0502

LINEARLY INTERPOLATED FDH EFFICIENCY SCORE FOR NONCONVEX FRONTIERS

JEONG, S.-O. and L. SIMAR

Linearly interpolated FDH efficiency score for nonconvex frontiers [*]

Seok-Oh Jeong[†]and Léopold Simar[‡]

December 30, 2004

Abstract

This paper address the problem of estimating the monotone boundary of a nonconvex set in a full nonparametric and multivariate setup. This is particularly useful in the context of productivity analysis where the efficient frontier is the locus of optimal production scenarios. Then efficiency scores are defined by the distance of a firm from this efficient boundary. In this setup, the Free Disposal Hull (FDH) estimator has been extensively used due to its flexibilty and because it allows nonconvex attainable production sets. However the nonsmoothness and discontinuities of the FDH is a drawback for conducting inference in finite samples. In particular, it is shown that the bootstrap of the FDH has poor performances and so is not useful in practice. Our estimator, the LFDH, is a linearized version of the FDH, obtained by linear interpolation of appropriate FDH-efficient vertices. It offers a continuous, smooth version of the FDH. We provide an algorithm for computing the estimator, and we establish its asymptotic properties. We also provide an easy way to approximate its asymptotic sampling distribution. The latter could offer bias-corrected estimator and confidence intervals of the efficiency scores. In a Monte-Carlo study, we show that these approximations works well even in moderate sample sizes and that our LFDH estimator outperforms, both in bias and in MSE the original FDH estimator.

Key words: Nonparametric frontiers; Nonconvex monotone boundaries; Efficiency; Bootstrapping FDH.

AMS Classification: 62G10, 62G20, 62G09, 62P20

[*] The research support from "Interuniversity Attraction Pole", Phase V (No. P5/24) from the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

[†] Institut de statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. E-mail: seokoh@stat.ucl.ac.be

[‡] Institut de statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. E-mail: simar@stat.ucl.ac.be

1 Introduction

This paper address the problem of estimating the monotone boundary of nonconvex set in a full nonparametric and multivariate setup. The problem found its sources in productivity analysis and efficiency measurements of firms. Foundations of the economic theory on productivity and efficiency analysis date back to the works of Koopmans (1951) and Debreu (1951) on activity analysis. Shephard (1970) proposes a modern formulation of the problem. Following these lines, we consider a production technology where the activity of the firms is characterized by a set of inputs $x \in \mathbb{R}_+^p$ used to produce a set of outputs $y \in \mathbb{R}_+^q$. In this framework the production set is the set of technically feasible combinations of (x, y). It is defined as

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

Assumptions are usually done on this set, such as free disposability of inputs and outputs, meaning that if $(x, y) \in \Psi$, then $(x', y') \in \Psi$, as soon as[1] $x' \geq x$ and $y' \leq y$. In some cases, convexity of Ψ is also assumed (see Shephard, 1970, for more details).

As far as efficiency of a firm is of concern, the boundaries of Ψ are of interest. The efficient boundary (frontier) of Ψ is the locus of optimal production scenarios (minimal achievable input level for a given output or maximal achievable output given the input). The Farrell-Debreu efficient frontier is defined in a "radial sense" and the efficiency scores for a given production scenario $(x, y) \in \Psi$, are defined as:

$$\text{Input oriented} \quad : \quad \theta(x, y) = \inf\{\theta \geq 0 \,|\, (\theta x, y) \in \Psi\} \tag{1.2}$$

$$\text{Output oriented} : \quad \lambda(x, y) = \sup\{\lambda \geq 1 \,|\, (x, \lambda y) \in \Psi\} \tag{1.3}$$

If (x, y) is inside Ψ, $\theta(x, y) \leq 1$ is the proportionate reduction of inputs a unit working at the level (x, y) should perform to achieve efficiency. If $\theta(x, y) = 1$, the unit is on the efficient frontier of Ψ. In the output direction, we have $\lambda(x, y) \geq 1$ represents the proportionate increase of outputs the unit operating at level (x, y) should attain to be considered as being efficient and if $\lambda(x, y) = 1$, the unit is on the efficient frontier.

In practice Ψ is unknown and so has to be estimated from a random sample of production units $\{(x_i, y_i) \,|\, i = 1, \ldots, n\}$, where we assume that $\text{Prob}((x_i, y_i) \in \Psi) = 1$ (referred in the literature as deterministic frontier models). So the problem is related to the problem of estimating the support of the random variable (x, y) where, for mathematical convenience, we will assume that Ψ is compact. The most popular nonparametric estimators are based

[1] From here and below inequalities between vectors $a, b \in \mathbb{R}^k$ have to be understood element by element. Writing $a \leq b$ means $a_i \leq b_i$, for $i = 1, \ldots, k$.

on envelopment ideas: we search for estimators of Ψ which envelops at best the observed data points. The statistical properties of these estimators are now well established (see *e.g.* Simar and Wilson, 2000, for a recent survey).

The most flexible nonparametric estimator, initiated by Deprins, Simar and Tulkens (1984), is the Free Disposal Hull (FDH) estimator. It is provided by the free disposal hull of the sample points:

$$\widehat{\Psi}_{\text{FDH}} = \left\{(x,y) \in \mathbb{R}_{+}^{p+q} \,|\, y \leq y_i,\ x \geq x_i, \quad i = 1, \ldots, n\right\}. \tag{1.4}$$

The FDH efficiency scores are obtained by plugging $\widehat{\Psi}_{\text{FDH}}$ in equations (1.2) and (1.3) in place of the unknown Ψ. The asymptotic properties of the resulting estimators are provided by Park, Simar and Weiner (2000). In summary, the error of estimation converges at a rate $n^{1/(p+q)}$ to a limiting Weibull distribution.

If we assume that Ψ is convex, the convex hull of $\widehat{\Psi}_{\text{FDH}}$ provides the Data Envelopment Analysis (DEA) estimator of Ψ, initiated by Farrell (1957) and popularized as linear programming estimator by Charnes, Cooper and Rhodes (1978). It is defined as

$$\begin{aligned}\widehat{\Psi}_{\text{DEA}} &= \{(x,y) \in \mathbb{R}_{+}^{p+q} \,|\, y \leq \sum_{i=1}^{n} \gamma_i y_i \,;\, x \geq \sum_{i=1}^{n} \gamma_i x_i \text{ for } (\gamma_1, \ldots, \gamma_n) \\ &\qquad \text{such that } \sum_{i=1}^{n} \gamma_i = 1 \,;\, \gamma_i \geq 0, i = 1, \ldots, n\}. \end{aligned} \tag{1.5}$$

It is the smallest free disposal convex set covering all the data points. When using $\widehat{\Psi}_{\text{DEA}}$ as the estimated attainable set, the asymptotic properties of resulting DEA efficiency scores have been investigated in Kneip, Park and Simar (1998), Kneip, Simar and Wilson (2003) and Jeong (2004). In summary, the error of estimation converges at a rate $n^{2/(p+q+1)}$ to a limiting nondegenerate distribution.

These nonparametric estimator are very popular and very attractive due to their flexibility. Under convexity assumption, the DEA estimator provides a piecewise linear continuous "convex" frontier estimate, but the less restrictive FDH estimator, allowing for nonconvex attainable sets provides a discontinuous boundary estimate (for $p = q = 1$, it is a "stair-case" monotone function) whereas the true unknown boundary is often assumed to be smooth (continuous and differentiable). As explained below, the nonsmoothness and the discontinuity of the FDH estimator make it very difficult to use for inference in finite sample and the bootstrap (even in its consistent version) fails to provide sensible practical solutions (poor finite sample properties).

The objective of this paper is to propose a smooth (linearized) version of the FDH estimator allowing nonconvex attainable set Ψ which address the main drawbacks of the

FDH estimator. Our resulting estimator appears indeed to behave better than the FDH in finite samples and its asymptotic distribution can be easily evaluated in practice. The paper is organized as follows. Section 2 summarizes the main properties of the FDH estimator then, Section 3 presents our linearized version of the FDH (LFDH) and the practical way for computing the corresponding efficiency scores. Section 4 analyzes the asymptotic properties of our estimator and suggests bias-correction and confidence intervals for the frontier and for the efficiency scores. Section 5 investigates the finite samples properties of our estimator and shows its superiority over the original FDH estimator. Section 6 concludes. In the Appendix, we show that the subsampling bootstrap is consistent for the FDH estimators but we indicate by some simulation its practical limitations in finite sample, advocating again for the use of our LFDH estimator.

2 The FDH estimator

Given a sample $\mathcal{X}_n = \{(x_i, y_i), i = 1, \cdots, n\} \subset \Psi$, the FDH estimate of Ψ was defined in (1.4). The resulting FDH estimators of the efficiency scores for a firm operating at the level (x, y) are defined by

$$\hat{\theta}_n(x,y) = \min\{\theta \geq 0 \,|\, (\theta x, y) \in \widehat{\Psi}_{\text{FDH}}\},$$
$$\hat{\lambda}_n(x,y) = \max\{\lambda \geq 1 \,|\, (x, \lambda y) \in \widehat{\Psi}_{\text{FDH}}\}.$$

One may easily verify that

$$\hat{\theta}_n(x,y) = \min_{i|Y_i \geq y} \max_{1 \leq k \leq p} \frac{x_i^{(k)}}{x^{(k)}}, \tag{2.1}$$

$$\hat{\lambda}_n(x,y) = \max_{i|X_i \leq x} \min_{1 \leq k \leq q} \frac{y_i^{(k)}}{y^{(k)}}. \tag{2.2}$$

where $a^{(k)}$ denotes the k-th component of the vector a.

