Econometric Theory, **25**, 2009, 1289–1318. Printed in the United States of America.
doi:10.1017/S0266466609090513

A BIAS-CORRECTED NONPARAMETRIC ENVELOPMENT ESTIMATOR OF FRONTIERS

LUIZA BĂDIN
Bucharest Academy of Economic Studies
and
"Gheorghe Mihoc–Caius Iacob" Institute of
Mathematical Statistics and Applied Mathematics
Romanian Academy

LÉOPOLD SIMAR
Institut de Statistique
Université Catholique de Louvain

In efficiency analysis, the production frontier is defined as the set of the most efficient alternatives among all possible combinations in the input-output space. The nonparametric envelopment estimators rely on the assumption that all the observations fall on the same side of the frontier. The free disposal hull (FDH) estimator of the attainable set is the smallest free disposal set covering all the observations. By construction, the FDH estimator is an inward-biased estimator of the frontier.

The univariate extreme values representation of the FDH allows us to derive a bias-corrected estimator for the frontier. The presentation is based on a probabilistic formulation where the input-output pairs are realizations of independent random variables drawn from a joint distribution whose support is the production 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. The method is also illustrated with a real data example. A comparison with the parametric stochastic frontier indicates that the parametric approach can easily fail under wrong specification of the model.

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

Luiza Bădin acknowledges research support from the EC project "Inco-Copernicus," IC15-CT98-1007. Léopold Simar also gratefully acknowledges research support from the "Interuniversity Attraction Pole" (IAP research network P5/24 and P6/03) of the Belgian government (Belgian Science Policy). Address correspondence to Léopold Simar, Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays, 20 B-1348 Louvain-la-Neuve, Belgium; e-mail: simar@stat.ucl.ac.be.

© 2009 Cambridge University Press 0266-4666/09 $15.00

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

as the set of the most efficient alternatives among all possible combinations in the input-output space. It could be characterized by the maximal achievable amount of outputs for a given level of the inputs or, conversely, by the minimum level of inputs allowing production of a given level of the outputs.

The basic economic idea 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 that yield at least y:

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

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

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

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 $\theta(x, y) = 1$, whereas a value $\theta(x, y) < 1$ indicates the radial reduction in inputs that producer (x, y) should perform 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:

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

Because Ψ is generally unknown, it has to be estimated from an independent and identically distributed (i.i.d.) sample of n firms $\mathcal{X} = \{(X_i, Y_i),\ i = 1, \ldots, n\}$. 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 $\tilde{y} \leq y$,

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

where the inequalities between vectors are defined element by element. The DEA estimator of Ψ is

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

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:

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

The efficiency of the arbitrary point (x,y) is then measured with respect to the boundary of the FDH 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 (2005), 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 $F(x|y) = P(X \le x | Y \ge y)$, 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:

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

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

where $\mathbb{I}(\cdot)$ 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 (2005).

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 cumulative distribution function (c.d.f.) $F_X(x)$ 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 $\mathcal{X} = \{(X_i, 1), i = 1, \ldots, n\}$ is a random sample of producers and $X_{(1)} \leq \cdots \leq X_{(n)}$ 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 an 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.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

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

$$T_n^{(k)} = \sum_{i=0}^{k} (-1)^i \binom{k+1}{i+1} X_{(n-i)}, \qquad (0 < k < n) \tag{1.17}$$

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:

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

which, for large n, is asymptotically equivalent to

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

In this paper we propose nonparametric estimators for the efficient frontier and for the efficiency scores that attempt to improve the FDH estimators in terms of bias using Cooke's ideas, without affecting the mean squared error. We denote our bias-corrected FDH estimator FDHbc. We will first introduce the idea by following the presentation of Cazals et al. (2002), where the minimum input frontier function is estimated for one input and several outputs; then using the Daraio and Simar (2005) extension, we generalize to the full multivariate case.

Deriving a bias-corrected nonparametric estimator of efficiency scores, without computational burden, is useful because, to the best of our knowledge, there are no real alternatives: (a) Park, Simar, and Weiner (2000) derive the asymptotic bias and variance of the FDH estimator that depend on unknown characteristics of the data generating process (DGP), but their Monte Carlo study indicates that the noise introduced by estimating these characteristics reduces the quality of the inference unless the sample size is very big (see the subsequent discussion of the curse of dimensionality); (b) Jeong and Simar (2006) have shown that the bootstrap based on subsampling is consistent for the FDH estimator, but the quality of the inference here relies on the size of the subsamples, and no practical rules have been proposed so far to select the appropriate subsample size.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

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 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 is the bias of the FDH efficiency estimator clearly reduced but also the mean squared error. We also consider several examples of simulated and real data for which we compare the FDH and FDHbc efficiency scores with those derived from estimating a parametric stochastic frontier. Section 5 concludes and an Appendix follows.

2. BIAS CORRECTION FOR THE FDH EFFICIENT FRONTIER

2.1. Estimation of a Minimum Input Function

To present the basic ideas of our bias-corrected estimator, we first analyze, as in Cazals et al. (2002), the estimation of a "input" 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

$$x^{\partial}(y) = \inf\{x|F(x|y) > 0\} = \inf\{x|S(x|y) < 1\} \tag{2.1}$$

and represents the lower boundary of the support of $F(x|y)$. As shown in Cazals et al. (2002, Thm. 2.1), this is a monotone function nondecreasing in y that 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 $\{X_1^y, \ldots, X_{n_y}^y\}$, where n_y is the number of such observations.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

Conditionally on $Y \geq y$, this can be considered a random sample of size n_y generated by the conditional distribution $F(x|y)$. Denote $X^y_{(1)} \leq \cdots \leq X^y_{(n_y)}$ the corresponding order statistics. The random variable $X^y_{(1)}$ is, in this univariate input case, the FDH estimator of the efficient level of input (corresponding to $\widehat{x}^{\partial}_{FDH}(y)$ as defined in (1.9)). It has the conditional distribution

$$P\left(X^y_{(1)} \leq x|Y \geq y\right) = 1-[1-F(x|y)]^{n_y} = 1-[S(x|y)]^{n_y} \tag{2.2}$$

and conditional mean

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

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

$$x^{\partial}(y) = \mathrm{E}\left(X^y_{(1)}|Y \geq y\right) - \int_{x^{\partial}(y)}^{\infty} [S(x|y)]^{n_y}\, dx. \tag{2.3}$$

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

$$\widehat{x^{\partial}_{n_y}}(y) = X^y_{(1)} - \int_{X^y_{(1)}}^{\infty} [\widehat{S}(x|y)]^{n_y}\, dx = X^y_{(1)} - \int_{X^y_{(1)}}^{X^y_{(n_y)}} [\widehat{S}(x|y)]^{n_y}\, dx,$$

where the empirical conditional survivor function is defined as

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

We thus obtain a new estimator of the minimum input function:

$$\widehat{x^{\partial}_{n_y}}(y) = X^y_{(1)} - \sum_{i=1}^{n_y-1} \left(1-\frac{i}{n_y}\right)^{n_y} \left(X^y_{(i+1)} - X^y_{(i)}\right), \tag{2.4}$$

where the first term is the FDH estimator of the frontier and the second term can be interpreted as a bias correction of the FDH estimator. The estimator can also be written as

$$\widehat{x^{\partial}_{n_y}}(y) = 2X^y_{(1)} - \sum_{i=1}^{n_y} \left[\left(1-\frac{i-1}{n_y}\right)^{n_y} - \left(1-\frac{i}{n_y}\right)^{n_y}\right] X^y_{(i)} \tag{2.5}$$

and will be called the FDHbc estimator of the minimum input function.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

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. For large n_y, the estimator defined in (2.5) is asymptotically equivalent to

$$\widehat{x^{\partial}_{n_y}}(y) = 2X^y_{(1)} - (e-1)\sum_{i=1}^{n_y} e^{-i} X^y_{(i)} \tag{2.6}$$

and coincides with the estimator proposed by Cooke (1979) for the case of univariate i.i.d. random variables.

