INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0146 Revised

DETECTING OUTLIERS IN FRONTIER MODELS: A SIMPLE APPROACH

L. SIMAR

http://www.stat.ucl.ac.be

Detecting Outliers in Frontier Models: A Simple Approach

Léopold SIMAR[1]
Institut de Statistique
Université Catholique de Louvain
Louvain-la-Neuve, Belgium
simar@stat.ucl.ac.be

Revised version
June 7, 2002

Abstract

In frontier analysis, most of the nonparametric approaches (DEA,FDH) are based on envelopment ideas which suppose that with probability one, all the observed units belong to the attainable set. In these "deterministic" frontier models, statistical theory is now mostly available (Simar and Wilson, 2000a). In the presence of super-efficient outliers, envelopment estimators could behave dramatically since they are very sensitive to extreme observations. Some recent results from Cazals, Florens and Simar (2002) on robust nonparametric frontier estimators may be used in order to detect outliers by defining a new DEA/FDH "deterministic" type estimator which does not envelop all the data points and so is more robust to extreme data points. In this paper we summarize the main results of Cazals, Florens and Simar (2002) and we show how this tool can be used for detecting outliers when using the classical DEA/FDH estimators or any parametric techniques. We propose a methodology implementing the tool and we illustrate through some numerical examples with simulated and real data. The method should be used in a first step, as an exploratory data analysis, before using any frontier estimation.

Keywords: Nonparametric frontier models, Robust estimators, Outliers.

[1] I am grateful for constructive comments from O. B. Olesen and P. W. Wilson on a previous version of the paper. Research support from "Projet d'Actions de Recherche Concertées" (No. 98/03–217) from the Belgian Government is acknowledged.

1 Introduction and Notations

In frontier analysis, most of the nonparametric approaches (DEA, FDH) are based on envelopment ideas which suppose that with probability one, all the observed units belong to the attainable set. Economic theory of firms (Koopmans, 1951, Debreu, 1951, Shephard, 1970), introduces the production set, where activity is described through a set of p inputs $x \in I\!R_+^p$ used to produce a set of q outputs $y \in I\!R_+^q$, is defined as the set of physically attainable points (x, y):

$$\Psi = \{(x, y) \in I\!R_+^{p+q} \mid x \text{ can produce } y\}. \tag{1.1}$$

Then the production process, which generates observations $\{(x_i, y_i)|i = 1, \dots, n\}$ is defined, *e.g.*, through the joint distribution of (X, Y) on $I\!R_+^p \times I\!R_+^q$ where, in deterministic frontier models, $\text{Prob}\,((X, Y) \in \Psi) = 1$.

The production set can be described by its sections: the input requirement sets is defined for all $y \in \Psi$ as $C(y) = \{x \in I\!R_+^p \mid (x, y) \in \Psi\}$, and the output requirement set is defined for all $x \in \Psi$ as $P(x) = \{y \in I\!R_+^q \mid (x, y) \in \Psi\}$.

The radial (input-oriented) efficiency boundary ("efficient frontier") is defined by

$$\partial C(y) = \{x \mid x \in C(y), \theta x \notin C(y)\ \forall\ 0 < \theta < 1\}. \tag{1.2}$$

The Farrell input measure of efficiency of a production unit working at level (x_0, y_0) is defined as

$$\theta(x_0, y_0) = \inf\{\theta \mid \theta x_0 \in C(y_0)\} = \inf\{\theta \mid (\theta x_0, y_0) \in \Psi\}. \tag{1.3}$$

Note that $\partial C(y) = \{x \mid \theta(x, y) = 1\}$. The same could be done in the output space where the radial efficient boundary is

$$\partial P(x) = \{y \mid y \in P(x), \lambda y \notin P(x)\ \forall\ \lambda > 1\}. \tag{1.4}$$

Then the Farrell output measure of efficiency for a production unit working at level (x_0, y_0) is given by

$$\lambda(x_0, y_0) = \sup\{\lambda \mid \lambda y_0 \in P(x_0)\} = \sup\{\lambda \mid (x_0, \lambda y_0) \in \Psi\}. \tag{1.5}$$

Here, $\partial P(x) = \{y \mid \lambda(x, y) = 1\}$. Note that the frontier of Ψ is unique, and $\partial C(y)$ and $\partial P(x)$ are two different ways of describing it. Different assumptions can be assumed on Ψ (*e.g.*, free disposability, convexity,etc.; see Shephard, 1970 for details).

The econometric problem arises since usually Ψ (and hence $\partial C(y)$ and $\partial P(x)$) are unknown, and therefore the efficiency measures $\theta(x_0, y_0)$ and $\lambda(x_0, y_0)$ have to be estimated. Several estimators of Ψ from a random sample of production units $\{(X_i, Y_i) \mid i = 1, \ldots, n\}$ are available: in this deterministic frontier framework, the most popular nonparametric estimators are the Free Disposal Hull (FDH) and the Data Envelopment Analysis (DEA). Both derive from the pioneering work of Farrell (1957). In summary,

$$\widehat{\Psi}_{FDH} = \Big\{(x, y) \in I\!\!R_+^{p+q} \mid y \le Y_i,\ x \ge X_i, \quad i = 1, \ldots, n\Big\}$$

is the free disposable hull of the sample proposed by Deprins, Simar and Tulkens (1984) and

$$\begin{aligned} \widehat{\Psi}_{DEA} \quad = \quad & \{(x, y) \in I\!\!R_+^{p+q} \mid y \le \sum_{i=1}^{n} \gamma_i Y_i \,;\, x \ge \sum_{i=1}^{n} \gamma_i X_i \,; \\ & \sum_{i=1}^{n} \gamma_i = 1 \,;\, \gamma_i \ge 0, i = 1, \ldots, n\} \end{aligned}$$

is the convex hull of $\widehat{\Psi}_{FDH}$, initiated by Farrell (1957) and popularized by Charnes, Cooper and Rhodes (1978). The estimators of the efficiency measures for a production unit working at level (x_0, y_0), $\hat{\theta}(x_0, y_0)$ and $\hat{\lambda}(x_0, y_0)$, are then obtained by plugging $\widehat{\Psi}_{FDH}$ or $\widehat{\Psi}_{DEA}$ in place of Ψ in the appropriate expressions. Today, statistical inference based on DEA/FDH-type estimators is available either by using asymptotic results (Kneip, Park and Simar, 1998, Gijbels, Mammen, Park and Simar (1999) and Park, Simar and Weiner, 2000) or by using the bootstrap (Simar and Wilson, 1998, 2000b). See Simar and Wilson (2000a) for a recent survey of available results.

Nonparametric deterministic frontier models are very appealing because they rely on few assumptions; but, by construction, they are quite sensitive to extreme values and to outliers. Detecting outliers is thus of primary importance: it is not an easy task in this multivariate setup. Most of the standard geometrical methods for detecting outliers are very computer intensive in multivariate set-ups and do not take the frontier aspects of the problem into account: we are mostly interested to detect super-efficient outliers which will be very influential to the efficiency measures $\hat{\theta}(x_0, y_0)$ and $\hat{\lambda}(x_0, y_0)$. Wilson (1993, 1995) proposed methods making use of influence functions to detect outliers in this framework but the methods become computationally prohibitive as the number of observations increases, particularly if one exhaustively examines the "masking effect" mentioned below in Section 4.

Recently Cazals, Florens and Simar (2002) (CFS hereafter) proposed nonparametric frontier estimators that are robust with respect to these extreme values. The CFS idea are based on the concept of "expected frontier of order-m", where m can be viewed as a "trimming" parameter of the frontier but, as shown below, m has also its own empirical economic interpretation. The properties of the order-m frontier and its relation to the frontier of Ψ are investigated in CFS. In particular, as explained below, for large values of m, the two frontiers coincide. A nonparametric estimator of the order-m frontier is very easy to derive and very fast to compute. Also it has remarkable statistical properties: no curse of dimensionality (standard $\sqrt{n}$-consistency) and asymptotic normality. Due to trimming nature of the order-m frontier, the estimator does not envelop all the observed data points, even for large m, and so it is more robust to outliers and/or to extreme values. The procedure proposed below to detect outliers is based on a sensitivity analysis relative to several values of m, so that m will also be viewed as a tuning parameter for outliers detection.

In this paper, we briefly summarize the definition and the properties of the expected frontier of order-m and explain how to derive a nonparametric estimator of it. Then we show how the procedure can be used to detect potential outliers in the data set. The approach is multivariate (multi-inputs and/or multi-outputs) and can be applied either to FDH or to DEA approaches or even to any parametric frontier model. The procedure is illustrated through some simulated and real examples. No "optimal" procedure nor "miracle" procedure can be defined to detect outliers in this difficult context, but our method is very easy and very fast to implement. Consequently, it certainly offers an appealing and useful methodology in the exploratory data analysis phase of any efficiency analysis with real data.

An "outlier" is an atypical observation or a data point outlying the cloud of data points. In the statistical literature, an outlier does not have a generally accepted, precise definition (Davies and Gather, 1993). Often it is referred to as an observation which appears to be inconsistent with the remainder of the set of data (Barnett and Lewis, 1995). There are many reasons why an observation might be atypical. An observation could be an outlier because it contains an error (bad coding, etc.), or because it arose from a different data generating process (DGP) than the others, or it might simply be a datum with low probability of being drawn from the same DGP (see also the discussion in Wilson, 1993). In deterministic frontier models (parametric or nonparametric), outlying points might be highly influential if they distort the enveloping estimator of the frontier of Ψ. So it is important to develop exploratory data analysis tools which allow detection of extreme values. Once a potential outlier is detected, a careful consideration must be given to determine why it is an outlier (see the discussion in Olesen *et al.*, 1996, page 343). The idea is to use outlier detectors as diagnostic tools, flagging a (hopefully) small set of observations for closer scrutiny.

In the framework here (deterministic frontiers), the concept of outlier is different from the concept of noise in the data, although some isolated data points, perturbed by noise, could be viewed as outliers. If the DGP contains a noise process, then procedures based on stochastic frontier models (parametric and/or nonparametric) would be more appropriate (for nonparametric approaches, see Hall and Simar, 2000 and Simar, 2002); in this case DEA/FDH are inconsistent.

In Section 2 we summarize the basic concepts and the CFS results. In Section 3, we explain how to implement estimators in a multivariate context. Section 4 proposes a methodology for detecting the outliers, while Section 5 illustrates the procedure with simulated as well as real data. Section 6 concludes. The Appendix provides a Matlab code that computes the estimator of the expected order-m frontier in the input and output directions.

2 Basic Concepts: Expected Frontier of Order-m

We define the basic concepts in a simple bivariate case (one-input, one output), in the input-orientated case. We also briefly indicate how to adapt the formulation to the output-oriented case. All the proofs are provided in CFS.

2.1 An other way for defining the efficient frontier of Ψ

The DGP is characterized by the distribution of the random point (X, Y) on Ψ. We focus here on how to define the frontier of Ψ. Consider, for a moment, the simplest case consisting of a one-dimensional frontier: suppose that every firm produces one unit of output, and that we are looking for the univariate, input-efficient frontier. Here we have only one random variable X, and we are interested in ϕ, the lower boundary of X. This unknown parameter can be defined as

$$\phi = \inf\{x | F_X(x) > 0\},$$

where $F_X(\cdot)$ is the distribution function of X. Equivalently, we could use the survivor function $S_X(x) = \text{Prob}(X \geq x) = 1 - F_X(x)$ to define ϕ:

$$\phi = \inf\{x | S_X(x) < 1\}. \tag{2.1}$$

In the bivariate case, where X is the input and Y is the output, we can also define the boundary of the support of (X, Y) in the input direction. It will be, in the input space, the lower level of input X attainable for a firm producing at least a given level of the output. This can also be characterized by a survivor function, but here, we will use the appropriate conditional survivor function of X, given $Y \geq y$:

$$S_c(x|y) = \text{Prob}(X \geq x|Y \geq y) = \frac{S(x,y)}{S_Y(y)}, \tag{2.2}$$

where $S(x,y) = \text{Prob}(X \geq x, Y \geq y)$, and $S_Y(y) = \text{Prob}(Y \geq y)$ is the marginal survivor function of Y ($S_Y(y) = S(0,y)$). The lower boundary of this conditional survivor function is thus defined, for any value of y by

$$\phi(y) = \inf\{x|S_c(x|y) < 1\}. \tag{2.3}$$

Figure 1 illustrates the concept of the input-frontier function.