From now on, for the presentation, we will only focus on the input orientation, but of course, all the results and properties are easily translated into the output oriented case.

Park, Simar and Weiner (2000) analyze the asymptotic properties of the estimators. They rely on some regularity assumptions on the Data Generating Process.

Assumption 1. $\theta(x, y)$ is continuously (partially) differentiable in $(x, y) \in \Psi$, and its partial derivatives are all nonzero for all $(x, y) \in \Psi$.

Assumption 2. (x_i, y_i)'s are iid with a density f which is continuous, and $f(x, y) > 0$ on Ψ and $f(x, y) = 0$ outside Ψ.

It is then shown that $n^{1/(p+q)}(\hat{\theta}_n(x, y) - \theta(x, y))$ has a limiting Weibull distribution $W(\mu_{xy}, p + q)$ where μ_{xy} is a parameter depending on the shape of the boundary and on the density f at the boundary point. The expression of μ_{xy} and a way for estimating it in finite sample is provided in Park, Simar and Weiner (2000). The curse of dimensionality is a particularly sensible issue in this setup, as shown by their simulations and makes the use of the estimated limiting Weibull distribution rather imprecise in small sample (say n smaller than 1000, when $p + q \geq 5$).

In fact, in addition to the curse of dimensionality shared by most of the nonparametric techniques, the nonsmoothness and the discontinuity of FDH estimator make its use very difficult for statistical inference in finite sample. An alternative for doing inference might be the bootstrap. In the Appendix A, we show indeed that a subsampling bootstrap provides a consistent approximation of the sampling distribution of FDH estimators. This bootstrap is very easy to implement and avoids the problem of estimating the unknown parameter of the limiting Weibull.

However, one may doubt on its usefulness in practice. In fact, from our small simulation study in the Appendix, it is verified that the subsampling bootstrap is not a good idea particularly for the confidence interval. The main reason is that the FDH estimate is determined by only one sample point, and each subsampling chooses this point too often. Also, the choice of the optimal subsample size remains an open and sensible question, as shown in our simulations.

Since under our assumptions the true frontier is continuous, we might hope to improve the performance of the FDH estimator by smoothing its corresponding frontier. This is the idea of the linearized free disposal hull (LFDH) estimator defined in the next section. As shown below, it turns out that indeed the LFDH estimator outperforms the FDH estimator.

3 The LFDH estimator

3.1 Main idea

Linear interpolation is certainly the simplest plan for smoothing the FDH frontier. The idea is to interpolate the vertices of the free disposal hull of a given data set to get the smoothed version of FDH estimate. In a two dimensional setup ($p = q = 1$), this would be easy to do, by drawing the polygonal line smoothing the staircase production frontier. But in multidimensional setups it is not straightforward to identify the points which are to be interpolated among the vertices.

In this section we propose an algorithm for doing this, which results in the linearly interpolated FDH (LFDH) efficiency scores. We will consider the estimation of $\theta(x_0, y_0)$ (or of $\lambda(x_0, y_0)$, for the output oriented case) for a given firm (x_0, y_0). A sketch of the idea is as follows:

Step 1: Identify the vertices of the free disposal hull built by the observations $\mathcal{X}_n$.

Step 2: Move the vertices to a convex surface along the direction parallel to x_0 (or y_0 for the output oriented case).

Step 3: Compute the convex hull of the moved points.

Step 4: Identify the vertices constituting the facet penetrated by the ray θx_0 (or λy_0 for the output oriented case).

Step 5: Interpolate them.

A precise algorithm for this idea shall be described in the next subsection. For the practical implementation, Step 2–3 seems rather ambiguous when both input and output variables are multidimensional. Hence we will translate the problem, by considering a new coordinate system, where the frontier will be described by a scalar function and multidimensional covariates. This idea is analogous to the idea developed in Kneip, Simar and Wilson (2003) when analyzing the properties of the DEA estimator.

For this, we need to build an orthonormal basis of a $(d-1)$ dimensional space $t^{\perp} = \{v \in \mathbb{R}^d \mid t'v = 0\}$ for a given vector $t \in \mathbb{R}^d$, where $'$ denotes the transpose of a vector or a matrix. An algorithm to obtain an orthonormal basis of $t^{\perp}$ can be given as follows:

ONB 1: Compute $S_j = \left(\sum_{i=1}^{j} t_i^2\right)^{1/2}$ for $j = 1, \cdots, d$.

ONB 2: Set $j = 1$.

ONB 3: If $S_j = 0$, then set

$$v_j = \Big(\underbrace{0, 0, \cdots, 0}_{j-1},\ 1,\ \underbrace{0, \cdots, 0}_{d-j}\Big)'.$$

Otherwise, set

$$v_j = \Big(\underbrace{t_1 c_j,\ t_2 c_j, \cdots,\ t_j c_j}_{j},\ -S_j/S_{j+1},\ \underbrace{0, \cdots, 0}_{d-j-1}\Big)',$$

where $c_j = t_{j+1}/(S_j S_{j+1})$.

ONB 4: If $j = d-1$ then STOP. Otherwise, set $j = j+1$ and goto [ONB 3].

One may easily verify that those v_j's obtained by the above algorithm are all orthogonal to the vector $t \in \mathbb{R}^d$, and that $\{v_j \,|\, j = 1, \cdots, d-1\}$ are linearly independent. In addition $v_j' v_j = 1$ holds for all $j = 1, \cdots, d-1$. Thus $\{v_j \,|\, j = 1, \cdots, d-1\}$ forms an orthonormal basis for the $(d-1)$-dimensional space $t^\perp$.

3.2 Definition of the LFDH estimator

Fix $(x_0, y_0) \in \mathbb{R}_+^p \times \mathbb{R}_+^q$ the point of interest. We are to define a linearly interpolated FDH (LFDH) efficiency score at (x_0, y_0) as an estimator of $\theta(x_0, y_0)$. For brevity we restrict the detailed presentation for the input efficiency scores. In a remark below we give the algorithm for the output orientation.

Let $\mathcal{X}_B$ be the set of the FDH-efficient vertices obtained by $\mathcal{X}_n$, i.e.

$$\mathcal{X}_B = \left\{ (x_i, y_i) \in \mathcal{X}_n \,|\, \hat\theta_n(x_i, y_i) = 1 = \hat\lambda_n(x_i, y_i),\ i = 1, \cdots, n \right\}, \tag{3.1}$$

and let n_B be the cardinality of $\mathcal{X}_B$. The problem is now to identify the points to be interpolated among the n_B points in $\mathcal{X}_B$.

Let $\{v_j \,|\, j = 1, \cdots, p-1\}$ be an orthonormal basis for $x_0^\perp$. Consider a transformation r_{x_0} from $\mathbb{R}_+^p$ to $\mathbb{R}^{p-1} \times \mathbb{R}_+$:

$$r_{x_0} : x \mapsto \left(x' v_1,\ x' v_2,\ \cdots,\ x' v_{p-1},\ \frac{x' x_0}{\sqrt{x_0' x_0}} \right). \tag{3.2}$$

Then $(r_{x_0}^{(1)}(x), \cdots, r_{x_0}^{(p-1)}(x))$ is the coefficient vector of x in the space spanned by $\{v_j \,|\, j = 1, \cdots, p-1\}$ and $r_{x_0}^{(p)}(x)$ is the distance between x and $x_0^\perp$. Therefore, it holds that $r_{x_0}(x_0) = (0, \cdots, 0, \sqrt{x_0' x_0})$. Moreover, r_{x_0} is a one-to-one transformation, and the following reciprocal relation holds

$$x = \sum_{j=1}^{p-1} r_{x_0}^{(j)}(x) v_j + r_{x_0}^{(p)}(x) \frac{x_0}{\sqrt{x_0' x_0}}.$$

When $p = 1$, we have $r_{x_0}(x) = x$ for all $x \in \mathbb{R}_+$.

Consider now a transformation $h_{x_0,y_0} : \mathbb{R}_+^p \times \mathbb{R}_+^q \mapsto \mathbb{R}^{p-1+q} \times \mathbb{R}_+$ which maps (x, y) to (z, u), where

$$\begin{aligned} z &= \left(r_{x_0}^{(1)}(x), \cdots, r_{x_0}^{(p-1)}(x), y^{(1)} - y_0^{(1)}, \cdots, y^{(q)} - y_0^{(q)} \right)', \\ u &= r_{x_0}^{(p)}(x). \end{aligned} \tag{3.3}$$

Note that $h_{x_0,y_0}(x_0, y_0) = (0, 0, \cdots, 0, \sqrt{x_0' x_0})$. See Figure 1 for a graphical illustration of the new coordinate system given by the transform h_{x_0,y_0}.