Remark 2.2. Surprisingly, our estimator is a symmetric version of the order-m minimum input function proposed by Cazals et al. (2002), relative to the FDH estimator. Indeed, the order-m estimator can be written as

$$\widehat{\varphi}_{m,n}(y) = X^y_{(1)} + \sum_{i=1}^{n_y-1} \left(1 - \frac{i}{n_y}\right)^m \left(X^y_{(i+1)} - X^y_{(i)}\right), \tag{2.7}$$

which can be compared to (2.4). Choosing the order $m = n_y$, both estimators 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 that 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 (2.5) can also be derived by using a conditional bootstrap argument. Consider as previously $\{X^y_1, \ldots, X^y_{n_y}\}$, a random sample generated by the conditional distribution $F(x|y)$, and $X^y_{(1)} \leq \cdots \leq X^y_{(n_y)}$, the associated order statistics.

Let $\{X^{*y}_1, \ldots, X^{*y}_{n_y}\}$ 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 $\{X^y_1, \ldots, X^y_{n_y}\}$. Let P_* be the associate resampling probability, which here is the uniform, and $X^{*y}_{(1)} \leq \cdots \leq X^{*y}_{(n_y)}$, the order statistics. We call the method "conditional bootstrap" because the original sample (from the "true world") is generated by a conditional distribution.

PROPOSITION 2.1. *The conditional bootstrap bias-corrected FDH estimator of the minimum input function coincides with the FDHbc estimator defined in (2.5). Consequently, the conditional bootstrap estimator of the bias admits an exact formula, and no simulations are required.*

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

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

$$X^y_{(1)} - \text{bias}(X^y_{(1)}) = X^y_{(1)} - \left[\text{E}(X^y_{(1)}|Y \geq y) - x^\partial(y)\right]. \tag{2.8}$$

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

$$\text{bias}_*(X^y_{(1)}) = \text{E}_*\left(X^{*y}_{(1)}|Y \geq y\right) - X^y_{(1)}, \tag{2.9}$$

where E_* is computed with respect to the resampling probability P_*. Now, a bias-corrected estimator of $x^\partial(y)$, based on the conditional bootstrap, can be defined as

$$\widehat{x^\partial}(y) = X^y_{(1)} - \text{bias}_*\left(X^y_{(1)}\right) = 2X^y_{(1)} - \text{E}_*\left(X^{*y}_{(1)}|Y \geq y\right). \tag{2.10}$$

The interesting feature of this bootstrap argument is that $\text{E}_*(X^{*y}_{(1)})$ can be computed analytically and does not require bootstrap resampling and Monte Carlo approximations. We have indeed

$$P_*\left(X^{*y}_{(1)} < X^y_{(i+1)}|Y \geq y\right) = 1 - \left(1 - \frac{i}{n_y}\right)^{n_y} \tag{2.11}$$

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

$$\text{E}_*\left(X^{*y}_{(1)}|Y \geq y\right) = \sum_{i=1}^{n_y} \left[\left(1 - \frac{i-1}{n_y}\right)^{n_y} - \left(1 - \frac{i}{n_y}\right)^{n_y}\right] X^y_{(i)}. \tag{2.12}$$

Therefore, replacing $\text{E}_*(X^{*y}_{(1)}|Y \geq y)$ in (2.10), we obtain the FDHbc estimator defined in (2.5) and an exact formula for the conditional bootstrap estimator of the bias:

$$\text{bias}_*\left(X^y_{(1)}\right) = \sum_{i=1}^{n_y-1} \left(1 - \frac{i}{n_y}\right)^{n_y} \left(X^y_{(i+1)} - X^y_{(i)}\right). \tag{2.13}$$

■

Remark 2.3. The conditional bootstrap based on uniform resampling gives formula (2.5). Adapting previous results from Loh (1984), it can be shown that using other resampling schemes, different than the uniform, one could derive other bias-corrected estimators with an exact formula for the bootstrap bias, provided that the bootstrap mean $\text{E}_*(X^{*y}_{(1)}|Y \geq y)$ can be derived analytically.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

3. MULTIVARIATE EXTENSION

In the full multivariate case, the frontier set cannot 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). Because $x^{\partial}(y) = \theta(x, y)x$, 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 for doing 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:

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

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

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

obtaining

$$\theta(x, y) = \mathrm{E}(\widehat{\theta}_{FDH}(x, y) \,|\, Y \ge y) - \int_{\theta(x,y)}^{\infty} [1 - F(ux|y)]^{n_y} du, \tag{3.1}$$

where the integral represents exactly the bias of the FDH estimator. Because the integral depends on the unknown $\theta(x, y)$, one cannot attempt to define an unbiased estimator for the efficiency score. A bias-corrected estimator can be defined by replacing the expectation $\mathrm{E}\left(\widehat{\theta}_{FDH}(x, y) | Y \ge y\right)$ 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:

$$\widehat{\theta}_{FDHbc}(x, y) = \widehat{\theta}_{FDH}(x, y) - \int_{\widehat{\theta}_{FDH}(x,y)}^{\infty} [1 - \widehat{F}_n(ux|y)]^{n_y} du. \tag{3.2}$$

In the multivariate case, we lose the nice formula (2.5) based on order statistics, which allowed an easy computation of our estimator. Instead, we have to compute

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

the FDH estimator $\widehat{\theta}_{FDH}(x, y)$ and compute the integral on the right-hand side of (3.2). But, because the integral is univariate, an appropriate numerical method can easily solve the problem, at no computational cost.

To establish asymptotic properties of $\widehat{\theta}_{FDHbc}(x, y)$ we need some assumptions on the production process and also 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 because 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 et al. (2000), where the asymptotic properties of the FDH estimator are given. We summarize them here.

Assumption 1. Ψ is a free disposal set.

Assumption 2. The input requirement set $X(y)$ is compact.

Assumption 3. The sample observations are realizations of i.i.d. random variables on Ψ with joint distribution $F(x, y)$ that has a continuous, strictly positive density at the frontier.