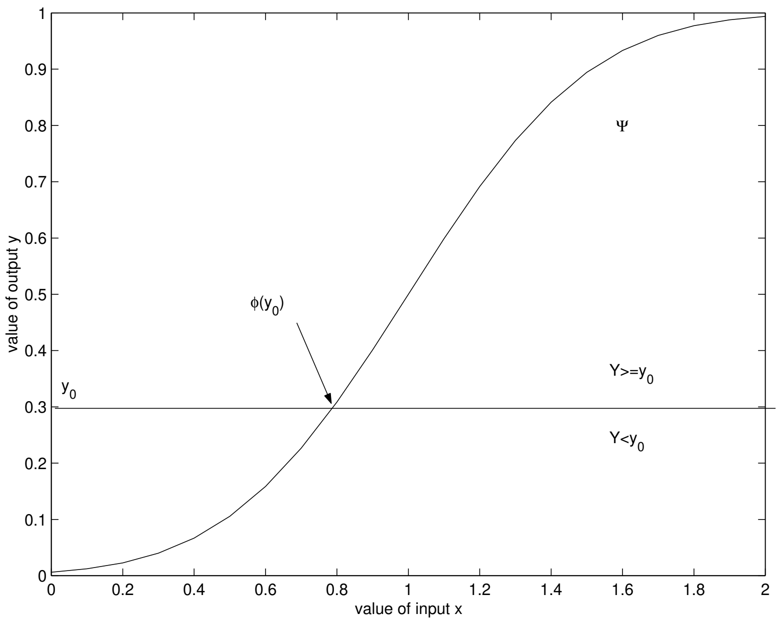

Figure 1: *Input-oriented frontier in bivariate case: for any value of* y_0, $\phi(y_0) = \inf\{x|S_c(x|y_0) < 1\}$.

For the output-oriented frontier, the same approach can be followed to describe the boundary of Ψ. In the bivariate framework, it can be defined for any level of the input x as the upper boundary of the conditional distribution function of Y, given $X \leq x$:

$$F_c(y \mid x) \;=\; \text{Prob}(Y \leq y \mid X \leq x) \tag{2.4}$$

$$= \frac{F(x,y)}{F_X(x)}, \tag{2.5}$$

where $F_X(x) = \text{Prob}(X \leq x)$. Then, for any value of x, the frontier of Ψ in the output direction is given by

$$\psi(x) = \sup\{y \mid F_c(y \mid x) < 1\}, \tag{2.6}$$

the maximum level of output attainable for any firm using less than the level x of input.

CFS prove that the frontier functions $\phi(y)$ and $\psi(x)$ are monotone, non-decreasing in their arguments. Moreover, $\phi(y)$ ($\psi(x)$) is the largest (smallest) monotone function which is smaller (larger) or equal to the efficient frontier $\partial C(y)$ ($\partial P(x)$). If the production set Ψ is free disposal (a quite reasonable assumption in practice), $\phi(y) = \partial C(y)$ and $\psi(x) = \partial P(x)$ which amounts to a reparametrization of the definition of the efficient frontier of Ψ.

In this new formulation, a natural nonparametric estimator of the frontiers in both directions is given by plugging the empirical analog of $S_c(x|y)$ or of $F_c(y|x)$ into the appropriate formula. Let

$$\widehat{S}_{c,n}(x \mid y) = \frac{\widehat{S}_n(x,y)}{\widehat{S}_{Y,n}(y)}, \tag{2.7}$$

where $\widehat{S}_n(x,y) = (1/n)\sum_{i=1}^n \mathbb{1}(x_i \geq x, y_i \geq y)$, and let

$$\widehat{F}_{c,n}(y \mid x) = \frac{\widehat{F}_n(x,y)}{\widehat{F}_{X,n}(x)}, \tag{2.8}$$

where $\widehat{F}_n(x,y) = (1/n)\sum_{i=1}^n \mathbb{1}(x_i \leq x, y_i \leq y)$. Then we have

$$\hat{\phi}_n(y) = \inf\{x \mid \widehat{S}_{c,n}(x \mid y) < 1\} \tag{2.9}$$

and

$$\hat{\psi}_n(x) = \sup\{y \mid \widehat{F}_{c,n}(y \mid x) < 1\}. \tag{2.10}$$

Note that the obtained estimators are the input- and output-oriented frontiers obtained by the FDH estimator. They share the same properties as the corresponding estimated functions, *i.e.*, monotonicity in their arguments.

2.2 The order-m frontiers of Ψ

Input-oriented case

Consider again the simplest univariate case, where firms produce one unit of output. The lower boundary of X, ϕ, is defined in (2.1). Suppose, that instead of focusing on the true lower boundary of X, we consider an alternative measure defined as follows. Consider a fixed integer $m \geq 1$. We define as the order-m lower boundary for X, the expected value of the minimum of m random variables $X^1, \ldots, X^m$ drawn from the distribution function of X, as

$$\phi_m = \mathrm{E}\left[\min(X^1, \ldots, X^m)\right] = \int_0^\infty [S_X(x)]^m dx. \tag{2.11}$$

So ϕ_m is the expected minimum achievable input-level among m firms drawn from the population of firms (all the firms producing here, one unit of output). The value of m is arbitrary and can be fixed to any desired level, but it is interesting to analyze the value of ϕ_m as a function of m. In particular, it is easy to prove that $\lim_{m\to\infty} \phi_m = \phi$ and that, for all finite m, $\phi_m \geq \phi$.

For any fixed value of $m \geq 1$, ϕ_m is an unknown parameter, but can be estimated easily from a sample of observed values $(x_1, \ldots, x_n)$. At this elementary stage, it is important to notice the difference between m and n: m is a "trimming" parameter fixed at any desired level defining the level of the benchmark, whereas, n is the sample size and so, there are no *a priori* relations between m and n. The idea of trimming is not new in statistics (most readers know the concept of a "trimmed mean" where the mean is computed after a part of the observations in both tails of the observations are deleted), its use here in boundary estimation is new.

Nonparametric estimators of ϕ and of ϕ_m based on a random sample of size n can be obtained by plugging the empirical survivor function of X into (2.1) and (2.11), respectively, yielding $\hat{\phi}_n$ and $\hat{\phi}_{m,n}$. The relations between ϕ and ϕ_m are reflected in their empirical counterparts: for all finite m, $\hat{\phi}_{m,n} \geq \hat{\phi}_n = x_{(1)}$, where $x_{(1)}$ is the first order statistic and $\lim_{m\to\infty} \hat{\phi}_{m,n} = \hat{\phi}_n$. Clearly, $\hat{\phi}_n \leq x_i, i = 1, \ldots, n$, but this is not true for the order-m frontier estimator, $\hat{\phi}_{m,n}$, even for large m. This is primarily due to the expectation in (2.11) and to the finiteness of m. If some observed points x_i remain below $\hat{\phi}_{m,n}$, even when m increases, this could indicate "potential" outliers. The main idea of this paper is to use order-m frontiers to detect outliers.

The bivariate extension is immediate. We first deal with the input-oriented case. Consider the input levels X of firms producing at least a given level y of output. The process generating the input levels of such firms can be characterized by the conditional distribution of X given $Y \geq y$. The minimum achievable level of input for these firms is given by the input-oriented frontier $\phi(y)$ defined in (2.3). Now consider a fixed integer $m \geq 1$. We define the (expected) order-m lower boundary of inputs X, for firms producing more than y, as the expected value of the minimum of m random variables $X^1, \ldots, X^m$ drawn from the conditional distribution function of X given $Y \geq y$. Formally, this (expected) frontier function of order-m is defined by

$$\phi_m(y) = \mathrm{E}\left[\min(X^1, \ldots, X^m) \mid Y \geq y\right] = \int_0^\infty [S_c(x \mid y)]^m \, dx. \tag{2.12}$$

For all value of y and for all finite m, $\phi_m(y) \geq \phi(y)$, and for all y, $\lim_{m\to\infty} \phi_m(y) = \phi(y)$.

From an economic perspective, $\phi_m(y)$ has its own interest. It does not provide the input-efficient frontier, but rather another reasonable benchmark value of the input for a firm producing a level y of output: it is the expected minimal value of input achievable among a fixed number of m firms drawn from the population of firms producing at least this level y of output. Being far above this order-m input-frontier is a clear indication of being input-inefficient, whereas, being near or below this level could indicate efficiency, or in some case, super-efficiency. The value of m is chosen at the desired level for fixing this benchmark level, but a sensitivity analysis with a few values of m could be helpful in evaluating the performance of a firm.

Of course, $\phi_m(y)$ is unknown, but it can be estimated non-parametrically by plugging the empirical survivor function into (2.12) to obtain

$$\hat{\phi}_{m,n}(y) = \widehat{\mathrm{E}}\left[\min(X^1, \ldots, X^m) \mid Y \geq y\right], \tag{2.13}$$

where $X^1, \ldots, X^m$ are m i.i.d. random variables generated by the empirical distribution of X given $Y \geq y$, whose survivor function is $\widehat{S}_{c,n}(x \mid y)$. So,

$$\hat{\phi}_{m,n}(y) = \int_0^\infty [\widehat{S}_{c,n}(x \mid y)]^m \, dx. \tag{2.14}$$

As before, the relations between the order-m frontier and the true frontier carry over to their estimators $\hat{\phi}_{m,n}(y)$ and $\hat{\phi}_n(y)$. The asymptotic behavior of $\hat{\phi}_{m,n}(y)$, when the sample size n increases, is investigated by CFS: in summary, $\hat{\phi}_{m,n}(y)$ achieves $\sqrt{n}$-consistency, is asymptotically unbiased, and normally distributed: $\mathcal{L}\left(\sqrt{n}(\hat{\phi}_{m,n}(y) - \phi_m(y))\right) \to N(0, \sigma^2(y))$. An expression for $\sigma^2(y)$ is given in CFS.

Remark 2.1 *As noted in CFS, for a fixed sample size, the value of $\sigma^2(y)$ increases with y since there are fewer observed points (x_i, y_i) where $y_i \leq y$, and consequently fewer points* *to estimate the conditional survivor function $\widehat{S}_{c,n}(x \mid y)$. This border effect often arises in nonparametric methods. Note that we have the same problem for the FDH estimator (see (2.7) and (2.9)), and to some extent for the DEA estimator (which is the convex closure of the FDH). This point is often ignored. In Park et al. (2000) the standard deviation of the FDH estimator at a point is shown to be proportional to the inverse of the probability of observing a point near the frontier at this point: again this probability is small and difficult to estimate for border points. Gijbels et al. (1999) obtain the same result for the DEA estimator (a formula is available only when $p = q = 1$). The practical consequences of this problem in the present setup are discussed in Section 4.*

Output-oriented case

For the output oriented case, a similar approach can be followed. Given a fixed integer $m \geq 1$, define for a given level of input x the (expected) order-m output-oriented frontier as the expected value of the maximum of m random variables $Y^1, \ldots, Y^m$ drawn from the conditional distribution function of Y given $X \leq x$:

$$\psi_m(x) = \mathrm{E}\left[\max(Y^1, \ldots, Y^m) \mid X \leq x\right] = \int_0^\infty \left(1 - [F_c(y \mid x)]^m\right) dy. \tag{2.15}$$

A nonparametric estimator of $\psi_m(x)$ is given by:

$$\hat{\psi}_{m,n}(x) = \widehat{\mathrm{E}}\left[\max(Y^1, \ldots, Y^m) \mid X \leq x\right], \tag{2.16}$$

which may be computed as

$$\hat{\psi}_{m,n}(x) = \int_0^\infty \left(1 - [\widehat{F}_{c,n}(y \mid x)]^m\right) dy. \tag{2.17}$$

Mutatis mutandis, this estimator achieves the same properties as in the input-oriented case. Here, the standard deviation of $\hat{\psi}_{m,n}(x)$ increases when x is small (see Remark 2.1 above).

We will describe in the next section how these estimators can be computed in a multivariate setup. But from now on, it should be clear that in a sample of points, if the input (output) level of a data point is far below (above) its corresponding order-m input (output)-frontier, even when m increases, this indicates a potential outlier. This is true regardless of which estimator is to be used in hte final analysis (FDH, DEA, or any parametric method).