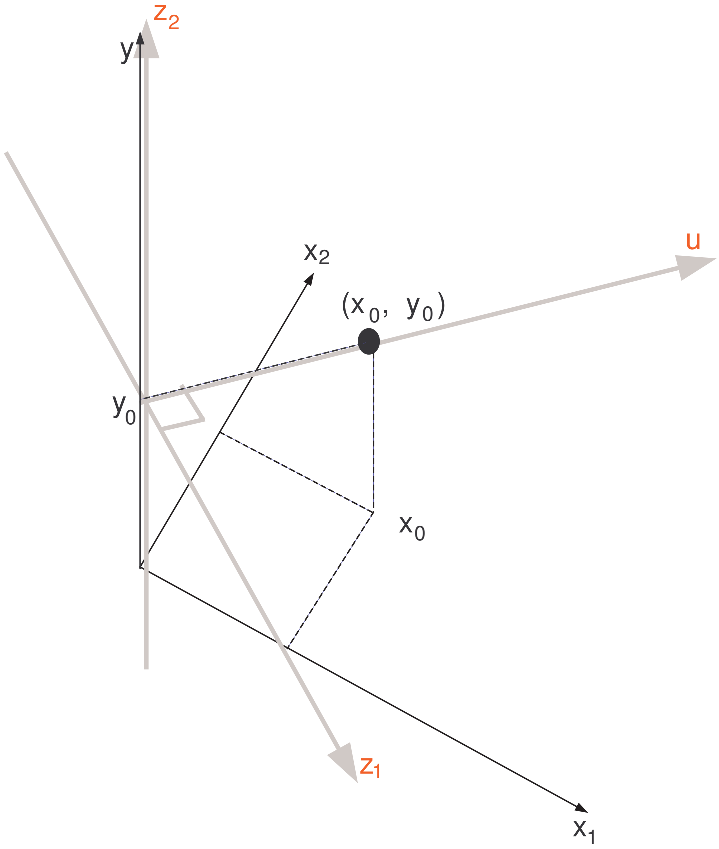

Figure 1: *An illustration of the new coordinate system* (z, u) *in the case of* $p = 2$ *and* $q = 1$.

Applying the transform in (3.3) to the points in $\mathcal{X}_B$, we get the set of transformed data in the new coordinate system (z, u):

$$\{(z_i, u_i) \,|\, (z_i, u_i) = h_{x_0,y_0}(x_i, y_i),\ (x_i, y_i) \in \mathcal{X}_B\}\,.$$

Now we are to identify the adjacent z_i's among them which forms a (smallest) simplex in z-space containing $z = 0$. Those adjacent z_i's must satisfy the property that the interior of the circumcircle determined by the adjacent z_i's contains no other z_i's. This is closely related to Delaunay triangulation (or tessellation) in computational geometry, see Barber, Dobkins and Huhdanpaa (1996) and Section 5.3 in O'Rourke (1998). A simple way to do this, suggested by Brown (1979), is as follows. By substituting u_i with $w_i = z_i'z_i$ for each $i = 1, \cdots, n_B$, we get the moved points $\{(z_i, w_i) \,|\, i = 1, \cdots, n_B\}$ laid on the strictly convex surface $u = z'z$ in the coodinate system (z, u). Next, compute the convex hull of (z_i, w_i)'s and identify the set of indices $\mathcal{I}$ defined by

$$\begin{aligned} \mathcal{I} \;\; &= \;\; \{i \,|\, (z_i, w_i) \text{ makes the facet of the convex hull containing the origin } (z = 0), \\ &\qquad\qquad\qquad\qquad i = 1, \cdots, n_B\}. \end{aligned} \tag{3.4}$$

That is, if we define $\gamma^* = (\gamma_1^*, \cdots, \gamma_{n_B}^*)$ by

$$\gamma^* = \arg\min_{\gamma} \left\{ \sum_{i=1}^{n_B} \gamma_i w_i \,\middle|\, \sum_{i=1}^{n_B} \gamma_i z_i = 0, \;\; \sum_{i=1}^{n_B} \gamma_i = 1, \;\; \gamma_i \geq 0, \;\; i = 1, \cdots, n_B \right\},$$

then we have $\mathcal{I} = \{i \,|\, \gamma_i^* > 0\}$. For a graphical illustration of the idea so far, see Figure 2.

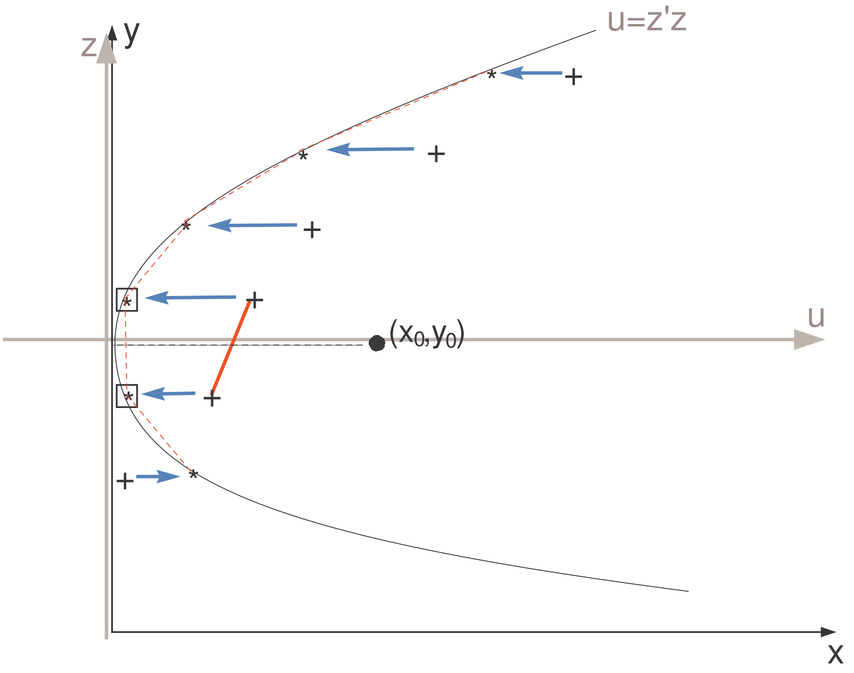

Figure 2: *Identifying the adjacent points which are to be interpolated. The crosses (+) represent* (z_i, u_i)*'s and the asterisks (*) represent the correspoding* (z_i, w_i)*'s.*

Finally, define the LFDH estimator $\tilde{\theta}_n(x_0, y_0)$ by

$$\begin{aligned} \tilde{\theta}_n(x_0, y_0) = \min\{\theta > 0 \,|\, \theta x_0 \geq \textstyle\sum_{i \in \mathcal{I}} \gamma_i x_i, y_0 \leq \sum_{i \in \mathcal{I}} \gamma_i y_i \\ \text{for some } \gamma_i \geq 0, i \in \mathcal{I} \text{ such that } \textstyle\sum_{i \in \mathcal{I}} \gamma_i = 1\}. \end{aligned} \tag{3.5}$$

Note that the proposed estimator coincides with a DEA estimator with reference set (sample points) given by the points $\{(x_i, y_i) \,|\, i \in \mathcal{I}\}$. We note also that it satisfies the free disposability assumption. Moreover, we are able to build the corresponding LFDH estimator of the attainable set Ψ by

$$\widehat{\Psi}_{\text{LFDH}} = \left\{ (x,y) \in \mathbb{R}_+^{p+q} \,\middle|\, x \geq \tilde{\theta}_n(t,s)t,\ y \leq s,\ (t,s) \in \widehat{\Psi}_{\text{FDH}} \right\}. \tag{3.6}$$

Remark 1. The algorithm for output LFDH efficiency score is given similarly. Define $y_0^\perp = \{v \in \mathbb{R}_+^q \,|\, y_0'v = 0\}$ and let $\{v_j \,|\, j = 1, \cdots, q-1\}$ be an orthonormal basis for $y_0^\perp$. Consider a transformation r_{y_0} from $\mathbb{R}_+^q$ to $\mathbb{R}^q$:

$$r_{y_0} : s \mapsto \left(s'v_1, \cdots, s'v_{p-1}, \frac{s'y_0}{\sqrt{y_0'y_0}} \right).$$

For each observation $(x_i, y_i) \in \mathcal{X}_B$, apply a transform $(x_i, y_i) \mapsto (z_i, w_i)$:

$$\begin{aligned} z_i &= \left(x_i^{(1)} - x_0^{(1)}, \cdots, x_i^{(p)} - x_0^{(p)}, r_{y_0}(y_i)^{(1)}, \cdots, r_{y_0}(y_i)^{(q-1)} \right)' \\ w_i &= z_i'z_i \end{aligned}$$

for $i = 1, \cdots, n_B$. Next, we construct a convex hull of the transformed data and identify the set of indices defined by

$$\mathcal{I} = \{i \,|\, (z_i, w_i) \text{ makes the facet of the convex hull containing } z = 0\}.$$

Hence we have got $\{(x_i, y_i) \in \mathcal{X}_B \,|\, i \in \mathcal{I}\}$ which are to be interpolated. Finally the LFDH estimator is given by

$$\begin{aligned} \tilde{\lambda}_n(x_0, y_0) = \max\{\theta \geq 1 \,|\, x_0 \geq \textstyle\sum_{i\in\mathcal{I}} \gamma_i x_i, \lambda y_0 \leq \sum_{i\in\mathcal{I}} \gamma_i y_i \\ \text{for some } \gamma_i \geq 0, i \in \mathcal{I} \text{ such that } \textstyle\sum_{i\in\mathcal{I}} \gamma_i = 1\}. \end{aligned} \tag{3.7}$$

Remark 2. LFDH estimator can be regarded as a rolling-ball estimator by Hall, Park and Turlach (2002) with variable ball size. LFDH determines the smoothing parameter, i.e. the ball size, automatically in such a way that it locally adapts to the shape of the given cloud of points.

