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

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

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

0604

NONPARAMETRIC CONDITIONAL EFFICIENCY MEASURES: ASYMPTOTIC PROPERTIES

S.-O. JEONG, B.U. PARK and L. SIMAR

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

Nonparametric Conditional efficiency measures: asymptotic properties

by

Seok-Oh Jeong

Department of Statistics, Hankuk University of Foreign Studies, South Korea.

Byeong U. Park

Department of Statistics, Seoul National University, South Korea.

Léopold Simar

Institut de statistique, Université catholique de Louvain, Belgium.

E-mail: simar@stat.ucl.ac.be Phone: +32-10-474308

March 1, 2006

Abstract

Daraio and Simar (2005a, b) developed a conditional frontier model which incorporates the environmental factors into measuring the efficiency of a production process. They also provided the corresponding nonparametric efficiency measures: conditional FDH estimator, conditional DEA estimator and conditional order-m estimator. The aim of this paper is to provide an asymptotic analysis of the first two estimators.

Keywords Frontier model, environmental variable, conditional frontier, conditional DEA, conditional FDH, asymptotic distribution

Introduction

Performance of any production unit is quantified by the efficiency measures, which is of the primary interest in productivity analysis. The efficiency of a producer is usually defined by its distance to the frontier built by the best production scenario. The production scenario is composed by two factors, that is, input factors and output factors. For example, labor and capital are most typical input factors, and profit is an output counterpart. Recently environmental factors are considered at the same time for assessing the performance of a production unit properly. Environmental variables are exogenous factors which are neither inputs nor outputs of a production process, but affect the performance of the production process. For this, several approaches have been developed: see Banker and Morey (1986), Adolphson, Cornia and Walters (1991), Fried, Lovell and Vanden Eeckaut (1993), McCarty and Yaisawarng (1993), Bhattacharyya, Lovell and Sahay (1997), Fried, Schmidt and Yaisawarng (1999), Daraio and Simar (2005a, b). Among them, Daraio and Simar (2005a, b), the most recent one, suggested a fully nonparametric approach for frontier models with environmental variables, which overcomes drawbacks of other approaches. They defined a conditional frontier model and a conditional efficiency measure, and then proposed the corresponding nonparametric estimators such as conditional FDH, conditional DEA and conditional order-m estimators. This paper aims at providing the asymptotic distributions of the first two estimators. The aymptotics of the order-m estimator was analyzed in Cazals, Florens and Simar (2002) and Park, Jeong and Lee (2006).

This paper is organized as follows. In section 1 we provide a quick review on the frontier model and the nonparametric estimators such as FDH and DEA which are used for analyzing efficiency in productivity analysis. In section 2 we summarize a probabilistic formulation developed in Daraio and Simar (2005a, b), which is useful for introducing a conditional argument when defining the data generating process of the production process. In section 3, the basic idea and definition of conditional frontier and the corresponding nonparametric estimators are presented. Section 4 is devoted to an asymptotic analysis of the conditional estimators. Finally, section 5 concludes.

1 Frontier Analysis

1.1 The model

Suppose that activities of production units are characterized by pairs of inputs $x \in R_+^p$ and outputs $y \in R_+^q$. The production set Ψ is defined by the set of all those technically feasible pairs of (x, y):

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

It is very common in economics to assume Ψ be free disposable, which means that it is always technically feasible to produce less output using more input. Precisely, a set Ψ is said to be free disposable if $(x, y) \in \Psi$ implies $(x', y') \in \Psi$ for any (x', y') such that $x' \geq x$ and $y' \leq y$. Throughout this paper, inequalities between vectors are to be understood as component-wise. Also, convexity is often assumed for the shape of the production set Ψ, i.e., it is assumed that if $(x, y) \in \Psi$ and $(x', y') \in \Psi$ then $(\alpha x + (1-\alpha)x', \alpha y + (1-\alpha)y') \in \Psi$ for any $\alpha \in [0,1]$.

When the output is scalar, we may define a frontier function $g(\cdot)$ which forms the roof of the production set Ψ:

$$g(x) = \sup\{y \mid (x, y) \in \Psi\}, \quad x \in R_+^p.$$

Then the production set Ψ is characterized by the frontier function in such a way that

$$\Psi = \{(x, y) \in R_+^p \times R_+^1 \mid y \leq g(x)\}.$$

Free disposability of Ψ implies that the frontier function $g(x)$ is monotone increasing in x, and convexity of Ψ entails that g is a concave function of x.