3 Multivariate Extensions

For handling the multivariate case and to provide results which are easier to read, the practical computations are made in terms of radial distances (Farrell efficiency measures) from a particular firm (x_0, y_0) to the frontier or to the order-m frontier.

3.1 Order-m efficiency scores

Let (x_0, y_0) be the point of interest in $\Psi \subset I\!R^p_+ \times I\!R^q_+$, and let $m \geq 1$ be a fixed integer. Now, we consider m (p-dimensional) random variables $X^1, \ldots, X^m$ drawn from the conditional distribution function of X given $Y \geq y_0$. Define the following random variable:

$$\tilde{\theta}_m(x_0, y_0) = \min_{i=1,\ldots,m} \left\{ \max_{j=1,\ldots,p} \left(\frac{X^{i,j}}{x_0^j} \right) \right\} \tag{3.1}$$

where $X^{i,j}$ (x_0^j) denotes the jth component of X^i (of x_0 respectively). This random variable measures the radial distance, in the input space, between the point x_0 and the free disposal hull of the random points $X^1, \ldots, X^m$ generated from the conditional distribution function of X given $Y \geq y_0$. The (expected) order-m input measure of efficiency of a point (x_0, y_0) is defined as

$$\theta_m(x_0, y_0) = \mathrm{E}\left[\tilde{\theta}_m(x_0, y_0) \mid Y \geq y_0 \right]. \tag{3.2}$$

It may be proven (see CFS for details) that $\lim_{m\to\infty} \theta_m(x_0, y_0) = \theta(x_0, y_0)$, the Farrell input measure of efficiency defined in (1.3). Also when $p = 1$ it is easy to show that $x_0\, \theta_m(x_0, y_0) = \phi_m(y_0)$, the order-$m$ input frontier defined above.

For the output orientation, we consider m (q-dimensional) random variables $Y^1, \ldots, Y^m$ generated from the conditional distribution of Y given $X \leq x_0$. Then define

$$\tilde{\lambda}_m(x_0, y_0) = \max_{i=1,\ldots,m} \left\{ \min_{j=1,\ldots,p} \left(\frac{Y^{i,j}}{y_0^j} \right) \right\}. \tag{3.3}$$

$\tilde{\lambda}_m(x_0, y_0)$ measures the radial distance, in the output space, between the point y_0 and the free disposal hull of the random points $Y^1, \ldots, Y^m$ generated from the conditional distribution function of Y given $X \leq x_0$. The (expected) order-m output measure of efficiency of a point (x_0, y_0) is now defined as

$$\lambda_m(x_0, y_0) = \mathrm{E}\left[\tilde{\lambda}_m(x_0, y_0) \mid X \leq x_0 \right]. \tag{3.4}$$

Here again, $\lim_{m\to\infty} \lambda_m(x_0, y_0) = \lambda(x_0, y_0)$, which is the Farrell output measure of efficiency defined in (1.5). When $q = 1$, we have $y_0\, \lambda_m(x_0, y_0) = \psi_m(y_0)$, the order-$m$ output frontier defined in Section 2 for the bivariate case.

3.2 Nonparametric estimation

As above, plug-in nonparametric estimators of the order-m frontiers are obtained by using empirical distribution functions in place of the unknown population distributions. We have

$$\hat{\theta}_{m,n}(x_0, y_0) \quad = \quad \widehat{\mathrm{E}}(\tilde{\theta}_m(x_0, y_0) \mid Y \geq y_0), \tag{3.5}$$

where the expectation $\widehat{\mathrm{E}}$ is taken with respect to the empirical conditional distribution of X, given $Y \geq y_0$. In multivariate setups, this involves a numerical integration which is easier to solve by Monte-Carlo approximation. The computational algorithm is very simple to implement and works as follows. For a given y_0, draw a random sample of size m with replacement among those x_i where $y_i \geq y_0$, and denote this sample by $(X_b^1, \ldots, X_b^m)$. Then compute

$$\tilde{\theta}_m^b(x_0, y_0) = \min_{i=1,\ldots,m} \left\{ \max_{j=1,\ldots,p} \left(\frac{X_b^{i,j}}{x_0^j} \right) \right\}.$$

Repeat this for $b = 1, \ldots, B$, where B is the number of Monte-Carlo replications (B is large). Then,

$$\hat{\theta}_{m,n}(x_0, y_0) = \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}_m^b(x_0, y_0). \tag{3.6}$$

By the law of large numbers, $\hat{\theta}_{m,n}(x_0, y_0)$ converges to $\widehat{\mathrm{E}}(\tilde{\theta}_m(x_0, y_) \mid Y \geq y_0)$, as $B \to \infty$. In order to appreciate the quality of the Monte-Carlo approximation (which can be tuned by an appropriate choice of B), it may be worthwhile to also compute the Monte-Carlo standard deviation of the approximation, *i.e.*,

$$\mathrm{STD_{MC}}(\hat{\theta}_{m,n}(x_0, y_0)) = \frac{1}{\sqrt{B}} \sqrt{\frac{\sum_{b=1}^{B} (\tilde{\theta}_m^b(x_0, y_0) - \hat{\theta}_{m,n}(x_0, y_0))^2}{B-1}}. \tag{3.7}$$

It should be clear, to avoid misunderstandings, that $\mathrm{STD_{MC}}$ **is not** the sampling standard deviation of our estimator $\hat{\theta}_{m,n}(x_0, y_0)$; it only gives the standard deviation of the approximation (3.6) when trying to evaluate (3.5). This MC-standard deviation can be made arbitrarily small by increasing B.

It can also be proven that for all value of the sample size n, the order-m frontier estimator converges to the FDH estimator of the frontier when $m \to \infty$:

$$\lim_{m\to\infty} \hat{\theta}_{m,n}(x_0, y_0) = \hat{\theta}_n(x_0, y_0) = \hat{\theta}_{FDH,n}(x_0, y_0) = \min_{i|y_i \geq y_0} \left\{ \max_{j=1,\ldots,p} \left(\frac{x_i^j}{x_0^j} \right) \right\}. \tag{3.8}$$

The same can be done for the order-m output efficiency measures. We have

$$\hat{\lambda}_{m,n}(x_0, y_0) \quad = \quad \widehat{\mathrm{E}}(\tilde{\lambda}_m(x_0, y_{)} \mid X \leq x_0), \tag{3.9}$$

where the expectation $\widehat{\mathrm{E}}$ is taken with respect to the empirical conditional distribution of Y, given $X \leq x_0$. The Monte-Carlo approximation works as follows. For a given x_0, draw a random sample of size m with replacement among those y_i such that $x_i \leq x_0$, and denote this sample by $(Y_b^1, \ldots, Y_b^m)$. Then compute

$$\tilde{\lambda}_m^b(x_0, y_0) = \max_{i=1,\ldots,m} \left\{ \min_{j=1,\ldots,p} \left(\frac{Y_b^{i,j}}{y_0^j} \right) \right\}.$$

Repeat this for $b = 1, \ldots, B$. Then

$$\hat{\lambda}_{m,n}(x_0, y_0) = \frac{1}{B} \sum_{b=1}^{B} \tilde{\lambda}_m^b(x_0, y_0). \tag{3.10}$$

The relation with the FDH estimator is, for any value of n, given by

$$\lim_{m\to\infty} \hat{\lambda}_{m,n}(x_0, y_0) = \hat{\lambda}_n(x_0, y_0) = \hat{\lambda}_{FDH,n}(x_0, y_0) = \max_{i|x_i \leq x_0} \left\{ \min_{j=1,\ldots,p} \left(\frac{y_i^j}{y_0^j} \right) \right\}. \tag{3.11}$$

As shown in the appendix, all these efficiency measures are very easy (and fast) to compute, even for large sample sizes.

4 Detecting Outliers: Practical Computations

4.1 A Methodology for Outliers Detection

Now we have a tool for detecting potential outliers among the data set $\mathcal{X} = \{(x_i, y_i) \mid i = 1, \ldots, n\}$. Any point $(x_0, y_0) \in \Psi \subset I\!R_+^p \times I\!R_+^q$ is a likely outlier when, even if m increases, its order-m input efficiency measure $\hat{\theta}_{m,n}(x_0, y_0)$ is greater than one. The same may be said for the output direction when the order-m output efficiency measure $\hat{\lambda}_{m,n}(x_0, y_0)$ is smaller than one.

For example, consider a point (x_0, y_0) such that $\hat{\theta}_{100,n}(x_0, y_0) = 1.5$. This production unit uses 50% less inputs (proportionate reduction) than the expectation of the minimum input level of 100 other firms drawn from the population and producing more than y_0 output. This point is a potential super-input-efficient point. The same could be said for the output direction; if, for example, $\hat{\lambda}_{100,n}(x_0, y_0) = 0.33$, the firm represented by (x_0, y_0) produces 3 times more output (in radial extension) than the expected value of the maximal level of output of 100 other firms drawn from the population of firms using less than x_0 inputs. It is a potential super-output-efficient outlier.

We present the basic ideas in the input-oriented case, but the same will also be done in the output-oriented case. In practice, both results will be useful for detecting outliers. Let us compute for each data point (x_i, y_i) its order-m input-efficiency score leaving out the observation (x_i, y_i) from the reference set. We denote this "leave-one-out" efficiency score by $\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ and the corresponding reference set by $\mathcal{X}^{(i)}$. We compute these scores for several reasonable values of m: we have to detect values of $\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ substantially larger than 1. We know that, for finite m, $\hat{\theta}^{(i)}_{m,n}(x_i, y_i) \geq \hat{\theta}^{(i)}_{FDH,n}(x_i, y_i)$ (at the Monte-Carlo precision for computing $\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$). In practice, for every "FDH-efficient" point, $\hat{\theta}^{(i)}_{m,n}(x_i, y_i) \geq 1$, but this does not automatically indicate a potential outlier: we must choose a threshold value, *e.g.* $(1 + \alpha)$; $\hat{\theta}^{(i)}_{m,n}(x_i, y_i) \geq 1 + \alpha$ then indicates observation i is an outlier.

So, both m and the threshold level must be chosen. There are no definite rules, but the two issues can be addressed through sensitivity analysis, taking into account that m has an empirical, economic interpretation and that we want to flag points clearly outside the cloud of other points (for α sufficiently large).

For each observation i, we compute $\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$, for, say, m=10, 25, 50, 75, 100, 150 (any other set of values for m could of course be chosen). Here we must realize that the sample is finite. When computing $\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ by (3.5), we estimate the conditional distribution function $S_c(x|Y \geq y_i)$ by its empirical counterpart. The latter will be achieved by looking at all points in the sample $\mathcal{X}^{(i)}$ with output value larger or equal to y_i; this number could be small (in particular, for points at the edge of the sample values it could even be equal to zero). Denote this number of points by $N_{input}(x_i, y_i)$. This value is useful because it indicates the number of points used to estimate a p-variate distribution function, and it indicates how close the point (x_i, y_i) is to the edge of the support of the data points.

Then we will report in a first table of results, for each data point (x_i, y_i), the values of

$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$, for the increasing values of m, along with their Monte-Carlo standard deviations (3.7), to assess the precision of the Monte-Carlo approximation (if too much imprecision, we increase B and redo the computations). We report also in the table the values of $N_{input}(x_i, y_i)$.

Mutatis mutandis, the same can be done in the output direction, providing a second table of results for the "leave-one-out" output-efficiency scores $\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ with their MC standard deviations and the values of $N_{output}(x_i, y_i)$ the number of sample points in $\mathcal{X}^{(i)}$ with input value smaller or equal to x_i.

Looking carefully through these two tables will help detect potential outliers: for extreme points, where $\hat{\theta}^{(i)}_{m,n}(x_i, y_i) \geq 1$, both the decrease of the order-m efficiencies as a function of m and the values of N_{input} are of interest; in the output table, we analyze the increase of the values of $\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ which are ≤ 1 as a function of m and the corresponding values of N_{output}. If even for large values of m, both $\hat{\theta}^{(i)}_{m,n}(x_i, y_i) \geq 1$ and $\hat{\lambda}^{(i)}_{m,n}(x_i, y_i) \leq 1$, indicate that (x_i, y_i) is a potential outlier. Then of course, this data point deserves closer scrutiny.

