I N S T I T U T D E
S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

0406

A BIAS CORRECTED NONPARAMETRIC ENVELOPMENT ESTIMATOR OF FRONTIERS

BADIN, L. and L. SIMAR

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

A Bias Corrected Nonparametric Envelopment Estimator of Frontiers

Luiza Bădin*
Department of Mathematics
Academy of Economic Studies
Piaţa Romană nr. 6
Bucharest, Romania

Léopold Simar†
Institut de Statistique
Université Catholique de Louvain
Voie du Roman Pays, 20
B-1348 Louvain-la-Neuve, Belgium

February 26, 2004

Abstract

In productivity analysis, the efficiency scores of economic producers are evaluated by measuring their distance toward a production frontier. This frontier is defined as the set of the most efficient alternatives among all possible combinations in the input-output space, and it has to be estimated from a random sample of firms. The nonparametric envelopment estimators rely on the assumption that all the observations fall on the same side of the frontier. Therefore any deviation from the frontier is due only to inefficiency. The Free Disposal Hull (FDH) estimator of the attainable set, introduced by Deprins, Simar and Tulkens (1984), is the smallest free disposal set covering all the observations. However, by construction, the FDH estimator is an inward-biased estimator of its theoretical correspondent.

In this paper we consider the univariate extreme values representation of the FDH estimators and we propose a bias corrected estimator for the efficient frontier or for the efficiency scores, based on order statistics and closely related to FDH. The presentation is based on a probabilistic formulation of the model where the input-output pairs are realizations of independent random variables drawn from a joint distribution whose support is the production attainable set. The bias corrected estimator shares the asymptotic properties of the FDH estimator. But in finite samples, Monte-Carlo experiments indicate that our bias corrected estimator reduces significantly not only the bias of the FDH estimator, but also its mean squared error,with no computational cost.

Keywords: frontier, nonparametric efficiency estimation, FDH, bias correction, bootstrap.

*Research support from the EC project "Inco-Copernicus", IC15-CT98-1007 is acknowledged.

†Research support from "Projet d'Actions de Recherche Concertées" (No. 98/03–217) and from the "Interuniversity Attraction Pole", (IAP research network nr P5/24) of the Belgian Government (Belgian Science Policy) are also gratefully acknowledged.

1 Introduction

In productivity analysis, the efficiency scores of economic producers are evaluated by measuring their distance toward a production frontier. This frontier is defined as the set of the most efficient alternatives among all possible combinations in the inputs-outputs space. It could be characterized by the maximal achievable amount of outputs for a given level of the inputs or conversely, by the minimal input-combination allowing to produce a given level of the outputs.

The basic economic ideas dates back to Debreu (1951) and Farrell (1957) who used radial distances for defining efficiency measures. The production set, *i.e.* the feasible input-output combinations, can be defined as:

$$\Psi = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\right\}, \tag{1.1}$$

where $x \in \mathbb{R}_+^p$ is the input vector required to produce the output vector $y \in \mathbb{R}_+^q$. In the input-oriented[1] approach, the goal is to achieve the maximum radial (proportional) reduction in inputs that still allows a producer to give the same output. Therefore, for any $y \in \mathbb{R}_+^q$ we can define the input requirement set $X(y)$ as the set of all input vectors which yield at least y:

$$X(y) = \{x \in \mathbb{R}_+^p \mid (x, y) \in \Psi\} \tag{1.2}$$

and the corresponding input-oriented Debreu-Farrell efficiency boundary or input efficient frontier:

$$\partial X(y) = \{x \in \mathbb{R}_+^p \mid x \in X(y),\ \theta x \notin X(y)\ \ \forall \theta \in (0, 1)\}. \tag{1.3}$$

Consequently, the Debreu-Farrell input-oriented measure of technical efficiency can be defined as:

$$\theta(x, y) = \inf\{\theta \mid \theta x \in X(y)\} = \inf\{\theta \mid (\theta x, y) \in \Psi\}. \tag{1.4}$$

A producer (x, y) operating on the frontier has an input efficiency score $\theta(x, y) = 1$, while a value $\theta(x, y) < 1$ indicates the radial reduction in inputs that producer (x, y) should perform in order to become input-efficient. This allows us to define for each level of output, the efficient level of input, by projecting (x, y) radially, on the frontier, in the input direction:

$$x^{\partial}(y) = \theta(x, y)x. \tag{1.5}$$

[1]In the presentation of the paper, we will follow the input-orientation, the presentation for the output-orientation is summarized in the Appendix.

Since Ψ is generally unknown, it has to be estimated from a *iid* sample of n firms $\mathcal{X} = \{(X_i,Y_i)\,,\ i = 1, ..., n\}$. Farrell (1957) suggested as an estimator of Ψ, the minimal convex set covering all the observations. This idea was formalized as a linear programming model by Charnes, Cooper and Rhodes (1978) and is known as Data Envelopment Analysis (DEA). The approach is thus based on envelopment techniques assuming the convexity of the production possibilities set and the free disposability of inputs and outputs, *i.e.* if $(x, y) \in \Psi$, then $(\tilde{x}, \tilde{y}) \in \Psi$ for all $(\tilde{x}, \tilde{y})$ such that $\tilde{x} \geq x$ and $\tilde{y} \leq y$, where the inequalities between vectors are defined element by element. The DEA estimator of Ψ is:

$$\widehat{\Psi}_{DEA} = \left\{ (x, y) \in \mathbb{R}_+^{p+q} \mid y \leq \sum_{i=1}^{n} \gamma_i Y_i,\ x \geq \sum_{i=1}^{n} \gamma_i X_i,\ \gamma_i \geq 0,\ i = 1, ..., n \right\}. \tag{1.6}$$

