On testing equality of distributions of technical efficiency scores

Simar, Leopold and Zelenyuk, Valentin

Institut de Statistique, Université Catholique de Louvain, Belgium, EERC, UPEG

14 December 2004

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

0434

ON TESTING EQUALITY OF DISTRIBUTIONS OF TECHNICAL EFFICIENCY SCORES

SIMAR, L. and V. ZELENYUK

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

On Testing Equality of Distributions of Technical Efficiency Scores

Léopold Simar [*] **and Valentin Zelenyuk** [**]

Abstract

The challenge of the econometric problem in production efficiency analysis is that the very efficiency scores to be analyzed are *unobserved*. Recently, statistical properties have been discovered for a class of estimators popular in the literature, known as data envelopment analysis (DEA) approach. This opens a wide range of possibilities for a well-grounded statistical inference about the true efficiency scores from their DEA-estimates. In this paper we investigate possibility of using existing tests for *equality of two distributions* for such a context. Considering statistical complications pertinent to our context, we consider several approaches to adapt the Li (1996) test to the context and explore their performance in terms of the size and the power of the test in various Monte Carlo experiments. One of these approaches showed good performance both in the size and in the power, thus encouraging for its wide use in empirical studies.

Keywords: Kernel Density Estimation and Tests, Bootstrap, DEA.

Last Revision: December 14, 2004

* Institut de Statistique, Université Catholique de Louvain, Belgium.
Mailing address: 20 voie du Roman Pays, B 1348, Louvain-la-Neuve, Belgium.
Tel. : +32-10-474308, fax : +32-10-473032. simar@stat.ucl.ac.be.

** Institut de Statistique, Université Catholique de Louvain, Belgium and UPEG at EROC and EERC of "Kyiv-Mohyla Academy", Kyiv (Kiev), Ukraine.
Mailing address: 20 voie du Roman Pays, B 1348, Louvain-la-Neuve, Belgium.
Tel. : +32-10-474311, fax : +32-10-473032. zelenyuk@stat.ucl.ac.be.

Acknowledgement:: We would like to thank participants of 8th EWEPA (Oviedo, Spain, 2003) and of NAPW2004 (Toronto, Canada) for helpful comments. Research support from the "Interuniversity Attraction Pole", Phase V (No.P5/24) from the Belgian Government (Belgian Science Policy) is acknowledged.

1. Introduction

A typical research question of many empirical studies in the area of efficiency and productivity analysis attempts comparing differences in efficiencies among various groups in an industry or region (e.g., private vs. public) or differences across time for the same group(s). One way to approach this type of questions is to use non-parametric kernel density estimator, which is becoming more and more popular in many applied areas, including empirical economics.

Some attempts of using the kernel density estimators (KDE) have already appeared in efficiency analysis (see, for example, Färe and Grosskopf (1996), Simar and Wilson (1998), Kumar and Russell (2002)). A natural way of inferring on differences in efficiency distribution in this case would be to employ statistical tests of equality of two densities, for example those based on the Hall (1984) central limit theorem (CLT) for degenerate U-statistics or its extensions. These include tests proposed by Mammen (1992), Anderson et al. (1994), Li (1996, 1999) and Fan and Ullah (1999).

The goal of this paper is to adapt one of such tests, in particular the one proposed by Li (1996), to the context of comparing distributions of efficiency scores. The challenge of the problem here is that the random variables (efficiency scores) whose distributions are compared are *unobserved*. Is it reliable to use their estimates from data envelopment analysis (DEA)—in order to test the equality of *true* distributions of *true* efficiency scores? Recent discovery of statistical properties of this popular estimator (which before was called deterministic) gives hopes for the positive answer, along with some warnings.

There are at least four major concerns pertinent to analysis of distributions of DEA-estimated *efficiency scores*. One is the issue of *bounded support* of the (random) efficiency scores distributed across population of given firms, with most of the mass often being near the bound. The other three concerns come directly from the fact that *estimated* (rather than the *true*) efficiencies are used. Namely, in a finite sample, the estimated efficiency scores are *biased* and *not independent*. These two problems vanish asymptotically—but with a rate of convergence that depends on the dimension of the DEA model, which is the fourth concern of our context. Thus, the *goal* of this paper is to address these concerns and carefully adapt an existing test to the context of DEA-based efficiency analysis. At the end, we suggest to researchers a reliable tool for comparing densities of true Farrell-type efficiency scores as well as to warn about potential problems with it.

2. Theoretical Model

Assume the technology of a firm k ($k = 1, \ldots, n$) is characterised by the set T,

$$T \equiv \{(x, y) \ : \ x \ can \ produce \ y\}, \tag{2.1}$$

where $x = (x_1, ..., x_N)' \in \Re_+^N$ denotes the vector of N inputs, while the vector of M outputs is denoted by $y = (y_1, ..., y_M)' \in \Re_+^M$. Equivalently, the technology can be characterised by the *output sets*

$$P(x) \equiv \{ y : \ (x, y) \in T\}, \qquad x \in \Re_+^N. \tag{2.2}$$

Throughout, we assume that the technology satisfies the usual regularity axioms of production theory, under which the Shephard (1970) *output* distance function $D_o : \Re_+^N \times \Re_+^M \to \Re_+^1 \cup \{\infty\}$, defined as

$$D_o(x, y) \equiv \inf\{\theta : \ y / \theta \in P(x)\} \tag{2.3}$$

gives a complete characterisation of the technology, in the sense that[1]

$$D_o(x, y) \leq 1 \quad \Leftrightarrow \quad y \in P(x). \tag{2.4}$$

This function is particularly convenient as a criterion for technical efficiency of any particular firm k since, roughly speaking, it gives a 'measure' (valued between 0 and 1) of a distance from a point y^k in $P(x^k)$ to the 'upper' boundary of $P(x^k)$, measured along the ray from the origin. Such efficiency criterion often appears in another form, as the output oriented *Debreu* (1951)-*Farrell* (1957) measure of *technical efficiency*, defined for all $y^k \in P(x^k)$ as[2]

$$TE(x^k, y^k) \equiv \max\{\theta : \ \theta y^k \in P(x^k)\} = 1 / D_o(x^k, y^k). \tag{2.5}$$

Formally, if we let the technological frontier be the 'upper' boundary of $P(x^k)$ defined as

---

[1] For a detailed discussion of the axioms and related properties see Färe and Primont (1995).

[2] We choose output orientation here, but the same could be done for the input orientation case.

$$\partial P(x^k) = \{ y \in \Re_+^M \quad : \quad y \in P(x^k), \quad \lambda y \notin P(x^k), \quad \forall \lambda \in (1, \infty) \} \tag{2.6}$$

then,

$$0 < D_o(x^k, y^k) < 1 \quad \Leftrightarrow \quad y^k \in P\ (x^k), \quad y^k \notin \partial P\ (x^k), \quad y^k \neq 0\,, \tag{2.7}$$

and thus we will call the particular observation (x^k, y^k) as *technically* inefficient, with inefficiency score given by (2.5) (or its reciprocal). Alternatively, (x^k, y^k) is called *technically* efficient (having efficiency score equal unity) if and only if

$$D_o(x^k, y^k) = 1 \quad \Leftrightarrow \quad y^k \in \partial P\ (x^k) \tag{2.8}$$

Finally, $D_o(x^k, y^k) = 0$, if and only if $y^k = 0$. Also, for further reference, we will denote the Farrell-type efficient 'reference' point for any $y \in P(x)$—the radial projection of y onto the technological frontier $\partial P(x)$ with $y^{\partial}(x)$. Formally,

$$y^{\partial}(x) = yTE(x, y) = y / D_o(x, y) \in \partial P(x) \tag{2.9}$$

There are also other measures of technical efficiency offered in efficiency literature. However, the Debreu-Farrell measure seems to have been the most popular—for its relative computational simplicity, intuitive interpretation, relationship to duality theory in economics and, most importantly, possession of various desirable mathematical properties (e.g., see Russell, 1990).

There are also various ways of estimating the *true* frontier of $P(x)$ and the *true* efficiency score $TE(x,y)$ for any $y \in P(x)$. Apparently, one of the most popular of them is the Data Envelopment Analysis (DEA) estimator, which we consider in the next section.

3. The DEA estimator

In this section we focus on the class of estimators of the Farrell (1957) type (in)efficiencies known under the general name—Data Envelopment Analysis (DEA). There are many variations in this class, all intending to estimate the technological *frontier* of some set-wise characterisation of the technology and then to compute a point estimate of efficiency scores for each observation, relative to the estimated frontier. While we will consider only the very basic, most commonly used DEA model (that takes output orientation, assumes variable returns to scale and free disposability of inputs and outputs), the extension to other variants shall be evident.