4.2 A semi-automatic warning procedure

For large values of n, a careful reading of the tables of results is not easy so we need some help in flagging potential outliers. It is never easy to derive a fully automatic procedure to detect outliers; this is particularly true in the present context because any tuning parameter or threshold value will depend on the underlying DGP in a very complicated way. For instance, a choice of a value for m and for the threshold value α activating our "flag" will not be the same in the input and the output orientations. The underlying conditional distributions used in both cases could indeed be completely different: for instance, homoscedastic output inefficiencies do not imply homoscedastic input inefficiencies (even in the constant return to scale case). The idea is to propose a simple procedure which seems to be reasonable, and we will show in the numerical illustrations that it performs well.

As noted earlier, we must first decide on some threshold values for the efficiency scores. This is achieved again through a sensitivity analysis. We propose to choose several reasonable threshold values distant from 1, such as $1 \pm \alpha$ where $\alpha \in \{0.20, 0.30, 0.40, 0.50\}$. Recall that we want to detect points which are really outlying the cloud of points, so super-efficiency of less than say, 10%, in this leave-one-out approach would not be a useful indicator of potential outlier. Of course, any other set of values for α could be chosen. Then we will plot the percentage of points in the sample $\mathcal{X}$ with $\hat{\theta}^{(i)}_{m,n}(x_i, y_i) \geq 1 + \alpha$, as a function of m, for the different chosen values of α. For the output oriented case, we plot the percentage of points with $\hat{\lambda}^{(i)}_{m,n}(x_i, y_i) \leq 1 - \alpha$. In practice[1], we take the Monte-Carlo standard deviations into account (because B is finite). For instance, in the input case, we consider points such that $\hat{\theta}^{(i)}_{m,n}(x_i, y_i) - 1.645 * \mathrm{STD}_{\mathrm{MC}}(\hat{\theta}^{(i)}_{m,n}(x_i, y_i)) \geq 1 + \alpha$, where 1.645 is the 0.95 percentile of the standard normal distribution[2]. These curves indicate the percentage of points outside the order-m frontier, as a function of m and of α.

By construction, all these curves should decrease when m increases, and if there are no outliers, they should converge (approximately, due to sampling imprecision) linearly to the percentage of points having a leave-on-out FDH efficiency score greater than 1. So any strong departure from linearity will indicate the potential existence of outliers: for instance, if the curves show an "elbow effect" (sharp negative slope, followed by a smooth decreasing slope), it indicates that the points remaining outside the order-m frontier, for this value of m and for the chosen threshold $(1 + \alpha)$, need closer analysis. Indeed, we need much larger values of m (eventually to ∞), to get these points inside (or closer to) the order-m frontier. Since the situation could be different in the input and in the output direction, we need to look at both pictures. In addition, all data points with order-m efficiency measures ≥ 1 for input (≤ 1 for output) and having small values of N_{input} (for N_{output}) should be flagged as being extreme. For these points, the value of the order-m efficiency in the other direction will confirm if they are potential outliers.

For choosing the value of (m, α), we must also realize that the percentage of points left outside the frontier has to be reasonable. Again no theoretical rules exist on how to fix an upper bound for the number of outliers. In the statistical literature an upper bound is sometimes given for the accepted proportion of outliers; for instance, Barnett and Lewis (1995) suggests $\sqrt{n}/n$ as a reasonable upper bound[3].

[1]Since we know the asymptotic sampling distribution of $\hat{\theta}_{m,n}$ and its standard deviation (see CFS) we could adjust for sampling variability, but except for $p = 1$ in the input-oriented case and for $q = 1$ in the output oriented case, the computation of the sampling standard deviations is rather intricate. In this exploratory data analysis, we only use the point estimates, corrected for their Monte-Carlo imprecisions.

[2]Although distributions of efficiencies are generally skewed, the central limit theorem can be used here, in the Monte-Carlo experiment: we approximate indeed a mean of a random variable by its empirical analog over B random replications. The correction $1.645 * \mathrm{STD}_{\mathrm{MC}}$ is just to adjust for the fact that we use finite values of B.

[3] We are assuming here that we only have one DGP and not a mixture of different DGPs. In the latter case the problem of outliers detection becomes a problem of cluster analysis or of discrimination between two or more DGPs. As pointed in the discussion by Ole Olesen, note that if our procedure flags too much potential outliers with values of N_{input} **and** N_{output} not too small, this might indicate the presence of clusters of points and so of more than one DGP. These points could be at the border of a cluster and detected as

With the help of these tuning values (m, α), it is easy to identify these extreme points by an appropriate computer program and then go back to the tables of results and to the data. Several values of (m, α) could also be tried.

Once outliers have been detected and confirmed, the analysis can be redone without these points, in order to avoid the "masking" effect, discussed by Wilson (1993, 1995). An outlier could mask another outlier (or several other outliers) situated near the first one. Our procedure should be more robust with respect to this masking effect (due to the expectation operator) but of course, the method is not foolproof[4]. The required computations are very fast and so the entire process can be repeated several times, by sequentially deleting outliers from one run to the next (see the appendix for more information on computing times).

As observed in the introduction, no "optimal" procedure nor "miracle" exist to detect outliers in this difficult context, but the method proposed here is very easy and very fast to implement. It cannot be a complete "automatic" procedure (we must inspect pictures, go back to the table of results, etc.). However, it certainly offers an appealing and useful methodology in the exploratory data analysis phase of any efficiency analysis with real data, and the semi-automatic warning procedure of potential outliers is particularly useful when n is large. The next section illustrates the ideas with numerical examples.

5 Numerical Illustrations

All computations were performed using the MATLAB code listed in the Appendix.

being extreme but they are not extreme in the full cloud of points. This idea will not be pursued here.

[4]In other words, in the presence of two nearby outliers (one masking the other), our procedure is more robust to detect both in the first run. The point is that here, we use an **expectation** rather than an extreme observed point to estimate the frontier. Let us illustrate the idea in the simplest univariate input oriented case: the classical envelopment estimator is $\min(X_1, \ldots, X_n)$, and the order-m estimator is $\widehat{E}[\min(X^1, \ldots, X^m)]$ where $(X^1, \ldots, X^m)$ is a i.i.d. random sample drawn from $(X_1, \ldots, X_n)$. Suppose we have, say, two outlying nearby too small values $X_{(1)}, X_{(2)}$. Only if $m \to \infty$, both estimators are the same. At the first run the min operator will get $X_{(1)}$, at the second run ($X_{(1)}$ is dropped out), it will get $X_{(2)}$ "masked" by $X_{(1)}$ at the first run, even if n is large. Whereas, for finite m, $\widehat{E}[\min(X^1, \ldots, X^m)]$ is less sensitive to the values of both $X_{(1)}$ and $X_{(2)}$: for reasonable values of n, both could be detected more easily as being outside the order-m boundary.

5.1 Example 1: bivariate case

The first example is a simulated one: we simulate a sample of $n = 100$ observations (x_i, y_i) according the DGP:

$$Y = X^{\beta} * \exp(-U), \tag{5.1}$$

where $\beta = 0.5$, X is uniform on (0,1) and U is an exponential with mean $\mu = 1/3$. The true average output-efficiency is $1/(\mu + 1) = 3/4$. Three outliers were added arbitrarily: units #(1, 2, 3), resulting in 103 observations shown in Figure 2.

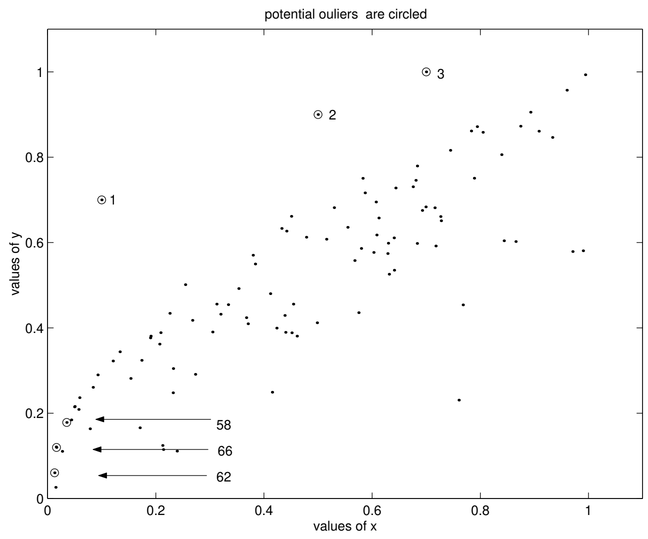

Figure 2: *Data set for Example 1 with 3 outliers: units #(1,2,3).*

To save space, we show in Tables 1 and 2 the order-m efficiencies of 14 units: the three outliers (unit #1, unit #2 and unit #3), some points detected as being potential outliers by our semi-automatic procedure, and other points chosen randomly.

Note first that the STD_{MC} are small, indicating that a value of B equal to 200, in this case, is large enough for our purpose. We see also that the 3 first units have clear extreme values for both order-m efficiency scores and that in the output direction unit #2 has only 4 other units with output larger than y_2. Unit #3 has the most extreme value for the output: $N_{input} = 0$ and $\hat{\theta}^{(3)}_{m,n}(x_3, y_3) = \infty$. The procedure confirms that units #(1, 2, 3) are outliers. Also, units #(58, 62, 65) are rather extreme in both directions. Unit #58 has the smallest observed input in the sample $N_{output} = 0$ and $\hat{\lambda}^{(3)}_{m,n}(x_{58}, y_{58}) = 0$. Note also that unit #61 is rather extreme in both directions but too a lesser extent (for $m = 150$, only

unit	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 10$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 25$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 50$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 75$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 100$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 150$	$\hat{\theta}^{(i)}_{FDH,n}(x_i, y_i)$
1	5.6741	5.2500	5.0747	5.0041	5.0000	5.0000	5.0000
	0.0407	0.0288	0.0168	0.0041	0.0000	0.0000	
	20	20	20	20	20	20	
2	1.4161	1.4000	1.4000	1.4000	1.4000	1.4000	1.4000
	0.0056	0.0000	0.0000	0.0000	0.0000	0.0000	
	4	4	4	4	4	4	
3	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
4	1.0506	0.9124	0.8705	0.8595	0.8579	0.8579	0.8579
	0.0130	0.0085	0.0044	0.0016	0.0000	0.0000	
	15	15	15	15	15	15	
5	0.6826	0.5041	0.3910	0.3230	0.3124	0.2603	0.2361
	0.0179	0.0151	0.0124	0.0103	0.0097	0.0058	
	67	67	67	67	67	67	
6	0.7494	0.5060	0.3435	0.2813	0.2101	0.2069	0.1940
	0.0234	0.0238	0.0194	0.0157	0.0071	0.0064	
	36	36	36	36	36	36	
58	5.9967	2.8815	1.8760	1.5627	1.3681	1.2949	1.2252
	0.4020	0.1615	0.0790	0.0491	0.0284	0.0188	
	101	101	101	101	101	101	
59	1.0936	0.8035	0.5396	0.4441	0.3530	0.2896	0.2604
	0.0259	0.0289	0.0266	0.0233	0.0174	0.0101	
	49	49	49	49	49	49	
60	0.5927	0.3915	0.2718	0.2320	0.1757	0.1634	0.1464
	0.0164	0.0180	0.0152	0.0132	0.0082	0.0063	
	40	40	40	40	40	40	
61	2.2698	1.6145	1.3062	1.2191	1.2095	1.1192	1.0771
	0.0678	0.0419	0.0222	0.0160	0.0150	0.0076	
	81	81	81	81	81	81	
62	3.8436	1.8396	1.4627	1.3594	1.3120	1.2693	1.2421
	0.2115	0.0684	0.0267	0.0191	0.0101	0.0044	
	92	92	92	92	92	92	
65	7.1394	3.7481	2.8314	2.6031	2.4688	2.3832	2.2423
	0.4058	0.1451	0.0684	0.0327	0.0250	0.0199	
	96	96	96	96	96	96	
102	0.6560	0.3302	0.2564	0.2340	0.2295	0.2196	0.2069
	0.0363	0.0133	0.0049	0.0025	0.0024	0.0017	
	93	93	93	93	93	93	
103	0.7040	0.7040	0.7040	0.7040	0.7040	0.7040	0.7040
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	1	1	1	1	1	1	