THEOREM 3.1. *Under Assumptions 1–3, for any* (x, y) *in the interior of* Ψ, $\widehat{\theta}_{FDHbc}(x, y)$ *shares the same asymptotic properties as the FDH estimator.*

Proof. Park et al. (2000) established the rate of convergence and derived the asymptotic distribution for the FDH estimator of efficiency score. Under Assumptions 1–3 we have

$$\mathcal{L}(n^{1/(p+q)}(\widehat{\theta}_{FDH}(x, y) - \theta(x, y))) \to Weibull(\mu_{x,y}, p+q) \quad \text{as } n \to \infty, \qquad \textbf{(3.3)}$$

where $\mu_{x,y}$ is a constant defined in the Appendix of Park et al. (2000) that depends on the DGP. 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 that dominates the point (x, y).

For the bias-corrected estimator we have

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

From Park et al. (2000) we know that $\widehat{\theta}_{FDH}(x, y) - \theta(x, y) = O_p(n^{-1/(p+q)})$. Therefore, to derive the result, we need the last term of the previous expression to be $o_p(1)$ as $n \to \infty$.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

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 2 and using a mean value theorem, the last term of the preceding expression becomes

$$n^{1/(p+q)} \int_{\widehat{\theta}_{FDH}(x,y)}^{\max_{i|Y_i \geq y} \left\{ \max_{j=1,\ldots,p} (X_i^j / x^j) \right\}} [1 - \widehat{F}_n(ux|y)]^{n_y} du$$

$$= n^{1/(p+q)} \left(\max_{i|Y_i \geq y} \left\{ \max_{j=1,\ldots,p} \frac{X_i^j}{x^j} \right\} - \widehat{\theta}_{FDH}(x, y) \right)$$

$$\times \exp\left[n_y \log(1 - \widehat{F}_n(\tilde{u}x|y)) \right],$$

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

$$n^{1/(p+q)} \left(\max_{i|Y_i \geq y} \left\{ \max_{j=1,\ldots,p} \frac{X_i^j}{x^j} \right\} - \widehat{\theta}_{FDH}(x, y) \right) \exp[-n S_Y(y) B], \tag{3.4}$$

because $n_y \asymp n S_Y(y)$, where $S_Y(\cdot)$ is the q-variate survivor function of Y and B is a positive quantity. As $n \to \infty$, $\left(\max_{i|Y_i \geq y} \left\{ \max_{j=1,\ldots,p} (X_i^j / x^j) \right\} - \widehat{\theta}_{FDH}(x, y) \right)$ is bounded because $X(y)$ is compact by Assumption 2, $n^{1/(p+q)}$ is dominated by $\exp[-n S_Y(y) B]$, and, consequently, $n^{1/(p+q)} \exp[-n S_Y(y) B] = o_p(1)$.

Therefore we have

$$n^{1/(p+q)} \left(\max_{i|Y_i \geq y} \left\{ \max_{j=1,\ldots,p} \frac{X_i^j}{x^j} \right\} - \widehat{\theta}_{FDH}(x, y) \right) \exp[-n S_Y(y) B] = o_p(1), \tag{3.5}$$

which leads to

$$\widehat{\theta}_{FDHbc}(x, y) - \theta(x, y) = O_p(n^{-1/(p+q)}), \tag{3.6}$$

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 samples we hope that the bias correction will improve its performance without increasing the mean squared error. The Monte Carlo experiments presented in Section 4 will confirm this fact.

4. NUMERICAL RESULTS

In this section we investigate the finite-sample performance of the FDHbc estimator in several Monte Carlo experiments. We also provide empirical comparisons of our method with FDH and with the maximum likelihood estimator (MLE) of efficiency, derived using a parametric stochastic frontier model. The comparison is made using both real and simulated data sets.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

4.1. Monte Carlo Experiments

We conducted several Monte Carlo experiments to analyze the performance of our bias-corrected estimator on convex and nonconvex technologies. Our data are simulated according to several models considered in Kneip, Simar, and Wilson (2008) and Simar (2007).

- Model 1: Convex technology with input orientation and $p = q = 1$. We generated the efficient level of input x_{eff} uniformly on [10, 20] and then perturbed it by the inefficiency $e^{0.2|\varepsilon|}$, with $\varepsilon \sim N(0, 2)$ in a first scenario and then with $\varepsilon \sim N(0, 4)$, the ε's 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 $n = 25, 50, 100, 200, 400, 800$, 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 compares the bias and the mean squared error (MSE) of the FDH estimator and the FDHbc estimator for the cases $\varepsilon \sim N(0, 2)$ (left) and $\varepsilon \sim N(0, 4)$ (right), respectively. The line "Ratio" represents the ratio 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 around 25–30%, and the reduction of the mean squared error is of the order of 50% depending on the sample size and the variance of the inefficiency term. So a significant improvement exists relative to the FDH estimator. We also notice, as expected, that the effect decreases when the sample size increases.
- Model 2: Convex technology with input orientation and $p = q = 2$. In this multivariate setup, we first draw the efficient level of input by generating $x_{1eff}, x_{2eff} \sim Uniform[10, 20]$ and $\omega \sim Uniform[(\pi/18), (8\pi/18)]$ 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 $\theta(x_0, y_0) = 0.6$. In this case, we simulated samples of size $n = 50, 100, 200, 400, 800$. As shown by Table 2, because of 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.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

TABLE 1. Results from Monte Carlo experiments with data simulated according to Model 1: $y = x_{eff}^{0.8} = x^{0.8}e^{-0.16|\varepsilon|}$

Panel A. $\varepsilon \sim N(0,2)$				Panel B. $\varepsilon \sim N(0,4)$			
n	Estimator	Bias	MSE	n	Estimator	Bias	MSE
25	$\widehat{x^{\partial}}_{FDH}(y_0)$	2.1432	4.6008	25	$\widehat{x^{\partial}}_{FDH}(y_0)$	3.1167	9.7297
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	1.5816	2.5111		$\widehat{x^{\partial}}_{FDHbc}(y_0)$	2.2862	5.2465
	Ratio	0.7380	0.5458		Ratio	0.7335	0.5392
50	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.5250	2.3289	50	$\widehat{x^{\partial}}_{FDH}(y_0)$	2.2351	5.0028
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	1.1252	1.2704		$\widehat{x^{\partial}}_{FDHbc}(y_0)$	1.6627	2.7739
	Ratio	0.7378	0.5455		Ratio	0.7439	0.5545
100	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.0248	1.0517	100	$\widehat{x^{\partial}}_{FDH}(y_0)$	1.4537	2.1161
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.7639	0.5854		$\widehat{x^{\partial}}_{FDHbc}(y_0)$	1.0659	1.1399
	Ratio	0.7454	0.5566		Ratio	0.7332	0.5387
200	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.6493	0.4223	200	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.9318	0.8696
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.4655	0.2177		$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.6574	0.4340
	Ratio	0.7169	0.5154		Ratio	0.7055	0.4991
400	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.5117	0.2622	400	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.7297	0.5331
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.3865	0.1499		$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.5523	0.3060
	Ratio	0.7555	0.5718		Ratio	0.7569	0.5740
800	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.3476	0.1210	800	$\widehat{x^{\partial}}_{FDH}(y_0)$	0.4842	0.2347
	$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.2601	0.0679		$\widehat{x^{\partial}}_{FDHbc}(y_0)$	0.3550	0.1264
	Ratio	0.7483	0.5610		Ratio	0.7331	0.5386