One fundamental assumption in most DEA estimations is that all firms have *access* to the *same* technology, denoted with T, or its sections, $P(x)$. This access is determined by certain physical laws and current understanding of these laws by the humanity. This assumption is needed to justify the estimation of *one frontier* for all the firms in a sample, often called the (estimated) *best practice frontier*. Although, all firms have the access to the same technology, this does not guarantee that all firms, in a particular point in time, are able to exploit this technology in the same way. That is, the conditions of the access to this best-practice technology and its utilisation might be different for each firm. Namely, some internal or external factors pertinent to each specific firm (e.g., principal-agent problems, government interventions, particular institutions, etc) may create additional restrictions (not always observed by researchers) on the accessibility or/and utilisation of the best practice technology at a given moment in time. Despite optimisation behaviour, such restrictions are likely to cause some firms or even distinct groups of them to be below the 'best-practice' frontier. The goal of the DEA estimator is to measure 'how far' each firm is from this frontier by assigning for each firm an estimate of its (in)efficiency score, which will represent all those restrictions on efficient use of the technology in a cumulative, residual fashion.

A closely related fundamental assumption of the DEA estimator is that all observed input-output combinations (x^k, y^k), $k = 1, \ldots, n$ are *feasible* under T. In other words, $Prob\{(x^k, y^k) \in T\} = 1, \quad k = 1, \ldots, n$. This assumption implicitly allows *no* measurement errors and *all* deviations from the frontier are assumed to be due to technical inefficiency. However, all the data (input, output) as well as the unobserved inefficiency levels are allowed to be random and follow some *unknown* distributions. Finally, we also assume the true technology set, T, is convex.

Here, we consider the most common DEA estimator that allows for variable returns to scale and free disposability of inputs and outputs. The *best practice frontier* under such assumptions is estimated by an empirical analogue of (2.6) as

$$\partial\hat{P}(x) = \{y \in \Re_+^M \quad : \quad y \in \hat{P}(x), \quad \lambda y \notin \hat{P}(x), \quad \lambda \in (1, \infty)\}, \tag{3.1}$$

where

$$\hat{P}(x) = \{y \in \Re_+^M \quad : \quad \sum_{k=1}^{n} z_k y_m^k \geq y_m, \quad m = 1, ..., M, \quad \sum_{k=1}^{n} z_k x_n^k \leq x_i, \quad i = 1, ..., N,$$

$$z_k \geq 0 \ , \quad k = 1, \ldots , n \ , \quad \sum_{k=1}^{n} z_k = 1 \ \} . \tag{3.2}$$

Notably, $\hat{P}(x)$ is the *smallest convex free-disposal hull* that fits the observed data and $\partial\hat{P}(x)$ is its 'upper' boundary, a '*piece-wise linear*' estimate of the true best-practice frontier of $P(x)$.

The DEA estimator of individual *technical* efficiency at a fixed point (x, y), is computed relative to this estimated frontier—as a solution to the following linear programming problem (LPP)

$$\hat{TE}(x, y) = \max_{\theta, z_1, \ldots, z_n} \{ \theta \quad : \sum_{k=1}^{n} z_k y_m^k \geq \theta y_m \ , \ m = 1, \ldots, M, \ \sum_{k=1}^{n} z_k x_n^k \leq x_i \ , \ i = 1, \ldots, N,$$

$$z_k \geq 0 \ , \quad k = 1, \ldots , n, \quad \sum_{k=1}^{n} z_k = 1 \} . \tag{3.3}$$

or

$$\hat{TE}(x, y) = \max_{\theta, z_1, \ldots, z_n} \{ \theta \quad : \quad y\theta \in \hat{P}(x) \}$$

The next section formally outlines the data generating process and summarises the resulting known statistical properties of this basic version of the DEA estimator.

4. Known Statistical Results for the DEA Estimator

First, it is obvious that $\hat{P}(x) \subseteq P(x)$, and thus $\partial\hat{P}(x)$ is a *downward biased* estimator of $\partial P(x)$. As a result, $\hat{TE}(x, y)$ is a downward biased estimator of $TE(x, y)$, i.e.,

$$1 \leq \hat{TE}(x, y) \leq TE(x, y) , \qquad \forall \ y \in \hat{P}(x) \tag{4.1}$$

The statistical asymptotic properties also have recently been discovered for the DEA estimator presented above. In particular, (3.2) is consistent and is a maximum-likelihood estimator of the frontier of $P(x)$, as shown by Korostelev et al (1995) for the univariate (one-output or one-input) case. More recently Kneip et al. (1998) have shown consistency and derived the rates of convergence of the DEA technical efficiency estimator (3.3) for the multi-output-multi-input case. Gijbels et al. (1999) have discovered the limiting distribution of DEA in the 1-input-1-output case and recently, Kneip et al. (2003) have unveiled it for the multi-output-multi-input case.

These statistical results require additional assumptions on the data generating process (DGP). Let us first represent the problem by using the *polar* coordinates of $y \in \Re_+^M$ defined by the modulus $\omega = \omega(y) \in \Re_+^1$, where $\omega(y) \equiv \sqrt{y'y}$, and the angle $\eta \equiv \eta(y) \in [0, \pi/2]^{M-1}$, where $\eta_m \equiv \arctan(y_{m+1} / y_1)$, if $y_1 > 0$ or $\eta_m \equiv \pi/2$, if $y_1 = 0$ for $m = 1, \ldots, M-1$. Now, we assume:[3]

A1. $\{(x^k, y^k) : k = 1,..., n\}$ are *independent* random variables on the *convex* technology set T. All observations $\{(x^k, y^k) : k = 1,..., n\}$ can be partitioned into L *sub-samples* (by some exogenous criterion) such that each sub-sample l ($l = 1, \ldots, L$) represents a distinct *sub-group* l ($l = 1, \ldots, L$) of interest that exists in the population (e.g., public vs. private firms in an industry).

A2. For all l ($l = 1,\ldots,L$), the inputs $x \in \Re_+^N$ has density $f_{x,l}(x)$, with compact support $\aleph \subseteq \Re_+^N$.

A3. For all l ($l = 1,\ldots,L$) and all $x \in \aleph$, the vector $\eta \equiv (\eta_1,..., \eta_{M-1})$ has a conditional p.d.f. $f_{(\eta|x),l}(\eta \mid x)$ on $[0, \pi/2]^{M-1}$ and the modulus ω has a conditional p.d.f. $f_{(\omega|\eta,x),l}(\omega \mid \eta, x)$.

A4. For all l ($l = 1,\ldots,L$), all $x \in \aleph$, and all $\eta \in [0, \pi/2]^{M-1}$ there exist constants $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\forall \omega \in [\omega(y^\partial(x)), \omega(y^\partial(x)) - \varepsilon_2]$, $f_{(\omega|\eta,x),l}(\omega \mid \eta, x) \geq \varepsilon_1$, $l = 1, \ldots, L$, where $y^\partial(x)$ was defined in (2.9).

A5. The technical efficiency measure $TE(x, y)$ is differentiable in both its vectors.

Since all the groups have access to the same technology T, the lower boundary of the support $f_{(\omega|\eta,x),l}(\omega \mid \eta, x)$ is the same for each group, $l = 1, \ldots, L$. So that,

$$\omega(y^\partial(x)) = \sup\{\omega \in \Re_+^1 \quad : \quad f_{(\omega|\eta,x),l}(\omega \mid \eta, x) > 0\} \tag{4.2}$$

and the relation between $\omega(y)$ and the technical efficiency at (x, y), $TE(x, y)$, is now

$$TE(x, y) = \omega(y^\partial(x)) / \omega(y). \tag{4.3}$$

---

[3] We have adapted assumptions of Kneip et al. (1998) to the context of analyzing groups (or sub-populations) within a population, i.e., group-wise heterogeneous population case (see also Simar and Zelenyuk, 2003).

Although the same efficiency measure (4.3) is applied for all firms in the population, the efficiency score it yields for particular points may have different densities $f_{(\omega|\eta,x),l}(\omega \mid \eta, x)$ for different sub-populations of firms l ($l = 1, \ldots, L$). From now on, the superscript l on the measure (4.3) will imply that this measure has been applied to a firm belonging to group l ($l = 1, \ldots, L$).

