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

0425

ASYMPTOTIC DISTRIBUTION OF DEA EFFICIENCY SCORES

JEONG S.-O.

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

Asymptotic distribution of DEA efficiency scores*

S.-O. Jeong†

September 8, 2004

Abstract

Data envelopment analysis (DEA) estimators have been widely used in productivity analysis. The asymptotic distribution of DEA estimator derived by Kneip, Simar and Wilson (2003) is too complicated and abstract for analysts to use in practice, though it should be appreciated in its own right. This paper provides another way to express the limit distribution of the DEA estimator in a tractable way.

KEYWORDS. Productivity analysis, frontier model, data envelopment analysis, DEA, efficiency scores, asymptotic distribution.

1 Introduction

Let Ψ be the production set, *i.e.* $\Psi = \{(x, y) \,|\, x \text{ can produce } y\}$. Suppose (x, y) represents a pair of input and output exhibited by a production unit during a period. In productivity analysis one is interested in measuring the efficiency of the production unit. The relative efficiency of this production unit can be measured by the (radial) distance from (x, y) to the frontier of the production set Ψ. But, since the production set is generally unknown, we have to estimate it using the observed pairs of input and output. A natural idea to estimate the production set is to envelop or wrap the observed data

* 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, Belgium (e-mail : seokoh@stat.ucl.ac.be)

points. In fact, when the production set is convex and free disposable, the data envelopment analysis (DEA) provides an optimal estimator in the minimax sense, see Korostelev, Simar and Tsybakov (1995). By measuring the distance from (x, y) to the frontier of DEA estimate of Ψ along a given direction of interest, we may obtain the DEA estimate of the efficiency score at (x, y).

Statistical properties of the DEA efficiency score are now available in the literature. Kneip, Park and Simar (1998) obtained the consistency and the convergence rate of DEA estimator of efficiency score in a very general setup. Gijbels, Mammen, Park and Simar (1999) derived the explicit formula for the asymptotic distribution of the DEA estimator when the inputs and the outputs are scalar. Jeong and Park (2004) extended this result to a general case with multiple inputs and scalar outputs. Kneip, Simar and Wilson (2003) derived the asymptotic distribution of DEA estimator in the case that both inputs and outputs are multidimensional. But the asymptotic representation in Kneip, Simar and Wilson (2003) is not manageable due to its abstract and complex expression. The aim of this paper is to provide a tractable way to express the asymptotic distribution of DEA estimator.

2 DEA Estimator

Let Ψ be the production set, *i.e.* the set of feasible pairs of input x and output y exhibited by production units during a period:

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

It is helpful in multidimensional situation to describe the production set by its sections. The input set for a given output level y, $X(y)$, is defined by the set of all possible inputs producing the output y:

$$X(y) = \{x \,|\, (x, y) \in \Psi\}.$$

On the other hand, the output set for a given input level x is defined by the set of all possible outputs from the input x:

$$Y(x) = \{y \,|\, (x, y) \in \Psi\}.$$

The production set Ψ is generally assumed to be closed and convex, so that $X(y)$ is closed and convex for all $y \in \mathbb{R}_+^q$ and $Y(x)$ is closed, convex and

bounded for all $x \in \mathbb{R}^p_+$. Hence the following radial efficiency measures are well-defined:

$$\theta(x,y) = \inf\{\theta > 0 \,|\, \theta x \in X(y)\} \equiv \inf\{\theta > 0 \,|\, (\theta x, y) \in \Psi\};$$
$$\lambda(x,y) = \sup\{\lambda \geq 1 \,|\, \lambda y \in Y(x) \equiv \sup\{\lambda \geq 1 \,|\, (x, \lambda y) \in \Psi\}.$$

Note that, for $(x,y) \in \Psi$, $0 < \theta(x,y) \leq 1$ and $\lambda(x,y) \geq 1$. If $\theta(x,y) = 1$ holds, then the (x,y) is considered as efficient in terms of input. Similarly $\lambda(x,y) = 1$ holds, the point is efficient in terms of output. From now on, we consider only input efficiency measure $\theta(x,y)$ to save space. All the consequent results are valid for the output efficiency measure $\lambda(x,y)$ but a few changes in notations.

In practice Ψ and $\theta(x,y)$ are not observable, and hence we have to estimate them from a sample of observations. Let $\mathcal{X}_n = \{(X_i, Y_i),\ i = 1, \ldots, n\}$ be an observed sample. Fix a point $(x,y) \in \mathbb{R}^{p+q}_+$ of interest.