4 Asymptotic distribution

4.1 Asymptotics of LFDH estimator

Recall the transformation $h_{x_0,y_0} : \mathbb{R}_+^p \times \mathbb{R}_+^q \mapsto \mathbb{R}^{p-1+q} \times \mathbb{R}_+$ given by

$$\begin{aligned} z_i &= \left(r_{x_0}(x_i)^{(1)}, \cdots, r_{x_0}(x_i)^{(p-1)}, y_i^{(1)} - y_0^{(1)}, \cdots, y_i^{(q)} - y_0^{(q)} \right)' \\ u_i &= r_{x_0}(x_i)^{(p)} \end{aligned} \tag{4.1}$$

for $i = 1, \cdots, n$. We denote the transformed dataset by $\widetilde{\mathcal{X}}_n$:

$$\widetilde{\mathcal{X}}_n = \{(z_i, u_i) \,|\, i = 1, \cdots, n\}.$$

In the new coordinate system (z, u), the attainable set Ψ is reexpressed as

$$G = \{(z, u) \in \mathbb{R}^{p-1+q} \times \mathbb{R}_+ \,|\, (z, u) = h_{x_0,y_0}(x, y),\ (x, y) \in \Psi\}. \tag{4.2}$$

And we can define the boundary of Ψ through its correspondent in the coodinate system (z, u):

$$g(z|x_0, y_0) = \inf\{u > 0 \,|\, (z, u) \in G\}. \tag{4.3}$$

Hence the set G can equivalently be represented by the function g as well:

$$G = \{(z, u) \in \mathbb{R}^{p-1+q} \times \mathbb{R}_+ \,|\, u \geq g(z|x_0, y_0)\}.$$

Furthermore, since the point of interest (x_0, y_0) is transformed by h_{x_0,y_0} into $(0, |x_0|)$, we have

$$\theta(x_0, y_0) = |x_0|^{-1} g(0|x_0, y_0), \tag{4.4}$$

where $|\cdot|$ denotes the Euclidean norm of a vector. See Figure 3 for a graphical illustration.

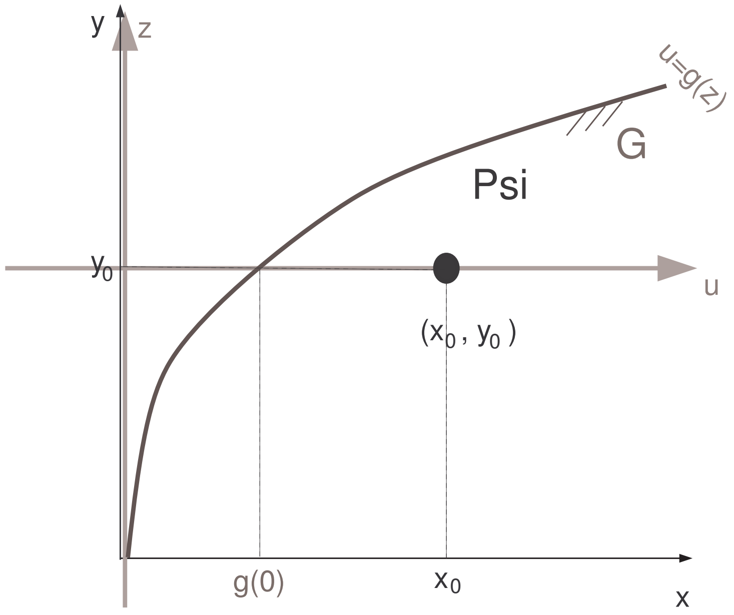

Figure 3: Ψ, G *and* g *in the case of* $p = 1$ *and* $q = 1$.

The following assumptions are the analogy in the coodinate system (z, u) of Assumption 1 and Assumption 2 in Section 2.

Assumption 1a. The function $g(z)$ is continuously differentiable in z, and $g_1(z)$ is nonsingular for all z, where $g_1(z)$ is the diagonal matrix having $(\partial/\partial z)g(z)$ as its diagonal elements.

Assumption 2a. (z_i, u_i)'s are iid with a density $\tilde{f}$ which is continuous and positive on G. In particular $\tilde{f}(z, g(z))$ is bounded away from zero for all z.

Define the LFDH estimator, $\hat{g}_{\mathrm{LFDH}}(z|x_0, y_0)$, of $g(z|x_0, y_0)$ for $z \in \mathbb{R}^{p-1+q}$ by the following procedure. Firstly, identify the vertices of the free disposal hull of $\mathcal{X}_n$ and apply the transform in (4.1) on the vertices to get $\{(z_i, u_i) \,|\, i = 1, \cdots, n_B\}$. Secondly, move $\{(z_i, u_i) \,|\, i = 1, \cdots, n_B\}$ to a convex surface along the u-axis. Thirdly, build the convex hull of the projected points and identify the set of indices $\mathcal{I}_z$ such that $\{(z_i, u_i) \,|\, i \in \mathcal{I}_z\}$ is involved in the facet of the convex hull computed at z. Note that $\mathcal{I}_0$ is identical to $\mathcal{I}$ in (3.5). Finally define the LFDH estimator of $g(z|x_0, y_0)$ by

$$\hat{g}_{\mathrm{LFDH}}(z|x_0, y_0) = \min \left\{ \textstyle\sum_{i \in \mathcal{I}_z} \gamma_i u_i \,\middle|\, \textstyle\sum_{i \in \mathcal{I}_z} \gamma_i z_i = z \text{ for some } \gamma_i \geq 0, \; i \in \mathcal{I}_z \right. \\ \left. \text{such that } \textstyle\sum_{i \in \mathcal{I}_z} \gamma_i = 1 \right\}. \tag{4.5}$$

Note that, in the coordinate system (z, u), $\hat{g}_{\mathrm{LFDH}}(z|x_0, y_0)$ is the highest piecewise linear surface below (z_i, u_i)'s. Moreover, it is easily verified that

$$\bar{\theta}_{z,n} = \min\{\theta > 0 \,|\, \theta x_0 = \textstyle\sum_{i \in \mathcal{I}_z} \gamma_i x_i, y_0 = \sum_{i \in \mathcal{I}_z} \gamma_i y_i \\ \text{for some } \gamma_i \geq 0, i \in \mathcal{I}_z \text{ such that } \textstyle\sum_{i \in \mathcal{I}_z} \gamma_i = 1\}$$

is equal to $|x_0|^{-1} \hat{g}_{\mathrm{LFDH}}(z|x_0, y_0)$.

Under Assumption 1a–2a one may prove that, as $n \to \infty$,

$$\bar{\theta}_{z=0,n} = \tilde{\theta}_n(x_0, y_0)$$

holds with probability tending to one, see Proposition 2 in Jeong and Park (2004). Hence it holds that with probability tending to one

$$\hat{g}_{\mathrm{LFDH}}(0|x_0, y_0) = |x_0| \cdot \tilde{\theta}_n(x_0, y_0). \tag{4.6}$$

Consequently, the asymptotic behaviour of $\tilde{\theta}_n(x_0, y_0)$ is equivalent to that of $\hat{g}_{\mathrm{LFDH}}(0|x_0, y_0)$. To simplify the notation, we will omit '$|x_0, y_0$' in $g(\cdot|x_0, y_0)$ and $\hat{g}_{\mathrm{LFDH}}(\cdot|x_0, y_0)$ from now on.

Consider an estimation of a continuous frontier function $g(z)$, $z \in \mathbb{R}^{p-1+q}$. Due to the complexity in multidimensional situation it is not easy to derive the explicit formula for the asymptotic distribution of $\hat{g}_{\rm LFDH}(z)$. So we suggest to follow the strategy used in Hwang, Park and Ryu (2002), Jeong (2004) and Jeong and Park (2004). Omitting the detailed proofs which are very similar to those in Jeong and Park (2004), we sketch the main steps to obtain the limit distribution as follows. Consider a linear transformation that takes (z_i, u_i) to

$$\begin{aligned} z_i^* &= n^{1/(p+q)} g_1(z)(z_i - z) \\ u_i^* &= n^{1/(p+q)}\{u_i - g(z)\}. \end{aligned}$$

Then (z_i^*, u_i^*) has as its frontier the surface with the equation

$$u^* = \mathbf{1}' z^* + o(1)$$

uniformly on any compact set of z^*, where $\mathbf{1} = (1, 1, \ldots, 1)'$. Moreover, for large n, the density in the new coordinate system (z^*, u^*) is approximated by $n^{-1}\|g_1(z)\|^{-1} f_0$ uniformly in the region

$$\left\{(z^*, u^*) \,\middle|\, |z^*| \le \varepsilon_n n^{1/(p+q)},\ \mathbf{1}'z^* \le u^* \le \mathbf{1}'z^* + \varepsilon_n n^{1/(p+q)}\right\}$$

for each sequence $\varepsilon_n \to 0$, where $\tilde{f}_0$ denotes the density at $(z, g(z))$ and $\|\cdot\|$ denotes the determinant of a matrix.