Importantly, note that A3 along with (4.3) implies the existence of a conditional (on (η, x)) density for $TE^l(x, y)$, $l = 1, \ldots, L$ (with the support $[1, \infty)$), which we will denote with $f_l(TE \mid \eta, x)$. Moreover, A4 along with (4.3), implies that $f_l(TE \mid \eta, x) \geq \varepsilon_1, \forall TE \in [1, 1 + \varepsilon_2]$, $l = 1, \ldots, L$. Finally, with assumptions A1-5, the DGP, denoted with $\wp = \wp(P(x), g_l(TE, \eta, x), l = 1, ..., L)$ is completely defined through the *joint* densities of (TE, η, x),

$$g_l(TE, \eta, x) = f_l(TE \mid \eta, x) f_{(\eta|x),l}(\eta \mid x) f_{x;l}(x), \qquad l = 1, \ldots, L, \tag{4.4}$$

each with the support $\Omega \equiv [1, \infty) \times [0, \pi/2]^{M-1} \times \aleph$. It is this DGP that we assume has generated our sample $\Xi_n = \{(x^k, y^k) : k = 1, ..., n\}$ of independent observations that are identically distributed *within* each sub-group l ($l = 1, \ldots, L$) but not necessarily *across* them.

These assumptions allow us to employ the Kneip et al., (1998) theorem to claim that the DEA estimator is consistent, and the rate of convergence is given by

$$\hat{TE}^l(x, y) - TE^l(x, y) = O_p(n^{-2/(M+N+1)}). \tag{4.5}$$

It is worth noting that the DGP defined above still assumes that *all* firms have *access* to the *same* technology, but the conditions of this access ("how easy it is to get to the frontier") are allowed to be different for different sub-groups. Such DGP can be well justified for considering many economic phenomena. Differences in regulation regimes, ownership structures, environments, various institutions, etc., pertinent to different groups within the population, may have *systematic* differences in economic *incentives* or just 'physical' capabilities of firms in different sub-groups (sub-populations) to reach the *frontier* of the *same* technology. For example, private firms are often expected to have different incentives for being closer to the frontier than state-owned firms. Also, firms under average-cost pricing regulation regime might be influenced by rather different incentives than firms

under the rate-of-return regulation, or/and unregulated firms. More of philosophical grounds for presence of inefficiency can be found in the seminal work of Liebenstein (1966, 1992) about his concept of "X-efficiency". The DEA thus can be viewed as an estimator of X-efficiency, based on the Debreu-Farrell type efficiency measure. In all the examples given above, the marginal densities that generate technical (in)efficiency (as well as densities generating inputs and outputs) for these firms might be different across sub-groups, while still having common best practice technology.

5. The Essence of the Li (1996) Test for Equality of Two Densities

Although the test of Li (1996, 1999) can be used for multivariate cases, our context is univariate and we would thus consider only the univariate version of the test in this paper. Suppose we are interested in comparing densities of two random variables, $U^A, U^Z \in \Re^1$ for which the random samples, $\{u^{A,k}: \ k = 1,..., n_A\}$ and $\{u^{Z,k}: \ k = 1,..., n_Z\}$, representing the two sub-groups in a population, A and Z, respectively, are available. Formally, letting f_l denoting density of distribution of a random variable U^l ($l = A, Z$), our null and alternative hypotheses would be:

H_0: $f_A(u^A) = f_Z(u^Z)$,
H_1: $f_A(u^A) \neq f_Z(u^Z)$, on a set of positive measure.

For testing such hypotheses, Mammen (1992), Anderson et al. (1994), Li (1996, 1999) and Fan and Ullah (1999), have considered the integrated square difference criterion

$$I \equiv \int \left(f_A(u) - f_Z(u)\right)^2 dt = \int \left(f_A{}^2(u) + f_Z{}^2(u) - 2 f_A(u) f_Z(u)\right) du$$

$$= \int f_A(u) dF_A(u) + \int f_Z(u) dF_Z(u) - \int f_A(u) dF_Z(u) - \int f_Z(u) dF_A(u) \qquad (5.1)$$

(which satisfies the property that $I \geq 0$ and $I = 0$ if and only if H_0 is true) and suggested slightly different empirical analogues for it as tests statistics. A particularly convenient statistic has been proposed by Li (1996, 1999) and extended by Fan and Ullah (1999). Using Hall (1984) CLT, Li (1996) have shown that a consistent and asymptotically normal estimator of (5.1) is obtained by replacing the unknown distribution functions $F_A(\cdot)$ and $F_Z(\cdot)$ in (5.1) with corresponding empirical distribution functions, $F_{A,n_A}(\cdot)$ and $F_{Z,n_Z}(\cdot)$,

$$F_{l,n_l}(u) \equiv \frac{1}{n_l}\sum_{k=1}^{n_l} I\left(u^{l,k} \le u\right), \quad l = A, Z$$

where $I\{X\}$ is the indicator function so that $I\{X\}=1$ if statement X is true, and zero otherwise, and the unknown densities in (5.1) are replaced with the non-parametric kernel density estimators, $\hat{f}_{A,n_A}(\cdot)$ and $\hat{f}_{Z,n_Z}(\cdot)$, where

$$\hat{f}_{l,n_l}(u) \equiv \frac{1}{n_l h_l}\sum_{k=1}^{n_l} K\left(\frac{u-u^{l,k}}{h_l}\right), \qquad l = A, Z. \tag{5.2}$$

Here, $h_l = h(n_l)$ is the smoothing parameter (bandwidth) obtained using some optimal criterion for each sub-group $l = A, Z$,[4] while K is an appropriate kernel function (e.g., Gaussian density) and u is a point at which we aim to estimate the density. The latter is often chosen to be the observed points, $\{u^{A,k}: \ k=1,...,n_A\}$ and $\{u^{Z,k}: \ k=1,...,n_Z\}$, so that the statistic becomes

$$\begin{aligned}\hat{I}_{n_A,n_Z,h} &= \int \hat{f}_A(u)dF_{A,n_A}(u) + \int \hat{f}_Z(u)dF_{Z,n_Z}(u) - \int \hat{f}_A(u)dF_{Z,n_Z}(u) - \int \hat{f}_Z(u)dF_{A,n_A}(u) \\ &= \frac{1}{h\ n_A n_A}\sum_{j=1}^{n_A}\sum_{k=1}^{n_A} \mathrm{K}\left(\frac{u^{A,j}-u^{A,k}}{h}\right) + \frac{1}{h\ n_Z n_Z}\sum_{j=1}^{n_Z}\sum_{k=1}^{n_Z} \mathrm{K}\left(\frac{u^{Z,j}-u^{Z,k}}{h}\right) \\ &\quad - \frac{1}{h\ n_A n_Z}\sum_{j=1}^{n_Z}\sum_{k=1}^{n_A} \mathrm{K}\left(\frac{u^{Z,j}-u^{A,k}}{h}\right) - \frac{1}{h\ n_Z n_A}\sum_{j=1}^{n_A}\sum_{k=1}^{n_Z} \mathrm{K}\left(\frac{u^{A,j}-u^{Z,k}}{h}\right)\end{aligned} \tag{5.3}$$

Further, Li (1996) notes that the statistic based on (5.3) with the 'diagonal' terms removed, i.e.,

$$\begin{aligned}\hat{I}^{nd}_{n_A,n_Z,h} &= \left\{\frac{1}{h\ n_A(n_A-1)}\sum_{j=1}^{n_A}\sum_{k\neq j,k=1}^{n_A} \mathrm{K}\left(\frac{u^{A,j}-u^{A,k}}{h}\right) + \frac{1}{h\ n_Z(n_Z-1)}\sum_{j=1}^{n_Z}\sum_{k\neq j,k=1}^{n_Z} \mathrm{K}\left(\frac{u^{Z,j}-u^{Z,k}}{h}\right)\right. \\ &\left. - \frac{1}{h\ n_A(n_Z-1)}\sum_{j=1}^{n_Z}\sum_{k\neq j,k=1}^{n_A} \mathrm{K}\left(\frac{u^{Z,j}-u^{A,k}}{h}\right) - \frac{1}{h\ n_Z(n_A-1)}\sum_{j=1}^{n_A}\sum_{k\neq j,k=1}^{n_Z} \mathrm{K}\left(\frac{u^{A,j}-u^{Z,k}}{h}\right)\right\}.\end{aligned} \tag{5.4}$$

[4] Note that under the null, the bandwidths is the same for both A and Z and, in practice, to estimate p-values, one can use the average, the maximum or, as in Li (1999) and in our study—the minimum of the two.