- Model 3: Nonconvex technology with output orientation and $p = q = 1$. Next we consider a nonconvex frontier model. Because we can no longer rely on the convexity assumption, the single alternative for nonparametric estimation of efficiency is the FDH estimator. We generated samples of size $n = 25, 50, 100, 200, 400, 800$, according to the following theoretical model from Simar (2007):

$$y = 10\frac{\exp(-3+10x)}{1+\exp(-3+10x)}\exp(-u), \tag{4.1}$$

with $X \sim Beta(1,2)$ and $U \sim Exp(\frac{1}{3})$. We have $\mathrm{E}(X) = \frac{1}{3}$ and $\mathrm{E}(U) = \frac{1}{3}$. The efficient level of output, $\widehat{y^{\partial}}_{FDH}(x_0)$ and $\widehat{y^{\partial}}_{FDHbc}(x_0)$, is estimated for the fixed point $(x_0, y_0) = (0.6, 4)$. The results in Table 3 are average results over 200 Monte Carlo simulations. In this case also, we notice that our

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

TABLE 2. Results from Monte Carlo experiments with data simulated from Model 2: The evaluated point is $x_0 = (20.69, 20.69)$, $y_0 = (5.59, 5.59)$

Panel A. $\varepsilon \sim N(0,2)$				Panel B. $\varepsilon \sim N(0,4)$			
n	Estimator	Bias	MSE	n	Estimator	Bias	MSE
50	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.3257	0.1061	50	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.4108	0.1689
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2802	0.0786		$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.3098	0.0962
	Ratio	0.8602	0.7401		Ratio	0.7542	0.5694
100	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2706	0.0732	100	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.3285	0.1079
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2409	0.0580		$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2817	0.0794
	Ratio	0.8901	0.7925		Ratio	0.8576	0.7357
200	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2278	0.0519	200	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2654	0.0704
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2051	0.0421		$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.2345	0.0550
	Ratio	0.9004	0.8108		Ratio	0.8836	0.7809
400	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.1926	0.0371	400	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.2211	0.0489
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1732	0.0300		$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1971	0.0389
	Ratio	0.8996	0.8095		Ratio	0.8915	0.7950
800	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.1640	0.0269	800	$\widehat{\theta}_{FDH}(x_0, y_0)$	0.1891	0.0358
	$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1473	0.0217		$\widehat{\theta}_{FDHbc}(x_0, y_0)$	0.1688	0.0285
	Ratio	0.8982	0.8069		Ratio	0.8927	0.7970

estimator performs well, providing around 30% bias reduction for the FDH estimator and more than 50% of reduction for the MSE.

We propose also in what follows, under three scenarios, a comparison between nonparametric and parametric approaches. We simulate the data under the models $y = f(x, \beta)\exp(-u)$. One of the most popular parametric stochastic models (see Aigner, Lovell, and Schmidt, 1977) for estimating efficiencies is based on the model $y = f(x, \beta)\exp(v - u)$ where $f(x, \beta)$ is appropriately specified (Cobb–Douglas, ...), $U \sim N^+(0, \sigma_u^2)$, and $V \sim N(0, \sigma_v^2)$. Individual efficiencies are then estimated through an estimate of $\mathrm{E}(u_i \mid \hat{\varepsilon}_i)$, where $\hat{\varepsilon}_i = \log(y_i) - \log(f(x_i, \hat{\beta}))$ (for details, see Jondrow, Lovell, Materov, and Schmidt, 1982). Note that these stochastic parametric models encompass the deterministic models analyzed in this paper because the latter correspond to the case $\sigma_v^2 = 0$ (no noise). In our Monte Carlo scenarios, we will consider three distributions of inefficiencies: the half normal, $N^+(0, \sigma_u^2 = 1)$, and two truncated normals, $N^+(1, \sigma_u^2 = 1)$ and $N^+(2, \sigma_u^2 = 1)$. We expect of course that the parametric approach will outperform the nonparametric approaches when the specification of the model is correct, but we will investigate how the parametric approach is robust to specification error either only for the stochastic part of the model or only for the frontier function compared to the FDH and FDHbc estimators.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

TABLE 3. Results from Monte Carlo experiments with data simulated from Model 3: The evaluated point is $(x_0, y_0) = (0.6, 4)$

n	Estimator	Bias	MSE
25	$\widehat{y^{\partial}}_{FDH}(x_0)$	1.3966	1.9545
	$\widehat{y^{\partial}}_{FDHbc}(x_0)$	0.9667	0.9396
	Ratio	0.6922	0.4808
50	$\widehat{y^{\partial}}_{FDH}(x_0)$	0.9765	0.9551
	$\widehat{y^{\partial}}_{FDHbc}(x_0)$	0.6992	0.4910
	Ratio	0.7161	0.5140
100	$\widehat{y^{\partial}}_{FDH}(x_0)$	0.6472	0.4196
	$\widehat{y^{\partial}}_{FDHbc}(x_0)$	0.4759	0.2275
	Ratio	0.7353	0.5421
200	$\widehat{y^{\partial}}_{FDH}(x_0)$	0.4127	0.1706
	$\widehat{y^{\partial}}_{FDHbc}(x_0)$	0.2946	0.0872
	Ratio	0.7139	0.5110
400	$\widehat{y^{\partial}}_{FDH}(x_0)$	0.2717	0.0739
	$\widehat{y^{\partial}}_{FDHbc}(x_0)$	0.1940	0.0378
	Ratio	0.7142	0.5112
800	$\widehat{y^{\partial}}_{FDH}(x_0)$	0.1996	0.0399
	$\widehat{y^{\partial}}_{FDHbc}(x_0)$	0.1428	0.0205
	Ratio	0.7154	0.5129

- Model 4: Nonparametric FDH and FDHbc versus stochastic parametric; Cobb–Douglas–type frontier with $p = q = 1$. Here we simulate our data from the true model $y = x^{0.8}\exp(-u)$, where $X \sim U[0, 2]$ and $U \sim N^{+}(0, \sigma_u^2 = 1)$ in a first case, $U \sim N^{+}(1, \sigma_u^2 = 1)$ in a second case, and finally $U \sim N^{+}(2, \sigma_u^2 = 1)$. Table 4 summarizes the mean results for the three cases. The figures in the tables are the average value of the output efficiency scores, averaged over the n observations. The columns with the heading $\lambda_{Farrell}$ are the true values of these averages. As expected, there is no hope to do better than the parametric MLE when the specifications of the parametric frontier and of the stochastic part of the model are correct (Table 4A). Note however that, as pointed out by Simar and Wilson (in press), in small to moderate sample sizes, the MLE can fail to estimate the inefficiencies (this happens when the ordinary least squares (OLS) residuals have the wrong skewness: this is the case for $n = 25$ in Table 4A). But the parametric approach is very sensitive to the correct specification of the model, even with a correct