Define $\kappa = \{\|g_1(z)\|/\tilde{f}_0\}^{1/(p+q)}$, and consider a new random sample from the uniform distribution on

$$\begin{aligned} \mathcal{B}_\kappa = \Big\{(z^*, u^*) \,\big|\, z^* \in [-(\kappa/2)n^{1/(p+q)}, (\kappa/2)n^{1/(p+q)}]^{p-1+q}, \\ \mathbf{1}'z^* \le u^* \le \mathbf{1}'z^* + \kappa n^{1/(p+q)} \Big\}. \end{aligned} \tag{4.7}$$

Note that the uniform density on this region is $n^{-1}\|g_1(z)\|^{-1}\tilde{f}_0$. Let $\hat{g}^*_{\rm LFDH}(\cdot)$ be the version of $\hat{g}_{\rm LFDH}(\cdot)$ obtained from the new sample. Then, according to the same arguments for Theorem 1 in Jeong and Park (2004), we have the following theorem.

Theorem 4.1. *Under Assumption 1a–2a, For $z \in \mathbb{R}^{p-1+q}$, $n^{1/(p+q)}\{\hat{g}_{\rm LFDH}(z) - g(z)\}$ and $\hat{g}^*_{\rm LFDH}(0)$ have the same limit distribution.*

Corollary 4.1. *Under Assumption 1–2, $n^{1/(p+q)}\{\tilde{\theta}_n(x_0, y_0) - \theta(x_0, y_0)\}$ and $\hat{g}^*_{\rm LFDH}(0)/|x_0|$ have the same limit distribution.*

Once the unknown parameters $\tilde{f}_0$ and $g_1(z)$ are determined, the limit distribution of LFDH estimator $\tilde{\theta}_n(x_0, y_0)$ can be simulated based on this result: we only need to simulate a large number of times the value $\hat{g}^*_{\rm LFDH}(0)$, each of which being computed fropm a random sample drawn from the uniform on $\mathcal{B}_\kappa$. Of course, $\tilde{f}_0$ and $g_1(z)$ are unknwon, but they can be easily and consistently estimated as follows.

For estimating $\tilde{f}_0$, we propose to use the histogram type version in Jeong and Park (2004):

$$\frac{\sum_{i=1}^{n} I\{(z_i, u_i) \in \mathcal{D}(\delta)\}}{n \cdot \mathrm{mes}(\mathcal{D}(\delta))} \tag{4.8}$$

where $\mathcal{D}(\delta)$ is the region in the coordinate system (z, u) defined by, for $\delta > 0$,

$$\mathcal{D}(\delta) = \big\{(z, u) \,\big|\, z \in [-\delta/2,\ \delta/2]^{p-1+q},\ \hat{g}_{\mathrm{LFDH}}(0) \le u \le \hat{g}_{\mathrm{LFDH}}(0) + \delta\big\}$$

and $\mathrm{mes}(A)$ denotes the Lebesgue measure of a set $A \in \mathbb{R}^{p+q}$. Its consistency is directly derived by the consistency of $\hat{g}_{\mathrm{LFDH}}$, as long as δ is chosen to satisfy $n\delta^{p+q} \to \infty$ as $n \to \infty$, see Theorem 2 in Gijbels, Mammen, Park and Simar (1999).

For estimating $g_1(z)$, we propose to use the slope of the facet involved in $\hat{g}_{\mathrm{LFDH}}(z)$, see Park (2001) and Hall and Park (2002). The slope can be easily obtained by computing the hyperplane in the coordinate system (z, u) which passes through the set of points

$$\{(z_i, u_i) \,|\, i \in \mathcal{I}_z\} \cup \{(z, \hat{g}_{\mathrm{LFDH}}(z))\}.$$

That is, the estimator of $\|g_1(z)\|$ is defined by the product of $(\beta_0, \beta_1, \cdots, \beta_{p-1+q})$ which is the solution of the system of equations below:

$$u_i = \beta_0 + \sum_{j=1}^{p-1+q} \beta_j z_i^{(j)}, \quad i \in \mathcal{I}_z. \tag{4.9}$$

Note that the cardinality of $\mathcal{I}_z$ is greater than or equal to 1 and less than or equal to $p + q$ by construction. In fact it is equal to $p + q$ with probability tending to one as $n \to \infty$, and hence the system (4.9) is nonsingular in probability. When the cardinality of $\mathcal{I}_z$ is less than $p + q$, we suggest to add the equation

$$\hat{g}_{\mathrm{LFDH}}(z) = \beta_0 + \sum_{j=1}^{p-1+q} \beta_j z^{(j)}$$

to (4.9) at first. If the singularity still exists, then, by the ascending order of $|z_i - z|$ for z_i's such that $(z_i, u_i) \in \widetilde{\mathcal{X}}_n \setminus \{(z_i, u_i) \,|\, i \in \mathcal{I}_z\}$, add the equation

$$\hat{g}_{\mathrm{LFDH}}(z_i) = \beta_0 + \sum_{j=1}^{p-1+q} \beta_j z_i^{(j)}$$

to (4.9) in turn until the system becomes nonsingular. We point out that, for estimating g_1, the proposed estimator does not require any smoothing parameter.

Having an estimate of the asymptotic distribution of the LFDH estimators, the next subsection suggests procedures for correcting for the bias and for constructing confidence intervals of the quantities of interest.

4.2 Bias-corrected estimator and confidence interval

By using the distribution of $\hat{g}^*_{\mathrm{LFDH}}(0)$, we may indeed quantify the bias of the LFDH estimator $\hat{g}_{\mathrm{LFDH}}(0)$ or $\tilde{\theta}_n(x_0, y_0)$. Let $\{\hat{g}^*_{\mathrm{LFDH},b}(0)\}_{b=1}^B$ be the set of B values of $\hat{g}^*_{\mathrm{LFDH}}(0)$, each of which is computed from a random sample from the uniform distribution on $\mathcal{B}_{\hat{\kappa}}$, see (4.7), where $\hat{\kappa}$ is an estimate κ which is the unknown in the large sample approximation in Theorem 4.1. Since the empirical distribution of $\{\hat{g}^*_{\mathrm{LFDH},b}(0)\}_{b=1}^B$ approximates the distribution of $\hat{g}^*_{\mathrm{LFDH}}(0)$, we may estimate the asymptotic mean of $n^{1/(p+q)}\{\hat{g}_{\mathrm{LFDH}}(0) - g(0)\}$ by

$$B^{-1}\sum_{b=1}^{B} \hat{g}^*_{\mathrm{LFDH},b}(0).$$

Thus, a bias corrected estimator of $g(0)$ is given by

$$\hat{g}_{\mathrm{LFDH}}(0) - n^{-1/(p+q)}B^{-1}\sum_{b=1}^{B} \hat{g}^*_{\mathrm{LFDH},b}(0).$$

Of course we get the bias-corrected version of $\tilde{\theta}_n(x_0, y_0)$ by

$$|x_0|^{-1} \cdot \left\{ \hat{g}_{\mathrm{LFDH}}(0) - n^{-1/(p+q)}B^{-1}\sum_{b=1}^{B} \hat{g}^*_{\mathrm{LFDH},b}(0) \right\}.$$

The empirical distribution of $\{\hat{g}^*_{\mathrm{LFDH},b}(0)\}_{b=1}^B$ also enables us to construct a confidence interval for $g(0)$. Let $\hat{q}_\alpha$ be the α-th quantile of the empirical distribution of $\{\hat{g}^*_{\mathrm{LFDH},b}(0)\}_{b=1}^B$. Then, $100(1-\alpha)\%$ confidence interval for $g(0)$ is given by

$$\left[\hat{g}_{\mathrm{LFDH}}(0) - n^{-1/(p+q)}\hat{q}_{1-\alpha/2},\ \hat{g}_{\mathrm{LFDH}}(0) - n^{-1/(p+q)}\hat{q}_{\alpha/2}\right],$$

and hence $100(1-\alpha)\%$ confidence interval for $\theta(x_0, y_0)$ is given by

$$\left[\tilde{\theta}_n(x_0, y_0) - n^{-1/(p+q)}|x_0|^{-1}\hat{q}_{1-\alpha/2},\ \tilde{\theta}_n(x_0, y_0) - n^{-1/(p+q)}|x_0|^{-1}\hat{q}_{\alpha/2}\right].$$

5 Numerical study

5.1 Sampling distribution of FDH and LFDH estimators

In this section, we compare the sampling distribution of FDH estimator and that of LFDH estimator by a simulation study. We performed 500 Monte Carlo experiments with the sample sizes $n = 100$ and $n = 400$ and we choose the simulation models used in the Section 4.1 in Park, Simar and Weiner (2000).

- Model 1: Simulation II in Park, Simar and Weiner (2000). $p = q = 2$.

- Model 2: Simulation III in Park, Simar and Weiner (2000). $p = 4$, $q = 1$.

Figure 4 depicts the sampling distributions of the FDH and LFDH estimates in all the Monte-Carlo scenarios. We clearly see in the figures the superior performance of LFDH estimators over FDH estimators: in each case, the sampling distribution is more concentrated (less variance) and the sampling distribution is shifted toward the true value of the estimated parameters. This is confirmed by analyzing the mean squared error of both parameters as provided in Table 1. The table demonstrates again the desirable properties of LFDH estimator over the original FDH: the bias is substantially smaller and the MSE are smaller by a factor 4–5.