Since the boundary of Ψ defines the locus of optimal production scenario, one may assess the efficiency of a given level of input and output by measuring its distance to the boundary of Ψ. Particularly when outputs are scalar, one may measure the efficiency by referring the frontier function g since the function g defines the boundary of Ψ. For example, the efficiency of a production unit $(x_0, y_0) \in \Psi \subset R_+^{p+1}$ can be measured by $g(x_0) - y_0$ or $g(x_0)/y_0$. However, when outputs are multiple, we cannot think of such a way to measure the efficiency. In that case it is convenient to define the efficiency scores in a radial way: given a level of input and output $(x_0, y_0) \in \Psi$,

(1) $\theta_0 = \theta(x_0, y_0) = \inf\{\theta > 0 \mid (\theta x_0, y_0) \in \Psi\}$

(2) $\lambda_0 = \lambda(x_0, y_0) = \sup\{\lambda > 1 \mid (x_0, \lambda y_0) \in \Psi\}$

Since θ_0 is the proportionate reduction of inputs for a production unit (x_0, y_0) to be technically efficient, it is called the input efficiency score. It is always less than or equal to one, and $\theta_0 = 1$ means that no proportionate reduction of inputs is available and (x_0, y_0) is efficient in terms of input-orientation. In parallel λ_0 is the technically

feasible proportionate increase of outputs for (x_0, y_0) to be efficient, and it is referred to the output efficiency score. It is always greater than or equal to one, and $\lambda_0 = 1$ indicates that (x_0, y_0) is efficient in terms of output-orientation. Note that θ_0 and λ_0 are the reciprocals of Shephard's input and output distance functions, respectively, see Shephard (1970).

By construction, both $\theta_0 x_0$ and $\lambda_0 y_0$ are laid on the boundary of Ψ. Therefore $\theta_0 x_0$ is the technically efficient input level for producing the output level y_0 among the input levels proportional to x_0. Similarly, $\lambda_0 y_0$ is the efficient output level produced by using input level x_0 among the output levels proportional to y_0.

1.2 Nonparametric estimation

Unfortunately, the production set Ψ is unknown in general. Hence we do not have any reference set for measuring efficiency in such a way as defined in the previous section. Instead, we observe a set of input and output levels performed by given production units:

$$S_n = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$$

which can be considered as a random sample drawn from a joint distribution (or density) of $(X, Y) \in R_+^{p+q}$ supported on a set D. We assume $\Psi \equiv D$ for technical convenience.

We are interested in estimating the frontier of the production set Ψ or efficiencies of a given production unit (x_0, y_0) based on S_n. Let us assume for the data generating process a deterministic frontier model for the identifiability. It means that no noise is allowed while observing S_n, which results in $P(S_n \subset \Psi) = 1$. Under this assumption, one is allowed to consider an idea of enveloping S_n in order to estimate Ψ. Among the existing methods for doing this, the data envelopment analysis (DEA) and the free disposal hull (FDH) estimators are the most popular nonparametric estimators.

Under the free disposability assumption on Ψ, Deprins, Simar and Tulkens (1984) proposed the FDH estimator defined as the minimal free disposable set which contains S_n:

$$\hat{\Psi}_{FDH} = \bigcup_{i=1}^{n} \{(x, y) \in R_+^{p+q} \mid x \geq X_i, y \leq Y_i\}.$$

Assuming convexity on Ψ as well as free disposability, the DEA estimator of Ψ is

defined as the smallest set containing S_n that are convex and free disposable:

$$\hat{\Psi}_{DEA} = \{(x,y) \in R_+^{p+q} \mid x \geq \sum_{i=1}^{n} \xi_i X_i,\ y \leq \sum_{i=1}^{n} \xi_i Y_i \ \textit{for some}\ \xi_i \geq 0, i = 1,2,\cdots,n$$

$$\textit{such that} \sum_{i=1}^{n} \xi_i = 1\}.$$

Farrell (1957) is considered as the first empirical study of DEA approach, and Charnes, Cooper and Rhodes (1978) popularized it by adopting a linear programming technique.

Using these estimates we can define corresponding efficiency scores of a production unit (x_0, y_0) as well, i.e.

$$\hat{\theta}(x_0, y_0) = \min\{\theta > 0 \mid (\theta x_0, y_0) \in \hat{\Psi}\},$$

$$\hat{\lambda}(x_0, y_0) = \max\{\lambda > 1 \mid (x_0, \lambda y_0) \in \hat{\Psi}\}.$$

When $\hat{\Psi} = \hat{\Psi}_{FDH}$, the resulting estimates of efficiency scores are the FDH efficiency scores. If $\hat{\Psi} = \hat{\Psi}_{DEA}$, then we get the DEA efficiency scores. Especially, it is easily seen that the FDH efficiency scores are re-expressed in an explicit form:

$$\hat{\theta}_{FDH}(x_0, y_0) = \min_{i: Y_i \geq y_0} \max_{1 \leq k \leq p} \frac{X_i^{(k)}}{x_0^{(k)}},$$

$$\hat{\lambda}_{FDH}(x_0, y_0) = \max_{i: X_i \leq x_0} \min_{1 \leq k \leq q} \frac{Y_i^{(k)}}{y_0^{(k)}}$$

where $a^{(k)}$ denotes the k-th component of a vector a. The DEA efficiency scores are expressed as

$$\hat{\theta}_{DEA}(x_0, y_0) = \min\{\theta > 0 \mid \theta x_0 \geq \sum_{i=1}^{n} \xi_i X_i,\ y_0 \leq \sum_{i=1}^{n} \xi_i Y_i \ \textit{for some}\ \xi_i \geq 0$$