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

TABLE 4. Cobb–Douglas Model 4

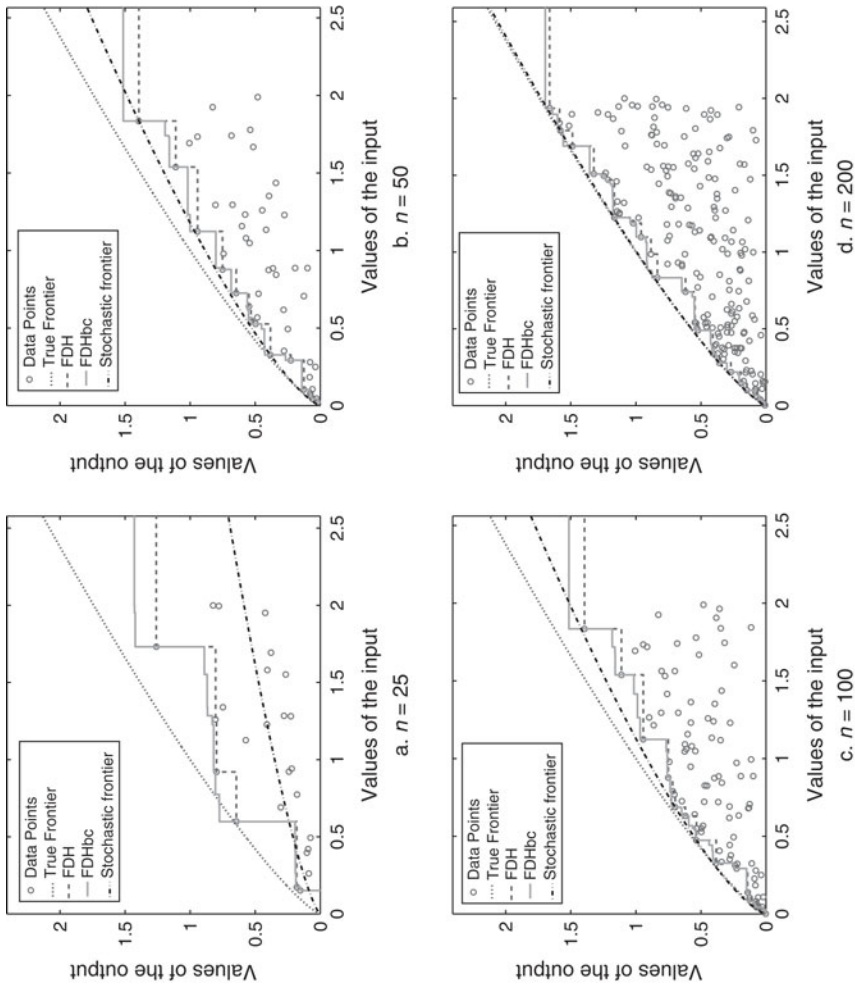
n	$\widehat{\lambda}_{FDH}(x,y)$	$\widehat{\lambda}_{FDHbc}(x,y)$	$\widehat{\lambda}_{MLE}(x,y)$	$\lambda_{Farrell}(x,y)$
A. Correct specification for MLE, normal–half normal case				
25	2.0375	2.2193	1.0000	3.2771
50	1.7846	1.9189	2.2861	2.5144
100	2.0454	2.1516	2.5317	2.6684
200	2.5411	2.6515	2.8857	2.9004
400	2.4578	2.5230	2.7656	2.8053
800	2.4534	2.5040	2.6475	2.6500
B. Correct frontier model but wrong specification for MLE: Normal–truncated normal case $N^+(1,\sigma_u^2=1)$, estimated by normal–half normal stochastic frontier model				
25	3.6759	4.1269	1.0000	7.1177
50	3.7941	4.1578	1.0000	6.6008
100	5.0266	5.4963	4.0704	6.3710
200	5.0968	5.4058	4.4434	6.1714
400	4.7873	5.0274	4.0695	5.5661
800	4.7936	4.9529	4.5341	5.4412
C. Correct frontier model but wrong specification for MLE: Normal–truncated normal case $N^+(2,\sigma_u^2=1)$ estimated by normal–half normal stochastic frontier model				
25	4.2193	4.7942	1.0000	12.0749
50	4.9737	5.6265	1.0000	14.0809
100	6.3217	6.8029	3.5111	14.7054
200	7.3730	8.1083	3.8555	12.2322
400	9.5753	10.3263	2.4921	13.9324
800	9.9437	10.5424	2.9322	13.4269

specification of the frontier function. We see indeed in Tables 4B and 4C that the MLE fails to estimate the inefficiencies (with a complete failure here for $n = 25, 50$). The nonparametric estimators behave much better in these two cases. Note also that in both cases the bias-corrected FDH is much better than the basic FDH estimator, mainly for small to moderate sample sizes.

The corresponding samples of sizes $n = 25, 50, 100, 200, 400, 800$ are also illustrated for the three cases in Figures 1–3, where the FDH, the FDHbc, and the MLE estimators of the frontier are displayed. The analysis of these pictures confirms the preceding comments.[3]

- Model 5: Nonparametric FDH and FDHbc versus stochastic parametric model; logit frontier with $p = q = 1$. Here we analyze an example where the parametric stochastic frontier model uses a wrong specification of the frontier itself (estimating a Cobb–Douglas frontier, whereas the frontier is a logit). We generate our data using Model 3 described previously, and we