Table 1: *Leave-One-Out order-m input efficiency measures for example 1: the first 3 units are outliers. For each unit, the first row is the efficiency score, the second row, the Monte-Carlo standard deviation ($B = 200$) and the third row is N_{input}, the number of points in $\mathcal{X}$, with output level greater or equal to y_i.*

unit	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 10$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 25$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 50$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 75$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 100$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 150$	$\hat{\lambda}^{(i)}_{FDH,n}(x_i, y_i)$
1	0.3873	0.4070	0.4133	0.4135	0.4139	0.4139	0.4139
	0.0026	0.0012	0.0004	0.0003	0.0000	0.0000	
	14	14	14	14	14	14	
2	0.6672	0.7239	0.7539	0.7648	0.7700	0.7750	0.7778
	0.0069	0.0036	0.0022	0.0017	0.0013	0.0007	
	55	55	55	55	55	55	
3	0.7289	0.7798	0.8235	0.8443	0.8620	0.8806	0.9000
	0.0060	0.0055	0.0051	0.0047	0.0043	0.0033	
	80	80	80	80	80	80	
4	0.8788	1.0063	1.0602	1.1242	1.1502	1.1699	1.1992
	0.0101	0.0109	0.0101	0.0087	0.0074	0.0059	
	62	62	62	62	62	62	
5	1.3436	1.5104	1.6448	1.7113	1.7262	1.7414	1.7527
	0.0161	0.0142	0.0123	0.0078	0.0064	0.0042	
	44	44	44	44	44	44	
6	1.0749	1.2299	1.3378	1.3876	1.4255	1.4572	1.4806
	0.0161	0.0144	0.0125	0.0111	0.0090	0.0060	
	56	56	56	56	56	56	
58	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
59	0.9723	1.1140	1.2023	1.2096	1.2576	1.2610	1.2736
	0.0136	0.0111	0.0088	0.0082	0.0043	0.0039	
	41	41	41	41	41	41	
60	1.2057	1.2891	1.3721	1.3952	1.4335	1.4716	1.5055
	0.0115	0.0098	0.0094	0.0091	0.0081	0.0061	
	76	76	76	76	76	76	
61	0.8445	0.8840	0.8977	0.8994	0.8998	0.8998	0.8998
	0.0045	0.0024	0.0009	0.0004	0.0000	0.0000	
	13	13	13	13	13	13	
62	0.6787	0.6796	0.6796	0.6796	0.6796	0.6796	0.6796
	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	
	5	5	5	5	5	5	
65	0.4982	0.4982	0.4982	0.4982	0.4982	0.4982	0.4982
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	2	2	2	2	2	2	
102	2.8720	3.6409	4.1102	4.1960	4.2174	4.2174	4.2174
	0.0818	0.0681	0.0331	0.0151	0.0000	0.0000	
	18	18	18	18	18	18	
103	0.8309	0.9141	0.9608	0.9754	0.9841	0.9949	1.0068
	0.0074	0.0047	0.0035	0.0027	0.0025	0.0019	
	102	102	102	102	102	102	

Table 2: *Leave-One-Out order-m output efficiency measures for example 1: the first 3 units are outliers. For each unit, the first row is the efficiency score, the second row, the Monte-Carlo standard deviation ($B = 200$) and the third row is N_{output}, the number of points in $\mathcal{X}$, with input level smaller or equal to x_i.*

10% of superefficiency) and its leave-one-out FDH input score is not far from one. Finally unit #103 deserves some comments: it is extreme in the output direction $N_{input} = 1$, but its output efficiency score moves rapidly close toward 1 when m increases, so it is not an outlier.

As a conclusion, units #(1, 2, 3, 58, 62, 65) seem to be outside the cloud of points. Since here we have two dimensions we can simply plot the observations and identify the outliers. This is done in Figure 2, which reveals why units #(58, 62, 65) were flagged as outliers: they have extremely low input level, where the true frontier increases very fast ($\psi(x) = \sqrt{x}$) and there are not so many points in this area.

The selection of the printed rows in Tables 1 and 2 was done by our semi-automatic warning procedure, plus some random rows, to compare. For comparison, we also need the curves showing the percentage of points outside the frontier as a function of m. This is done in Figure 3.

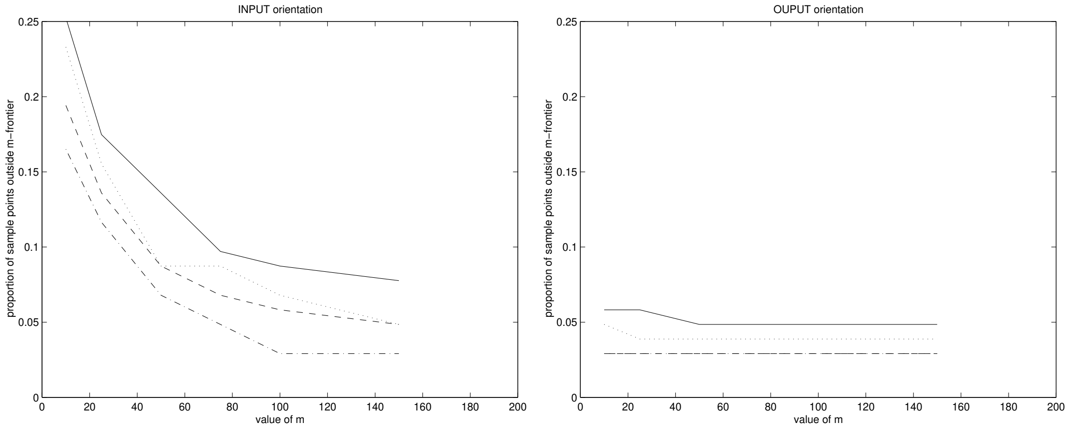

Figure 3: *Percentages of points outside the order-m frontier as a function of m and of the threshold value α for example 1 with 3 outliers. Solid line is for $\alpha = 0.20$, dotted for $\alpha = 0.30$, dashed for $\alpha = 0.40$ and dash-dotted for $\alpha = 0.50$. Left panel: input-oriented, right panel output oriented.*

The differences in the shape of the two curves should be noticed: this stresses the asymmetry of the treatment of inputs and outputs. This is due to our DGP where the Xs are uniformly generated but the Ys are generated conditionally on X according to the efficiency model (5.1). Here we realize that we have many more points outside the order-m frontier in the input direction than in the output direction, even for large values of m: since we want to flag points being outside the frontier in both directions, it will mainly be the output curves that will dominate in our choice of the tuning parameters.

The analysis of the right panel of Figure 3 suggests (elbow effect) in the output direction to select the values $m = 25$ and $\alpha = 0.20$. Looking to the left panel, we see this choice is very conservative with respect to the input-oriented case. This choice provides a reasonable number of flagged points: the procedure identifies automatically units #(1, 2, 3, 58, 62, 65) as extreme points. An automatic test on the values of N_{input} and of N_{output} for units having efficiency score greater or equal to 1 identifies units #(3, 58) as particularly extreme at least in one direction and are indeed flagged as potential outliers. So we see how the semi-automatic procedure helped to select a small number of rows in the full tables of results needing closer scrutiny.

The same analysis was performed on the same cloud of 100 points, without the three outliers units #(1,2,3). To save place, we do not reproduce the tables of results: the order of magnitude of the efficiency scores are the same, because the order-m concept is rather robust to outliers since the frontier does not envelop all the points. A careful analysis of the tables concludes exactly as above. The semi-automatic procedure is based on the analysis of Figure 4 (same peculiar shapes as above). The points flagged for the values $m = 25$, $\alpha = 0.20$ are as above units #(58, 62, 65), pointing here units #(58, 103) as being extremes when looking to N_{input} and of N_{output}; however unit #103 is not detected as a potential outlier, at this level of the tuning parameters.

We repeated the same exercice, in two dimensions (because we can see the clouds of points), with many other simulated data sets with convex and non-convex attainable set and we obtained roughly the same qualitative results. The conclusion is that effective outliers are detected but that, in a conservative approach some other extreme points (with small values for N_{input} or N_{output}) on the border of the support of the cloud of points in each variable direction could also be flagged by the procedure.

5.2 Example 2: multivariate simulated case

Here we simulate a data set of $n = 100$ points with $p = 2$ inputs and $q = 2$ outputs, according the scenario proposed in Park, Simar and Weiner (2000, page 866). Here the function describing the efficient frontier is given by:

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

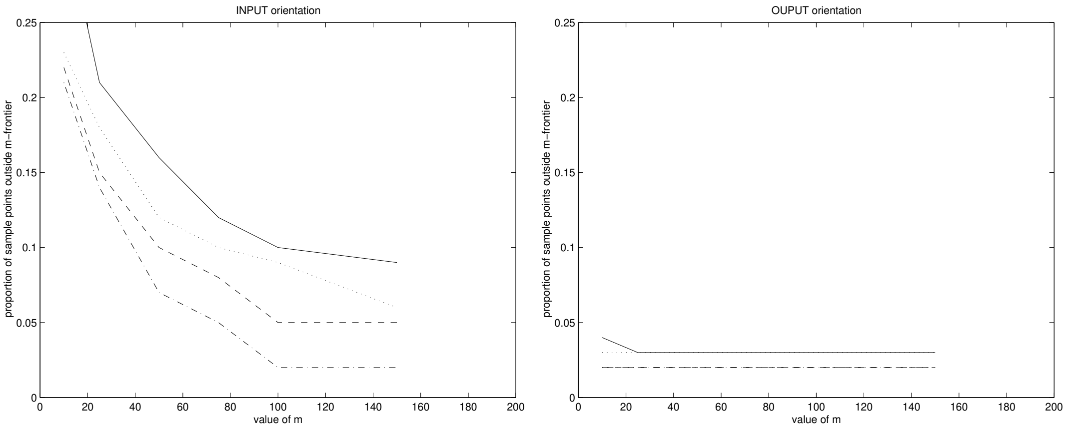

Figure 4: *Percentages of points outside the order-m frontier as a function of m and of the threshold value α for example 1 without outliers. Solid line is for $\alpha = 0.20$, dotted for $\alpha = 0.30$, dashed for $\alpha = 0.40$ and dash-dotted for $\alpha = 0.50$. Left panel: input-oriented, right panel output oriented.*

where y^j, (x^j), denotes the jth component of y, (of x), for $j = 1, 2$. We draw X_i^j independent uniforms on $(1, 2)$ and $\tilde{Y}_i^j$ independent uniform on $(0.2, 5)$. Then the generated random rays in the output space are characterized by the slopes $S_i = \tilde{Y}_i^2/\tilde{Y}_i^1$. Finally, the generated random points on the frontier are defined by:

$$
\begin{aligned}
Y_{i,eff}^1 &= \frac{1.0845(X_i^1)^{0.3}(X_i^2)^{0.4}}{S_i + 1} \\
Y_{i,eff}^2 &= 1.0845(X_i^1)^{0.3}(X_i^2)^{0.4} - Y_{i,eff}^1.
\end{aligned}
$$

We chose, as above, the efficiencies generated by $\exp(-U_i)$ where U_i are drawn from an exponential with mean $\mu = 1/3$. So that finally $Y_i = Y_{i,eff} * \exp(-U_i)$.

Here also, we add 3 outliers in the output space as follows. We define an outlier at $X_1 = (1.5, 1.5)$ with a slope $S_1 = 1$ in the output space and $\widehat{\lambda}_{FDH,n}(X_1, Y_1) = 0.6$, at $X_2 = (1.25, 1.75)$ with a slope $S_2 = 1/2$ and $\widehat{\lambda}_{FDH,n}(X_2, Y_2) = 0.6$ and finally for $X_3 = (1.75, 1.25)$ with a slope $S_3 = 2$ and $\widehat{\lambda}_{FDH,n}(X_3, Y_3) = 0.6$, where $\widehat{\lambda}_{FDH,n}(x, y)$ is the FDH efficiency score computed with the reference set given by the $n = 100$ points generated above. So the points are outside the FDH frontier estimated from the 100 “regular” points . Of course in this multivariate setup, we expect to have many FDH-efficient points. The outliers are again units #(1, 2, 3).

The results are shown in Tables 3 and 4. Again, a subset of the tables are printed to save space. This subset has been chosen by our semi-automatic warning procedure, plus some additional units chosen at random. The flagged rows were detected for a selected value of (m, α) chosen by looking to Figure 5.