ASSUMPTION 1. Ψ is convex and free disposable, *i.e.*

(a) If $(x_1, y_1) \in \Psi$ and $(x_2, y_2) \in \Psi$, then

$$(\alpha x_1 + (1-\alpha)x_2,\ \alpha y_1 + (1-\alpha)y_2) \in \Psi \text{ for all } \alpha \in [0,1],$$

(b) If $(x,y) \in \Psi$, then $(\tilde{x}, \tilde{y}) \in \Psi$ for $\tilde{x} \geq x$ and $\tilde{y} \leq y$.

Under Assumption 1, a natural estimator for Ψ is the smallest convex and free disposable set contaning the observed sample $\mathcal{X}_n$, which is called the data envelopment analysis (DEA) estimator in the literature. Precisely, the DEA estimator $\widehat{\Psi}_{\text{DEA}}$ for Ψ is defined by:

$$\widehat{\Psi}_{\text{DEA}} = \{(x,y) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+ \,|\, x \geq \sum_{i=1}^n \xi_i X_i,\ y \leq \sum_{i=1}^n \xi_i Y_i \ \text{ for some } (\xi_1, \ldots, \xi_n)$$
$$\text{such that } \sum_{i=1}^n \xi_i = 1,\ \xi_i \geq 0,\ i = 1, \ldots, n\}.$$

And the DEA (input) efficiency score is given by

$$\hat{\theta}_{\text{DEA}}(x,y) = \min\{\theta > 0 \,|\, \theta x \geq \sum_{i=1}^n \xi_i X_i,\ y \leq \sum_{i=1}^n \xi_i Y_i \ \text{ for some } (\xi_1, \ldots, \xi_n)$$
$$\text{such that } \sum_{i=1}^n \xi_i = 1,\ \xi_i \geq 0,\ i = 1, \ldots, n\}.$$

3 Main Results

3.1 Limit distribution

To derive the limit distribution of the DEA efficiency score we need the following assumptions:

ASSUMPTION 2. (a) (X_i, Y_i)'s are iid with a density f having its support $\mathcal{D} \subset \Psi$, and $f(x, y) = 0$ for $(x, y) \notin \mathcal{D}$. (b) The density f is continuous on $\mathcal{D}$ and $f(\theta(x, y)x, y) > 0$ for all $(x, y) \in \mathcal{D}$.

ASSUMPTION 3. (a) For (x, y) in the interior of $\mathcal{D}$, $\theta(\cdot, \cdot)$ is twice continuously differentiable in a sufficiently small neighborhood of (x, y). (b) The Hessian matrix of $\theta(\cdot, \cdot)$ at (x, y) is positive definite.

Fix a point $(x_0, y_0) \in \mathbb{R}_+^{p+q}$ of interest, and let $\{v_j \,|\, j = 1, \cdots, p-1\}$ be an orthonormal basis for $x_0^\perp = \{t \in \mathbb{R}^p \,|\, t'x_0 = x_0't = 0\}$. Consider a transformation r_{x_0} from $\mathbb{R}_+^p$ to $\mathbb{R}^p$:

$$r_{x_0} : t \mapsto \left(t'v_1,\ t'v_2,\ \cdots,\ t'v_{p-1},\ \frac{t'x_0}{\sqrt{x_0'x_0}} \right).$$

Note that $(r_{x_0}^{(1)}, \cdots, r_{x_0}^{(p-1)})$ is the coefficient vector in the new coordinate system span$[v_1, \cdots, v_{p-1}]$ and $r_{x_0}^{(p)}$ is the distance between t and $x_0^\perp$. r_{x_0} is a one-to-one transformation, and $r_{x_0}(x_0) = \left(0, \cdots, 0, \sqrt{x_0'x_0}\right)'$. Moreover, in case of $p = 1$, we have $r_{x_0}(t) = t$ for all $t \in \mathbb{R}_+$. For each observation $(x_i, y_i) \in \mathcal{X}_n$, apply a transform $(x_i, y_i) \mapsto (z_i, w_i)$:

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

for $i = 1, \cdots, n$. Then, in the new coordinate system, (z_i, w_i)'s are laid on the region

$$\left\{ (z, w) \in \mathbb{R}^{p-1+q} \times \mathbb{R}_+ \,|\, w \geq g(z) \right\}$$

where $g(\cdot) = g(\cdot | x_0, y_0)$ is a convex function, see Lemma 1(b) in Kneip, Simar and Wilson (2003). Note that

$$\theta(x_0, y_0) = \frac{g(\mathbf{0})}{\sqrt{x_0'x_0}}.$$

Define the convex hull estimator of g at $z \in \mathbb{R}^{p-1+q}$ as

$$\hat{g}_{\text{conv}}(z) = \min \left\{ \sum_{i=1}^{n} \xi_i w_i \,\middle|\, \sum_{i=1}^{n} \xi_i z_i = z \ \text{ for some } (\xi_1, \ldots, \xi_n) \right.$$
$$\left. \text{such that } \sum_{i=1}^{n} \xi_i = 1,\ \xi_i \geq 0,\ i = 1, \ldots, n \right\}.$$

Then it can be easily checked that, with probability tending to 1,

$$\hat{\theta}_{\text{DEA}}(x_0, y_0) = \frac{\hat{g}_{\text{conv}}(\mathbf{0})}{\sqrt{x_0' x_0}}.$$

Hence we have the following lemma.

Lemma 1 *Under Assumption 1–3, the limit distribution of*

$$n^{2/(p+q+1)} \{\hat{\theta}_{\text{DEA}}(x_0, y_0) - \theta(x_0, y_0)\}$$

is the same as that of

$$n^{2/(p+q+1)} (x_0' x_0)^{-1/2} \{\hat{g}_{\text{conv}}(\mathbf{0}) - g(\mathbf{0})\}.$$

Now the problem is to derive the limit distribution of the convex hull estimator for a convex function g on $\mathbb{R}^{p-1+q}$. This is a good news, because the limit distribution of convex hull estimators is already available in Jeong and Park (2004). Omitting the detailed proofs which are very similar to those in Jeong and Park (2004), we describe the way to obtain the limit distribution of $n^{2/(p+q+1)} \{\hat{g}_{\text{conv}}(\mathbf{0}) - g(\mathbf{0})\}$ in the followings. Consider a linear transformation taking (z_i, w_i) to

$$\widetilde{z_i} = n^{1/(p+q+1)} \|g_2\|^{1/2} z_i$$
$$\widetilde{w_i} = n^{2/(p+q+1)} \{w_i - g_0 - g_1' z_i\}$$

where $\|\cdot\|$ is the determinant of a matrix,

$$g_0 = g(\mathbf{0}),$$

$$g_1 = \nabla g(\mathbf{0}),$$

and

$$g_2 = \frac{1}{2}\nabla^2 g(\mathbf{0}).$$

In the new coordinate system $(\widetilde{z}, \widetilde{w})$, the transformed data points now have as their frontier the surface with the equation

$$\widetilde{w} = \widetilde{z}'\widetilde{z} + o(1)$$

uniformly on any compact set of $\widetilde{z}$. The density $\tilde{f}$ of the transformed sample points is approximated by $n^{-1}\|g_2\|^{-1/2} f_0$ uniformly in the region

$$\left\{(\widetilde{z}, \widetilde{w}) \,|\, \sqrt{\widetilde{z}'\widetilde{z}} \le \varepsilon_n n^{1/(p+q+1)},\ \widetilde{z}'\widetilde{z} \le \widetilde{w} \le \varepsilon_n n^{2/(p+q+1)}\right\}$$

for any sequence $\varepsilon_n \to 0$. Define $\kappa = (\|g_2\|/f_0^2)^{1/(p+q+1)}$. Consider a new sample from the uniform distribution on the region

$$\left\{(\widetilde{z}, \widetilde{u}) \,|\, \widetilde{z} \in \left[-(\sqrt{\kappa}/2)n^{1/(p+q+1)}, (\sqrt{\kappa}/2)n^{1/(p+q+1)}\right]^{p+q-1} \right.$$
$$\left. \text{and } \widetilde{z}'\widetilde{z} \le \widetilde{u} \le \widetilde{z}'\widetilde{z} + \kappa n^{2/(p+q+1)} \right\}$$

on which the uniform density equals $n^{-1}\kappa^{-(p+q+1)/2} = n^{-1}\|g_2\|^{-1/2} f_0$. Let $\widetilde{g}_{\mathrm{conv}}$ be the version of $\widehat{g}_{\mathrm{conv}}$ obtained by the new sample.

Lemma 2 *Under Assumption 1–3,* $\widetilde{g}_{\mathrm{conv}}(\mathbf{0})$ *and* $n^{2/(p+q+1)}\{\hat{g}_{\mathrm{conv}}(\mathbf{0}) - g(\mathbf{0})\}$ *have the same limit distribution.*

By Lemma 1 and Lemma 2 we finally have the following theorem.