has performed better (in most Monte Carlo experiments) than the one based on the centred version of (5.3), which may cause size distortions for finite samples. Letting $\lambda_n = n_A / n_Z$, and assuming $\lambda_n \to \lambda, as\ n_A \to \infty$, where $\lambda \in (0, \infty)$ is a constant, Li (1996) shows that after appropriate 'standardisation', the limiting distribution of (5.4) is standard normal,

$$\hat{J}^{nd}_{n_A, n_Z, h} \equiv \frac{n_A h^{1/2} \hat{I}^{nd}_{n_A, n_Z, h}}{\hat{\sigma}_{\lambda, h}} \quad \xrightarrow{d} \quad N(0,1) \tag{5.5}$$

where $\hat{\sigma}_{\lambda,h}$ is a consistent estimator of $\sigma_\lambda^2 = 2\left(\int (f_A(u) - \lambda f_Z(u))^2 du\right)\left(\int \mathrm{K}^2(u) du\right)$, obtained as

$$\hat{\sigma}^2_{\lambda,h} = 2\left\{ \frac{1}{h\ n_A^2} \sum_{j=1}^{n_A} \sum_{k=1}^{n_A} \mathrm{K}\left(\frac{u^{A,j} - u^{A,k}}{h} \right) + \frac{\lambda_n^{\ 2}}{h\ n_Z^2} \sum_{j=1}^{n_Z} \sum_{k=1}^{n_Z} \mathrm{K}\left(\frac{u^{Z,j} - u^{Z,k}}{h} \right) \right.$$
$$\left. - \frac{\lambda_n}{h\ n_A n_Z} \sum_{j=1}^{n_Z} \sum_{k=1}^{n_A} \mathrm{K}\left(\frac{u^{Z,j} - u^{A,k}}{h} \right) - \frac{\lambda_n}{h\ n_A n_Z} \sum_{j=1}^{n_A} \sum_{k=1}^{n_Z} \mathrm{K}\left(\frac{u^{A,j} - u^{Z,k}}{h} \right) \right\} \left[\int \mathrm{K}^2(u) du \right]. \tag{5.6}$$

6. Adapting the Li Test to the DEA Context

In this section we will be interested in comparing distributions of *efficiency scores* for two sub-groups, A and Z, in an economic system (industry, country, etc). Let us first check the satisfaction of regularity conditions used in Li (1996). According to A1-A5, the *true* technical efficiency scores in each sub-group, $\{TE^{A,k} : k = 1, ..., n_A\}$ and $\{TE^{Z,k} : k = 1, ..., n_Z\}$, are distributed *independently* (and identically within each sub-group) with densities $f_A(\cdot)$ and $f_Z(\cdot)$, respectively. (Contemporaneous correlation is allowed between the two samples.) In real world, however, the true independent efficiency scores are unobserved but estimated with DEA, which are independent only asymptotically, while in finite samples they are dependent (by construction) with unknown structure of dependency. It is thus not clear whether the original Li-test would perform well in this context. The assumption A1-A5 also ensure that $f_A(\cdot)$ and $f_Z(\cdot)$ are continuous and bounded in $\Re^1$ and have some distribution functions $F_A(\cdot)$ and $F_Z(\cdot)$, respectively. For simplicity, we will also use the Gaussian kernel to estimate these densities, which is a non-negative second order kernel as suggested (but not required) by Li (1996).

Since construction of the test statistic essentially involves formulas for kernel density estimator, the special issues pertinent to the kernel density estimation in DEA context, might be important in the Li-test context as well. The first is the issue of *bounded support* in density estimation, since it is known that the standard kernel density estimator is inconsistent at the boundary. This can easily be circumvented by using the Schuster (1985)-Silverman (1986) reflection method, which provides the consistent estimator [5]

$$\hat{f}^{\mathrm{R}}(u) \equiv \begin{cases} \dfrac{1}{nh_{\mathrm{R}}} \sum_{k=1}^{n} \left[K\left(\dfrac{u_i - u}{h_{\mathrm{R}}} \right) + K\left(\dfrac{(2b - u_i) - u}{h_{\mathrm{R}}} \right) \right], & u \geq b \\ 0, & \textit{otherwise}. \end{cases} \tag{6.1}$$

A natural question is whether one also needs to use this (or other) method for adapting the Li-test to the DEA context. In his Monte Carlo experiments, Li (1996) have considered application of his test to the case of Uniform density and obtained good performance. Our case is somewhat more severe than the uniform: most of the mass might be very close to the bound and we know that the estimator for such case can give very poor fit. However, note that for symmetric kernels, the estimation procedure (6.1) is equivalent to estimating $2\hat{f}(u)$ using (5.2) from the original data $\{u_1,...,u_n\}$, on the bounded support, but with a bandwidth selected from the *reflected* data $\{u_1,..., u_n, 2b-u_1,..., 2b-u_n\}$, which we denote as h_{R}. Hence, the reflected analogue of (5.3), will be twice as much as (5.3), given h_{R}, i.e.,

$$\hat{I}^{\mathrm{R}}_{n_A,n_Z,h_{\mathrm{R}}} \equiv \hat{I}^{nd,\mathrm{R}}_{n_A,n_Z,h_{\mathrm{R}}} + \hat{I}^{d,\mathrm{R}}_{n_A,n_Z,h_{\mathrm{R}}} = 2\hat{I}^{nd}_{n_A,n_Z,h_{\mathrm{R}}} + 2\hat{I}^{d}_{n_A,n_Z,h_{\mathrm{R}}}, \tag{6.2},$$

where $\hat{I}^{d,\mathrm{R}}_{n_A,n_Z,h_{\mathrm{R}}}$ is part of $\hat{I}^{\mathrm{R}}_{n_A,n_Z,h_{\mathrm{R}}}$, involving only diagonal terms of it. Using the Hall (1984) theorem, similarly as Li (1996), we get that (for a univariate case)

$$\hat{J}^{\mathrm{R},nd}_{n_A,n_Z} \equiv \frac{n_A h_{\mathrm{R}}^{1/2} \hat{I}^{nd,\mathrm{R}}_{n_A,n_Z}}{\hat{\sigma}_{\lambda,\mathrm{R},h_{\mathrm{R}}}} = \frac{\sqrt{2} n_A h_{\mathrm{R}}^{1/2} \hat{I}^{nd}_{n_A,n_Z,h_{\mathrm{R}}}}{\hat{\sigma}_{\lambda,h_{\mathrm{R}}}} \quad \xrightarrow{d} \quad N(0,1) \tag{6.3}$$

where we have also used the fact that $\hat{\sigma}^2_{\lambda,\mathrm{R},h_{\mathrm{R}}} = 2\hat{\sigma}^2_{\lambda,h_{\mathrm{R}}}$. Thus, the statistic based on reflection method is essentially the same as the original Li (1996) test, with the difference

[5] An alternative approach would be to use special 'boundary' kernels.

being a factor of $\sqrt{2}$ and the fact that the bandwidth used in estimation of statistic is obtained from the data with reflection rather than the original data. As we will see in Monte Carlo experiments, (6.3) does not improve upon the original Li (1996) statistic in the sense that both of them have similar performance (e.g., bootstrap estimated sizes of the test are not significantly different from each other and from the true sizes.) This suggests that despite the fact that reflection method is useful in density estimation *per se*, it is essentially not needed for adapting the original Li-test to the case of comparing the densities with bounded support (with a lot of mass near the bound).

The second is the issue of using not the true random variables whose densities we want to compare, but their estimates which, in finite samples, are downward-biased and dependent.[6] The Li (1996) test is *asymptotic* in nature, so the problem of the finite-sample bias and dependency of DEA-estimated efficiency scores is mainly an empirical problem. For example, it might be that due to high sampling variation or noise from the DEA estimation, researchers may run into the *type-I* (incorrectly reject the true Ho) and *type-II* (failing to reject the incorrect Ho) errors more often than they would under the (unrealistic) situation when the true efficiency scores are known. If so, such inference may bring researchers to misleading conclusions and incorrect policy recommendations. It also might be that since the bias problem influences, in the same fashion, the estimation of *both* densities under comparison, then the overall 'damage' might happen to be not so significant for the test itself, especially under the null. Experimental evidence through Monte Carlo experiments is clearly needed here.

Overall, even in large samples, and especially in finite samples, the performance of the original Li-test adapted to the DEA context using the (first order) Normal approximation might be far from desirable. As a result, carefully designed bootstrap may be needed to improve its performance. In fact, Li (1999), using the true random variable, give some Monte Carlo evidence that the bootstrap provides better inference than the use

---

[6] The CLT of Hall (1984) was developed for independent processes, thus the tests based upon it, including the Li (1996) test, require independence of observations *within* each sample (though, allow for correlation between the samples). The assumption of independence was weakened to some extent with the theorem of Fan and Li (1996), who extended the Hall's CLT to the *absolutely regular* processes and made it possible for Fan and Ullah (1999) to extend the Li (1996) test to the class of such processes. Definition of absolutely regular processes which Fan and Li (1996) use is given for the context of strictly stationary processes and it is not clear at this point whether it can be extended to dependency in a cross-sectional framework of DEA, where the structure of dependency is unknown. The encouraging fact is that since DEA is consistent

of normal tables. This shall not be surprising, given the asymptotically pivotal nature of the statistic. In the next section, we will consider several bootstrap algorithms in the aim of suggesting the most reliable one for empirical research.

