of professors and assistant professors (for the year 2007) and as output the total number of publications (between the years 2008 to 2011). Figure 8 shows on the x-axis the inputs and on the y-axis the outputs of all US universities of the sample. It is clear from this figure that some universities behave differently in comparison to the majority of universities. In particular, Harvard university has a huge number of publications compared to

the rest of universities while some other institutions such as Caltech, MIT and Stanford for instance seem to behave slightly differently with respect to the rest of universities.

Figure 8: Inputs (aggregated number of Professors and Assistant professors) and outputs (number of publications) for the 124 US universities, among which the three analysed institutions.



As in the section dedicated to simulations, the methodology for identifying influential points is applied to each of the three universities and for three different directions (the input, output and median directions). Once influential universities have been detected for the three institutions being analysed, a robust efficiency measurement of the true efficiency score may be computed. The results are shown in Table 3. From this table, one can observe that Harvard University is influential for the Washington University regardless of the direction used to assess its efficiency while Harvard University has no influence at all on the efficiency score of the Colorado State University for instance. Once again, this proves that removing outlying observations from the sample is not ideal given that those same observations can prove to be useful in the efficiency measurement of other universities. Moreover, this table allows fair comparisons between the three universities given that all of the influential universities for each of the three institutions have been identified. By doing so, the researcher aims at estimating the true efficiency measurement related to the three universities. It can be noticed that among these universities, the Washington University is the most efficient for the three directions considered. The Colorado State University is more efficient than the Texas University for the median direction whereas it is performing less efficiently for both the input and output orientations. Although simple, this example has proved to be informative and allows to understand in what way some universities can influence the efficiency measurement of others.

Multivariate As a second illustration, we consider a situation with two inputs and two outputs ($p = 2$ and $q = 2$). In addition to the aggregated number of Professors and Assistant Professors, the total of research expenses in dollars (between the years 2006

University	**Direction** *Input* Outliers	$\widehat{\theta}_{\alpha,n}(x,y)$	*Output* Outliers	$\widehat{\lambda}_{\alpha,n}(x,y)$	*Median* Outliers	$\widehat{\delta}_{\alpha,n}(x,y)$
Colorado	Caltech	0.66	MIT Georgia	1.48	Caltech	0.22
Texas	MIT	0.76	Harvard Stanford	1.68	-	0.3
Washington	Harvard	0.89	Harvard	1.07	Harvard	0.19

Table 3: Influential universities and robust efficiency estimator for the input, output and median directions for three US universities with one input (aggregated number of Professors and Assistant Professors) and one output (number of publications).

and 2007) is considered as an additional input. The output of the research production is measured by the total number of publications as well as the total number of citations (both between the years 2008 and 2011). More information about those variables can be found in *Bruffaerts et al. (2013b)*. Because of the curse of dimensionality ($p = q = 2$), it is to be expected that many universities take part in the construction of the estimated production frontier and can therefore be themselves outlying for other universities. The methodology for identifying influential points is applied and results are shown in Table 4. As can be observed from this table, irrespective of the direction used, the Colorado State University has no influential point. The Texas University is itself an influential university for both the output and median orientations while this institution has no influential point for the input orientation. The same can be observed for the Washington University. For the three orientations, the Texas University is the most efficient university among the three institutions while the Colorado State University is the least efficient. Compared to the previous analysis in which Harvard University was influential for the Washington University (for all orientations) and for the Texas University (for the output orientation), none of the three university is influenced by Harvard University. This is mainly due to the introduction of the research expenses as an input in the analysis. Harvard University indeed invests much money in research but is not making the best out of this input in comparison to some other universities which have a relatively high number of publications and citations for a moderate investment in research expenses. This example shows that the choice of the inputs and outputs used to define the production process are very important and has a clear impact on the observations that are influential to the efficiency measurement of universities.