$$\textit{such that} \sum_{i=1}^{n} \xi_i = 1\};$$

$$\hat{\lambda}_{DEA}(x_0, y_0) = \max\{\lambda \geq 1 \mid x_0 \geq \sum_{i=1}^{n} \xi_i X_i,\ \lambda y_0 \leq \sum_{i=1}^{n} \xi_i Y_i \ \textit{for some}\ \xi_i \geq 0$$

$$\textit{such that} \sum_{i=1}^{n} \xi_i = 1\}.$$

1.3 Statistical inference

Statistical inference on these efficiency scores is fully available. Park, Simar and Weiner (2000) showed that the FDH efficiency scores have the Weibull limit distribution. Kneip, Park and Simar (1996) proved the consistency of DEA efficiency scores in a quite general setup and obtained its rate of convergence. Gijbels et al. (1999) derived the explicit form of the limit distribution of the DEA estimator when the input and output variables are all scalar. Methods for approximating the sampling distribution of the DEA estimator in a general multidimensional setup were investigated by Kneip, Simar and Wilson (2003), Jeong (2004), Jeong and Park (2006). Jeong and Simar (2006) proposed a hybrid version of FDH and DEA, say LFDH, which is defined by interpolating the vertices of FDH. For a general review of statistical inference on nonparametric frontier models, see Simar and Wilson (2000), Park, Jeong and Lee (2006).

2 Probabilistic formulation of frontier models

In section 1.2, we pointed out that the production set Ψ can be identified by the support of the density of (X,Y). Precisely,

$$\Psi = \{(x,y) \in R_+^{p+q} \mid f_{XY}(x,y) > 0\}$$

where f_{XY} is the joint density of (X,Y). Define a probability function H_{XY} by

$$H_{XY}(x,y) = P(X \le x, Y \ge y).$$

Then, we may also assume the identity

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

which implies free disposability of Ψ. If the conditional probability

$$H_{X|Y}(x \mid y) = P(X \le x \mid Y \ge y)$$

exists, we may consider the following decomposition:

$$H_{XY}(x,y) = H_{X|Y}(x \mid y) S(y),$$

where S_Y denotes the survival function of Y, i.e. $S_Y(y) = P(Y \ge y)$. Likewise, conditioning on X, we have

$$H_{XY}(x,y) = H_{Y|X}(y \mid x) F_X(x)$$

if $H_{Y|X}(y \mid x) = P(Y \ge y \mid X \le x)$ exists, where F_X is the distribution function of X,

i.e., $F_X(x) = P(X \le x)$.

Now suppose the following identities for Ψ hold:

$$\Psi = \{(x,y) \in R_+^{p+q} \mid H_{X|Y}(x \mid y) > 0\} = \{(x,y) \in R_+^{p+q} \mid H_{Y|X}(y \mid x) > 0\}.$$

Then, together with (1) and (2), given a production unit (x_0, y_0) we have

(3) $$\theta(x_{0,}y_0) = \inf\{\theta > 0 \mid H_{X|Y}(\theta x_0 \mid y_0) > 0\};$$

(4) $$\lambda(x_{0,}y_0) = \sup\{\lambda > 0 \mid H_{Y|X}(\lambda y_0 \mid x_0) > 0\}.$$

Interestingly, replacing $H_{X|Y}$ and $H_{Y|X}$ by their corresponding empirical versions

(5) $$\hat{H}_{X|Y}(x \mid y) = \frac{\sum_{i=1}^{n} I(X_i \le x, Y_i \ge y)}{\sum_{i=1}^{n} I(Y_i \ge y)};$$

(6) $$\hat{H}_{Y|X}(y \mid x) = \frac{\sum_{i=1}^{n} I(X_i \le x, Y_i \ge y)}{\sum_{i=1}^{n} I(X_i \le x)}$$

in (3) and (4), we obtain the FDH efficiency scores $\hat{\theta}_{FDH}(x_0, y_0)$ and $\hat{\lambda}_{FDH}(x_0, y_0)$.

3 Conditional frontier model

3.1 Introduction

While comparing production units by assessing their efficiency measures, we are to have in mind environmental factors that might cause the difference in efficiency. Such environmental factors affect the production process indeed, but they are not under the control of production managers. Hence understanding how those environmental factors make the difference in efficiency is quite important for productivity analysis. For a detailed discussion on this topic, see the references cited in Introduction. In this section we introduce a conditional approach suggested by Daraio and Simar (2005a, b).

3.2 The model

For brevity we confine attention to the input-orientation case from now on. The output-orientation case can be treated in a very similar way. Extending the probabilistic formulation in section 2, Daraio and Simar (2005a, b) considered a general model that involves an environmental variable. Let $Z \in R^r$ denote the environmental variable. The basic idea of Daraio and Simar (2005a, b) is that, when the environmental variable takes the value of $Z = z_0$, the conditional distribution of (X, Y) given $Z = z_0$ still defines the data generating process which takes into account the exogenous environment represented by Z.