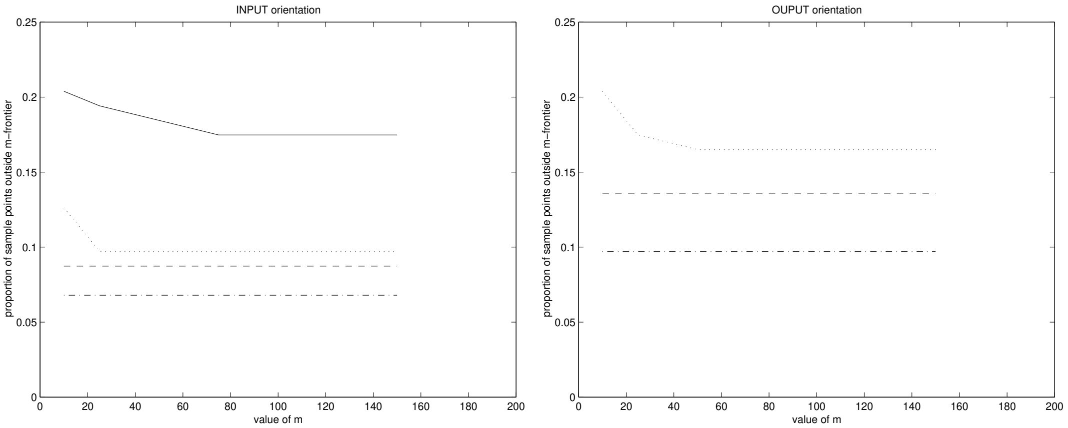

Figure 5: *Percentages of points outside the order-m frontier as a function of* m *and of the threshold value* α *for example 2 with 3 outliers. Solid line is for* $\alpha = 0.20$*, dotted for* $\alpha = 0.30$*, dashed for* $\alpha = 0.40$ *and dash-dotted for* $\alpha = 0.50$*. Left panel: input-oriented, right panel output oriented.*

The pictures suggests choosing $m = 50$ and $\alpha = 0.30$: this identifies as potential outliers the units#(1, 2, 3, 26). A more conservative choice is $m = 25$, $\alpha = 0.30$ which flags, in addition, the unit #85. Also, a test on the values of N_{input} and N_{output} for efficient points identifies the points #(32, 43, 72, 80, 84, 103) as being extreme in at least one direction. The other units in Tables 3 and 4 are chosen at random for comparison.

By looking more carefully at these tables, it appears that units #(1, 2, 3, 26) and, to a smaller extent, unit #85 are potential outliers. Of course, here, in a 4-dimensional framework, we have much more extreme points, with order-m input efficiency scores equal to ∞ for units # (1, 2, 3, 84), but unit #84, is not so extreme in the output direction (less than 7% superefficient for $m = 100$) and so is not detected as a potential outlier. For the order-m output efficiencies, units #(32, 43, 72, 80, 103) have a score equal to zero. But units #(32, 80) are not so extreme in the input direction (10% of super efficiency), however, units #(43, 72, 103) deserves a warning althoug much below the threshold of 30% of super efficiency chosen above. As a conclusion, we would stay with our 5 potential outliers #(1, 2, 3 , 26, 85) and 3 warnings #(43, 72, 103).

As for example 1, we redid the same analysis without the 3 outliers, and the results confirm what is written above. The semi-automatic warning procedure is based on Figure

unit	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 10$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 25$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 50$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 75$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 100$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 150$	$\hat{\theta}^{(i)}_{FDH,n}(x_i, y_i)$
1	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
2	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
3	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
4	1.1938	1.1634	1.1530	1.1517	1.1508	1.1508	1.1508
	0.0036	0.0020	0.0008	0.0006	0.0000	0.0000	
	14	14	14	14	14	14	
5	1.0376	1.0200	1.0195	1.0195	1.0195	1.0195	1.0195
	0.0028	0.0005	0.0000	0.0000	0.0000	0.0000	
	6	6	6	6	6	6	
6	0.7868	0.7803	0.7794	0.7794	0.7794	0.7794	0.7794
	0.0012	0.0003	0.0000	0.0000	0.0000	0.0000	
	9	9	9	9	9	9	
26	1.4356	1.4086	1.4064	1.4061	1.4061	1.4061	1.4061
	0.0047	0.0015	0.0004	0.0000	0.0000	0.0000	
	8	8	8	8	8	8	
32	1.2083	1.1420	1.1137	1.1078	1.1059	1.1054	1.1054
	0.0073	0.0043	0.0019	0.0011	0.0005	0.0000	
	22	22	22	22	22	22	
43	1.2912	1.2298	1.2112	1.2017	1.2007	1.1955	1.1942
	0.0067	0.0030	0.0015	0.0010	0.0010	0.0005	
	56	56	56	56	56	56	
44	0.8862	0.8804	0.8804	0.8804	0.8804	0.8804	0.8804
	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	
	4	4	4	4	4	4	
72	1.2733	1.2647	1.2636	1.2636	1.2636	1.2636	1.2636
	0.0014	0.0004	0.0000	0.0000	0.0000	0.0000	
	8	8	8	8	8	8	
73	0.7262	0.6961	0.6823	0.6782	0.6778	0.6774	0.6774
	0.0032	0.0025	0.0014	0.0006	0.0004	0.0000	
	18	18	18	18	18	18	
80	1.1284	1.1273	1.1273	1.1273	1.1273	1.1273	1.1273
	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	
	3	3	3	3	3	3	
84	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
85	1.4147	1.4147	1.4147	1.4147	1.4147	1.4147	1.4147
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	2	2	2	2	2	2	
86	0.8926	0.8663	0.8448	0.8420	0.8376	0.8353	0.8353
	0.0033	0.0028	0.0018	0.0016	0.0010	0.0000	
	25	25	25	25	25	25	
102	0.7235	0.6845	0.6663	0.6616	0.6610	0.6604	0.6602
	0.0041	0.0027	0.0012	0.0003	0.0003	0.0001	
	32	32	32	32	32	32	
103	1.3085	1.2464	1.1924	1.1760	1.1630	1.1630	1.1608
	0.0089	0.0078	0.0054	0.0039	0.0015	0.0015	
	26	26	26	26	26	26	

Table 3: *Leave-One-Out order-m input efficiency measures for example 2: the first 3 units are outliers. For each unit, the first row is the efficiency score, the second row, the Monte-Carlo standard deviation ($B = 200$) and the third row is N_{input}, the number of points in $\mathcal{X}$, with output level greater or equal to y_i.*

unit	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 10$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 25$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 50$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 75$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 100$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 150$	$\hat{\lambda}^{(i)}_{FDH,n}(x_i, y_i)$
1	0.5637	0.5935	0.5983	0.5997	0.5999	0.6000	0.6000
	0.0038	0.0014	0.0005	0.0001	0.0001	0.0000	
	24	24	24	24	24	24	
2	0.5345	0.5804	0.5950	0.5993	0.5991	0.6000	0.6000
	0.0055	0.0028	0.0011	0.0004	0.0004	0.0000	
	20	20	20	20	20	20	
3	0.5464	0.5819	0.5947	0.5989	0.5995	0.6000	0.6000
	0.0045	0.0021	0.0009	0.0004	0.0003	0.0000	
	24	24	24	24	24	24	
4	0.8853	0.9537	0.9807	0.9831	0.9837	0.9837	0.9837
	0.0081	0.0048	0.0013	0.0006	0.0000	0.0000	
	15	15	15	15	15	15	
5	0.8610	0.9353	0.9649	0.9715	0.9726	0.9732	0.9732
	0.0084	0.0049	0.0021	0.0008	0.0002	0.0000	
	28	28	28	28	28	28	
6	1.0562	1.2143	1.3085	1.3902	1.4169	1.4701	1.5619
	0.0146	0.0155	0.0148	0.0141	0.0137	0.0116	
	89	89	89	89	89	89	
26	0.1071	0.1071	0.1071	0.1071	0.1071	0.1071	0.1071
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	1	1	1	1	1	1	
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
43	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
44	0.8255	0.9145	0.9780	1.0074	1.0203	1.0397	1.0493
	0.0085	0.0081	0.0064	0.0052	0.0042	0.0026	
	56	56	56	56	56	56	
72	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
73	1.2441	1.4034	1.4786	1.5004	1.5231	1.5440	1.5590
	0.0144	0.0095	0.0058	0.0049	0.0039	0.0026	
	96	96	96	96	96	96	
80	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	
84	0.8151	0.8827	0.9148	0.9235	0.9294	0.9310	0.9328
	0.0070	0.0043	0.0021	0.0013	0.0006	0.0005	
	71	71	71	71	71	71	
85	0.6547	0.6832	0.7001	0.7039	0.7052	0.7059	0.7062
	0.0037	0.0023	0.0012	0.0007	0.0005	0.0003	
	23	23	23	23	23	23	
86	1.2385	1.4027	1.4636	1.4841	1.5005	1.5013	1.5015
	0.0157	0.0098	0.0059	0.0043	0.0002	0.0001	
	52	52	52	52	52	52	
102	1.5044	1.7604	1.8580	1.9482	2.0689	2.1094	2.1917
	0.0243	0.0229	0.0214	0.0198	0.0161	0.0134	
	82	82	82	82	82	82	
103	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0	0	0	0	0	0	

Table 4: *Leave-One-Out order-m output efficiency measures for example 2: the first 3 units are outliers. For each unit, the first row is the efficiency score, the second row, the Monte-Carlo standard deviation* ($B = 200$) *and the third row is* N_{input}*, the number of points in* $\mathcal{X}$*, with input level smaller or equal to* x_i*.*

6, which is very similar to Figure 5. For the values $m = 25$, $\alpha = 0.30$ the units #(26, 85) are flagged as potential ouliers. We do not reproduce the tables in the interest of space.

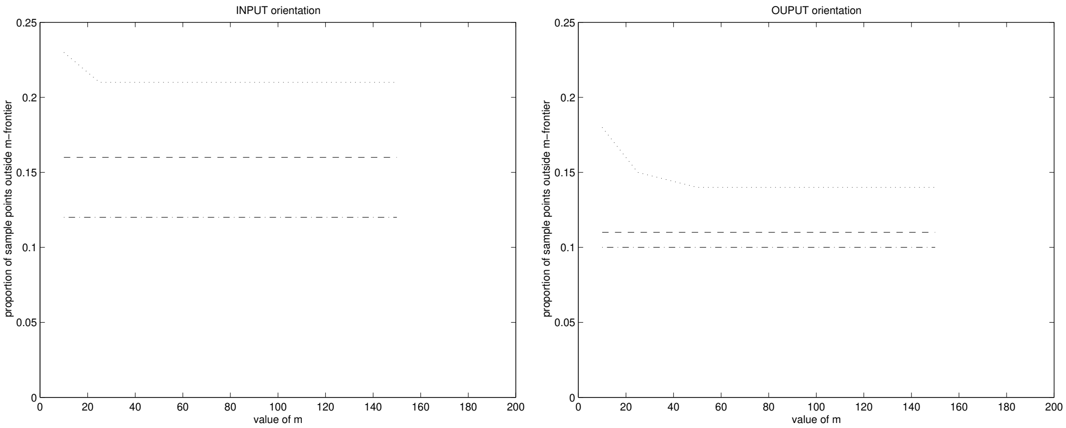

Figure 6: *Percentages of points outside the order-m frontier as a function of* m *and of the threshold value* α *for example 2 without the 3 outliers. Solid line is for* $\alpha = 0.20$*, dotted for* $\alpha = 0.30$*, dashed for* $\alpha = 0.40$ *and dash-dotted for* $\alpha = 0.50$*. Left panel: input-oriented, right panel output oriented.*

As a conclusion, the semi-automatic procedure helps effectively to build tables of results with a moderate number of rows. Of course, in a multidimensional framework, the number of warnings increases with the number of dimensions when the sample size is fixed.

In order to investigate the influence of the sample size on our procedure we redid the full example 2 with a basic reference sample of size $n = 1000$, adding the same outliers according the same process. Figure 7 identifies $m = 25$ and $\alpha = 0.30$ again as a reasonable choice of tuning parameters. Units #(1, 2, 3, 449) are identified as being potential outliers. In addition, 9 additional points are detected as being extreme in at least one dimension, although none of them are detected as potential outliers. A detailed analysis of the tables confirms this[5].

When redoing the analysis without the 3 outliers, the results are mostly confirmed: only two units #(128, 449) are detected as outliers over the 1000 data points (value of the tuning parameters $m = 25$, $\alpha = 0.30$). The procedure is clearly useful when sample size is large.

[5] The full tables of results are available on request to the author at simar@stat.ucl.ac.be.