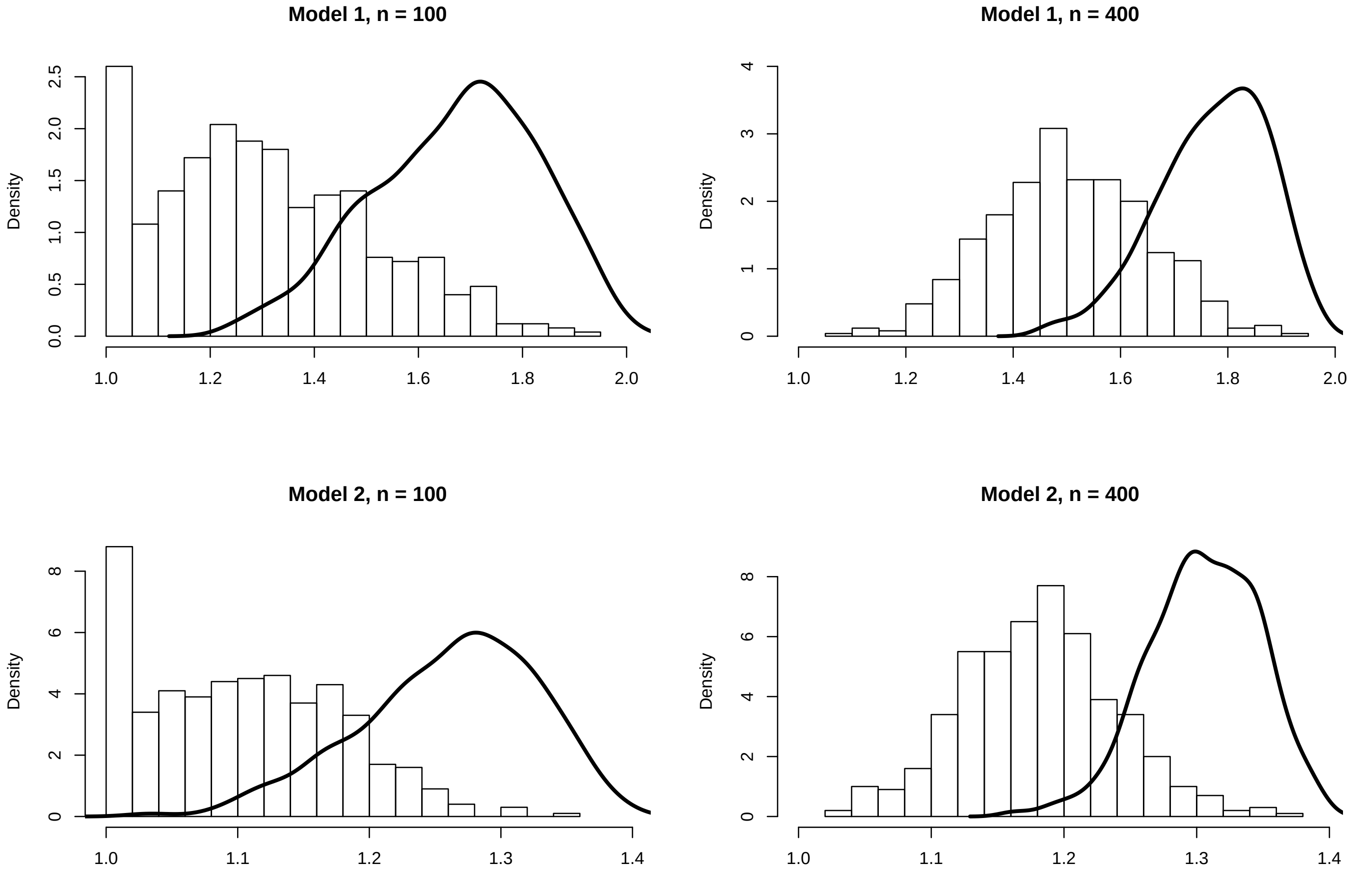

Figure 4: *Comparison of the sampling distribution of FDH (histogram) and that of LFDH (solid line) estimates. Model 1:* $\theta = 2.0027$. *Model 2:* $\theta = 1.4161$.

		Mean (s.e.)		MSE ($\times 10^{-1}$)	
	n	FDH	LFDH	FDH	LFDH
Model 1	100	1.3021 (0.0092)	1.6711 (0.0071)	5.3316	1.3535
($\theta = 2.0027$)	400	1.5039 (0.0065)	1.7728 (0.0046)	2.6965	0.6357
Model 2	100	1.1039 (0.0033)	1.2608 (0.0030)	1.0297	0.2858
($\theta = 1.4161$)	400	1.1806 (0.0026)	1.3053 (0.0019)	0.5886	0.1403

Table 1: MSE comparisons of FDH estimator and LFDH estimator.

5.2 Limit distribution vs. empirical distribution

We conducted another simulation study to evaluate the accuracy of our large sample approximation of the sampling distribution as described in Section 4 by Theorem 4.1. Under Model 2 in the previous subsection, 500 Monte Carlo experiments were done with the sample sizes $n = 400$ and $n = 1000$. We compared the empirical distribution of $n^{1/(p+q)}\{\hat{g}_{\mathrm{LFDH}}(z) - g(z)\}$ and the simulated distribution of $\hat{g}^*_{\mathrm{LFDH}}(0)$. Note that 2000 Monte Carlo simulations were done for the latter.

Figure 5 depicts the simulated survival function of $-\hat{g}^*_{\mathrm{LFDH}}(0)$ and the empirical survival function of $-n^{1/(p+q)}\{\hat{g}_{\mathrm{LFDH}}(z) - g(z)\}$, where we verify that our large sample approximation works quite well even with the moderate sample sizes $n = 400$ and $n = 1000$ in a 5 dimensional setup.

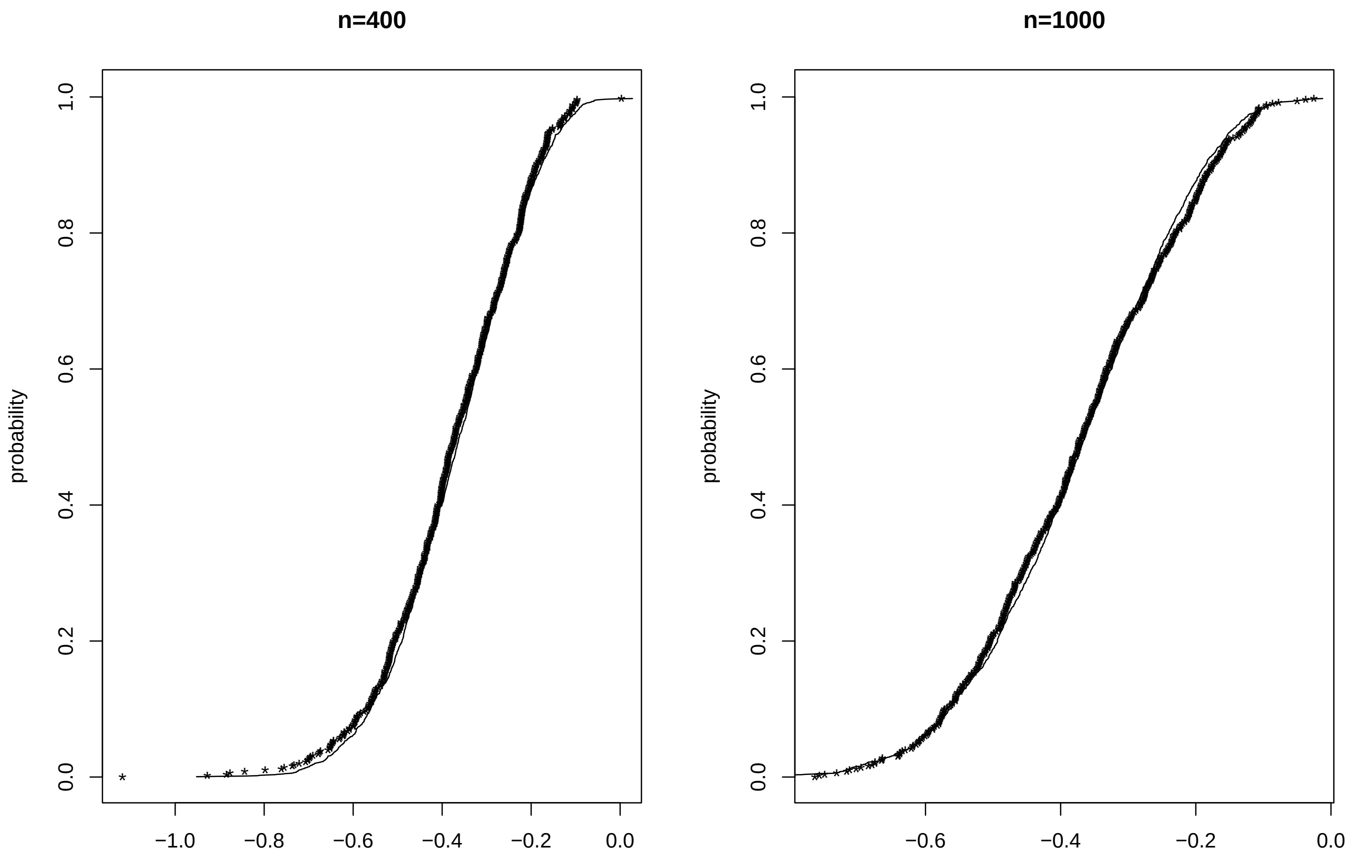

Figure 5: *Comparison of the simulated limit distribution (solid) and the empirical distribution (dotted) of LFDH estimates.*

6 Conclusion

In this paper we have proposed a new nonparametric estimator of monotone frontiers which allows for nonconvex sets below the frontier. This is particularly useful in the context of productivity analysis where the FDH estimator has been extensively used. However the nonsmoothness and discontinuities of the FDH is a drawback for conducting inference in finite samples. The bootstrap, even in a consistent version of it, does not provide a useful alternative (poor finite sample properties).