By relaxing the convexity assumption, on the physically attainable set, Deprins, Simar and Tulkens (1984) proposed the Free Disposal Hull (FDH) estimator, which is the smallest free disposal set covering all the observations:

$$\widehat{\Psi}_{FDH} = \left\{ (x, y) \in \mathbb{R}_+^{p+q} \mid y \leq Y_i, x \geq X_i \right\}. \tag{1.7}$$

The efficiency of the arbitrary point (x, y) is then measured with respect to the boundary of the free disposal hull of the sample. The FDH estimator of the input efficiency score is thus:

$$\widehat{\theta}_{FDH}(x, y) = \inf \left\{ \theta > 0 \mid (\theta x, y) \in \widehat{\Psi}_{FDH} \right\}, \tag{1.8}$$

and the efficient level of inputs for a firm operating at the level (x, y) is defined by

$$\widehat{x}^{\partial}_{FDH}(y) = \widehat{\theta}_{FDH}(x, y)x. \tag{1.9}$$

Recently Daraio and Simar (2003), extending ideas of Cazals, Florens and Simar (2002), proposed a probabilistic formulation of the model and of the Farrell efficiency scores, with a natural nonparametric estimator. Here the production process is described by the joint distribution of (X, Y) on $\mathbb{R}_+^p \times \mathbb{R}_+^q$, whose support is Ψ and the observations are drawn independently from this distribution. It is shown that the input oriented technical efficiency for a fixed point $(x, y) \in \Psi$ can also be defined in terms of the p-variate conditional distribution $F(x|y) = P(X \leq x|Y \geq y)$, as follows:

$$\theta(x, y) = \inf\{\theta \mid F(\theta x|y) > 0\}. \tag{1.10}$$

Under free disposability of the inputs and of the outputs, this is equivalent to the Debreu-Farrell input efficiency score.

A natural nonparametric estimator of $\theta(x,y)$, based on the random sample $\mathcal{X}$, is obtained by applying the plug-in principle, replacing $F(x|y)$ by its empirical version:

$$\widehat{F}_n(x|y) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \leq x, Y_i \geq y)}{\sum_{i=1}^n \mathbb{1}(Y_i \geq y)},$$

where $\mathbb{1}(\cdot)$ is the indicator function. The nonparametric estimator of the input efficiency score $\theta(x,y)$ is then:

$$\widehat{\theta}(x,y) = \inf\{\theta|\widehat{F}_n(\theta x|y) > 0\} \equiv \widehat{\theta}_{FDH}(x,y), \tag{1.11}$$

where the last identity was noted by Daraio and Simar (2003).

By looking to (1.11), it is clear that the FDH estimator of the input efficiency score can be computed through a "sample minimum":

$$\widehat{\theta}_{FDH}(x,y) = \min_{i|Y_i \geq y} \left\{ \max_{j=1,\ldots,p} \frac{X_i^j}{x^j} \right\}. \tag{1.12}$$

where a^j denotes the jth component of a vector a.

This representation as sample extreme value is the most intuitive in the simplest case of a univariate frontier. Suppose indeed that each firm produces one unit of output using input X, where X is a random variable with cdf $F_X(x)$ with support (ϕ, ∞). Then ϕ can be interpreted as the lowest level of input required to produce one unit of output:

$$\phi = \inf\{x|F_X(x) > 0\}. \tag{1.13}$$

Also, ϕ represents the input efficient frontier to be estimated:

$$\phi = x^{\partial}(x,1) = \theta(x,1)x. \tag{1.14}$$

If $\mathcal{X} = \{(X_i,1), i = 1,...,n\}$ is a random sample of producers and $X_{(1)} \leq ... \leq X_{(n)}$ are the order statistics of the input, then ϕ can be estimated by the sample minimum $X_{(1)}$:

$$\widehat{\phi} = X_{(1)}. \tag{1.15}$$

The input efficiency score of producer $(x,1)$ is now estimated by:

$$\widehat{\theta}(x,1) = X_{(1)}/x, \tag{1.16}$$

which coincides with the FDH estimator defined in (1.12).

However, by construction, FDH estimators are biased and this can be easily observed in the univariate frontier case: although the sample minimum is the closest observation to ϕ,

it is always larger than ϕ. In what concerns the estimators of the attainable set, we have $\widehat{\Psi}_{FDH} \subseteq \Psi$, thus $\widehat{\theta}_{FDH}$ is upward-biased estimator of the true input efficiency score.

Therefore, the problem of efficiency estimation is connected with the classical problem of estimating the bounds of random variables. Several authors who studied this issue proposed point estimators for ϕ that attempt to improve the extreme order statistics in terms of the bias, without affecting the mean squared error.

For the twin problem of estimating the upper bound of the support of a distribution, Robson and Whitlock (1964) applied Quenouille's bias reduction technique and obtained the family of estimators:

$$T_n^{(k)} = \sum_{i=0}^{k} (-1)^i \left(\begin{array}{c} k+1 \\ i+1 \end{array} \right) X_{(n-i)}, \quad (0 < k < n) \tag{1.17}$$

with a bias of order $n^{(-k-1)}$, but with the mean squared error increasing with k. Yet, the bias adjustment involving only the second largest observation does not affect the mean squared error, i.e. $T_n^{(1)} = X_{(n)} + \left(X_{(n)} - X_{(n-1)}\right)$ and $X_{(n)}$ are subject to the same mean squared error. In terms of the lower bound, the statistic:

$$W_n^{(1)} = X_{(1)} - \left(X_{(2)} - X_{(1)}\right) = 2X_{(1)} - X_{(2)} \tag{1.18}$$

partially corrects $X_{(1)}$ for bias and has the same mean squared error as $X_{(1)}$.