Given $Z = z_0$, let Ψ_{z_0} be the conditional production set, i.e.

$$\Psi_{z_0} = \left\{(x, y) \in R_+^{p+q} \mid f_{XY|Z}(x, y \mid z_0) > 0\right\} \tag{7}$$

where $f_{XY|Z}(x, y \mid z)$ denotes the conditional density of (X, Y) given $Z = z$. Putting

$$H_{X|YZ}(x \mid y, z) = P(X \le x \mid Y \ge y, Z = z),$$

we assume the following identity as in the previous section:

$$\Psi_{z_0} = \left\{(x, y) \in R_+^{p+q} \mid H_{X|YZ}(x \mid y, z_0) > 0\right\}. \tag{8}$$

Let (x_0, y_0, z_0) be a triple of input, output and environmental factor levels of a production unit. As in (3) and (4), the conditional efficiency score of a production unit (x_0, y_0, z_0) is defined by

$$\theta(x_0, y_0 \mid z_0) = \inf\left\{\theta > 0 \mid (\theta x_0, y_0) \in \Psi_{z_0}\right\} = \inf\left\{\theta > 0 \mid H_{X|YZ}(\theta x_0 \mid y_0, z_0) > 0\right\}. \tag{9}$$

3.3 Nonparametric estimation

With slight abuse of notation, let S_n denote a set of i.i.d. copies of $(X, Y, Z) \in R_+^p \times R_+^q \times R^r$:

$$S_n = \{(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \cdots, (X_n, Y_n, Z_n)\}.$$

Given a production unit (x_0, y_0, z_0), we are to estimate the conditional efficiency score $\theta(x_0, y_0 \mid z_0)$ using S_n.

Given $h>0$ such that $h\to 0$ and $nh^r\to\infty$ as $n\to\infty$, let $I(z_0,h)$ be the set of indices defined by $I(z_0,h)=\left\{i \,\middle|\, \|Z_i-z_0\|\le h/2\right\}$, where $\|a\|$ is the value of the norm for a vector a. Then we have an empirical version of $H_{X|YZ}(\cdot\,|\,\cdot,\cdot)$ given by

$$\hat{H}_{X|YZ}(x\,|\,y,z)=\frac{\sum_{i=1}^{n} I\left(X_i\le x, Y_i\ge y, \|Z_i-z\|\le h/2\right)}{\sum_{i=1}^{n} I\left(Y_i\ge y, \|Z_i-z\|\le h/2\right)}=\frac{\sum_{i\in I(z_0,h)} I\left(X_i\le x, Y_i\ge y\right)}{\sum_{i\in I(z_0,h)} I\left(Y_i\ge y\right)}. \tag{10}$$

The conditional FDH estimator is then obtained by plugging this empirical version of $H_{X|YZ}(\cdot\,|\,\cdot,\cdot)$ into (9):

$$\hat{\theta}_{FDH}(x_0,y_0\,|\,z_0)=\min\left\{\max_{1\le k\le p}\frac{X_i^{(k)}}{x_0^{(k)}}\,\middle|\, Y_i\ge y_0, i\in I(z_0,h)\right\}. \tag{11}$$

This is a version of the FDH estimator obtained only by the points taking its Z values in the neighborhood of z_0. Along this line, the conditional DEA estimator is given by

$$\hat{\theta}_{DEA}(x_0,y_0\,|\,z_0)=\min\left\{\theta>0\,|\,\theta x_0\ge\sum_{i\in I(z_0,h)}\xi_i X_i,\ y_0\le\sum_{i\in I(z_0,h)}\xi_i Y_i \ \textit{for some}\ \xi_i\ge 0 \right. \tag{12}$$
$$\left. \textit{such that}\ \sum_{i\in I(z_0,h)}\xi_i=1\right\}.$$

4 Statistical analysis of conditional FDH and DEA estimators

Rigorously speaking, the conditional FDH and DEA estimators in (11) and (12) do not target $\theta(x_0,y_0\,|\,z_0)$ in (9), but

$$\theta^h(x_0,y_0\,|\,z_0)=\inf\left\{\theta>0\,|\,(\theta x_0,y_0)\in\Psi_{z_0}^h\right\}$$

where

$$\Psi_{z_0}^h=\left\{(x,y)\in R_+^{p+q}\,|\,f_{XY|Z}^h(x,y\,|\,z_0)>0\right\}=\left\{(x,y)\in R_+^{p+q}\,|\,H_{X|YZ}^h(x\,|\,y,z_0)>0\right\},$$

$f_{XY|Z}^h(\cdot,\cdot\,|\,z)$ is the conditional density of (X,Y) given that $\|Z-z\|\le h/2$, and

$$H^h_{X|YZ}(x \mid y,z) = P\left(X \le x \mid Y \ge y, \|Z - z\| \le h/2\right).$$

Hence, we need the following conditions for a proper statistical analysis of the conditional FDH and DEA estimators.

Assumption 1F Both Ψ_{z_0} and $\Psi^h_{z_0}$ are free disposable.

Assumption 1D Both Ψ_{z_0} and $\Psi^h_{z_0}$ are free disposable and convex in R^{p+q}_+.