Our estimator, the LFDH, is a linearized version of the FDH, obtained by linear interpolation of the appropriate FDH-efficient vertex in the observed sample. It offers a continuous, smooth version of the FDH. We provide an algorithm for computing the estimator, and we establish its asymptotic properties. We also provide a easy way to approximate its asymptotic sampling distribution. The latter could offer bias-corrected estimator and confidence intervals of the efficiency scores.

In a Monte-Carlo study, we show that these approximations works well even in moderate sample sizes and that our LFDH estimator outperforms, both in bias and in MSE the original FDH estimator. Also, the gap between the asymptotic distribution and the empirical distribution diminishes faster as n increases.

The approach presented here could be extended if smoother version of the FDH would be wanted. We could use spline interpolation passing through the appropriate FDH-efficient vertices. However, we doubt the practical gain of such more elaborate smoothing procedure.

Finally, it should be noticed that the algorithm proposed above might be adapted to provide linear interpolation of emprical distribution functions of multidimensional random variables.

A Appendix: Subsampling FDH estimators

A.1 The consistency

We know that in this boundary estimation setup, the naive bootstrap (sampling with replacement, samples of size n from the original sample $\mathcal{X}_n$) is inconsistent (see Simar and Wilson, 2000 and the references reported there). Subsampling is a way to solve the problem. Let m be the size of each subsample such that $m/n \to 0$ and $m\to \infty$ as $n\to \infty$, that is, the subsample $\mathcal{X}^*_{n,m} = \{(X^*_i, Y^*_i), i = 1, \cdots, m\}$ is drawn randomly, without replacement, from the original sample $\mathcal{X}_n$. Let $\hat{\theta}^*_{n,m} = \hat{\theta}^*_{n,m}(x_0, y_0)$ be the FDH estimate of $\theta(x_0, y_0)$ obtained from the subsample $\mathcal{X}^*_{n,m}$. Write $\theta_0 = \theta(x_0, y_0)$ and $\hat{\theta}_n = \hat{\theta}_n(x_0, y_0)$ for brevity.

Theorem A.1. *Under the assumptions AI–AIII in Park et al. (2000), as $n\to \infty$ we have*

$$\sup_{z>0} \left| \Pr\left\{ n^{1/(p+q)} \left(\hat{\theta}_n/\theta_0 - 1 \right) \leq z \right\} \right.$$
$$\left. - \Pr\left\{ m^{1/(p+q)} \left(\hat{\theta}^*_{n,m}/\hat{\theta}_n - 1 \right) \leq z \,\middle|\, \mathcal{X}_n \right\} \right| \xrightarrow{p} 0. \tag{A.1}$$

Proof. Define $\mathrm{NW}(x, y) = \{(x', y') \in \mathrm{I\!R}^{p+q}_+ \,|\, x' < x, y' > y\} \cap \Psi$, and the events

$$E_n = \left[\mathrm{NW}(x^\partial_0(1 + n^{-1/(p+q)}z), y_0) \cap \mathcal{X}_n = \phi \right],$$
$$E^*_{n,m} = \left[\mathrm{NW}(x^\partial_0(1 + m^{-1/(p+q)}z), y_0) \cap \mathcal{X}^*_{n,m} = \phi \right].$$

Then, we may easily prove that for $z > 0$

$$\left| \Pr\left\{ n^{1/(p+q)} \left(\hat{\theta}_n/\theta_0 - 1 \right) > z \right\} - \Pr\left\{ E_n \right\} \right| \to 0;$$
$$\left| \Pr\left\{ m^{1/(p+q)} \left(\hat{\theta}^*_{n,m}/\theta_0 - 1 \right) > z \,\middle|\, \mathcal{X}_n \right\} - \Pr\left\{ E^*_{n,m} \,\middle|\, \mathcal{X}_n \right\} \right| \xrightarrow{p} 0$$

as $n\to \infty$. Since $m^{1/(p+q)}(\hat{\theta}_n \theta_0 - 1) \xrightarrow{p} 0$, it suffices to show that

$$\sup_{z>0} \left| \Pr\left\{ E_m \right\} - \Pr\left\{ E^*_{n,m} \,|\, \mathcal{X}_n \right\} \right| \xrightarrow{p} 0 \tag{A.2}$$

as $n\to \infty$.

Let $N_{n,m} = \binom{n}{m}$ be the number of subsamples, indexed by $j = 1, \cdots, N_{n,m}$, available from $\mathcal{X}_n$. And let $\mathcal{X}^*_{n,m,j}$ be the jth subsample, and $E^*_{n,m,j}$ be the event $E^*_{n,m}$ corresponding to $\mathcal{X}^*_{n,m,j}$. Then

$$\Pr\left\{ E^*_{n,m} \,|\, \mathcal{X}_n \right\} = \frac{1}{N_{n,m}} \sum_{j=1}^{N_{n,m}} I\left\{ E^*_{n,m,j} \right\},$$

which is a U-statistic with the mean $\Pr\{E_m\}$. Hence, by Hoeffding's inequality (see Serfling, 1980, Theorem A, p.201), $\left|\Pr\left\{E^*_{n,m}\,|\,\mathcal{X}_n\right\} - \Pr\{E_m\}\right| \xrightarrow{p} 0$. The uniformity in z of the convergence (A.2) is given by Polya's Theorem (see Serfling, 1980, p.18). □

The preceding argument is for subsampling without replacement, but asymptotically, subsampling with replacement does not make any difference.

A.2 Subsampling FDH in action

When the sample size n or the subsample size $m = n^\kappa$ is not large enough and the point of interest (x_0, y_0) is close to the frontier of the free disposal hull of the original sample $\mathcal{X}_n$, it may happen very often that the point (x_0, y_0) is not laid in the free disposal hull of the subsample $\mathcal{X}^*_{n,m}$, which may cause a problem to define the efficiency score and deteriorate the quality of resulting bootstrap distribution. But note that, for any positive real number a, $\theta(ax, y) = a^{-1}\theta(x, y)$, $\hat{\theta}_n(ax, y) = a^{-1}\hat{\theta}_n(x, y)$ and $\hat{\theta}^*_{n,m}(ax, y) = a^{-1}\hat{\theta}^*_{n,m}(x, y)$, so that we have

$$\hat{\theta}_n(ax, y)/\theta(ax, y) = \hat{\theta}_n(x, y)/\theta(x, y), \quad \hat{\theta}^*_{n,m}(ax, y)/\hat{\theta}_n(ax, y) = \hat{\theta}^*_{n,m}(x, y)/\hat{\theta}_n(x, y)$$

for all $a > 0$. By choosing a large enough for (ax_0, y_0) to be laid in the free disposal hull of each subsample and then computing $\hat{\theta}^*_{n,m}(ax_0, y_0)/\hat{\theta}_n(ax_0, y_0)$ instead of $\hat{\theta}^*_{n,m}(x_0, y_0)/\hat{\theta}_n(x_0, y_0)$, we can avoid the above technical difficulty. If the value of y_0 is too large (if $y_0 \geq, \neq \max\{y_j | (x_j, y_j) \in \mathcal{X}^*_{n,m}\}$), such value a does not exist. In such a case, we suggest to add the point of interest (x_0, y_0) to the bootstrap sample $\mathcal{X}^*_{n,m}$, this does not alter the asymptotic properties of the bootstrap. This trick is applicable to subsampling any other radial efficiency scores such as the DEA efficiency score as analyzed by Kneip, Simar and Wilson (2003).

Another practical difficulty is due to the fact that $N_{n,m}$ is very large in general. That is, it may be very difficult in practice to consider all possible $\mathcal{X}^*_{n,m,j}$, $j = 1, \ldots, N_{n,m}$. Therefore we consider the following Monte Carlo algorithm : Draw randomly a set of numbers $\{J_1, \cdots, J_B\}$ from $\{1, \cdots, N_{n,m}\}$ with or without relacement. Write $d^*_{n,m,b} = m^{1/(p+q)}(\hat{\theta}^*_{n,m,b}/\hat{\theta}_n - 1)$, where $\hat{\theta}^*_{n,m,b}$ is the value for $\hat{\theta}^*_{n,m}$ from the subsample $\mathcal{X}^*_{n,m,J_b}$, $b = 1, \ldots, B$. Then, the empirical distribution of $\{d^*_{n,m,b}, b = 1, \cdots, B\}$ approximates the exact sampling distribution of $n^{1/(p+q)}(\hat{\theta}_n/\theta - 1)$ given $\mathcal{X}_n$ as $B \to \infty$. For example, we can estimate the bias of FDH estimator by this approximation resulting the following bias corrected estimator:

$$\hat{\theta}_n \cdot \left\{1 + n^{-1/(p+q)} B^{-1} \sum_{b=1}^{B} d^*_{n,m,b}\right\}^{-1}.$$

Moreover we can construct a $100 \times (1-\alpha)\%$ confidence interval for θ_0 as follows:

$$\left[\hat{\theta}_n/(1+n^{-1/(p+q)}c^*_{1-\alpha/2,m}),\ \hat{\theta}_n/(1+n^{-1/(p+q)}c^*_{\alpha/2,m})\right]$$

where $c^*_{\alpha,m}$ denotes the α-th sample quantile of $\{d^*_{n,m,b}, b = 1, \cdots, B\}$.