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press



https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

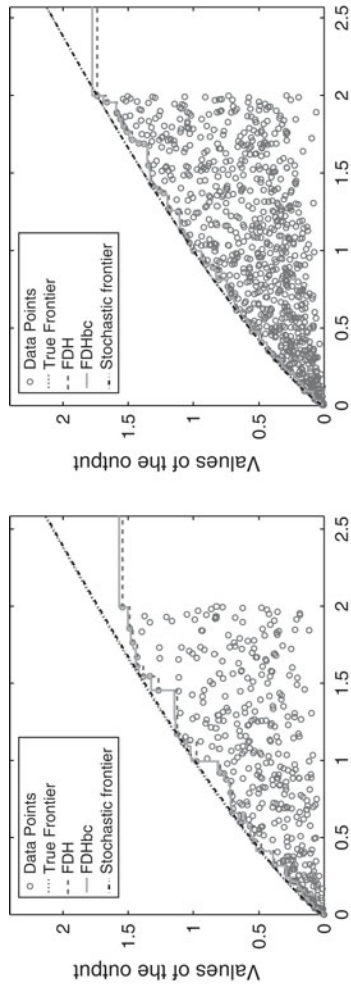
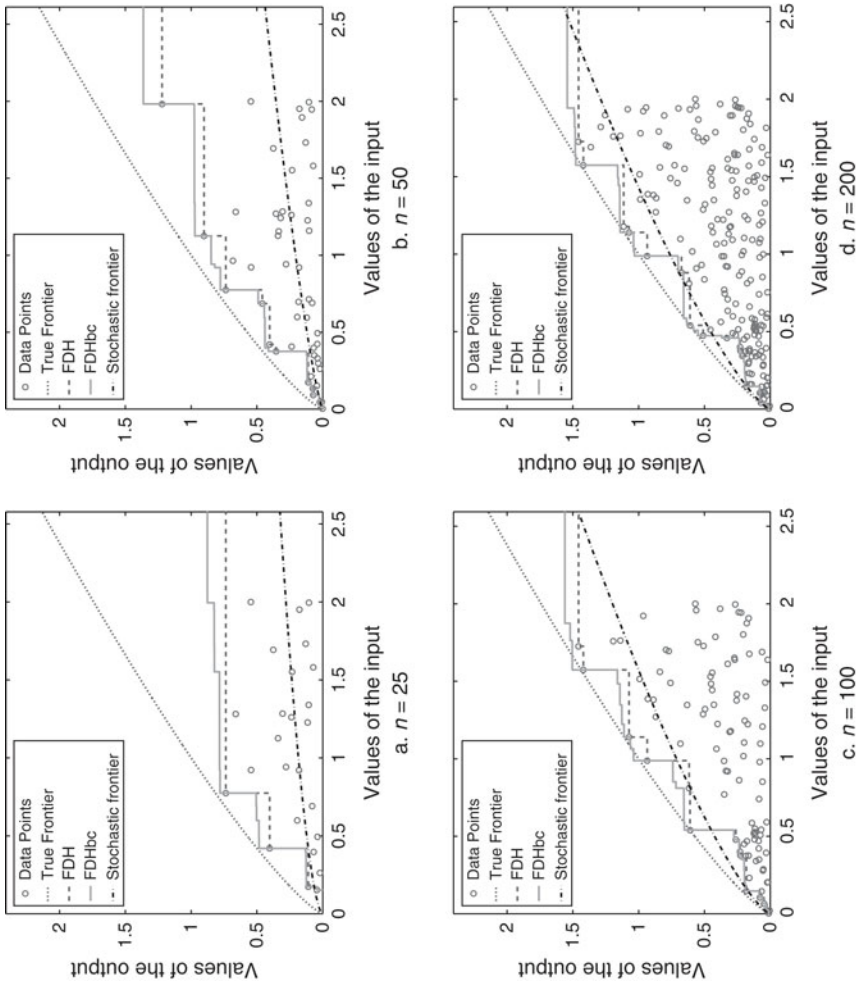


FIGURE 1. Comparison of FDH, FDHbc, and stochastic frontier model under correct specification for the distribution of efficiency: normal–half normal case.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press



https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

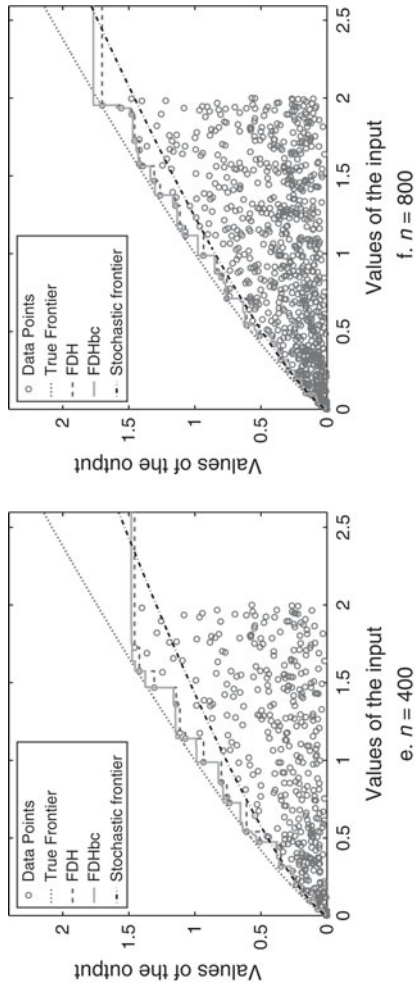
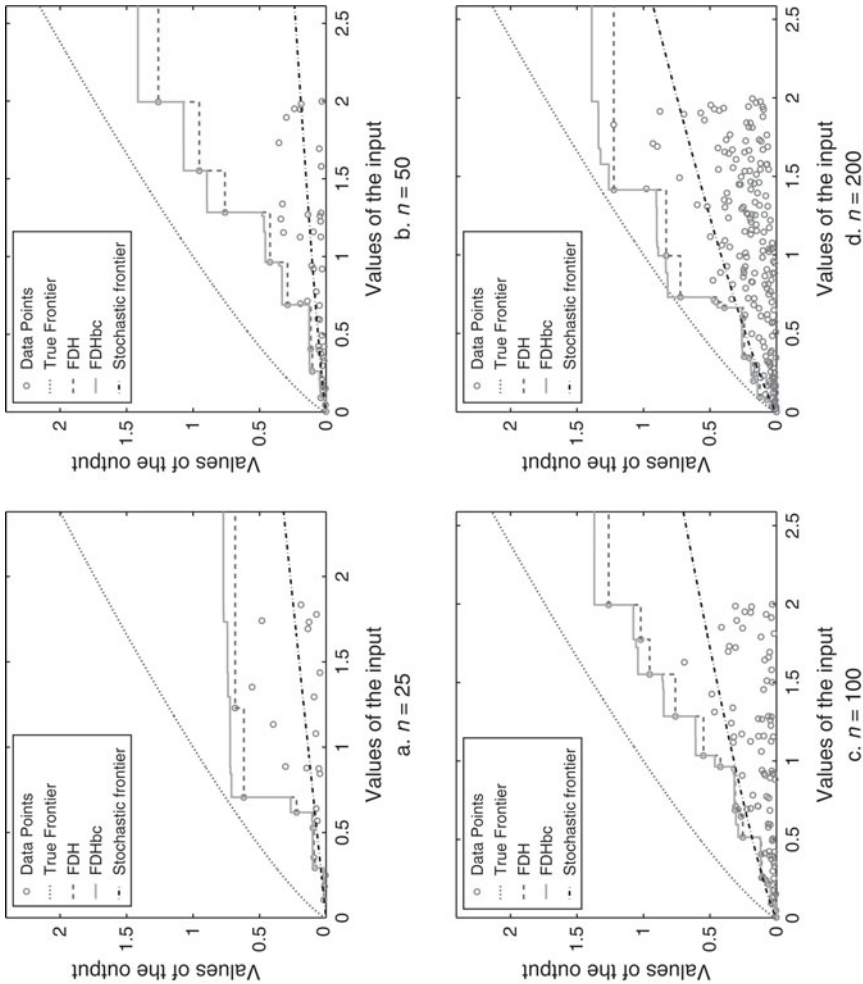