Cooke (1979) also discussed this issue and proposed the following estimator for ϕ :

$$\widehat{\phi}_n = 2X_{(1)} - \sum_{i=1}^{n} \left[\left(1 - (i-1)/n\right)^n - (1 - i/n)^n \right] X_{(i)}, \tag{1.19}$$

which, for large n, is asymptotically equivalent with:

$$\widetilde{\phi}_n = 2X_{(1)} - (e-1) \sum_{i=1}^{n} e^{-i} X_{(i)}. \tag{1.20}$$

In this paper we propose nonparametric estimators for the efficient frontier and for the efficiency scores which attempt to improve the FDH estimators in terms of bias using Cooke's ideas, without affecting the mean squared error. We will first introduce the idea by following the presentation of Cazals, Florens and Simar (2002), where the minimum input frontier function is estimated for one input and several outputs, then using the Daraio and Simar (2003) extension, we generalize to the full multivariate case.

The paper is organized as follows. Section 2 presents the idea for constructing the bias corrected estimator of the input efficient frontier and introduces a "conditional" bootstrap that permits derivation of the same estimator using bootstrap arguments. Section 3 introduces the full multivariate extension with the asymptotic properties of our bias corrected

estimator. Section 4 provides numerical results from Monte Carlo simulations in order to investigate the finite sample properties of the estimator and its significant improvement relative to the FDH estimator. The results show indeed that not only the bias of the FDH efficiency estimator is clearly reduced but also the mean squared error. The paper ends with a section giving conclusions.

2 Bias Correction for the FDH Efficient Frontier

2.1 Estimation of a minimum input function

In order to present the basic ideas of our bias-corrected estimator, we first analyze, as in Cazals, Florens and Simar (2002), the estimation of a "cost" function[2], *i.e.*, a minimum univariate input function for several outputs. Let us consider the production process defined by the continuous joint distribution of a random vector (X, Y) on $\mathbb{R}_+ \times \mathbb{R}_+^q$, where X represents a scalar input used to produce an output vector Y. The production possibilities set Ψ is the support of the input-output pairs and we assume here that Ψ is free disposal.

The joint distribution of (X, Y) is characterized by the joint survivor function:

$$S(x, y) = P(X \geq x, Y \geq y)$$

and we consider the conditional distribution of X given $Y \geq y$

$$F(x|y) = P(X \leq x|Y \geq y),$$

whose survivor function is:

$$S(x|y) = P(X \geq x|Y \geq y).$$

The input efficient frontier, is thus defined as:

$$x^{\partial}(y) = \inf\{x|F(x|y) > 0\} = \inf\{x|S(x|y) < 1\} \tag{2.21}$$

and represents the lower boundary of the support of $F(x|y)$. As shown in Cazals, Florens and Simar (2002) (Theorem 2.1), this is a monotone function, nondecreasing in y which coincides, under free disposability, to the efficient frontier as defined in (1.5).

Consider now the observations in the random sample $\mathcal{X}$ such that $Y_i \geq y$, we denote this sample by $\left\{X_1^y, ..., X_{n_y}^y\right\}$, where n_y is the number of such observations. Conditionally on $Y \geq y$, this can be considered as a random sample of size n_y generated by the conditional distribution $F(x|y)$. Denote $X_{(1)}^y \leq ... \leq X_{(n_y)}^y$ the corresponding order statistics. The

[2]For the case of a "production" function (one output and several inputs), see the Appendix.

random variable $X_{(1)}^{y}$ is, in this univariate input case, the FDH estimator of the efficient level of input (corresponding to $\widehat{x}_{FDH}^{\partial}(y)$ as defined in (1.9)). It has the conditional distribution

$$P\left(X_{(1)}^{y} \leq x | Y \geq y\right) = 1 - \left[1 - F\left(x|y\right)\right]^{n_y} = 1 - \left[S\left(x|y\right)\right]^{n_y} \tag{2.22}$$

and conditional mean

$$\begin{aligned} E\left(X_{(1)}^{y} | Y \geq y\right) &= \int_{x^{\partial}(y)}^{\infty} x\, d(1 - \left[S\left(x|y\right)\right]^{n_y}) = -\int_{x^{\partial}(y)}^{\infty} x\, d\left[S\left(x|y\right)\right]^{n_y} \\ &= x^{\partial}(y) + \int_{x^{\partial}(y)}^{\infty} \left[S\left(x|y\right)\right]^{n_y} dx, \end{aligned}$$

because $S\left(x|y\right) = 1$ for any $x \leq x^{\partial}(y)$ and $\lim_{x\to\infty} x\left[S\left(x|y\right)\right]^{n_y} = 0$, by assuming the existence of the expectation. Therefore, we obtain

$$x^{\partial}(y) = E\left(X_{(1)}^{y} | Y \geq y\right) - \int_{x^{\partial}(y)}^{\infty} \left[S\left(x|y\right)\right]^{n_y} dx. \tag{2.23}$$

Replacing in the right hand side of the previous formula, the unknowns $E\left(X_{(1)}^{y} | Y \geq y\right)$ and $x^{\partial}(y)$ by $X_{(1)}^{y}$ (the FDH estimator of the frontier), and the unknown survivor function by its empirical correspondent, we can define the following estimator:

$$\widehat{x_{n_y}^{\partial}}(y) = X_{(1)}^{y} - \int_{X_{(1)}^{y}}^{\infty} \left[\widehat{S}\left(x|y\right)\right]^{n_y} dx = X_{(1)}^{y} - \int_{X_{(1)}^{y}}^{X_{(n_y)}^{y}} \left[\widehat{S}\left(x|y\right)\right]^{n_y} dx,$$

where the empirical conditional survivor function is defined as:

$$\widehat{S}\left(x|y\right) = \begin{cases} 1 & , \text{ if } x < X_{(1)}^{y} \\ \frac{n_y - i}{n_y} & , \text{ if } X_{(i)}^{y} \leq x < X_{(i+1)}^{y}, \text{ for } 1 \leq i \leq n_y - 1 \\ 0 & , \text{ if } x \geq X_{(n_y)}^{y}. \end{cases}$$