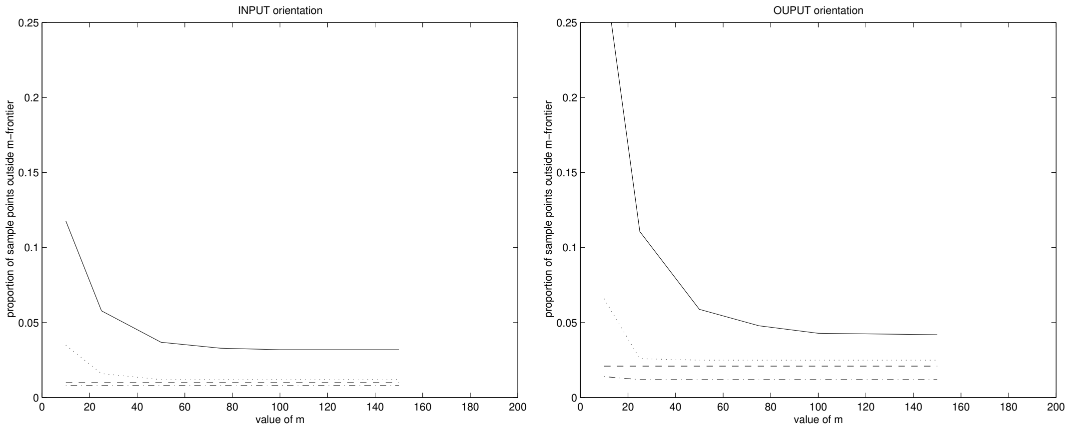

Figure 7: *Percentages of points outside the order-m frontier as a function of m and of the threshold value α for example 2 without the 3 outliers. Solid line is for $\alpha = 0.20$, dotted for $\alpha = 0.30$, dashed for $\alpha = 0.40$ and dash-dotted for $\alpha = 0.50$. Left panel: input-oriented, right panel output oriented.*

5.3 Example 3: multivariate real data

Her we examine real data in a multivariate setting: the data on Program Follow Through (PFT), an experimental education program administered in US schools are reported by Charnes, Cooper and Rhodes (1981). There are 5 inputs and 3 outputs for 70 schools. This data set has also been analyzed by Wilson (1993). Note that here in a 8-dimensional space, there is no room for doing inference, since the sample size is definitely too small for drawing inference in a FDH/DEA framework. So no definite conclusions can be drawn from this example, because all the estimators are flawed by large sampling variances. For instance, note that 64 of the 70 units are FDH-efficient: this indicates that in a 8 dimensional space, with a full non-parametric approach, 70 observations are too few to get sensible results. Wilson's results are based on convexity and of course we have no means to test, in a so small sample size, if the technology is convex. Our approach can handle convex and non convex sets. So we use this popular data set here just for illustration purposes and to check if our procedure provides some warnings on the most extreme points.

A sampled part of the tables of results is provided in Tables 5 and 6, where we show a selection of 28 units, among them, are the potential outliers detected by our "semi-automatic" procedure and those detected in Wilson (1993). The tables include also the 6 units #(4, 13, 31, 37, 53, 66), which **were not** FDH-efficient with the full data set. In order to save place in the tables we do not reproduce the value of the Monte-Carlo standard deviations (most of them were of an order much less than 0.01), even if, the semi-automatic procedure uses these values explicitely to decide if the order-m efficiency score is above (or below) the chosen threshold value.

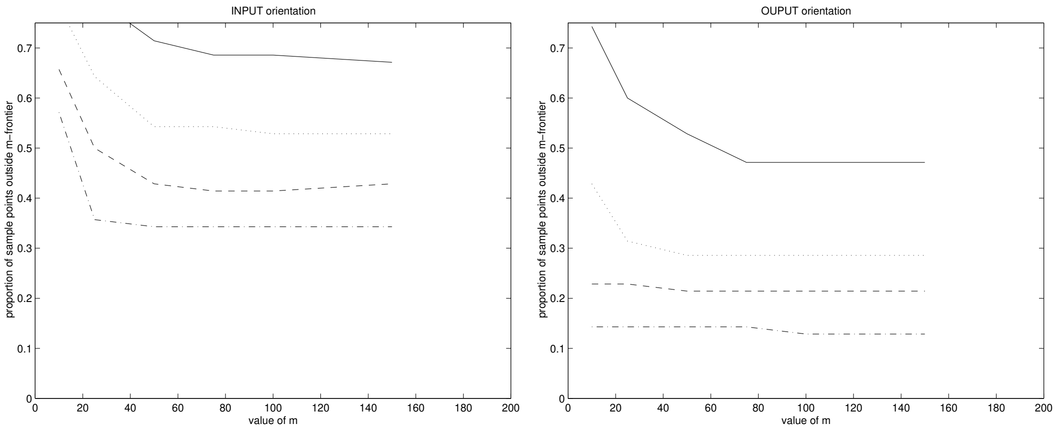

Figure 8: *Percentages of points outside the order-m frontier as a function of* m *and of the thershold value* $1 + \alpha$ *for the PFT data, example 3. Solid line is for* $\alpha = 0.20$*, dotted for* $\alpha = 0.30$*, dashed for* $\alpha = 0.40$ *and dash-dotted for* $\alpha = 0.50$*.*

The percentages of points outside the m-frontier for selected threshold values of α are shown in Figure 8. It appears that, as expected, these percentages are very high. So the idea here is to detect the most extreme points with small values for m and large values of α: here we select $m = 25$ and $\alpha = 0.5$ because we have several elbow effect at $m = 25$. The chosen value for $\alpha = 0.50$ will point out the 14% of the 70 units being the most extreme (note that here $\sqrt{n}/n \approx 0.12$).

This identifies the units #(5, 14, 32, 38, 44, 48, 58, 59, 62, 69). If we want to be more conservative and we chose the elbow at $m = 25$ and $\alpha = 0.40$, we identify, in addition the units #(15, 17, 49, 52, 56, 68). But this makes a total of 16 units over the sample of 70: this is too much in this small sample situation with a large dimensional space. Wilson's procedure identifies 3 groups of units as being potential outliers, #(44, 59), #(33, 35, 66, 67) and #(1, 50, 54). However these tables deserve some comments.

1. The 6 FDH non-efficient units #(4, 13, 31, 37, 53, 66) have indeed, as expected, their input (or their output) order-m efficiency score which is less or equal to 1 (larger or equal to 1) when m increases.

2. The most extreme points in both directions in the tables are indeed #(5, 14, 32, 38,

unit	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 10$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 25$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 50$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 75$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 100$	$\hat{\theta}^{(i)}_{m,n}(x_i, y_i)$ $m = 150$	$\hat{\theta}^{(i)}_{FDH,n}(x_i, y_i)$
1	1.7583	1.7583	1.7583	1.7583	1.7583	1.7583	1.7583
	2	2	2	2	2	2	
4	1.1456	1.0433	0.9934	0.9700	0.9593	0.9535	0.9511
	43	43	43	43	43	43	
5	2.2787	1.8529	1.6910	1.6346	1.6195	1.5979	1.5793
	64	64	64	64	64	64	
13	1.0447	0.9953	0.9883	0.9869	0.9867	0.9866	0.9866
	20	20	20	20	20	20	
14	1.9479	1.7054	1.6320	1.6109	1.6028	1.6005	1.5975
	59	59	59	59	59	59	
15	4.0746	3.4319	3.0624	2.9230	2.8396	2.7613	2.7273
	44	44	44	44	44	44	
17	2.0115	1.6784	1.5802	1.5397	1.5169	1.5076	1.5000
	43	43	43	43	43	43	
31	1.1637	1.0159	0.9761	0.9550	0.9546	0.9485	0.9454
	47	47	47	47	47	47	
32	1.9226	1.5438	1.4214	1.3504	1.3130	1.2769	1.2500
	69	69	69	69	69	69	
33	1.5352	1.5352	1.5352	1.5352	1.5352	1.5352	1.5352
	2	2	2	2	2	2	
35	1.8333	1.8333	1.8333	1.8333	1.8333	1.8333	1.8333
	2	2	2	2	2	2	
37	1.3465	1.1259	1.0403	1.0145	1.0138	1.0010	1.0000
	43	43	43	43	43	43	
38	2.2771	2.0168	1.9424	1.9221	1.9013	1.8935	1.8817
	60	60	60	60	60	60	
44	3.5097	3.5097	3.5097	3.5097	3.5097	3.5097	3.5097
	1	1	1	1	1	1	
48	3.2183	2.3049	1.9179	1.7644	1.7606	1.7229	1.7044
	60	60	60	60	60	60	
49	2.6536	2.2963	2.2388	2.2243	2.2204	2.2173	2.2143
	45	45	45	45	45	45	
50	1.1668	1.1000	1.0938	1.0938	1.0938	1.0938	1.0938
	9	9	9	9	9	9	
52	1.5066	1.5000	1.5000	1.5000	1.5000	1.5000	1.5000
	5	5	5	5	5	5	
53	1.4803	1.2371	1.1163	1.0582	1.0134	0.9924	0.9872
	37	37	37	37	37	37	
54	1.8891	1.8838	1.8838	1.8838	1.8838	1.8838	1.8838
	2	2	2	2	2	2	
56	3.8348	3.2193	2.7981	2.6725	2.6009	2.5475	2.5161
	44	44	44	44	44	44	
58	2.4656	2.1321	1.9937	1.9250	1.9019	1.9019	1.8908
	33	33	33	33	33	33	
59	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	0	0	0	0	0	0	
62	3.0023	2.3935	2.2156	2.1181	2.0646	2.0188	1.9810
	61	61	61	61	61	61	
66	1.2216	1.1152	1.0338	1.0032	0.9895	0.9839	0.9761
	36	36	36	36	36	36	
67	1.2585	1.1967	1.1794	1.1723	1.1711	1.1706	1.1706
	23	23	23	23	23	23	
68	2.8354	2.4874	2.3596	2.3382	2.3321	2.3291	2.3291
	18	18	18	18	18	18	
69	2.9957	2.4762	2.2437	2.1621	2.1215	2.0904	2.0604
	64	64	64	64	64	64	

Table 5: *Leave-One-Out order-m input efficiency measures for the PFT data (example 3). For each unit, the first row is the efficiency score, the second row, the number of points in* $\mathcal{X}$*, with output level greater or equal to* y_i*.*

unit	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 10$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 25$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 50$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 75$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 100$	$\hat{\lambda}^{(i)}_{m,n}(x_i, y_i)$ $m = 150$	$\hat{\lambda}^{(i)}_{FDH,n}(x_i, y_i)$
1	0.6318	0.7336	0.7697	0.7981	0.8085	0.8198	0.8284
	54	54	54	54	54	54	
4	0.9256	0.9871	1.0134	1.0164	1.0192	1.0198	1.0204
	21	21	21	21	21	21	
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0	0	0	0	0	0	
13	0.8864	0.9875	1.0153	1.0251	1.0308	1.0331	1.0339
	35	35	35	35	35	35	
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0	0	0	0	0	0	
15	0.5373	0.5373	0.5373	0.5373	0.5373	0.5373	0.5373
	2	2	2	2	2	2	
17	0.5740	0.5745	0.5745	0.5745	0.5745	0.5745	0.5745
	3	3	3	3	3	3	
31	1.0680	1.1268	1.1413	1.1453	1.1453	1.1456	1.1456
	17	17	17	17	17	17	
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0	0	0	0	0	0	
33	0.6891	0.7307	0.7463	0.7486	0.7507	0.7519	0.7527
	55	55	55	55	55	55	
35	0.7256	0.7801	0.7946	0.8004	0.8034	0.8051	0.8060
	37	37	37	37	37	37	
37	0.9746	1.0661	1.0788	1.0797	1.0797	1.0797	1.0797
	11	11	11	11	11	11	
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0	0	0	0	0	0	
44	0.4233	0.4533	0.4708	0.4765	0.4781	0.4822	0.4839
	52	52	52	52	52	52	
48	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0	0	0	0	0	0	
49	0.5656	0.5702	0.5702	0.5702	0.5702	0.5702	0.5702
	4	4	4	4	4	4	
50	0.6517	0.7054	0.7348	0.7473	0.7494	0.7525	0.7533
	39	39	39	39	39	39	
52	0.5212	0.5891	0.6275	0.6397	0.6443	0.6487	0.6487
	28	28	28	28	28	28	
53	0.8720	0.9968	1.0464	1.0658	1.0644	1.0672	1.0672
	20	20	20	20	20	20	
54	0.5926	0.6604	0.7052	0.7339	0.7379	0.7557	0.7609
	52	52	52	52	52	52	
56	0.5691	0.5691	0.5691	0.5691	0.5691	0.5691	0.5691
	2	2	2	2	2	2	
58	0.4370	0.4370	0.4370	0.4370	0.4370	0.4370	0.4370
	2	2	2	2	2	2	
59	0.3703	0.4238	0.4716	0.4919	0.4953	0.5176	0.5319
	69	69	69	69	69	69	
62	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0	0	0	0	0	0	
66	0.8874	0.9507	0.9827	0.9957	0.9963	1.0010	1.0021
	27	27	27	27	27	27	
67	0.7593	0.8553	0.9229	0.9362	0.9523	0.9523	0.9542
	32	32	32	32	32	32	
68	0.5383	0.5523	0.5545	0.5547	0.5547	0.5547	0.5547
	12	12	12	12	12	12	
69	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0	0	0	0	0	0	

Table 6: *Leave-One-Out order-m output efficiency measures for the PFT data (example 3). For each unit, the first row is the efficiency score, the second row, the number of points in* $\mathcal{X}$, *with input level smaller or equal to* x_i.