Theorem 1 *Under Assumption 1–3,*

$$n^{2/(p+q+1)}\{\hat{\theta}_{\mathrm{DEA}}(x_0, y_0) - \theta(x_0, y_0)\}$$

and

$$(x_0' x_0)^{-1/2}\widetilde{g}_{\mathrm{conv}}(\mathbf{0})$$

have the same limit distribution.

Note that, once κ has been determined, the distribution of $\widetilde{g}_{\text{conv}}(\mathbf{0})$ can be simulated by Monte Carlo method. Based on the simulated distribution, we may define a bias-corrected estimator and a confidence interval for $\theta(x_0, y_0)$, see Section 3 in Jeong and Park (2004) for details. In the next subsection we discuss the estimation of the unknown κ.

3.2 Estimation of parameters

Recall $\kappa = (\|g_2\|/f_0^2)^{1/(p+q+1)}$. Hence we are to estimate g_2 and f_0 to get the estimate of κ using transformed data (z_i, w_i)'s. For this, we naturally use the analogues of the estimators in Jeong and Park (2004). Consider the hypercube

$$\mathcal{C}(\mathbf{0}, \delta) = \left[-\frac{\delta}{2},\ \frac{\delta}{2}\right]^{p-1+q}$$

in $\mathbb{R}^{p-1+q}$ for some $\delta > 0$. Let

$$\mathcal{D}(\mathbf{0}, \delta) = \{(z, w) \,|\, z \in \mathcal{C}(\mathbf{0}, \delta),\ \hat{g}_{\text{conv}}(z) \le w \le \hat{g}_{\text{conv}}(\mathbf{0}) + \delta\}.$$

Define the estimator of f_0 by

$$\widehat{f_0} = \frac{\sum_{i=1}^{n} I[(z_i, w_i) \in \mathcal{D}(\mathbf{0}, \delta)]}{n\mu(\mathcal{D}(\mathbf{0}, \delta))}$$

where $\mu(\cdot)$ is the Lebesgue measure in $\mathbb{R}^{p+q}$. Now we discuss the estimation of g_2. Take $h > 0$ and define

$$\mathcal{X}_b(\mathbf{0}, h) = \{(z_i, \hat{g}_{\text{conv}}(z_i)) \,|\, z_i \in \mathcal{C}(\mathbf{0}, h)\} \cup \{(\mathbf{0}, \hat{g}_{\text{conv}}(\mathbf{0}))\}.$$

Fit the second order polynomial regression surface with the points in $\mathcal{X}_b(\mathbf{0}, h)$ by the least square method to get

$$\breve{g}(z, h) = \breve{g_0} + \breve{g_1}'z + z'\breve{g_2}z.$$

Then the matrix $\breve{g_2}$ is used for the estimator of g_2.

Using $\widehat{\kappa} = \left(\|\breve{g_2}\| \Big/ \widehat{f_0}^{\,2}\right)^{1/(p+q+1)}$ to simulate the distribution of $\widetilde{g}_{\text{conv}}(\mathbf{0})$, we may construct the bias-corrected estimator and the confidence intervals in the same way as in Jeong and Park (2004). In the next section we investigated the finite sample properties of the bias-corrected estimator and the confidence interval by a simulation study.

TABLE 4.1 *Summary of the simulation results*

	Mean squared error		*Coverage probability*	
(δ, h)	*DEA*	*bias-corrected*	*90%*	*95%*
(5.0, 5.0)	4.848×10^{-3}	2.702×10^{-4}	90.9%	94.1%

4 Numerical Study

To validate our large sample approximation in practice, we evaluated the finite sample performances of the bias-corrected estimator and the confidence interval defined in the previous section by a simulation study with the same data generating process (DGP) as in Kneip, Simar and Wilson (2003):

- $p = q = 2$.
- $x_{1e} \sim \text{Uniform}[10, 20]$, $x_{2e} \sim \text{Uniform}[10, 20]$.
- $y_1 = x_{1e}^{0.4} x_{2e}^{0.4} \cos\omega$, $y_2 = x_{1e}^{0.4} x_{2e}^{0.4} \sin\omega$, $\omega \sim \text{Uniform}[\frac{1}{9}\frac{\pi}{2}, \ \frac{8}{9}\frac{\pi}{2}]$.
- $x_1 = x_{1e} e^{0.2|\varepsilon|}$, $x_2 = x_{2e} e^{0.2|\varepsilon|}$, $\varepsilon \sim \text{N}(0, 1)$.