Assumption 2 As $n \to \infty$ it holds that

$$\theta^h(x_0, y_0 \mid z_0) - \theta(x_0, y_0 \mid z_0) = \begin{cases} o\left((nh^r)^{-1/(p+q)}\right) & \textit{for conditional FDH}; \\ o\left((nh^r)^{-2/(p+q+1)}\right) & \textit{for conditional DEA}. \end{cases}$$

Note that free disposability of Ψ_{z_0} and $\Psi^h_{z_0}$ in Assumption 1F and 1D is a direct consequence of the monotonicity of $H_{X|YZ}$ and $H^h_{X|YZ}$, respectively. Since the cardinality of $I(z_0, h)$ is proportional to nh^r, we may expect that the convergence rate of the conditional FDH and DEA estimator are $(nh^r)^{-1/(p+q)}$ and $(nh^r)^{-2/(p+q+1)}$, respectively. By virtue of Assumption 2, when we investigate the sampling distribution of the conditional FDH and DEA estimator, we only need to consider the limit behavior of the deviations

$$(nh^r)^{1/(p+q)} \left\{\hat{\theta}_{FDH}(x_0, y_0 \mid z_0) - \theta^h(x_0, y_0 \mid z_0)\right\}$$

and

$$(nh^r)^{2/(p+q+1)} \left\{\hat{\theta}_{DEA}(x_0, y_0 \mid z_0) - \theta^h(x_0, y_0 \mid z_0)\right\}$$

instead of

$$(nh^r)^{1/(p+q)} \left\{\hat{\theta}_{FDH}(x_0, y_0 \mid z_0) - \theta(x_0, y_0 \mid z_0)\right\}$$

and

$$(nh^r)^{2/(p+q+1)}\left\{\hat{\theta}_{DEA}(x_0,y_0\,|\,z_0)-\theta(x_0,y_0\,|\,z_0)\right\},$$

respectively.

In order to make the conditional FDH and DEA well-defined, it should be guaranteed that $\{(X_i,Y_i,Z_i)\,|\,i\in I(z_0,h)\}$ is, of course, not empty. Moreover, for proper asymptotic analysis, we need sufficiently many Z_i around z_0, which is endorsed by the following condition:

Assumption 3 Z has a continuous marginal density f_Z such that $f_Z(z_0)$ is bounded away from zero.

Proposition 1 *Given any finite integer* $C\geq 0$, *let* $E_{n,C}$ *be the event that the cardinality of* $I(z_0,h)$ *is less than or equal to* C. *Then, under Assumption 3,* $P(E_{n,C})$ *tends to zero as* $n\to\infty$.

Proof.

$$\begin{aligned}
P(E_{n,C}) &= \sum_{k=0}^{C} P\left(\textit{The cardinality of } I(z_0,h)=k\right)\\
&= \sum_{k=0}^{C}\frac{n!}{k!(n-k)!}P\left(\|Z-z_0\|\leq h/2\right)^k\left\{1-P\left(\|Z-z_0\|\leq h/2\right)\right\}^{n-k}\\
&\leq M\cdot n^C\left\{1-P\left(\|Z-z_0\|\leq h/2\right)\right\}^n \quad \text{for a constant } M>0\\
&= M\cdot n^C\exp\left\{-nh^r f_Z(z_0)\right\}\cdot\left\{1+o(1)\right\}=o(1). \qquad \blacksquare
\end{aligned}$$

This proposition ensures that we are provided sufficiently many data points in $\Psi_{z_0}^h$ as the sample size grows. Therefore an asymptotic analysis of the conditional FDH and DEA estimators can be justified. Next we investigate the sampling distribution of the conditional FDH and DEA estimators. For this we assume additionally:

Assumption 4 (X,Y,Z) has a joint density $f_{XYZ}(\cdot,\cdot,\cdot)$, and it is continuous on its support.

Assumption 5 For z in a neighborhood of z_0, the conditional density $f_{XY|Z}(\cdot,\cdot\,|\,z)$ of $(X,Y)\,|\,Z=z$ exists and it satisfies $f_{XY|Z}(\theta(x,y\,|\,z)x,y\,|\,z)>0$ for all (x,y,z) in a neighborhood of (x_0,y_0,z_0). Moreover, $f^h_{XY|Z}(\cdot,\cdot\,|\,z)$ converges to $f_{XY|Z}(\cdot,\cdot\,|\,z)$ as $n\to\infty$.

Assumption 6F $\theta(\cdot,\cdot\,|\,z_0)$ and $\theta^h(\cdot,\cdot\,|\,z_0)$ are continuously differentiable in a neighborhood of (x_0,y_0), and the elements of the vector of their first partial derivatives at (x_0,y_0) are all nonzero.

Assumption 6D $\theta(\cdot,\cdot\,|\,z_0)$ and $\theta^h(\cdot,\cdot\,|\,z_0)$ are twice continuously differentiable in a neighborhood of (x_0,y_0), and their Hessian matrices evaluated at (x_0,y_0) are positively definite.