A.3 Simulation study

We conducted some simulation studies following the settings in Section 5 in order to investigate the finite sample performances of the subsampling (with replacement) FDH estimators. From $M = 500$ Monte Carlo experiments, we computed the MSE of the bias-corrected estimator and the coverage probabilities of the confidence interval obtained from subsampling, which are summarized in Table 2 and Table 3. We considered the subsample sizes $m = n^\kappa$ for various κ in $0 < \kappa \leq 1$, and $B = 2000$ was used for each subsampling.

As is seen from the simulation results, while the subsampling worked fairly well for the bias-correction, the subsampling for building confidence intervals shows very poor performances. Indeed, the coverage probabilities obtained by the subsampling are not generally close to the nominal level even in the cases of the sample size of $n = 1000$ which is not small at all. These features are mainly due to the nature of FDH estimator rather than that of subsampling. Note that the subsampling approximates the continuous sampling distribution distribution by a discrete distribution, and that the value of FDH estimate is completely determined by only 'one' point of the FDH vertices of the data. In this situation the approximate discrete distribution tends to give too much probability mass on the point. Therefore, the approximation of the sampling distribution would have very poor accuracy particularly in the tail when the (sub)sample size is not large enough, which inherently gives the poor coverage accuracy of the confidence interval with large nominal probability such as 90%, 95% and 99%. Hence we may expect that the coverage probabilities would be very poor especially for the small values of κ, which is observed in Table 3. In the bias-correction we may expect the similar thing by the same reason as the above. By construction, the bias would be over-calibrated by the subsampling when κ is small (since large $d^*_{n,m,b}$'s would have large probability masses in the subsampling distribution), which results in the over-correction, see Table 2. The effect, however, is less crucial than that for the confidence interval, since it involves the average of the $d^*_{n,m,b}$'s not the tail distribution of the $d^*_{n,m,b}$'s.

	Model 1 ($\theta = 2.0027$)			
	$n = 100$		$n = 1000$	
κ	Mean (S.E.)	MSE	Mean (S.E.)	MSE
FDH	1.2892 (0.0085)	5.4480	1.6173 (0.0053)	1.6274
0.50	2.2612 (0.0253)	3.8720	2.1415 (0.0106)	0.7547
0.55	2.1827 (0.0243)	3.2625	2.0464 (0.0100)	0.5230
0.60	2.0724 (0.0228)	2.6383	1.9707 (0.0096)	0.4721
0.65	1.9549 (0.0211)	2.2473	1.9128 (0.0093)	0.5095
0.70	1.8266 (0.0190)	2.1093	1.8668 (0.0090)	0.5851
0.75	1.7374 (0.0175)	2.2361	1.8262 (0.0087)	0.6857
0.80	1.6558 (0.0161)	2.4910	1.7899 (0.0084)	0.8021
0.85	1.5814 (0.0148)	2.8658	1.7575 (0.0080)	0.9208
0.90	1.5228 (0.0136)	3.2277	1.7274 (0.0076)	1.0481
0.95	1.4731 (0.0126)	3.5974	1.7003 (0.0072)	1.1740
1.00	1.4286 (0.0116)	3.9693	1.6758 (0.0068)	1.2963

	Model 2 ($\theta = 1.4161$)			
	$n = 100$		$n = 1000$	
κ	Mean (S.E.)	MSE	Mean (S.E.)	MSE
FDH	1.0992 (0.0034)	1.0621	1.2267 (0.0020)	0.3787
0.50	1.3627 (0.0095)	0.4763	1.4524 (0.0037)	0.0818
0.55	1.3618 (0.0094)	0.4699	1.4215 (0.0036)	0.0650
0.60	1.3566 (0.0092)	0.4601	1.3835 (0.0034)	0.0694
0.65	1.3480 (0.0090)	0.4488	1.3498 (0.0033)	0.0976
0.70	1.3320 (0.0086)	0.4403	1.3239 (0.0032)	0.1356
0.75	1.3140 (0.0081)	0.4329	1.3048 (0.0031)	0.1715
0.80	1.2902 (0.0075)	0.4425	1.2899 (0.0030)	0.2039
0.85	1.2604 (0.0068)	0.4759	1.2775 (0.0029)	0.2339
0.90	1.2308 (0.0061)	0.5307	1.2662 (0.0028)	0.2628
0.95	1.2031 (0.0055)	0.6029	1.2561 (0.0026)	0.2900
1.00	1.1762 (0.0048)	0.6917	1.2471 (0.0024)	0.3154

Table 2: *Comparison of FDH estimator and its bias corrected versions. The values for MSE are multiplied by* 10^1.

	Model 1					
	$n = 100$			$n = 1000$		
κ	90%	95%	99%	90%	95%	99%
0.50	0.350	0.352	0.352	0.250	0.290	0.290
0.55	0.406	0.408	0.408	0.370	0.450	0.450
0.60	0.496	0.498	0.498	0.544	0.552	0.552
0.65	0.570	0.574	0.576	0.650	0.660	0.672
0.70	0.648	0.678	0.688	0.692	0.718	0.742
0.75	0.712	0.732	0.770	0.680	0.724	0.778
0.80	0.698	0.756	0.818	0.634	0.694	0.780
0.85	0.632	0.714	0.818	0.578	0.630	0.734
0.90	0.566	0.654	0.784	0.504	0.570	0.660
0.95	0.480	0.584	0.730	0.420	0.482	0.584
1.00	0.394	0.490	0.654	0.306	0.352	0.482

	Model 2					
	$n = 100$			$n = 1000$		
κ	90%	95%	99%	90%	95%	99%
0.50	0.552	0.552	0.552	0.280	0.360	0.360
0.55	0.554	0.554	0.554	0.432	0.502	0.502
0.60	0.564	0.564	0.564	0.672	0.674	0.674
0.65	0.588	0.588	0.588	0.810	0.818	0.822
0.70	0.632	0.632	0.632	0.806	0.858	0.892
0.75	0.660	0.664	0.664	0.694	0.792	0.890
0.80	0.682	0.708	0.708	0.540	0.634	0.810
0.85	0.698	0.740	0.764	0.450	0.520	0.644
0.90	0.670	0.768	0.816	0.350	0.406	0.528
0.95	0.576	0.710	0.832	0.268	0.308	0.408
1.00	0.444	0.600	0.802	0.176	0.210	0.304

Table 3: *Coverage probabilities of the confidence interval by subsampling FDH estimator*

References

Barber, C. B., Dobkin, D. P. and Huhdanpaa, H. (1996). The Quickhull algorithm for convex hulls, *The ACM Transactions on Mathematical Software* **22**, 469–483.

Brown, D. F. (1979). Voronoi diagrams from convex hulls, *Information Processing Letters* **9**, 223–228.

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

Debreu, G. (1951), The coefficient of ressource utilization,*Econometrica*, 19:3, 273-292.

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

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

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

Hall, P. and Park, B. U. (2002). New methods for bias correction at endpoints and boundaries, *The Annals of Statistics* **30**, 1460–1479.

Hall, P., Park, B. U. and Turlach, B. (2002). Rolling-ball method for estimating the boundary of the support of a point-process intensity, *Annales de l'Institut Henri Poincare-Probabilities et Statistiques* **38**, 959–971.

Hwang, J. H., Park, B. U. and Ryu, W. (2002). Limit theorems for boundary function estimators, *Statistics and Probability Letters* **59**, 353–360.

Jeong, S.-O. (2004). Asymptotic distribution of DEA efficiency scores, *Journal of the Korean Statistical Society*, to appear.

Jeong, S.-O. and Park, B. U. (2004). Large sample approximatiopn of the limit distribution of convex-hull estimators of boundaries, *Scandinavian Journal of Statistics*, to appear.

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

Kneip, A, L. Simar and P.W. Wilson (2003), Asymptotics for DEA Estimators in Nonparametric Frontier Models, Discussion paper #0317, Institut de Statistique, UCL, Belgium (http://www.stat.ucl.ac.be).

Koopmans, T.C.(1951), An analysis of production as an efficient combination of activities, in Koopmans, T.C. (ed) *Activity Analysis of Production and Allocation*, Cowles Commission for Research in Economics, Monograph 13, John-Wiley, New-York.

O'Rourke, J. (1998). *Computational Geometry in C*, Second Edition, Cambridge University Press.

Park, B. U. (2001). On estimating the slope of increasing boundaries, *Statistics and Probability Letters* **52**, 69–72.

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

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