7. Bootstrapping the 'adapted to DEA context Li-test'

It is now known that the naïve bootstrap is inconsistent for the individual DEA-estimates (e.g., see Simar and Wilson, 2000 for details). On the other hand, Kneip et al. (2003) had proven consistency property of the *sub-sampling* bootstrap—when the sub-sample size is smaller than the original sample size. They also provide Monte Carlo evidence that earlier suggested *smooth* bootstrap is an approximation of the consistent sub-sampling bootstrap. In principle, we could use the two-stage approach: first correct for the bias in the DEA efficiency estimates using one of the consistent versions of bootstrap for DEA and then, at the second stage, use these bias-corrected estimates in the Li-test for comparing the true densities of the true efficiency scores. Analysing such approach under various Monte-Carlo scenarios is very computer-time-consuming, even with modern computers (just one bootstrap bias-correction procedure may take several hours, since it requires LP optimisation for each efficiency score). Instead, we will use (slightly modified) sample of original DEA estimates for using in the Li-statistic, which in turn we bootstrap to obtain p-values. Results from various Monte Carlo scenarios yield evidence of good performance of such algorithms under moderate dimensions of DEA model, requiring much less computer time than the tests that would involve bias correction.

Specifically, we will analyse two bootstrap alternatives. The first one, call it *Algorithm I*, is based on computation and bootstrapping the Li-statistic using the sample of DEA-estimates *trimmed* from those that are equal to unity, i.e., just ignore the 'spurious ones' for the sake of the test. The second one, call it *Algorithm II*, is based on computation and bootstrapping the Li-statistic using the sample of DEA-estimates where those equal to unity are '*smoothed*' away from the bound by adding a small noise, say within 5%-quantile of the empirical distribution of $\hat{TE}(x^k, y^k)$ ignoring those equal unity, but of order smaller than the noise of estimation (suggested by the rate of convergence). Formally, the smoothing is made by

---

estimator, the dependency vanishes asymptotically (but, again, the rate of convergence is decreasing with the dimension of the DEA model).

$$\hat{TE}^{*}(x^{k}, y^{k}) = \begin{cases} \hat{TE}(x^{k}, y^{k}) + \varepsilon^{k}, & \hat{TE}(x^{k}, y^{k}) = 1 \\ \hat{TE}(x^{k}, y^{k}), & otherwise \end{cases} \tag{7.1}$$

where $\varepsilon^{k} = Uniform(0, \ \min\{n^{-2/(M+N+1)}, a\})$, a is the α-quantile of the empirical distribution of $\hat{TE}(x^{k}, y^{k})$ ignoring those equal unity. For the sake of reader's convenience, the complete description of both bootstrap algorithms for Li-statistic in DEA context is given in the next table.

< Insert Table 1 here >

The obtained *bootstrap* estimates of DEA-adapted Li-statistic can then be used to estimate the *p-value* of the test, by using

$$\hat{p} - value = \frac{1}{B}\sum_{b=1}^{B} I\left\{\hat{J}^{*nc,b}_{n_A,n_Z} > \hat{J}^{nc}_{n_A,n_Z}\right\}, \tag{7.2}$$

8. Monte-Carlo Investigation of the Size of the Test

The goal of this section is to illustrate the methods described above on some examples where we know the 'truth' and thus can get a feeling of the performance of the proposed techniques.

8.1. Simulated Example 1: Does Reflection Improve the Size of the Test?

In efficiency analysis, researchers often expect (or assume) efficiency distributions are skewed so that most of the mass is close to efficient bound (unity in our case) with a diminishing tail towards higher inefficiency—manifesting economic agent's tendency to strive for full efficiency or a level near it. A typical example for such a distribution can be constructed from normally distributed variable, truncated at unity. Here, we assume that the *true* technical efficiency is $TE^{l,k} = 1 + u^{l,k}$, where $u^{l,k} \sim N^{+}(\mu_l, \sigma_l^2)$, $l = A, Z$, where $\mu_A = \mu_Z = 0$ and $\sigma_A = \sigma_Z = 1$. For this example, assume that the true (realisations) of efficiency scores are known and our goal is to see if, in such ideal conditions, the application of the reflection principle to the original Li-test would bring any improvement in estimating the empirical size of the test. We consider four sample sizes: $n_A = n_Z = \{20, 50,$

100, 200} with MC = 1000 and B =400 (which would allow also comparing our results to those of Li, 1999; For real data, we would recommend B larger than 1000). We estimate four values of the true (nominal) sizes: $\{0.01, 0.05, 0.10, 050\}$.[7]

From Table 2, one can see that the bootstrap estimated sizes of the *original* Li-test applied to 'observed' (in simulation) true efficiency scores are already very close to the true sizes, statistically insignificantly different from them, and so is the test that accounts for boundary issue via the reflection principle. As a result, in our future investigations we will not worry about the reflection issue.[8]

< Insert Table 2 here >

8.2. Example 2: Algorithm I vs. Algorithm II in Different Dimensions

Now we consider the main problem of the paper: the case when the true efficiency scores are *not* observed but estimated via DEA. And, the goal of the experiment is to see what approach performs better: (i) trimming the estimates that equal to unity, (Algorithm I), or (ii) smoothing them according to (7.1), with $\alpha = 5\%$ (Algorithm II). DEA is a non-parametric estimator and the shape of technology does not matter much for it, as long as it satisfy the regularity assumptions on DEA (especially convexity). What matters the most, however, is the dimension of the DEA problem (number of inputs and outputs in DEA specification), as suggested by the rate of convergence (4.5). So we will consider only one technology type, Cobb-Douglass, but in different dimensions—to investigate an impact of curse of dimensionality. Specifically, assume that the true technology frontiers are characterised by the Shephard's output distance function of the following forms,

$$D_o(x, y^*) = y^* / (x_1^{0.3} x_2^{0.5}) \tag{8.1}$$

$$D_o(x, y^*) = y^* / (x_1^{0.1} x_2^{0.15} x_3^{0.2} x_4^{0.05}) \tag{8.2}$$

$$D_o(x, y^*) = y^* / (x_1^{0.1} x_2^{0.15} x_3^{0.2} x_4^{0.05} x_5^{0.25} x_6^{0.07} x_7^{0.08}) \tag{8.3}$$

---

[7] The 95% Monte-Carlo confidence intervals (with 1000 replications) for these values would be approximately {(0.0037, 0.0163), (0.0362, 0.0638), (0.081, 0.119), (0.4684, 0.0316)}, respectively.

[8] The investigation of the empirical power of the test have also suggested that the power function of the test without reflection is slightly above that with reflection.

where y^* is the technically efficient output level, given input levels $(x_1, x_2)'$, which we assume are both coming from *Uniform*(*0,1*) distribution for each sub-group. The 'observed' output y (for each group) is obtained as $y^{l,k} = y^* / TE^{l,k}$. Here, we assume that the null is true, in particular, $\mu_A = \mu_Z = 0$ and $\sigma_A = \sigma_Z = 0.3$, thus $E(TE^{l,k}) \approx 1.24$, and consider four sub-sample sizes: $n_A = n_Z = \{20, 50, 100, 200\}$, with MC = 1000 and B = 400.

In Figure 1, we produce three curves for a typical Monte-Carlo replications for this scenario and in Figure 2 we have $\mu = 0.3$ and $\sigma = 0.3$. For both cases we have $n = 50$ and technology (8.1). The dotted curve is the true density of the true efficiency scores. The dash-dotted curve is the corresponding estimated density when the true efficiency scores are observed, but their density is estimated via (6.1). Finally, the dashed curve is estimated density when the true efficiency scores are unobserved but their density is estimated via (6.1) from the DEA-estimates (except those equal unity)—the situation faced in reality by practitioners. We use Gaussian kernel and Sheather and Jones (1991) method for estimating h. One can clearly see the empirical issue raised in our paper: the problem of using the DEA-estimates in place of the true but unknown efficiency scores.

<Insert Figure 1 Here>
<Insert Figure 2 Here>

The 'double' estimation (of efficiency and of density) produces density estimate that may look quite differently than the true density. In particular, the downward bias of the DEA-estimates results in 'too much' mass allocated near the bound. The density estimation based on the smoothed DEA estimates via (7.1) was allocating even more mass close to the bound (figures for which are not presented for the sake of space). This is of course only one replication presented that seemed typical to us. In some draws the fit was better, in others it was worse. Increase in sample size certainly tended to improve the fit, while increase in the dimension tended to worsen it. All this naturally raises questions and perhaps serious concerns about reliability of Li-test for the case of comparing distributions of efficiency scores that are unobserved but estimated via DEA, which we now analyse.