The next theorem is the conditional version of the sampling distribution of the FDH estimator provided by Park, Simar and Weiner (2000):

Theorem 1 *Under Assumption 1F, 2-5 and 6F, the conditional FDH efficiency score* $\hat{\theta}_{FDH}(x_0,y_0\,|\,z_0)$ *in (11) has the Weibull limit distribution.*

Proof. Let c_{NW} be a positive constant and

$$NW^h_{z_0}(x,y)=\Psi^h_{z_0}\cap\left\{(u,v)\in R^{p+q}\,|\,u\le x,v\ge y\right\}.$$

For $t'=(nh^r)^{-1/(p+q)}t>0$,

$$
\begin{aligned}
&P\left(\hat{\theta}(x_0, y_0 \mid z_0) - \theta^h(x_0, y_0 \mid z_0) \geq t'\right) \\
&= P\left(\textit{No pair of } (X_i, Y_i) \in NW^h_{z_0}(t' x_0, y_0), \|Z_i - z_0\| \leq h/2\right) \\
&= \left\{1 - P\left((X,Y) \in NW^h_{z_0}(t' x_0, y_0) \;\middle|\; \|Z - z_0\| \leq h/2\right) P\left(\|Z - z_0\| \leq h/2\right)\right\}^n \\
&= \left\{1 - (t')^{p+q} c_{NW} f_{XY|Z}(x_0, y_0 \mid z_0) \cdot h^r f_Z(z_0)\right\}^n \times \{1 + o(1)\} \\
&= \exp\left\{-t^{p+q} c_{NW} f_{XYZ}(x_0, y_0, z_0)\right\} \times \{1 + o(1)\}
\end{aligned}
$$

■

Next, we present a large sample approximation procedure for the sampling distribution of the conditional DEA estimator. We point out that the following procedure is merely an extension of the procedure in Jeong (2004) based on a conditional argument. Let $P(x_0)$ be a $p \times (p-1)$ matrix whose columns form an orthonormal basis for $x_0^{\perp} = \left\{x \in R^p \mid x^T x_0 = x_0^T x = 0\right\}$. Consider the transformation which maps $x \in R^p$ to $(u^T, w)^T \in R^{p-1} \times R$:

$$
u = P(x_0)^T x; \quad w = \frac{x_0^T x}{|x_0|},
$$

where $|a|$ denotes the Euclidean norm of a vector a. This transform is one-to-one and its inverse is given by

$$
x = P(x_0)u + w\frac{x_0}{|x_0|}.
$$

Lemma 1 *Let g_0 and g_0^h be the functions defined by*

$$
g_0(u,v) = \inf\left\{w > 0 \;\middle|\; \left(P(x_0)u + w\frac{x_0}{|x_0|}, v + y_0\right) \in \Psi_{z_0}\right\},
$$

$$
g_0^h(u,v) = \inf\left\{w > 0 \;\middle|\; \left(P(x_0)u + w\frac{x_0}{|x_0|}, v + y_0\right) \in \Psi^h_{z_0}\right\}.
$$

Then we have

$$\theta(x_0, y_0 \mid z_0) = |x_0|^{-1} g_0(0_{p-1}, 0_q),$$

$$\theta^h(x_0, y_0 \mid z_0) = |x_0|^{-1} g_0^h(0_{p-1}, 0_q),$$

where 0_n *denotes the* n*-vector with all elements being zero. Moreover, under Assumption 1D, 2 and 6D, both* g_0 *and* g_0^h *are convex in* (u, v) *and it follows that*

$$g_0(0_{p-1}, 0_q) - g_0^h(0_{p-1}, 0_q) = o\left(\left(nh^r\right)^{-2/(p+q+1)}\right).$$

By this lemma, estimation of $\theta^h(x_0, y_0 \mid z_0)$ reduces to that of the convex function g_0^h at 0_{p-1+q}. Now consider the transformed data

$$S_n' = \left\{ (U_i, V_i, W_i, Z_i) \,\middle|\, \begin{pmatrix} U_i \\ V_i \end{pmatrix} = \begin{pmatrix} P(x_0)^T X_i \\ Y_i - y_0 \end{pmatrix}, W_i = \frac{x_0^T X_i}{|x_0|}, (X_i, Y_i, Z_i) \in S_n, i \in I(z_0, h) \right\}.$$

By the definition of g_0^h, (U_i, V_i, W_i) satisfies $W_i \geq g_0^h(U_i, V_i)$ for $i \in I(z_0, h)$. Hence, g_0^h can be estimated from the transformed data S_n'.

Lemma 2 *Define*

$$\hat{g}_0^h(u, v) = \min\left\{ \sum_{i \in I(z_0, h)} \xi_i W_i \,\middle|\, u = \sum_{i \in I(z_0, h)} \xi_i U_i, v = \sum_{i \in I(z_0, h)} \xi_i V_i \text{ for some } \xi_i \geq 0 \right.$$
$$\left. \text{such that } \sum_{i \in I(z_0, h)} \xi_i = 1 \right\}.$$

Then, with probability tending to one, it follows that

$$\hat{\theta}_{DEA}(x_0, y_0 \mid z_0) = |x_0|^{-1} \hat{g}_0^h(0_{p-1}, 0_q).$$

Thus, by Lemmas 1 and 2, for the sampling distribution of $\hat{\theta}_{DEA}(x_0, y_0 \mid z_0)$, we may investigate that of $\hat{g}_0^h(0_{p-1}, 0_q)$ instead.