FIGURE 2. Comparison of FDH, FDHbc, and stochastic frontier model under wrong specification for the distribution of efficiency $N^+(1, \sigma_u^2 = 1)$.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press



https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

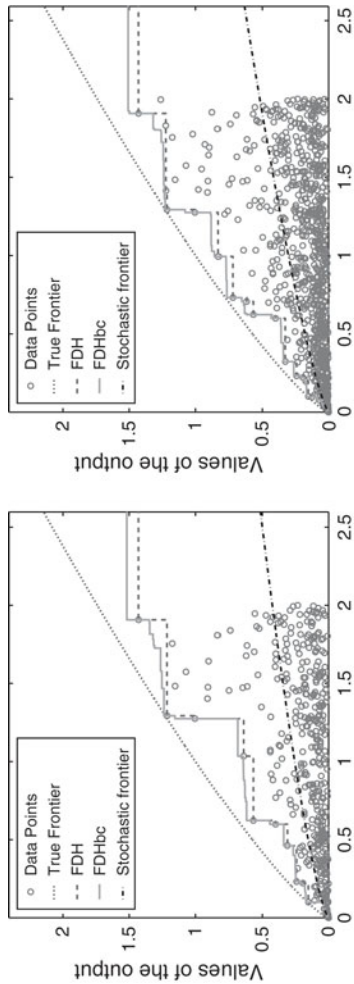
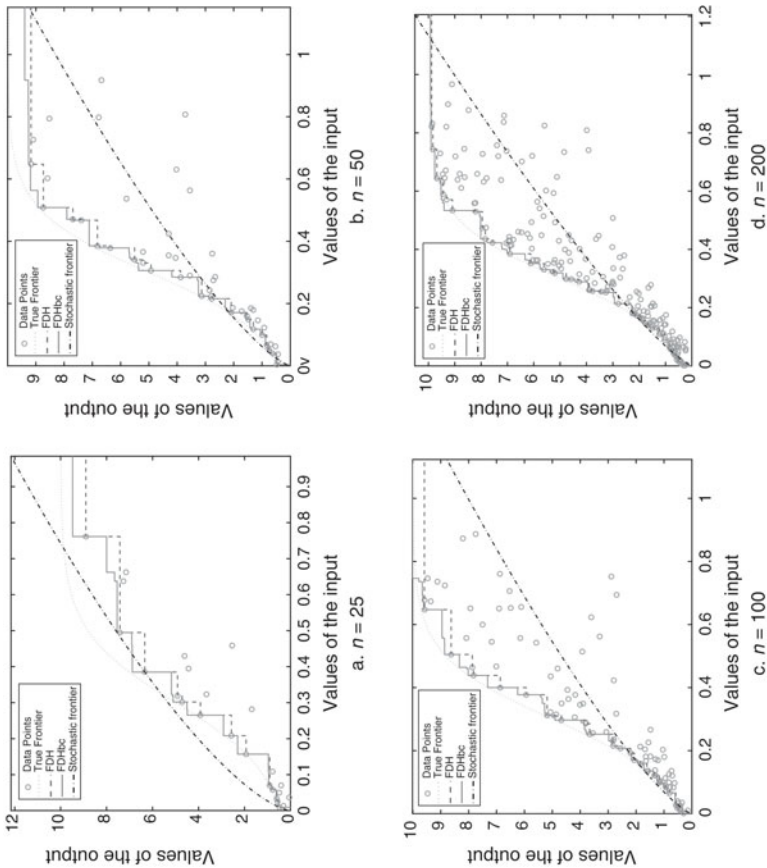


FIGURE 3. Comparison of FDH, FDHbc, and stochastic frontier model under wrong specification for the distribution of efficiency $N^+(2, \sigma_u^2 = 1)$.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press



https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

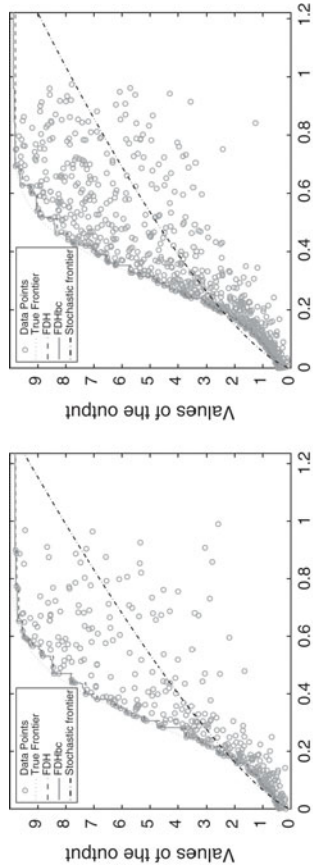


FIGURE 4. Comparison of FDH, FDHbc, and stochastic frontier model under wrong specification of the model.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

TABLE 5. Logit model 5 estimated by normal–half normal Cobb–Douglas–type frontier

n	$\widehat{\lambda}_{FDH}(x,y)$	$\widehat{\lambda}_{FDHbc}(x,y)$	$\widehat{\lambda}_{MLE}(x,y)$	$\lambda_{Farrell}(x,y)$
25	1.6914	1.8042	2.8251	1.9427
50	1.2241	1.2805	1.0000	1.3777
100	1.3416	1.3931	1.0000	1.4848
200	1.3245	1.3574	1.0000	1.4267
400	1.4188	1.4410	1.0000	1.5088
800	1.3940	1.4093	1.0000	1.4521

estimate it by a Cobb–Douglas normal–half normal stochastic frontier model. The results for random samples of sizes $n = 25, 50, 100, 200, 400, 800$ are reported in Table 5 and the corresponding pictures in Figure 4 allow us to compare the different estimators. As expected, the MLE estimator behaves dramatically badly under wrong specification of the frontier model: it is unable to detect any inefficiencies in most of the cases.

4.2. Real Data Example

To illustrate the behavior of the FDHbc estimator on real data, we use a data set from the NBER Manufacturing Productivity Database.[4] The database covers 408 manufacturing industries in the United States from 1958 to 1996. For analysis we use the data of year 1995, four inputs (Total payroll, Total cost of materials, Cost of electric and fuels, Total real capital stock), and an aggregated output (Total value added).

We report the results for the first 30 units of the sample in Table 6, where the mean in the last row of the table is computed over all 408 observations. For MLE estimation we used the normal–half normal stochastic frontier model. As we see from the first two columns of the table, the FDHbc estimators of the output efficiency score are less optimistic than the biased FDH estimators. For some units the difference is important (e.g., 18% for unit 4). With this data set, the MLE estimation (using a normal–half normal Cobb–Douglas model) fails to detect any inefficiencies for the reasons explained previously (wrong skewness of OLS residuals; for details, see Simar and Wilson, in press). Of course in this real data example we do not know what is the truth, but nonparametric estimates are more robust to specification errors, and our FDHbc tries to correct for the inherent bias of the FDH.