University	Direction					
	Input		*Output*		*Median*	
	Outliers	$\widehat{\theta}_{\alpha,n}(x,y)$	Outliers	$\widehat{\lambda}_{\alpha,n}(x,y)$	Outliers	$\widehat{\delta}_{\alpha,n}(x,y)$
Colorado	-	0.9	-	1.38	-	0.02
Texas	-	1	Texas*	0.61	Texas* Illinois	-0.4
Washington	-	0.98	Los Angeles Washington*	0.73	Los Angeles Washington*	-0.33

Table 4: Influential universities and robust efficiency estimator for the input, output and median directions for three US universities with two inputs (aggregated number of Professors and Assistant Professors and research expenses) and two outputs (number of publications and citations). A "*" indicates that the university being analysed is itself an influential university.

6 Conclusion

The directional distance function provides a very flexible way of measuring efficiency as it includes the two most famous efficiency orientations which are the input and output ones. This general orientation allows to measure for any direction the distance from a certain DMU to the frontier of the production set. Thanks to the relationship between the hyperbolic and directional orientations, the robustness properties of the directional efficiency estimator have been found. Those properties shed light on the importance of the trimming parameter α.

When performing efficiency analysis, it is crucial to account for the possible presence of outlying points in the sample. In the case of the directional orientation, outliers are defined with respect to the production unit being analysed as well as on the direction used to assess efficiency. The outlying points depend crucially on the DMU being analysed and on the direction that is used to assess its efficiency. Hence, different outliers can be found for different DMUs and different directions. For a given production unit and a given direction, one needs to have an idea of the number of outlying points to estimate the frontier of the production set. Once this has been done, an adequate value for the quantile can be chosen (*i.e.* $\alpha = 1 - k/n$). The data-driven methodology proposed here provides a way of detecting anomalous observations with respect to each DMU of the sample and to the direction that was chosen. The example on universities has shown the interest in identifying atypical observations. Finally, let us insist that the detection of outliers should be used as a first step in any efficiency analysis in nonparametric frontier models.

Efficiency analysis is widely used in many economic applications. The proposed method to detect outliers takes into account the concerns of the applied researcher whose goal is to benchmark in a "fair way" a production unit with respect to all other units. The method takes into account the fact that outliers are DMU specific as well as direction

specific. This provides the applied researcher with an easy and simple tool to detect outliers in this totally nonparametric and multidimensional setting.

Appendix

Proof of Proposition 3.1. By definition of the quantile based hyperbolic type of distance:

$$\gamma_\alpha(x,y) = R^\alpha_{xy}(H_{XY}) = \sup\{\gamma > 0 \mid H_{XY}(\gamma^{-1}x, \gamma y) > 1-\alpha\}.$$

From *Bruffaerts et al. (2013)*, the influence function of the hyperbolic type of estimator has the following form:

$$IF\left((x_0,y_0), R^\alpha_{xy}, H_{XY}\right) = \frac{\alpha - \mathbb{1}\left(Z_{xy}(x_0,y_0) \leq \gamma_\alpha(x,y)\right)}{f_{Z_{xy}(X,Y)}(\gamma_\alpha(x,y))},$$

where

$$Z_{xy}(x_0,y_0) = \min\left\{\min_{1\leq k\leq p} \frac{x^{(k)}}{x_0^{(k)}}, \min_{1\leq l\leq q} \frac{y_0^{(l)}}{y^{(l)}}\right\}.$$

The cdf and pdf of the variable $Z_{xy}(X,Y)$ are given respectively as:

$$\begin{aligned} F_{Z_{xy}(X,Y)}(\omega) &= 1 - H_{XY}(\omega^{-1}x, \omega y) = 1 - J(\omega), \\ f_{Z_{xy}(X,Y)}(\omega) &= F'_{Z_{xy}(X,Y)}(\omega) = -J'(\omega). \end{aligned}$$