Let ∇ denote the partial differential operator. Along the lines of Jeong (2004), consider the canonical transform on $\{(U_i, V_i, W_i) \mid i \in I(z_0, h)\}$ given by: for $i \in I(z_0, h)$

$$\begin{pmatrix} U_i^* \\ V_i^* \end{pmatrix} = \left(nh^r\right)^{1/(p+q+1)} \left(\frac{1}{2} G_{2,0}^h\right)^{1/2} \begin{pmatrix} U_i \\ V_i \end{pmatrix}$$

$$W_i^* = \left(nh^r\right)^{2/(p+q+1)} \left\{ W_i - g_{0,0}^h - g_{1,0}^{h\ T} \begin{pmatrix} U_i \\ V_i \end{pmatrix} \right\}$$

where $g_{0,0}^h = g_0^h(0_{p-1}, 0_q)$, $g_{1,0}^h = \nabla g_0^h(0_{p-1}, 0_q)$, and $G_{2,0}^h = \nabla^2 g_0^h(0_{p-1}, 0_q)$.

Lemma 3 *Let* $Conv(\cdot,\cdot \mid z_0, h)$ *be the lower boundary of the convex hull built by* $\{(U_i^*, V_i^*, W_i^*) \mid i \in I(z_0, h)\}$*:*

$$Conv(u^*, v^* \mid z_0, h) = \min \left\{ \sum_{i \in I(z_0,h)} \xi_i W_i^* \;\middle|\; u^* = \sum_{i \in I(z_0,h)} \xi_i U_i^*, v^* = \sum_{i \in I(z_0,h)} \xi_i V_i^* \right.$$

$$\left. \text{for some } \xi_i \geq 0 \text{ such that } \sum_{i \in I(z_0,h)} \xi_i = 1 \right\}.$$

Then, with probability tending to one it follows that

$$Conv(0_{p-1}, 0_q \mid z_0, h) = \left(nh^r\right)^{2/(p+q+1)} \left\{ \hat{g}_0^h(0_{p-1}, 0_q) - g_0^h(0_{p-1}, 0_q) \right\}.$$

By combining Lemmas 1, 2 and 3, we may show that $\left(nh^r\right)^{2/(p+q+1)} \left\{ \hat{\theta}_{DEA}(x_{0,} y_0 \mid z_0) - \theta(x_{0,} y_0 \mid z_0) \right\}$ and $|x_0|^{-1} Conv(0_{p-1}, 0_q \mid z_0, h)$ have the same limit distribution as $n \to \infty$. Note that, however, the sampling distribution of $Conv(0_{p-1}, 0_q \mid z_0, h)$ is not yet at hand. Next we present a procedure for the large sample approximation of the distribution of $Conv(0_{p-1}, 0_q \mid z_0, h)$.

Note that $\{(U_i^*, V_i^*, W_i^*) \mid i \in I(z_0, h)\}$ has the new lower boundary $w^* = g_0^{h^*}(u^*, v^*)$

in the coordinate system (u^*, v^*, w^*) such that

$$g_0^{h*}(u^*, v^*) = u^{*T}u^* + v^{*T}v^* + o(1)$$

uniformly for (u^*, v^*) in any compact set contained in $R^{p-1} \times R^q$, as $n \to \infty$.

Write f_0^{h*} for the conditional density of (U_i^*, V_i^*, W_i^*), in the coordinate system (u^*, v^*, w^*), given $\|Z - z_0\| \le h/2$. Then, via the change of variable technique, we have by Assumptions 4 and 5

$$\sup' \left| nh^r \det\left(G_{2,0}^h / 2\right)^{1/2} f_0^{h*}(u^*, v^*, w^*) - f_0^h \right| \to 0$$

for any $\varepsilon_n \downarrow 0$, where sup' denotes the supremum over (u^*, v^*, w^*) such that $\left|(u^*, v^*)\right| \le \left(nh^r\right)^{1/(p+q+1)} \varepsilon_n$ and $u^{*T}u^* + v^{*T}v^* \le w^* \le \left(nh^r\right)^{2/(p+q+1)} \varepsilon_n$, f_0^h is the conditional density of (U, V, W) at the boundary point $(0_{p-1}, 0_q, g_{0,0}^h)$ given $\|Z - z_0\| \le h/2$ which equals $f_{XY|Z}^h\left(x_0, y_0 \middle| z_0\right)$, and $\det(A)$ denotes the determinant of a matrix A.

Now we are ready to describe the procedure to simulate the limit distribution of $Conv(0_{p-1}, 0_q \mid z_0, h)$. Define $\kappa = \left\{\left(f_0^h\right)^2 \det\left(G_{2,0}^h / 2\right)\right\}^{1/(p+q+1)}$ and

$$B_\kappa = \left\{ (u^*, v^*, w^*) \middle| (u^*, v^*) \in \left[-\left(\sqrt{\kappa/2}\right)\left(nh^r\right)^{1/(p+q+1)}, \left(\sqrt{\kappa/2}\right)\left(nh^r\right)^{1/(p+q+1)} \right]^{p-1+q} \right.$$
$$\left. and\ u^{*T}u^* + v^{*T}v^* \le w^* \le u^{*T}u^* + v^{*T}v^* + \kappa\left(nh^r\right)^{1/(p+q+1)} \right\}.$$