The results of our size investigation, presented in Table 3, are quite interesting and encouraging. For the case of two inputs, results presented in Table 3 suggest that (under the null) the estimated sizes for both algorithms are in most cases insignificantly different from the true (nominal) sizes. Only once, for the sample size of 200, the estimated size of

Algorithm I was significantly different from its nominal value of 0.01. When we increase dimension to 4 and 7 inputs, we sometimes could not obtain the bootstrapped p-values for Algorithm I. This is because the number of observations on the frontier (to be deleted in Algorithm I) increases with the dimension, resulting in too few observations or even in estimated variance being very close to zero for some bootstrap samples. The estimated sizes for Algorithm II for these cases, however, were mostly insignificantly different from the nominal sizes. In case when it was possible to compute the p-values for both Algorithm I and II, their performance was very similar. On this basis, we conclude that the two bootstrap algorithms, I and II, for the Li (1996, 1999) test adapted to the DEA context are reliable in terms of size of the test, with Algorithm II being more robust to the curse of dimensionality problem and, in our experiments, always had correct size at 5% and 10% levels, which are those most commonly used in practice.

<Insert Table 3 here>

9. Monte-Carlo Investigation of the Power of the Test

The goal of this section is to investigate the power of the test (i.e., probability of correctly rejecting the false null hypothesis) in different dimensions of DEA model. We will compare the power of test based on Algorithm II *vs.* the power of the test based on the *true* efficiency scores, which are known in Monte-Carlo study and can be considered as a benchmark for comparison. In particular, we investigate the power of the test in the case of different modes/means (before truncation) of efficiency distributions. Here we take the set-up of simulation example 1, but assume $\mu_A = 0, \sigma_A = 0.4$, while $\mu_Z = \delta, \delta \in \{-1,-0.9,..., 0.9, 1\}$ and $\sigma_Z = 0.4$. So, when $\delta = 0$, the null hypothesis is true. We consider three sample sizes: $n_A = n_Z = \{20, 50, 100\}$, with MC = 1000 and B = 400. For the insight onto the impact of curse of dimensionality we present results for several technologies. In particular, for the case of 20 observations in each group we use (8.1), (8.2) and two more technologies,

$$D_o(x, y^*) = y^* / (x_1^{0.1} x_2^{0.15} x_3^{0.2}) \tag{9.1}$$

and

$$D_o(x, y^*) = y^* / (x_1^{0.1} x_2^{0.15} x_3^{0.2} x_4^{0.05} x_5^{0.25}), \tag{9.2}$$

while for the case of 50 and 100 observations in each group we use technologies defined in (8.1), (8.2), (8.3) and (9.2).

First of all, it must be quite intuitive that the power functions for algorithms II shall outperform that of I, unless too much noise is added. This is because algorithm II *always* uses more observations than algorithm I (which ignores observations equal unity), especially for large dimensions, and this was confirmed in all our simulations (not presented). Thus, we only present the power functions for algorithm II, where DEA-estimates are obtained under different dimensions and compare them to the case when the true (realizations) of efficiency scores are used in the original Li (1996, 1999) test.

For convenience, we present the results by plotting the estimated power functions (for nominal size of 0.05) in Figures 3-5, while numerical values are reported in the table in the appendix. Remarkably, although the size of the test was very good relative to the true (nominal) size even for 20 observations in each sample, the estimated power functions suggest that the power of the test for such a small sample as 20 is quite low, especially when the dimensions of the DEA model is high. Note that the power functions for 20 observations in each group are very asymmetric with a flatter left tail—warning about reliability of the test conclusions based on such small samples, even for the case when the *true* efficiency scores are used (in original Li test).

The power function for the algorithm II based on 2-input-1-output DEA-estimates of efficiency mimics the asymmetric shape of that based on *true* efficiencies, getting quite close to this 'ideal' when the two distributions have different modes (after truncation), but worse for the case when the two distributions being different are having the same (unity) mode. This pattern is repeated for higher dimensions as well, such that the higher the dimension the worse is the power—manifesting the curse of dimensionality problem of the DEA estimator. Again, this problem is especially clear when the two distributions have the same mode. The case of 5-input-1-output case with 20 observations in each group yields the power function that is virtually flat—alerting about the danger of misleading impression about equality of distributions from such large dimension relative to such small samples.

<Insert Figure 3 here>

Observing and comparing Figures 4 and 5 to Figure 3 brings some optimistic feeling. The pattern is similar, but the larger the sample size the less asymmetry and the steeper are the power functions. The curse of dimensionality problem reduces with increase of the sample, as it is expected to be. For example, for the case of 100 observations, the power function for the test based on estimates in the 2-input-1-output

case becomes almost identical with the power function for the case when the true efficiency scores had been used. Even the power function from the estimates in the 4-input case also becomes much closer to that based on the true efficiency scores—when we compare 100 observations case with 50 observations case and certainly relative to 20 observations case. For 5-input model, the performance of the test improves for 50 observations and becomes reasonably good for 100 observations case (but the left tail still lags behind).

<Insert Figure 4 here>

<Insert Figure 5 here>

Importantly, the power function from the estimates in the 7-input case is quite unsatisfactory even for 100 observations case (very flat). The intuition for this is in the fact that high dimension of DEA model relative to the sample size leads to many observations being on the estimated frontier (due to being 'unique in their own way') in both groups and thus smoothing such estimates according to mechanism (7.1) 'homogenises' a big chunk of the two samples. So the resulting smoothed samples look much more the same than their populations are and the test statistic thus would have less power than it should when the true efficiency scores are (unrealistically) observed.

An issue in itself deserving special attention is the asymmetry of the power functions, which luckily vanishes when the sample size increases. The reason for this asymmetry is in the nature of the distributions analysed. Namely, we deal with distributions truncated on one side, highly asymmetric and, as can be observed from the figures, the true alternative hypothesis that presumes mode greater than unity in one distribution vs. unity-mode for another is more empirically identifiable than when comparison is between two asymmetric distributions with unity modes. (Additional Monte Carlo experiments for comparison of symmetric distributions confirmed this conjecture.) In other words, the test has greater power in distinguishing distribution with what Simar and Zelenyuk (2003) called as tendency for 'pathological inefficiency' (non-unity mode) vs. distribution where firms are having tendency to full efficiency, represented by (unique) unity mode.

10. Conclusion

The goal of this paper was to provide researchers with a reliable tool for testing equality of distributions of Farrell-type *efficiency* scores between groups of decision making units (plants, firms, etc) within a population (industry, country, region, etc). We have considered two major specifics pertinent to analysis of distributions of DEA-estimated efficiency scores and investigated ways to incorporate them to testing procedure. One is the issue of *bounded support* of the distribution of efficiency scores. The other is the usage of *estimated* rather than the *true* efficiencies, which are then used to *estimate* the *true* densities of *true* efficiency scores. Such estimates are biased and not independent—problem that vanishes asymptotically, but with a rate of convergence that depends on DEA dimension.

We incorporated this knowledge into considering various algorithms for testing equality of densities based on Li (1996, 1999) test. We demonstrated that the reflection method is unnecessary here. We also considered two algorithms that handle the problem of 'spurious mass at the bound': Algorithm I simply ignores the boundary estimates and Algorithm II smoothes such estimates by adding uniform noise of order smaller than the speed of convergence of the DEA estimator. Limited Monte Carlo evidence suggests that both algorithms have good size (insignificantly different from nominal one). However, Algorithm II was more *robust* to the increase in dimension.

Finally, the results of investigation of the *power* of the test suggest that, for relatively small dimensions of DEA model (e.g., 2 or 3 inputs and 1 output for 50 observations in each group), the power is quite good—close to the 'ideal' case, when *true* efficiency scores are used in testing. However, the curse of dimensionality problem (of the DEA estimator) is indeed a problem here: When the dimension is high relative to the sample size (e.g., 5 inputs, 1 output, for 20 observations in each group) then the power of the test is quite low. Very low power (although correct size) was also identified for 7-input-1 output case for 50 (and to some extent even for 100) observations in each group, especially when compared distributions in both groups have unique unity mode (no pathological inefficiency).

Overall, we conclude that given no abuse with the dimension of DEA model relative to the sample size, the Li (1996) test, adapted via our Algorithm II, is a reliable tool for testing equality of distributions of unknown but DEA-estimated efficiency scores. For the sake of brevity we had not presented an application here, but an interested reader is referred to recent applications of this method to Henderson and Zelenyuk (2004) and Zelenyuk and Zheka (2004).

References

Anderson, N., Hall, P. and Titterington, D.M. (1994), "Two Sample Test statistics for Measuring Discrepancies between Two Multivariate Probability Density Functions Using Kernel-based Density Estimates," *Journal of Multivariate Analysis*, 50, 41-54.

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

Efron, B. (1979), "Bootstrap methods: another look at the jackknife," *Annals of Statistics* 7, 1-26.

Farrell, M.J. (1957), "The Measurement of Productive Efficiency," *Journal of Royal Statistical Society*, Series A, General, 120, part 3, 253-281.