Given the link between the hyperbolic and directional orientations, we have that:

$$\begin{aligned} \delta_\alpha(x,y;d_x,d_y) &= \log(\gamma_\alpha(x^*,y^*)) \\ &= \log(\sup\{\gamma > 0 \mid H_{X^*Y^*}(\gamma^{-1}x^*, \gamma y^*) > 1-\alpha\}) \\ &= \log(R^\alpha_{x^*y^*}(H_{X^*Y^*})) \end{aligned}$$

which enables us to write $W^\alpha_{xy;d_x,d_y}(H_{XY}) = \log(R^\alpha_{x^*y^*}(H_{X^*Y^*}))$.
Using the chain rule of calculus and the previous result, we obtain the following influence function for the directional type of estimator:

$$\begin{aligned} IF\left((x_0,y_0), W^\alpha_{xy;g_x,g_y}, H_{XY}\right) &= \frac{IF\left((x_0,y_0), R^\alpha_{x^*y^*}, H_{X^*Y^*}\right)}{\gamma_\alpha(x^*,y^*)} \\ &= \frac{\alpha - \mathbb{1}\left(Z_{x^*y^*}(x_0,y_0) \leq \gamma_\alpha(x^*,y^*)\right)}{\gamma_\alpha(x^*,y^*)\ f_{Z_{x^*y^*}(X^*,Y^*)}(\gamma_\alpha(x^*,y^*))} \end{aligned}$$

where

$$Z_{x^*y^*}(x_0,y_0) = \min\left\{\min_{1\leq k\leq p} \frac{x^{*(k)}}{x_0^{(k)}}, \min_{1\leq l\leq q} \frac{y_0^{(l)}}{y^{*(l)}}\right\},$$

which has the following cdf and pdf:

$$\begin{aligned} F_{Z_{x^*y^*}(X^*,Y^*)}(\omega) &= 1 - H_{X^*Y^*}(\omega^{-1}x^*, \omega y^*) = 1 - \phi(\omega), \\ f_{Z_{x^*y^*}(X^*,Y^*)}(\omega) &= F'_{Z_{x^*y^*}(X^*,Y^*)}(\omega) = -\phi'(\omega). \end{aligned}$$

□

References

Y. Aragon, A. Daouia, and C. Thomas-Agnan (2005). Nonparametric frontier estimation : A conditional quantile-based approach. *Cambridge University Press, Econometric Theory*, 21:358-389.

A. Azzalini (1985). A class of distributions which includes the normal ones. *Scandinivian Journal of Statistics*, 12:171-178.

A. Azzalini (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46:199-208.

S.X. Chen (1999) Probability density function estimation using gamma kernels. *Ann. Inst. Statist. Math.*, 52:471-480.

S.X. Chen (2000) Beta kernel estimators for density functions. *Computational Statistics and Data Analysis*, 31:131-145.

C. Bruffaerts, B. De Rock and C. Dehon (2013a). The robustness of the hyperbolic efficiency estimator. *Computational Statistics and Data Analysis*, 57(1):349-363.

C. Bruffaerts, B. De Rock and C. Dehon (2013b). The research efficiency of US universities: a nonparametric frontier modelling approach. *Working paper*.

C. Cazals, J. P. Florens and L. Simar (2002). Nonparametric frontier estimation : A robust approach. *Journal of Econometrics*, 106:1-25.

R.G. Chambers, Y.H. Chung and R. Färe (1996). Benefit and distance functions. *Journal of Economic Theory*, 70:407-419.

R.G. Chambers, Y.H. Chung and R. Färe (1998). Profit, directional distance functions and Nerlovian efficiency. *Journal of Optimization Theory and Applications*, 98:351-364.

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

T. Coelli, A. Estache, S. Perelman and L. Trujilo (2003). A primer on efficiency measurement for utilities and transport regulators. *World Bank Institute*.

A. Daouia, J.P. Florens and L. Simar (2010). Frontier estimation and extreme values theory. *Bernoulli*, 16(4):1039-1063.