We thus obtain our bias corrected estimator of the minimum input function:

$$\widehat{x_{n_y}^{\partial}}(y) = X_{(1)}^{y} - \sum_{i=1}^{n_y - 1} \left(1 - \frac{i}{n_y}\right)^{n_y} \left(X_{(i+1)}^{y} - X_{(i)}^{y}\right), \tag{2.24}$$

where the first term is the FDH estimator and the second term is the bias correction. The estimator can also be written as:

$$\widehat{x_{n_y}^{\partial}}(y) = 2X_{(1)}^{y} - \sum_{i=1}^{n_y} \left[\left(1 - \frac{i-1}{n_y}\right)^{n_y} - \left(1 - \frac{i}{n_y}\right)^{n_y}\right] X_{(i)}^{y}. \tag{2.25}$$

For large n_y, it is asymptotically equivalent to

$$\widehat{x_{n_y}^{\partial}}(y) = 2X_{(1)}^{y} - (e-1) \sum_{i=1}^{n_y} e^{-i} X_{(i)}^{y}, \tag{2.26}$$

which is the estimator proposed by Cooke (1979) for the case of univariate *iid* random variables. The asymptotic properties of our estimator will be derived in the full multivariate case in Section 3 and the finite sample behavior will be investigated by Monte-Carlo experiments in Section 4.

Remark 2.1. *Amazingly, our estimator is a symmetric version of the order-m minimum input function proposed by Cazals, Florens and Simar (2002), relative to the FDH estimator. Indeed, the order-m estimator can be written as:*

$$\widehat{\varphi}_{m,n}(y) = X^y_{(1)} + \sum_{i=1}^{n_y-1} \left(1 - \frac{i}{n_y}\right)^m \left(X^y_{(i+1)} - X^y_{(i)}\right), \tag{2.27}$$

which can be compared to (2.24). Choosing the order $m = n_y$, both estimator are symmetric with respect to the FDH estimator $X^y_{(1)}$: with a minus sign to correct for the bias and with a plus sign for the order-m which provides an estimator that does not envelop all the data points.

2.2 A "conditional" bootstrap version of the bias corrected estimator

We show in this section that the bias corrected estimator (2.25) can also be derived by using a "conditional" bootstrap argument. Consider as previously $\left\{X^y_1, ..., X^y_{n_y}\right\}$ a random sample generated by the conditional distribution $F(x|y)$ and $X^y_{(1)} \leq ... \leq X^y_{(n_y)}$, the order statistics.

Let $\left\{X^{*y}_1, ..., X^{*y}_{n_y}\right\}$ be a bootstrap sample from the empirical conditional distribution of X given $Y \geq y$. Such a bootstrap sample can be obtained by sampling with replacement from $\left\{X^y_1, ..., X^y_{n_y}\right\}$. Let P_* be the associate resampling probability, which here is the uniform, and $X^{*y}_{(1)} \leq ... \leq X^{*y}_{(n_y)}$, the order statistics.

The FDH estimator of the efficient level of input is $X^y_{(1)}$ and in the bootstrap world, it is replaced by $X^{*y}_{(1)}$. Since $X^y_{(1)}$ is a biased estimator of $x^\partial(y)$, a correction for bias involves computing:

$$X^y_{(1)} - bias(X^y_{(1)}) = X^y_{(1)} - [E(X^y_{(1)}|Y \geq y) - x^\partial(y)]. \tag{2.28}$$

However, the right-hand side depends on unknown quantities hence it can not be computed in practice. In the bootstrap world, $bias(X^y_{(1)})$ can be estimated by

$$bias_*(X^y_{(1)}) = E_*(X^{*y}_{(1)}|Y \geq y) - X^y_{(1)}, \tag{2.29}$$

where E_* is computed with respect to the resampling probability P_*.

Now, a bias corrected estimator of $x^{\partial}(y)$ can be defined as:

$$\widehat{x^{\partial}}(y) = X^y_{(1)} - bias_*(X^y_{(1)}) = 2X^y_{(1)} - E_*(X^{*y}_{(1)}|Y \geq y). \tag{2.30}$$

The interesting thing with this bootstrap argument is that $E_*(X^{*y}_{(1)})$ can be computed analytically and does not require bootstrap resampling and Monte Carlo approximations. We have indeed:

$$P_*(X^{*y}_{(1)} < X^y_{(i+1)}|Y \geq y) = 1 - \left(1 - \frac{i}{n_y}\right)^{n_y} \tag{2.31}$$

for $i = 1, ..., n_y - 1$ and where in the bootstrap world, we condition also on the observed values of (X_i, Y_i). Consequently:

$$E_*(X^{*y}_{(1)}|Y \geq y) = \sum_{i=1}^{n_y} \left[\left(1 - \frac{i-1}{n_y}\right)^{n_y} - \left(1 - \frac{i}{n_y}\right)^{n_y}\right] X^y_{(i)}. \tag{2.32}$$

Therefore, replacing $E_*(X^{*y}_{(1)}|Y \geq y)$ in (2.30), we obtain the same bias corrected estimator as in the previous section. We called the method "conditional" bootstrap, since the original sample (from the "true world") is generated by a conditional distribution.