44, 48, 58, 59, 62, 69), the points detected as potential outliers by our semi-automatic procedure.

3. Of course, many other points would also be flagged by chosing $\alpha < 0.4$, but we know it is not reasonable in this illustration (too much points).

4. The points flagged by Wilson's procedure are quite different, except for units #(44, 59) which are really extreme (highest input order-m scores with $N_{input} = 1$ and 0 respectively). As pointed above the comparison is difficult because Wilson's method is based on influence function arguments and incorporate a convexity assumption. In addition, Wilson's method is not direction-specific, as is the method here (even if one can look in several direction sequentially). So it is not a surprise that the list of observations flagged by the two methods are different. But anyway, we can check how our procedure warns these points:

   (a) Units #(1, 33, 35, 54) : they are not so extreme in the output direction, although 20%-25% super efficient even for $m = 150$. But with these values of tuning parameters, many other points (6) would also have been detected in the full tables.

   (b) Unit #50: it is not so extreme in the input direction, only 10% super efficient even with $m = 25$.

   (c) Unit #66 is not really extreme in either direction.

   (d) Unit #67 is not so extreme in the output direction.

5. Note that in the input direction only one unit (#59) has no other unit having a larger level of output and that in the output direction (here $p = 3$), 6 units have no other units with smaller input (here $q = 5$).

As a conclusion, in this example, we show how our semi-automatic procedure allows to points the most extreme outlying points, but of course, the sample size is too small regarding the dimension to draw some definite conclusions. A closer analysis of the flagged units would be useful to understand why they are outlying the cloud of data points.

It should also be noticed that different outlier detection methods never flag the same observations always. The point is to use several methods. Here a new, easy-to-compute method is offered to augment those that already exist. Given that there are so few methods suitable for productivity settings (as opposed, for instance, to the regression setting, where there are many more methods), there is certainly room for another method, especially one that does not involve large computational burdens.

6 Conclusions

The concept of order-m frontier, introduced by Cazals, Florens and Simar (2002) is a useful concept of frontier which is easy to estimate, with nice statistical properties and certainly robust to outliers or extreme data points. In this paper we describe in details and in a comprehensive way all the steps for computing the order-m input and output efficiency measures in the general multi-output, multi-input framework.

Then we have shown how this tool can be useful, in an exploratory data analysis phase of any efficiency analysis of firms, with real data, in order to detect any potential outliers.

This method, should be used in a first step, before performing any frontier estimation. This is true for DEA, FDH techniques, but also any parametric techniques using the deterministic approach.

Appendix

A MATLAB Code for Computing Order-m Efficiency measures.

The following code is written for MATLAB. It computes for a fixed point x_0, y_0 the input and the output order-m efficiency measures. the formulae are given in Section 3 by expression (3.6) and (3.10). The input arguments of the function are $xk = x_0, yk = y_0$, the data matrices for the reference set: (X, Y), the value of m and the value of B, to tune the precision of the Monte-Carlo approximations: usually, $B = 200$ provide already sensible approximations.

The output arguments of the function are

$$\text{effmk} = [\text{effmk}(1), \text{effmk}(2)] = \left[\hat{\theta}_{m,n}(x_0, y_0), \hat{\lambda}_{m,n}(x_0, y_0)\right]$$
$$\text{stdeffk} = [\text{stdeffk}(1), \text{stdeffk}(2)] = [\text{STD}_{\text{MC}}(\text{effmk}(1)), \text{STD}_{\text{MC}}(\text{effmk}(2)]$$

All the statements starting with a % are comments lines: so we have less than 30 effective statements in this code. The function "eff_m" uses an other function "resample" which is also provided below.

To give an idea on how the program is fast on a Pentium III, 450 Mghz, we provide some computing times for some of the examples of Section 5. The computation of the order-m efficiencies, for $m = 10, 25, 50, 75, 100, 150$, plus the computation of the FDH efficiency scores (all the 7 efficiencies togheter, both input and output oriented), for all the n units, is very fast. Table 7 shows how this computing time varies in function of n, and $p + q$. The dimension of the problem has very weak impact on the computing time and the sample size has only a linear effect. Clearly, the effect of the choice of B, the number of Monte-Carlo replications in computing the order-m efficiencies is also linear (not provided in Table 7, were $B = 200$).

sample size n	$p = 1$ and $q = 1$	$p = 2$ and $q = 2$
100	82	97
500	392	514
1000	785	1014

Table 7: *Computing time, in seconds, for producing the full tables of results for the n observations, for $m = 10, 25, 50, 75, 100, 150$ and the FDH efficiency scores. All scores in both direction (input and output). The value of B is set to 200.*

The full tables of results for the PFT example ($n = 70$ and $(p, q) = (3, 5)$) were obtained in 80 seconds.

```
%***************************************************************
function [effmk,stdeffk]=eff_m(xk,yk,X,Y,m,B)
%
%     ===========================
%     order-m efficiency measures
%     ===========================
%  written by L. SIMAR, july 8, 2001
%
%         IN
%         ------
%            X      : Matrix of input(s)  (n x p)
%            Y      : Matrix of output(s) (n x q)
%            xk,yk  : coordinate of the reference point
%            m      : ORDER of the frontier
%            B      : number of Monte Carlo replication
%
%         OUT
%         ------
%            effmk  : efficiency vector of ORDER-m,
%                     vector (1 x 2)=(m-input, m-output)
%            stdeffk: Monte-Carlo stand. dev. of estimates,
%                     vector (1 x 2)
%
%  This function uses the function 'resample'
%
[n,p] = size(X);
[n,q] = size(Y);%
%
% define the sample where the m iid units will be drawn
%-------------------
%  INPUT orientation
ykv=ones(n,1)*yk;
flagy=(Y>=ykv);
flagy=all(flagy,2);
%
XM=X(flagy,:);
[nxm,pxm]=size(XM);
if nxm==0
   disp('order-m input frontier not available, yk is too large')
   break
end
%--------------------
%  OUTPUT orientation
xkv=ones(n,1)*xk;
flagx=(X<=xkv);
flagx=all(flagx,2);
%
YM=Y(flagx,:);
[nym,qym]=size(YM);
if nym==0
   disp('order-m frontier not available, xk is too small')
   break
end
%
% start Monte-Carlo loop
%
thetab=[];
for b=1:B
   XMb=resample(XM,m);% 'resample' is another matlab function
   YMb=resample(YM,m);
   xkv=ones(m,1)*xk;
   ykv=ones(m,1)*yk;
```

```
   ratioxk=XMb./xkv;
   ratioyk=YMb./ykv;
   I_thetak=min(max(ratioxk,[],2),[],1);
   O_thetak=max(min(ratioyk,[],2),[],1);
   thetab=[thetab;[I_thetak  O_thetak]];
end
% end of the Monte-Carlo loop
%
effmk=mean(thetab);
stdeffk=std(thetab)/sqrt(B);
%
%
if effmk(1)==Inf
   stdeffk(1)=0;
end
% end of the function eff_m
%**************************************************************


%**************************************************************
function xb=resample(x,m)
%
% written by L. SIMAR, july 2001
%
% xb is a resample with replacement from  a matrix x: (n x k)
% the entire ROW of x is drawn at each step
% m can be smaller, equal or larger than n
%
[n,k]=size(x);
sample=floor(n*rand(m,1)+1);
xb=x(sample,:);
%
% end of the function resample
%**************************************************************
```

References

[1] Barnett, V. and T. Lewis (1995), *Outliers in Statistical Data*, Chichester, Wiley.

[2] Cazals, C., Florens, J.P. and L. Simar (2002), Nonparametric frontier estimation: a robust approach, *Journal of Econometrics*, 106, 1–25.

[3] Charnes, A., Cooper, W.W. and E. Rhodes (1978), Measuring the inefficiency of decision making units. *European Journal of Operational Research*, 2, 429–444.

[4] Charnes, A., Cooper, W.W. and E. Rhodes (1981), Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through, *Management Science* 27, 668–697.

[5] Christensen, L. and R. Greene (1976), Economies of scale in U.S. electric power generation. *Journal of Political Economy*, 84, 653–667.

[6] Davies, P.L. and U. Gather (1993), The identification of multiple outliers, *Journal of the American Statistical Association*, 88, 423, 782–801.

[7] Debreu, G. (1951), The coefficient of resource utilization, *Econometrica* 19(3), 273–292.

[8] Deprins, D., Simar, L. and H. Tulkens (1984), Measuring labor inefficiency in post offices. In *The Performance of Public Enterprises: Concepts and measurements*. M. Marchand, P. Pestieau and H. Tulkens (eds.), Amsterdam, North-Holland, 243–267.

[9] Farrell, M.J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A*, 120, 253–281.

[10] Gijbels, I., Mammen, E., Park, B.U. and L. Simar (1999). On estimation of monotone and concave frontier functions. *Journal of the American Statistical Association*, 94, 445, 220–228.

[11] Greene, W.H. (1990). A gamma-distributed stochastic frontier model. *Journal of Econometrics*, 46, 141–163.

[12] Hall, P., and L. Simar (2000), Estimating a change point, boundary or frontier in the presence of observation errors, Discussion paper #0012, Institut de Statistique,

UCL, Louvain-la-Neuve, Belgium (http://www.stat.ucl.ac.be), forthcoming, June 2002, in *Journal of the American Statistical Association*

[13] Kneip, A., Park, B.U. and L. Simar (1998). : A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, 14, 783–793.

[14] Koopmans, T.C. (1951), An Analysis of Production as an Efficient Combination of Activities, in *Activity Analysis of Production and Allocation*, ed. by T.C. Koopmans, Cowles Commission for Research in Economics, Monograph 13. New York: John-Wiley and Sons, Inc.

[15] Olesen, O.B., Petersen, N.C. and C.A.K. Lovell, eds. (1996), Summary of Workshop Discussion, *Journal of Productivity Analysis* 7, 341–345.

[16] Park, B. Simar, L. and Ch. Weiner (2000), The FDH Estimator for Productivity Efficiency Scores : Asymptotic Properties, *Econometric Theory*, Vol 16, 855-877.

[17] Shephard, R.W. (1970), *Theory of Cost and Production Function*. Princeton: Princeton University Press.

[18] Simar, L. (2002), How to improve the performances of DEA/FDH estimators in the presence of noise?, manuscript, Institut de Statistique, UCL, Belgium.

[19] Simar, L., and P.W. Wilson (1998), Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* 44(11), 49–61.

[20] Simar, L., and P.W. Wilson (2000a), Statistical inference in nonparametric frontier models: The state of the art, *Journal of Productivity Analysis* 13, 49–78.

[21] Simar, L., and P.W. Wilson (2000b), A General Methodology for Bootstrapping in Nonparametric Frontier Models, *Journal of Applied Statistics*,27 (6), 779–802.

[22] Wilson, P. W. (1993), Detecting outliers in deterministic nonparametric frontier models with multiple outputs, *Journal of Business and Economic Statistics* 11, 319–323.

[23] Wilson, P. W. (1995), Detecting influential observations in data envelopment analysis, *Journal of Productivity Analysis* 6, 27–45.