A. Daouia, L. Gardes and S. Girard (2012). Nadaraya's Estimates for Large Quantiles and Free Disposal Support Curves. In: Exploring Research Frontiers in Contemporary Statistics and Econometrics, A Festschrift for Léopold Simar. Springer-Verlag Berlin Heidelberg, pp. 1-22.

A. Daouia and I. Gijbels (2011). Robustness and inference in nonparametric partial frontier modeling. *Journal of Econometrics*, 161:147-165.

A. Daouia and A. Ruiz-Gazen (2006). Robust Nonparametric frontier estimators : qualitative robustness and influence function. *Statistica Sinica*, 16:1233-1253.

C. Daraio and L. Simar (2014). Directional Distances and their Robust versions: Computational and Testing Issues. to appear in *European Journal of Operational Research*.

A. Delaigle and I. Gijbels (2006). Data-driven boundary estimation in deconvolution problems. *Computational Statistics and Data Analysis*, 50(8):1965-1994.

D. Deprins, L. Simar and H. Tulkens (1984). Measuring Labor Inefficiency in Post Offices, in: M. Marchand, P. Pestiau, H. Tulkens (Eds.). The Performance of Public Enterprises: Concepts and Measurements, North-Holland, Amsterdam.

D.L. Donoho, P.J. Huber (1983). *The notion of breakdown point. In: Bickel, P.J., Doksum, K.A., Hodges Jr., J.L., (Eds), A Festschrift for Erich L. Lehmann.* Wadsworth, Belmont, CA, pp. 157-184.

A. Emrouznejad, B.R. Parker and G. Tavares (2008). Evaluation of research in efficiency and productivity: a survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences*, 42(3):151-157.

R. Färe, S. Grosskopf and C. A. K. Lovell (1985). *The Measurement of Efficiency of Production.* Boston: Kluwer-Nijhoff Publishing.

R. Färe and S. Grosskopf (2000). Theory and application of directional distance functions. *Journal of Productivity Analysis*, 13, 93-103.

R. Färe and S. Grosskopf (2004). New directions: Efficiency and productivity. Boston: Kluwer-Nijhoff Publishing.

R. Färe, S. Grosskopf and D. Margaritis (2008). Efficiency and productivity: Malmquist and more. In: Fried, H., Lovell, C.A. K., Schmidt, S. (Eds.), The Measurement of Productive Efficiency, 2nd ed. Oxford University Press..

M. J. Farrell (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society A*, 120:253-281.

S. Gattoufi, M. Oral and A. Reisman (2004). Data envelopment analysis literature: a bibliography update (1951-2001). *Journal of Socio-Economic Planning Sciences*, 38:159-229.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel (1986). Robust Statistics: The Approach Based on Influence Functions. *Wiley*.

P. Hall and L. Simar (2002). Estimating a Changepoint, Boundary, or Frontier in the Presence of Observation Error. *Journal of the American Statistical Association*, 97(458):523.

R. J. Karunamuni and T. Alberts (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, 2:191-212.

B. Park, L. Simar and C. Weiner (2000). The FDH estimator for productivity efficiency scores: Asymptotic properties. *Econometric Theory*, 16, 855-877.

P. Rousseeuw anc C. Croux (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88:1273-1283.

L. Simar (2003). Detecting Outliers in Frontier Models: A Simple Approach. *Journal of Productivity Analysis*, 20:391-424.

L. Simar and A. Vanhems (2012). Probabilistic characterization of directional distances and their robust versions. *Journal of Econometrics*, 166:342-354.

D.C. Wheelock and P.W. Wilson (2008). Non-parametric, unconditional quantile estimation for efficiency analysis with an application to Federal Reserve Check Processing Operations. *Journal of Econometrics*, 145(1-2): 209-225.

P.W. Wilson (1993). Detecting Outliers in Deterministic Nonparametric Frontier Models with Multiple Outputs. *Journal of Business and Economic Statistics*, 11: 319-323.

P.W. Wilson (1995). Detecting Influential Observations in Data Envelopment Analysis. *Journal of Productivity Analysis*, 6: 27-45.