3 Multivariate Extension

In the full multivariate case, the frontier set can not be described by a function. But for a fixed point $(x, y) \in \Psi$, we can describe the frontier; *e.g.*, through $x^{\partial}(y)$, the efficient level of the inputs for this level of outputs y, as defined in (1.5). Since $x^{\partial}(y) = \theta(x, y)\, x$, the frontier, for any (x, y), can be characterized either by $x^{\partial}(y)$ or by $\theta(x, y)$. In the full multivariate setup, the latter is the easier to do inference.

The nonparametric FDH estimator,$\widehat{\theta}_{FDH}(x, y)$, was defined in (1.12). It is then easy to derive the following conditional distribution:

$$\begin{aligned}
P(\widehat{\theta}_{FDH}(x, y) \leq u \,|\, Y \geq y) &= 1 - P(\min_{i=1,...,n_y} \{\max_{j=1,...,p} \frac{X^j_i}{x^j}\} > u \,|\, Y \geq y) \\
&= 1 - [P(\max_{j=1,...,p} \frac{X^j_i}{x^j} > u \,|\, Y \geq y)]^{n_y} \\
&= 1 - [1 - P(\max_{j=1,...,p} \frac{X^j_i}{x^j} \leq u \,|\, Y \geq y)]^{n_y} \\
&= 1 - [1 - P(X \leq ux \,|\, Y \geq y)]^{n_y} \\
&= 1 - [1 - F(ux|y)]^{n_y}.
\end{aligned}$$

Following the same idea as in the univariate case, we first derive the expectation of $\widehat{\theta}_{FDH}(x,y)$ with respect to the conditional distribution above,

$$\begin{aligned} E(\widehat{\theta}_{FDH}(x,y)\,|\,Y \geq y) &= -\int_{\theta(x,y)}^{\infty} u\, d[1-F(ux|y)]^{n_y} \\ &= \theta(x,y) + \int_{\theta(x,y)}^{\infty} [1-F(ux|y)]^{n_y}\, du, \end{aligned}$$

obtaining:

$$\theta(x,y) = E(\widehat{\theta}_{FDH}(x,y)\,|\,Y \geq y) - \int_{\theta(x,y)}^{\infty} [1-F(ux|y)]^{n_y}\, du, \tag{3.33}$$

where the integral represents exactly the bias of the FDH estimator. Since the integral depends on the unknown $\theta(x,y)$, one can not attempt to define an unbiased estimator for the efficiency score. A bias corrected estimator can be defined by replacing the expectation $E(\widehat{\theta}_{FDH}(x,y)|Y \geq y)$ with the FDH estimator itself and applying the plug-in principle to estimate the integral. So we obtain our bias corrected estimator of the efficiency score:

$$\widehat{\theta}_{FDHbc}(x,y) = \widehat{\theta}_{FDH}(x,y) - \int_{\widehat{\theta}_{FDH}(x,y)}^{\infty} [1-\widehat{F}_n(ux|y)]^{n_y} du. \tag{3.34}$$

In the multivariate case, we loose the nice formula (2.25) based on order statistics, which allowed an easy computation of our estimator. Instead, we have to compute the FDH estimator $\widehat{\theta}_{FDH}(x,y)$ and to compute the integral in the right hand side of (3.34). But, since the integral is univariate, an appropriate numerical method can easly solve the problem, at no computational cost.

In order to establish asymptotic properties of $\widehat{\theta}_{FDHbc}(x,y)$ we need some assumptions on the production process as well as regularity conditions on the production possibilities set. Note that, for the case $p=1$, a similar result can be derived for the bias corrected estimator of the efficient level of input since for a given point (x,y), $\widehat{x}^{\partial}(y) = \widehat{\theta}_{FDHbc}(x,y)\, x$. The regularity assumptions are the same as those proposed in Park, Simar and Weiner (2000), where the asymptotic properties of the FDH estimator are given. We summarize them below.

ASSUMPTION A1. *Ψ is a free disposal set.*

ASSUMPTION A2. *The input requirement set $X(y)$ is compact.*

ASSUMPTION A3. *The sample observations are realizations of iid random variables on Ψ with joint distribution $F(x,y)$ which has a continuous, strictly positive density at the frontier.*

Theorem 3.1. *Under assumptions A1–A3, for any (x, y) in the interior of Ψ, $\theta_{FDHbc}(x, y)$ shares the same asymptotic properties as the FDH estimator.*

Proof. Park, Simar and Weiner (2000) established the rate of convergence and derived the asymptotic distribution for the FDH estimator of efficiency score. Under assumptions A1–A3 we have:

$$\mathcal{L}(n^{\frac{1}{p+q}}(\widehat{\theta}_{FDH}(x,y) - \theta(x,y))) \rightarrow Weibull(\mu_{x,y}, p+q) \quad as\ n \rightarrow \infty, \tag{3.35}$$

where $\mu_{x,y}$ is a constant defined in the Appendix of Park, Simar and Weiner (2000) that depends on the data generating process. As shown there, we have

$$P(X \leq (\theta(x,y) + h)x, Y \geq y) = \mu_{x,y}\, h^{p+q} + O(h^{p+q+1}),$$

so that, for small $h > 0$, $\mu_{x,y}$ is proportional to the probability of observing a firm dominating a point near the frontier below (x, y). Therefore, $\mu_{x,y}$ will be larger (smaller), if $F(x, y)$ provides more (less) mass in the neighborhood of the frontier which dominates the point (x, y).