Let $\lfloor a \rfloor$ be the nearest integer to $a \in R$. Consider a new random sample $\left\{(U_i^u, V_i^u, W_i^u) \middle| i = 1, 2, \cdots, \lfloor nh^r \rfloor\right\}$ which is generated from the uniform distribution on B_κ. The uniform density is equal to $\left(nh^r\right)^{-1} \kappa^{-(p+q+1)/2} = \left(nh^r\right)^{-1} \det\left(G_{2,0}^h / 2\right)^{-1/2} f_0^h$. Let $Conv^u(\cdot, \cdot \mid z_0, h)$ be the version of $Conv(\cdot, \cdot \mid z_0, h)$ built by the new sample

$$\left\{(U_i^u, V_i^u, W_i^u) \,\middle|\, i = 1, 2, \cdots, \lfloor nh^r \rfloor \right\}.$$

Lemma 4 *Under Assumption 1D, 2, 5 and 6D,* $Conv(0_{p-1}, 0_q \mid z_0, h)$ *and* $Conv^u(0_{p-1}, 0_q \mid z_0, h)$ *have the same limit distribution.*

Theorem 2 *Suppose Assumption 1D, 2, 5 and 6D hold. For* $z > 0$

$$P\left(\left(nh^r\right)^{2/(p+q+1)} \left\{\hat{\theta}_{DEA}(x_0, y_0 \mid z_0) - \theta(x_0, y_0 \mid z_0)\right\} \le z\right) \to F(z)$$

as $n \to \infty$, *where*

$$F(z) = \lim_{n\to\infty} P\left(\left|x_0\right|^{-1} Conv^u(0_{p-1}, 0_q \mid z_0, h) \le z\right).$$

5 Concluding remarks

In this paper, we analyzed the asymptotics of the conditional FDH and DEA estimators. We established consistency of those estimators and obtained their proper limit distributions. By means of these results, we are able to correct their biases and construct confidence intervals for use in practice. However, as is typically observed in nonparametric function estimation problems, these procedures require additional information that depends on unknown quantities. In particular, a further statistical inference with the conditional FDH estimator based on its asymptotic properties may suffer from a severe departure of its finite sample properties from the asymptotic results, which was already pointed out in Park, Simar and Weiner (2000), Jeong and Simar (2006), and others. To avoid this problem, it is natural to consider a bootstrapping idea. For example, for the choice of the bandwidth h, one may use the minimizer of a consistent bootstrap approximation of $E\left(\hat{\theta}(x_0, y_0 \mid z_0) / \theta(x_0, y_0 \mid z_0) - 1\right)^2$. Any detailed study on this is left for future research.

Notes

Byeong U. Park's work was supported by SRC/ERC program of MOST/KOSEF (R11-2000-073-00000). Léopold Simar gratefully acknowledges the research support from the "Interuniversity Attraction Pole", Phase V (No. P5/24) from the Belgian Science Policy.

References

Adolphson, D. L., G. C. Cornia, and L. C. Walters. (1991). "A unified framework for classifying DEA models," Operational Research 90, 647-657.

Banker, R. D. and R. C. Morey. (1986). "Efficiency analysis for exogenously fixed inputs and outputs," Operations Research 34, 513-521.

Bhattacharyya, A., C. A. K. Lovell, and P. Sahay. (1997). "The impact of liberalization on the productive efficiency of Indian commercial banks," European Journal of Operational Research 98, 332-347.

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

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

Daraio, C. and L. Simar. (2005a). "Introducing environmental variables in nonparametric frontier models: a probabilistic approach," Journal of Productivity Analysis 24, 93-121.

Daraio, C. and L. Simar. (2005b). "Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach," Discussion Paper #0502, Institut de Statistique, UCL, Belgium (http://www.stat.ucl.ac.be).

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

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

Fried, H. O., C. A. K. Lovell, and P. Vanden Eeckaut. (1993). "Evaluating the performance of U. S. credit unions," Journal of Banking and Finance 17, 251-265.

Fried, H. O., S. S. Schmidt, and S. Yaisawarng. (1999). "Incorporating the operating environment into a nonparametric measure of technical efficiency," Journal of Productivity Analysis 12, 249-267.

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

Jeong, S. -O. (2004). "Asymptotic distribution of DEA efficiency scores," Journal of the Korean Statistical Society 33, 449-458.

Jeong, S. -O., and B. U. Park. (2006). "Large sample approximation of the limit distribution of convex-hull estimators of boundaries," Scandinavian Journal of Statistics 33, 139-151.

Jeong, S.-O., and L. Simar. (2006). "Linearly interpolated FDH efficiency score for nonconvex frontiers," Journal of Multivariate Analysis, in print.

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).

McCarty, T. and S. Yaisawarng. (1993). "Technical efficiency in New Jersey school districts." In H. O. Fried, C. A. K. Lovell, and S. S. Schmidt (eds.), The Measurement of Productivity Efficiency: Techniques and Applications. New York:Oxford University Press.

Park, B. U., S. –O. Jeong, and Y. K. Lee. (2006). "Nonparametric estimation of production efficiency," to appear in the volume in honor of Peter Bickel's 65th birthday.

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

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

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.