Fan, Y. and A. Ullah (1999) "On Goodness-of-fit Tests for Weakly Dependent Processes Using Kernel Method." *Journal of Nonparametric Statistic*s 11(1–3), 337–60.

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.

Hall, P. (1984), "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," *Annals of Statistics*, 14, 1-16.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansions*, Springler, New York.

Kneip, A., L. Simar and P. Wilson (2003a), "Asymptotics for DEA Estimators in Non-parametric Frontier Models", *Discussion Paper #0317, Institut de Statistique, Université Catholique de Louvain*, Belgium

Kneip, A., B. Park and L. Simar (1998), "A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores", *Econometric Theory* 14, 783-793.

Korostelev,A., Simar,L. and Tsybakov, A.B.(1995), "On estimation of monotone and convex boundaries," *Publ. Statist. Univ. Paris* XXXIX 1, 3 –18.

Li, Q. (1996), "Nonparametric Testing of Closeness between Two Unknown Distribution Functions," *Econometric Reviews* 15, 261-274.

Li, Q. (1999), "Nonparametric Testing the Similarity of Two Unknown Density Functions: Local Power and Bootstrap Analysis," *Nonparametric Statistics* 11, 189-213.

Liebenstein, H. (1966). "Allocative Efficiency vs. 'X-Efficiency'," *American Economic Review* 56, 392-415.

Leibenstein, H. and S. Maital (1992). "Empirical Estimation and Partitioning of X-Inefficiency: A Data-Envelopment Approach," *American Economic Review* 82, 428-433.

Mammen, E. (1992), *When Does Bootstrap Work? Asymptotic Results and Simulations*, Springler, New York.

Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press.

Park, B. Simar, L. and Ch. Weiner (2000), "The FDH Estimator for Productivity Efficiency Scores: Asymptotic Properties", *Econometric Theory* 16, 855-877.

Russell (1990), "Continuity of Measures of Technical Efficiency," *Journal of Economic Theory* 51, 255-267.

Schuster, E. (1985), "Incorporating Support Constraints into Nonparametric Estimators of Densities," *Communications in Statistics*, *Theory and Methods*, 14, 1123-1136.

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

Sheather, S. and M. Jones (1991). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Series B, 53, 683-90.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Simar, L. (1992), "Estimating Efficiencies from Frontier Models with Panel Data : a Comparison of Parametric, Non-parametric and Semi-Parametric Methods with Bootstrapping", *Journal of Productivity Analysis* 3, 167-203.

Simar, L. and P. Wilson (1998), "Sensitivity of efficiency scores : How to bootstrap in Nonparametric frontier models", *Management Science* 44(1), 49-61.

Simar L. and P. Wilson (2000 a), "A General Methodology for Bootstrapping in Nonparametric Frontier Models", *Journal of Applied Statistics* 27, 779-802.

Simar L. and P. Wilson (2000 b), "Statistical Inference in Nonparametric Frontier Models: The State of the Art," *Journal of Productivity Analysis* 13, 49-78.

Simar, L. and V. Zelenyuk (2003), "Statistical Inference for Aggregates of Farrell-type Efficiencies," *Discussion Paper #0324 of Institute of Statistics*, *University Catholique de Louvain*, Belgium.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

APPENDIX

Table A1. Power of the Test: based on True vs. DEA-estimates in different dimension

	$n_A = n_Z = 20$					$n_A = n_Z = 50$				
δ	True	Tech. (8.1)	Tech. (9.1)	Tech. (8.2)	Tech. (9.2)	True	Tech. (8.1)	Tech. (8.2)	Tech. (9.2)	Tech. (8.3)
-1.0	0.691	0.440	0.223	0.116	0.068	0.984	0.955	0.640	0.177	0.065
-0.9	0.592	0.363	0.174	0.102	0.047	0.963	0.927	0.587	0.179	0.055
-0.8	0.553	0.345	0.188	0.096	0.055	0.935	0.892	0.531	0.137	0.052
-0.7	0.450	0.278	0.148	0.085	0.062	0.856	0.780	0.442	0.146	0.063
-0.6	0.359	0.231	0.110	0.065	0.043	0.754	0.688	0.352	0.130	0.057
-0.5	0.268	0.181	0.087	0.063	0.061	0.620	0.559	0.299	0.111	0.067
-0.4	0.206	0.137	0.080	0.066	0.059	0.455	0.391	0.198	0.095	0.047
-0.3	0.155	0.101	0.063	0.062	0.048	0.275	0.267	0.155	0.081	0.051
-0.2	0.101	0.060	0.059	0.067	0.043	0.165	0.135	0.084	0.075	0.039
-0.1	0.049	0.043	0.052	0.045	0.055	0.086	0.090	0.060	0.062	0.070
0.0	0.042	0.038	0.039	0.044	0.054	0.044	0.048	0.049	0.056	0.053
0.10	0.057	0.062	0.059	0.051	0.054	0.082	0.074	0.054	0.058	0.049
0.20	0.117	0.104	0.086	0.053	0.045	0.234	0.208	0.130	0.055	0.061
0.30	0.223	0.166	0.129	0.068	0.055	0.464	0.461	0.274	0.116	0.050
0.40	0.347	0.315	0.206	0.096	0.058	0.740	0.673	0.508	0.194	0.058
0.50	0.591	0.483	0.334	0.153	0.069	0.941	0.892	0.736	0.346	0.081
0.60	0.784	0.661	0.460	0.208	0.080	0.991	0.984	0.913	0.545	0.066
0.70	0.900	0.787	0.608	0.289	0.094	1.000	1.000	0.971	0.753	0.124
0.80	0.971	0.900	0.755	0.435	0.123	1.000	1.000	0.997	0.896	0.159
0.90	0.994	0.942	0.858	0.553	0.151	1.000	1.000	0.999	0.971	0.237
1.00	1.000	0.978	0.927	0.674	0.239	1.000	1.000	0.999	0.988	0.374

Notes: MC = 1000, B = 400, Gaussian kernel is used with the bandwidth selected via Silverman (1986) rule.
* - indicates that the difference from the nominal size is significant at 5% level (not significant at 1% level).
Numbers for sample size 100 in each group is not presented for the sake of space (available upon request).

Table and Figures To Be Inserted into the Text

Table 1. Algorithms of the Bootstrap for Li-test for the Context of Comparing Distributions of Efficiency Scores Estimated via DEA.

1. For each observation in the sample $\Xi_n = \{(x^k, y^k) : k = 1, ..., n\}$ compute $\hat{TE}(x, y)$ (via DEA) thus obtaining some sequence of estimated efficiency scores $\{\hat{TE}^k : k = 1,..., n\}$.
2. For Algorithm I, trim the sample of original efficiency scores from those equal unity. Alternatively, for Algorithm II, smooth the estimates of original efficiency scores according to (7.1). Split the sample estimates into the two sub-samples of DEA estimates, representing each group, A and Z, thus obtaining:

$$\{\hat{TE}^{*A,k} : k = 1, ..., s_A\} \quad \text{(A1)}$$

$$\{\hat{TE}^{*Z,k} : k = 1, ..., s_Z\} \quad \text{(A2)}$$

   where $s_l < n_l$ for Algorithm I and $s_l = n_l$ for Algorithm II ($l = 1, \ldots, L$).
3. Estimate the Li (1996) test statistic (5.5) using the data (A1) and (A2), and bandwidth $h^* = \min\{h^*_A, h^*_Z\}$, where h^*_A, *and* h^*_Z are obtained using some optimal rule applied to (A1) and (A2), respectively.
4. Resample from the largest sub-sample out of (A1) or (A2), in order to obtain the bootstrap analogues of (A1) and (A2), call them

$$\{\hat{TE}_b^{**A,k} : k = 1, ..., n_A\} \quad \text{(A3)}$$

$$\{\hat{TE}_b^{**Z,k} : k = 1, ..., n_Z\} \quad \text{(A4)}$$

5. Estimate the *bootstrapped* Li-test statistic using data (A3) and (A4), and $h_b^{**} = \min\{h_{b,A}^{**}, h_{b,Z}^{**}\}$, where $h_{b,A}^{**}$, *and* $h_{b,Z}^{**}$ are obtained using the some optimal rule (same as in 3) to (A3) and (A4), respectively.

Repeat the steps 4-5: $b = 1, \ldots, B$ times—to obtain B *bootstrap* estimates of the Li-statistic that will mimic the distribution of the original estimate of the Li-statistic under the null hypothesis.