For the bias corrected estimator we have:

$$\begin{aligned} n^{\frac{1}{p+q}}[\widehat{\theta}_{FDHbc}(x,y) - \theta(x,y)] &= n^{\frac{1}{p+q}}[\widehat{\theta}_{FDH}(x,y) - \theta(x,y)] \\ &- n^{\frac{1}{p+q}} \int_{\widehat{\theta}_{FDH}(x,y)}^{\infty} [1 - \widehat{F}_n(ux|y)]^{n_y}\, du. \end{aligned}$$

In order to derive the result, we need the last term of the previous expression to be $o_p(1)$ as $n \rightarrow \infty$. If (x, y) is the point with the largest output value from the sample, then $\widehat{F}_n(ux|y) = 1$, $n_y = 1$ and $\widehat{\theta}_{FDHbc}(x, y) = \widehat{\theta}_{FDH}(x, y) = 1$. Otherwise, we have $0 < \widehat{F}_n(ux|y) < 1$.

Under assumption A2 and using a mean value theorem, the integral from the preceding expression becomes:

$$\begin{aligned} &\int_{\widehat{\theta}_{FDH}(x,y)}^{\max_{i|Y_i \geq y}\{\max_{j=1,\ldots,p} \frac{X_i^j}{x^j}\}} [1 - \widehat{F}_n(ux|y)]^{n_y} du \\ &\qquad = \left(\max_{i|Y_i \geq y} \{ \max_{j=1,\ldots,p} \frac{X_i^j}{x^j} \} - \widehat{\theta}_{FDH}(x,y) \right) \exp[n_y \log(1 - \widehat{F}_n(\tilde{u}x|y))], \end{aligned}$$

where $\widehat{\theta}_{FDH}(x, y) \leq \tilde{u} \leq \max_{i|Y_i \geq y}\{\max_{j=1,\ldots,p} \frac{X_i^j}{x^j}\}$. Moreover, the right hand side term can be approximated by

$$\left(\max_{i|Y_i \geq y} \{ \max_{j=1,\ldots,p} \frac{X_i^j}{x^j} \} - \widehat{\theta}_{FDH}(x,y) \right) \exp[-nS_Y(y)B], \tag{3.36}$$

since $n_y \asymp nS_Y(y)$, where $S_Y(\cdot)$ is the q-variate survivor function of Y and B is a positive quantity. As $n \to \infty$, $\left(\max_{i|Y_i \geq y}\{\max_{j=1,\ldots,p} \frac{X_i^j}{x^j}\} - \widehat{\theta}_{FDH}(x,y)\right)$ is bounded and $\exp[-nS_Y(y)B] = o_p(1)$.

Therefore, we have:

$$n^{\frac{1}{p+q}}[\widehat{\theta}_{FDHbc}(x,y) - \theta(x,y)] = n^{\frac{1}{p+q}}[\widehat{\theta}_{FDH}(x,y) - \theta(x,y)] + o_p(1) \tag{3.37}$$

as $n \to \infty$ and the conclusion follows. □

In other words, this theorem tells us that our bias corrected estimator is asymptotically equivalent to the FDH estimator but for finite sample we hope that the bias correction will improve its performance without increasing the mean squared error, as for the Cooke's estimator. The Monte-Carlo experiment of the next section will confirm this fact.

4 Monte Carlo Experiments

We conducted several Monte Carlo experiments in order to analyze the performance of our bias corrected estimator (input oriented case), for technologies with $p = q = 1$ and $p = q = 2$. We simulated our data according to the models considered in Kneip, Simar and Wilson (2003).

- Model 1 ($p = q = 1$)

  We generated the efficient level of input x_{eff} uniformly on [10,20], then perturbed it by the inefficiency $e^{0.2|\varepsilon|}$, with $\varepsilon \sim N(0,2)$ and $\varepsilon \sim N(0,4)$, being independent on the input. The input observation is computed as $x = x_{eff}e^{0.2|\varepsilon|}$. The output level is $y = x_{eff}^{0.8} = x^{0.8}e^{-0.16|\varepsilon|}$, which defines the true DGP. The fixed point considered is $(x_0, y_0) = (20.69, 7.5)$ with an input efficiency score $\theta(x_0, y_0) = 0.6$. We simulated samples of size $n = 50, 100, 200, 400, 800$ and we estimated the efficient input by the average over 200 Monte Carlo simulations of the FDH estimator and the bias corrected estimator FDHbc. Table 1 and Table 2 compare the bias and the mean squared error (MSE) of the FDH estimator and the FDHbc estimator for the cases $\varepsilon \sim N(0,2)$ and $\varepsilon \sim N(0,4)$ respectively. The line "RATIO" represents the ratios FDHbc/FDH in terms of bias and mean squared error. The numerical results show that the bias corrected estimator improves the FDH input efficiency estimator, reducing its bias and the mean squared error. The bias of the FDH estimator is reduced with around 25–30% and the reduction of the mean squared error is of the order of 50–60% depending on the sample size and the variance of the inefficiency term. So a significant improvement relative to the FDH estimator. We see also, as expected, that the effect decreases when the sample size increases.

- Model 2 ($p = q = 2$)

  In this multivariate setup, we first draw the efficient level of input, by generating $x_{1eff}, x_{2eff} \sim Uniform[10, 20]$ and $\omega \sim Uniform[\frac{\pi}{18}, \frac{8\pi}{18}]$ for computing the output components:

$$y_1 = x_{1eff}^{0.4} x_{2eff}^{0.4} \cdot \cos\omega$$
$$y_2 = x_{1eff}^{0.4} x_{2eff}^{0.4} \cdot \sin\omega.$$

  The input components are computed as $x_1 = x_{1eff}^{0.2|\varepsilon|}$, $x_2 = x_{2eff}^{0.2|\varepsilon|}$, where $\varepsilon \sim N(0, 2)$ and $\varepsilon \sim N(0, 4)$ as for Model 1. The input efficiency score is estimated for the fixed point $x_0 = (20.69, 20.69)$, $y_0 = (5.59, 5.59)$, with a true efficiency score $\theta(x_0, y_0) = 0.6$ as before. As shown by Table 3 and Table 4, due to the curse of dimensionality, the improvement in terms of bias reduction is less evident in this case. However, we still achieve at least 10% bias reduction and 20–40% MSE reduction, at an insignificant computational cost.