Under this DGP, 1000 Monte Carlo experiments were done with the sample size $n = 100$. On each experiment we took $x_0 = (20.69, 20.69)$ and $y_0 = (5.59, 5.59)$ for the point where the efficiency is measured. The true efficiency score at this point is 0.6. We estimated the mean squared error of the DEA estimator and that of the bias-corrected estimator. And the coverage probabilities of the confidence intervals at the nominal levels 90% and 95% were also computed. The results are summarized in Table 4.1, which proclaims that the proposed large sample approximation really works even with a small sample size of $n = 100$ in the $p + q = 4$ dimensional space. Figure 4.1 depicts the simulated distribution of DEA estimator (thin solid) and its bias-corrected version (thick dashed), which clearly shows that the bias-correction based on the proposed approach is valid. We point out that that the accurate estimation of κ is crucial for the proposed approach and that it is sensitive to the choice of smoothing parameters δ and h. We do not go further on this issue, which is left for a future work.

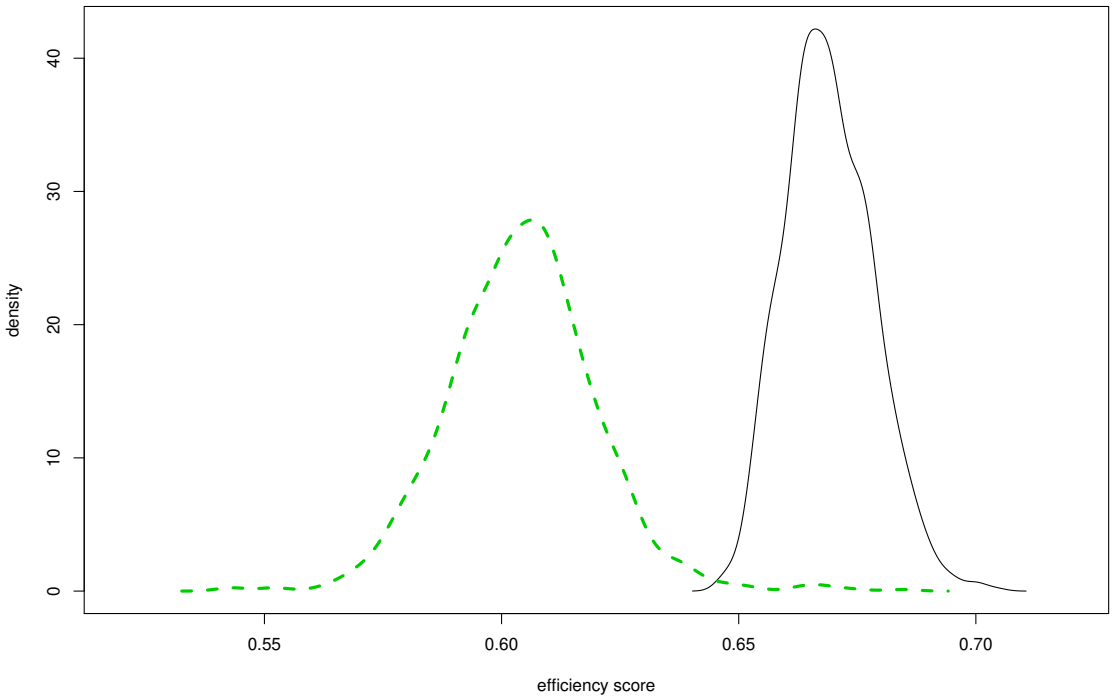

FIGURE 4.1 *The simulated distribution of DEA estimator (thin solid) and its bias-corrected version (thick dashed). The true efficiency score is equal to* 0.6.

References

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.

JEONG, S.-O. AND PARK, B. U. (2004). Limit distribution of convex-hull estimators of boundaries, Discussion Paper 0424, Institut de statistique, Université catholique de Louvain, Belgium.

KNEIP, A., PARK, B. U. AND SIMAR, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores, *Econometric Theory*, **14**, 783–793.

Kneip, A., Simar, L. and Wilson, P. (2003). Asymptotics for DEA estimators in non-parametric frontier models, Discussion Paper 0317, Institut de statistique, Université catholique de Louvain.

Korostelev, A. P., Simar, L. and Tsybakov, A. B. (1995). On estimation of monotone and convex boundaries, *Publications de l'Institut de statistique des Universités de Paris*, **XXXIX**, 3–18.