5. CONCLUSIONS

We proposed in this paper a simple bias-corrected nonparametric estimator for the efficient frontier and the efficiency score of a producer, which improves the

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

TABLE 6. Results for the first 30 units of a random sample of $n = 408$ observations from the NBER Manufacturing Productivity Database; output-oriented analysis

Units	$\widehat{\lambda}_{FDH}(x,y)$	$\widehat{\lambda}_{FDHbc}(x,y)$	$\widehat{\lambda}_{MLE}(x,y)$	$\widehat{y^{\partial}}_{FDH}(x,y)$	$\widehat{y^{\partial}}_{FDHbc}(x,y)$	n_x
1	2.4989	2.9245	1.0000	22,264.1000	26,055.1026	354
2	1.5032	1.5543	1.0000	9,371.9000	9,690.7103	325
3	1.3991	1.4793	1.0000	14,448.3000	15,275.8371	356
4	1.0000	1.1844	1.0000	180.6000	213.9078	7
5	1.5536	1.6261	1.0000	5,782.2000	6,051.7551	239
6	1.9640	2.3422	1.0000	5,782.2000	6,895.5483	161
7	2.3539	2.8598	1.0000	5,782.2000	7,024.9266	158
8	3.5058	4.1092	1.0000	22,264.1000	26,096.5012	335
9	1.3107	1.4892	1.0000	5,782.2000	6,569.3721	163
10	3.7831	4.6164	1.0000	22,264.1000	27,168.4148	299
11	1.6389	1.7971	1.0000	2,050.8000	2,248.7123	102
12	1.4447	1.6352	1.0000	5,782.2000	6,544.8633	158
13	1.5379	1.5797	1.0000	6,023.4000	6,187.2038	263
14	1.1606	1.2020	1.0000	5,782.2000	5,988.5289	223
15	2.8912	3.5304	1.0000	5,782.2000	7,060.4003	131
16	1.0000	1.1681	1.0000	8,922.1000	10,422.0996	185
17	3.3536	4.2143	1.0000	1,600.0000	2,010.6509	29
18	2.4521	3.0092	1.0000	5,782.2000	7,096.0584	117
19	1.0000	1.1389	1.0000	4,075.8000	4,641.9399	132
20	1.7633	2.0870	1.0000	5,782.2000	6,843.6197	138
21	1.1697	1.2121	1.0000	5,782.2000	5,992.0214	249
22	1.7809	2.0294	1.0000	22,264.1000	25,371.0168	345
23	1.0000	1.0306	1.0000	6,376.3000	6,571.2315	238
24	1.2095	1.3022	1.0000	1,600.0000	1,722.6816	69
25	2.8129	3.2561	1.0000	1,600.0000	1,852.0417	45
26	2.9619	3.4298	1.0000	1,600.0000	1,852.8013	51
27	2.3320	2.5754	1.0000	2,050.8000	2,264.7960	79
28	1.0236	1.0403	1.0000	6,023.4000	6,121.2861	241
29	1.3245	1.4422	1.0000	2,050.8000	2,233.1635	89
30	1.6322	1.9374	1.0000	1,600.0000	1,899.2104	27
Mean	1.6292	1.8374	1.0000	5,493.4936	6,155.6028	137.8

FDH estimator in terms of the bias. Our presentation focuses on input orientation, but similar results are easy to derive in the output-oriented case (see the Appendix). In one input, the multioutput case, we provide a formula based on order statistics for computing the estimator. We also define a conditional bootstrap that allows derivation of the same formula for the estimator and hence does not require bootstrap replications for computational purposes. Moreover, using the ideas proposed by Cazals et al. (2002) and developed by Daraio and Simar

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

(2005), 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 et al. (2000), and 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. These experiments show also that the FDHbc has better finite-sample mean squared error than the FDH. This is not surprising because the inherent bias of the FDH dominates its variance as already pointed out in Park et al. (2000). The comparisons with MLE presented in the previous section prove that although the nonparametric approach requires large samples to provide reliable estimators in a multivariate setup, the parametric approach can easily fail under wrong specification of the model. We know from Simar and Wilson (in press) that the drawbacks of MLE can even be more severe even if the parametric model is correctly specified.

NOTES

1. In the presentation of the paper, we will follow the input orientation; the presentation for the output orientation is summarized in the Appendix.
2. For the case of a "production" function (one output and several inputs), see the Appendix.
3. Other similar Monte Carlo experiments, with higher dimension for x, confirmed our comments but without the possibility of drawing pictures. The results are not displayed to save space.
4. National Bureau of Econometric Research, Cambridge, MA, U.S.A. This database is available at http://www.nber.org/nberces/nbprod96.htm.

REFERENCES

Aigner, D.J., C.A.K. Lovell, & P. Schmidt (1977) Formulation and estimation of stochastic frontier models. *Journal of Econometrics* 6, 21–37.

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

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

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

Daraio, C. & L. Simar (2005) Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *Journal of Productivity Analysis* 24, 93–121.

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

Deprins, D., L. Simar, & H. Tulkens (1984) Measuring labor inefficiency in post offices. In M. Marchand, P. Pestieau, & H. Tulkens (eds.), *The Performance of the Public Enterprises: Concepts and Measurements*, pp. 243–267. North–Holland.

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

Jeong, S.O. & L. Simar (2006) Linearly interpolated FDH efficiency score for nonconvex frontiers. *Journal of Multivariate Analysis* 97, 2141–2161.

Jondrow, J., C.A.K., Lovell, I.S. Materov, & P. Schmidt (1982) On the estimation of technical inefficiency in stochastic frontier production models. *Journal of Econometrics* 19, 233–238.

Kneip, A., L. Simar, & P.W. Wilson (2008) Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory* 24, 1663–1697.

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

Loh, W. (1984) Estimating an endpoint of a distribution with resampling methods. *Annals of Statistics* 12, 1543–1550.
Park, B., L. Simar, L. & C. Weiner (2000) The FDH estimator for productivity efficiency scores: Asymptotic properties. *Econometric Theory* 16, 855–877.
Robson, D.S. & J.H. Whitlock (1964) Estimation of a truncation point. *Biometrika* 51, 33–39.
Simar, L. (2007) How to improve the performances of DEA/FDH estimators in the presence of noise. *Journal of Productivity Analysis* 28, 183–201.
Simar, L. & P.W. Wilson (in press) Estimation and inference in cross-sectional stochastic frontier models. *Econometric Reviews*.

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 $\{Y_1^x, \ldots, Y_{n_x}^x\}$, 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 $F(y|x)$. Let $Y_{(1)}^x \leq \cdots \leq Y_{(n_x)}^x$ 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

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

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

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

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:

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

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press

Because the empirical conditional distribution function is defined as

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

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

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

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

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

Adapting formula (3.2) 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 et al. (2000), namely,

$$\mathcal{L}(n^{1/(p+q)}(\lambda(x, y) - \widehat{\lambda}_{FDH}(x, y))) \to \mathit{Weibull}(v_{x,y}, p+q) \quad \text{as } n \to \infty, \tag{A.11}$$

where $v_{x,y}$ is such that

$$P(X \le x, \qquad Y \ge (\lambda(x, y) - h)y) = v_{x,y}\, h^{p+q} + O(h^{p+q+1}),$$

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

https://doi.org/10.1017/S0266466609090513 Published online by Cambridge University Press