5 Conclusions

We proposed in this paper a simple bias corrected estimator for the efficient frontier, which improves the FDH estimator in terms of the bias. Our presentation focuses on input orientation, but the results are easy to derive in the output oriented case (see the appendix). In one input, multi-output case, we provide a formula based on order statistics for computing the estimator. We also define a "conditional" bootstrap which allows deriving the same formula for the estimator and hence does not require bootstrap replications for computational purposes. Moreover, using the ideas proposed by Cazals, Florens and Simar (2002) and developed by Daraio and Simar (2003), a full multivariate extension is also presented. The computation is very fast in both univariate and multivariate situations. Our bias corrected estimator shares the asymptotic properties of the FDH estimator as derived in Park, Simar and Weiner (2000). So the correction will be suitable only for finite sample situations. As the Monte-Carlo experiments show, the correction is more evident and significant for small samples and also when the inefficiency term has a larger variance. Note that for the unit with the largest output from the sample, the FDHbc estimator coincides with the FDH.

Appendix

The Output Oriented Approach

We first present the problem of estimating a production function, so that here, the production process is defined by the joint distribution of the random vector (X, Y) on $\mathbb{R}_+^p \times \mathbb{R}_+$, with support Ψ:

$$F(x, y) = P(X \leq x, Y \leq y)$$

and consider the conditional distribution of Y given $X \leq x$

$$F(y|x) = P(Y \leq y | X \leq x),$$

The output efficient frontier, can be defined as:

$$y^{\partial}(x) = \sup \{y | F(y|x) < 1\} \tag{A.1}$$

and represents the upper boundary of the support of $F(y|x)$.

Now consider the observations $\left\{Y_1^x, ..., Y_{n_x}^x\right\}$, where n_x is the number of observations in $\mathcal{X}$ with $X \leq x$. It can be viewed as a random sample generated by the conditional distribution $F(y|x)$ and let $Y_{(1)}^x \leq ... \leq Y_{(n_x)}^x$ be the order statistics. The random variable $Y_{(n_x)}^x$, is the FDH estimator of the efficient level of output and has conditional distribution

$$P\left(Y_{(n_x)}^x \leq y \,|\, X \leq x\right) = [F(y|x)]^{n_x} \tag{A.2}$$

and conditional mean

$$E\left(Y_{(n_x)}^x \,|\, X \leq x\right) = \int_0^{y^{\partial}(x)} y \, d([F(y|x)]^{n_x}) = y^{\partial}(x) - \int_0^{y^{\partial}(x)} [F(y|x)]^{n_x} \, dy,$$

because $F(y|x) = 1$ for any $y \geq y^{\partial}(x)$. Therefore, we have

$$y^{\partial}(x) = E\left(Y_{(n_x)}^x \,|\, X \leq x\right) + \int_0^{y^{\partial}(x)} [F(y|x)]^{n_x} \, dy. \tag{A.3}$$

Replacing the unknown expectation and $y^{\partial}(x)$ by $Y_{(n_x)}^x$ and the distribution function by its empirical correspondent, we obtain the following bias corrected estimator of the output efficient frontier:

$$\widehat{y_{n_x}^{\partial}}(x) = Y_{(n_x)}^y + \int_0^{Y_{(n_x)}^x} [\widehat{F}_n(y|x)]^{n_x} \, dy.$$

Since the empirical conditional distribution function is defined as:

$$\widehat{F}_n\left(y|x\right) = \begin{cases} 0 & , \text{ if } y < Y^x_{(1)} \\ \frac{i}{n_x} & , \text{ if } Y^x_{(i)} \leq y < Y^x_{(i+1)}, \text{ for } 1 \leq i \leq n_x - 1 \\ 1 & , \text{ if } y \geq Y^x_{(n_x)}, \end{cases}$$

we obtain

$$\widehat{y^{\partial}_{n_x}}(x) = Y^x_{(n_x)} + \sum_{i=1}^{n_x-1} \left(\frac{i}{n_x}\right)^{n_x} \left(Y^x_{(i+1)} - Y^x_{(i)}\right), \tag{A.4}$$

or equivalently,

$$\widehat{y^{\partial}_{n_x}}(x) = 2Y^y_{(n_x)} - \sum_{i=1}^{n_x-1} \left[\left(1 - \frac{i}{n_x}\right)^{n_x} - \left(1 - \frac{i+1}{n_x}\right)^{n_x}\right] Y^x_{(n_x-i)}. \tag{A.5}$$

In the full multivariate case, we focus on the Debreu-Farrell output efficiency score, which can be characterized as

$$\lambda(x,y) = \sup\{\lambda | S(\lambda y|x) > 0\}, \tag{A.6}$$

where $S(y|x) = P(Y \geq y|X \leq x)$. Here, for each level of input x, the efficient level of output, characterizing the frontier set is given by:

$$y^{\partial}(x) = \lambda(x,y)\, y \tag{A.7}$$

A nonparametric estimator for the output efficiency score $\lambda(x,y)$ is

$$\widehat{\lambda}(x,y) = \sup\{\lambda | \widehat{S}_n(\lambda y|x) > 0\}, \tag{A.8}$$

where $\widehat{S}_n(y|x)$ is the empirical version of $S(y|x)$, defined by:

$$\widehat{S}_n(y|x) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \leq x, Y_i \geq y)}{\sum_{i=1}^n \mathbb{1}(X_i \leq x)}.$$

As in the input oriented case, $\widehat{\lambda}(x,y)$ defined in (A.8) coincides to the FDH estimator of output efficiency score

$$\widehat{\lambda}_{FDH}(x,y) = \max_{i|X_i \leq x} \{\min_{j=1,\ldots,q} \frac{Y^j_i}{y^j}\}. \tag{A.9}$$

Adapting formula (3.34) accordingly, we obtain the bias corrected estimator for the output efficiency score of the point (x,y):

$$\widehat{\lambda}_{FDHbc}(x,y) = \widehat{\lambda}_{FDH}(x,y) + \int_0^{\widehat{\lambda}_{FDH}(x,y)} [1 - \widehat{S}_n(uy|x)]^{n_x}\, du, \tag{A.10}$$

Here the bias corrected estimator shares the asymptotic properties of the FDH estimator established in Park, Simar and Weiner (2000), namely,

$$\mathcal{L}(n^{\frac{1}{p+q}}(\lambda(x,y) - \widehat{\lambda}_{FDH}(x,y))) \rightarrow Weibull(\nu_{x,y}, p+q) \quad as\ n \rightarrow \infty, \tag{A.11}$$

where $\nu_{x,y}$ is such that:

$$P(X \leq x, Y \geq (\lambda(x,y) - h)y) = \nu_{x,y}\, h^{p+q} + O(h^{p+q+1}),$$

so that, for small $h > 0$, $\nu_{x,y}$ is proportional to the probability of observing a firm dominating a point near the frontier above (x, y).

References

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

[2] Charnes, A., Cooper, W.W., Rhodes, E. (1978), Measuring the inefficiency of decision making units, *European Journal of Operational Research* 2(6), 429-444.

[3] Cooke, P.J. (1979), Statistical inference for bounds of random variables, *Biometrika* 66, 367-374.

[4] Daraio, C., Simar, L., (2003), Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach, Discussion Paper #0313, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, to appear in *Journal of Productivity Analysis*.

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

[6] Deprins, D., Simar, L., Tulkens, H. (1984), Measuring Labor Inefficiency in Post Offices, in *The Performance of the Public Enterprises: Concepts and Measurements*, eds. Marchand, M., Pestieau, P. and Tulkens H., Amsterdam, North Holland, 243-267.

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

[8] Kneip, A., Simar, L., Wilson, P.W., Asymptotics for DEA estimators in non-parametric frontier models, Discussion Paper #0317, Institut de statistique, UCL, Louvain-la-Neuve, Belgium.

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

[10] Robson, D.S., Whitlock, J.H. (1964), Estimation of a truncation point, *Biometrika* 51, 33-39.

n	Estimator	BIAS	MSE
50	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.3595	1.8508
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.9298	0.8681
	RATIO	0.6839	0.4691
100	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.0726	1.1521
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.7892	0.6251
	RATIO	0.7358	0.5426
200	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.7127	0.5088
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.5093	0.2606
	RATIO	0.7147	0.5122
400	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.5375	0.2893
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.4148	0.1726
	RATIO	0.7718	0.5966
800	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.3581	0.1284
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.2615	0.0686
	RATIO	0.7303	0.5346

Table 1: Results from Monte Carlo experiments with data simulated according to Model 1: $y = x_{eff}^{0.8} = x^{0.8}e^{-0.16|\varepsilon|}$, with $\varepsilon \sim N(0, 2)$.

n	Estimator	BIAS	MSE
50	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.9730	3.8989
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	1.3712	1.8886
	RATIO	0.6950	0.4844
100	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.5739	2.4805
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	1.1797	1.3963
	RATIO	0.7496	0.5629
200	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.0395	1.0826
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.7583	0.5776
	RATIO	0.7295	0.5336
400	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.7617	0.5810
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.5808	0.3384
	RATIO	0.7625	0.5824
800	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.5183	0.2690
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.3825	0.1468
	RATIO	0.7380	0.5459

Table 2: Results from Monte Carlo experiments with data simulated according to Model 1: $y = x_{eff}^{0.8} = x^{0.8}e^{-0.16|\varepsilon|}$, with $\varepsilon \sim N(0, 4)$.

n	Estimator	BIAS	MSE
50	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.3257	0.1061
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2802	0.0786
	RATIO	0.8602	0.7401
100	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2706	0.0732
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2409	0.0580
	RATIO	0.8901	0.7925
200	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2278	0.0519
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2051	0.0421
	RATIO	0.9004	0.8108
400	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.1926	0.0371
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1732	0.0300
	RATIO	0.8996	0.8095
800	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.1640	0.0269
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1473	0.0217
	RATIO	0.8982	0.8069

Table 3: Results from Monte Carlo experiments with data simulated from Model 2 with $\varepsilon \sim N(0, 2)$.

n	Estimator	BIAS	MSE
50	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.4108	0.1689
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.3098	0.0962
	RATIO	0.7542	0.5694
100	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.3285	0.1079
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2817	0.0794
	RATIO	0.8576	0.7357
200	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2654	0.0704
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2345	0.0550
	RATIO	0.8836	0.7809
400	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2211	0.0489
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1971	0.0389
	RATIO	0.8915	0.7950
800	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.1891	0.0358
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1688	0.0285
	RATIO	0.8927	0.7970

Table 4: Results from Monte Carlo experiments with data simulated from Model 2 with $\varepsilon \sim N(0, 4)$.