Table 2. Size of the Test: With and Without Using Reflection Method

	Without reflection				With reflection			
n_A , n_Z	0.01	0.05	0.10	0.50	0.01	0.05	0.10	0.50
20	0.012	0.045	0.093	0.509	0.018*	0.056	0.102	0.502
50	0.012	0.063	0.107	0.493	0.015	0.058	0.097	0.512
100	0.013	0.052	0.094	0.498	0.015	0.043	0.097	0.517
200	0.014	0.048	0.089	0.494	0.013	0.050	0.093	0.493

Notes: MC = 1000, B = 400, Gaussian kernel is used with the bandwidth selected via Silverman (1986) rule.
* - indicates that the difference from the nominal size is significant at 5% level (not significant at 1% level).

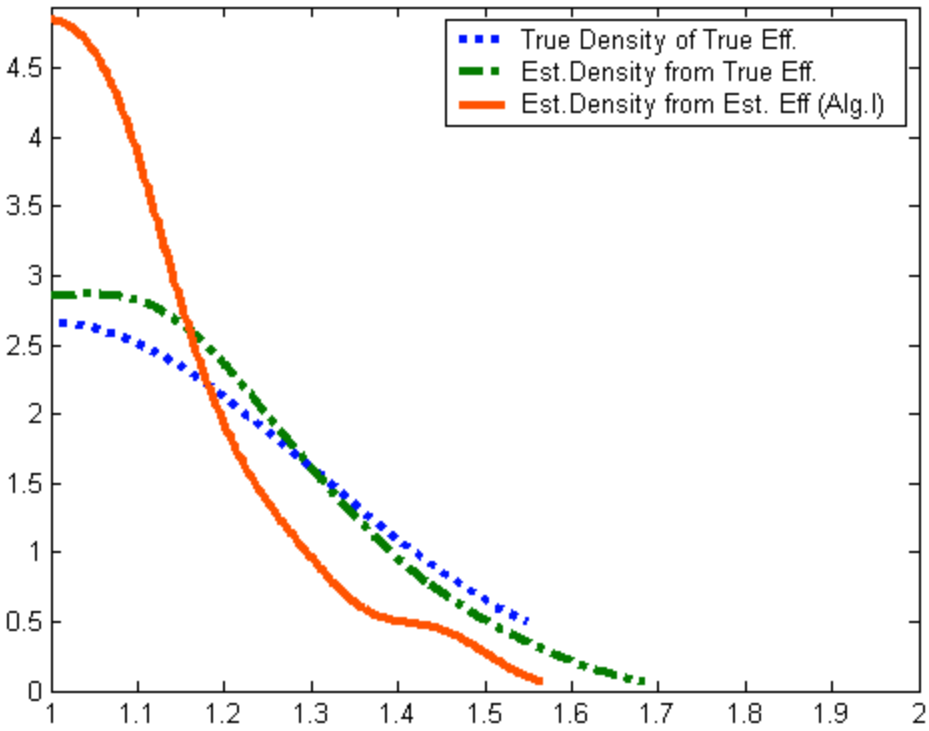

Figure 1. True and Estimated Densities for a Simulated Data, $1+|N(0,0.3)|$.

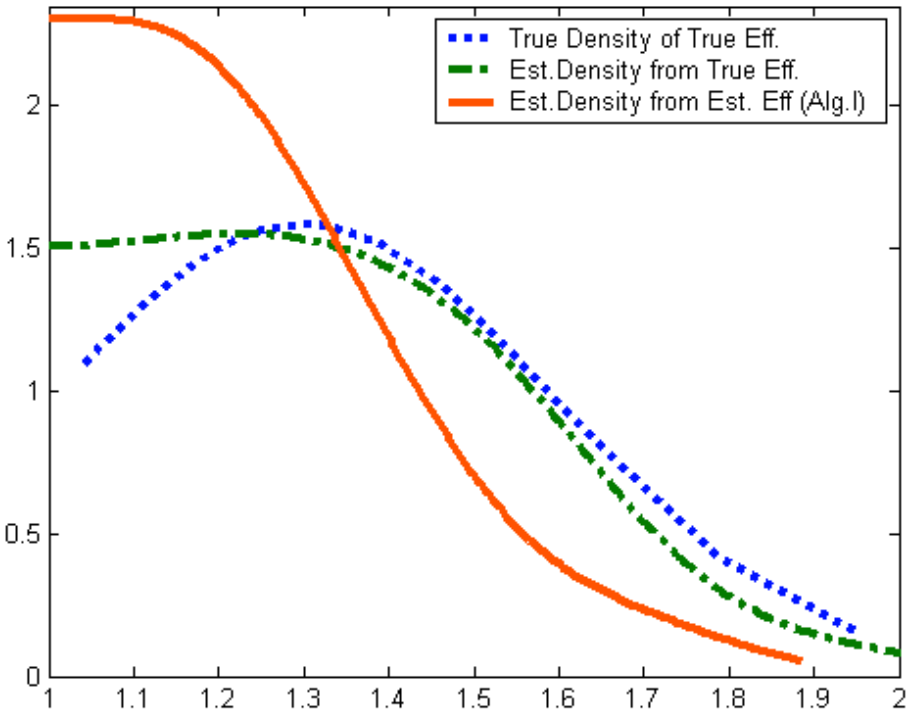

Figure 2. True and Estimated Densities for a Simulated Data, $1+N^{+}(0.3,0.3)$.

Table 3. Size of the Test: Algorithm I vs. Algorithm II against Dimension

$n_A = n_Z$	Algo-rithm	Technology (8.1)				Technology (8.2)				Technology (8.3)			
		0.01	0.05	0.10	0.50	0.01	0.05	0.10	0.50	0.01	0.05	0.10	0.50
20	I	0.008	0.038	0.104	0.522	NA	NA	NA	NA	NA	NA	NA	NA
	II	0.011	0.044	0.098	0.485	0.011	0.050	0.089	0.460*	0.009	0.052	0.099	0.505
50	I	0.013	0.047	0.092	0.492	0.004	0.035*	0.085	0.494	NA	NA	NA	NA
	II	0.007	0.047	0.091	0.501	0.012	0.044	0.093	0.463*	0.017*	0.052	0.093	0.484
100	I	0.006	0.046	0.095	0.501	0.014	0.038	0.084	0.486	0.010	0.045	0.082	0.494
	II	0.009	0.047	0.100	0.514	0.015	0.048	0.085	0.497	0.018*	0.060	0.108	0.474
200	I	0.018*	0.062	0.115	0.496	0.007	0.042	0.094	0.051	0.008	0.044	0.089	0.459*
	II	0.015	0.051	0.106	0.479	0.012	0.060	0.097	0.050	0.017*	0.053	0.100	0.512

Notes: MC = 1000, B = 400, Gaussian kernel is used with the bandwidth selected via Silverman (1986) normal rule.
* - indicates that the difference from the nominal size is significant at 5% level (but not significant at 1% level).

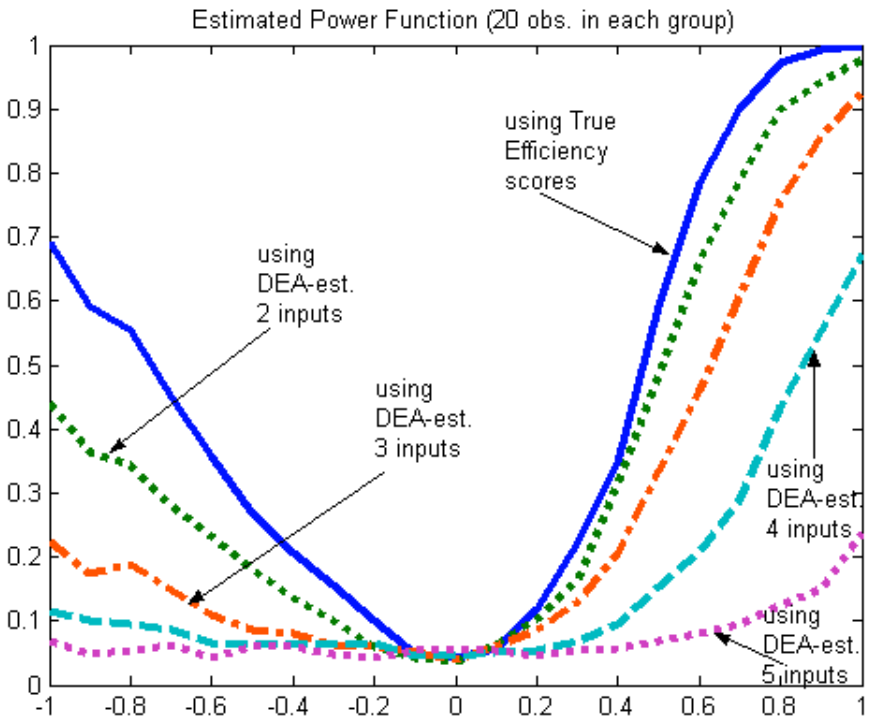

Figure 3. Estimated Power Function (20 observations in each group).

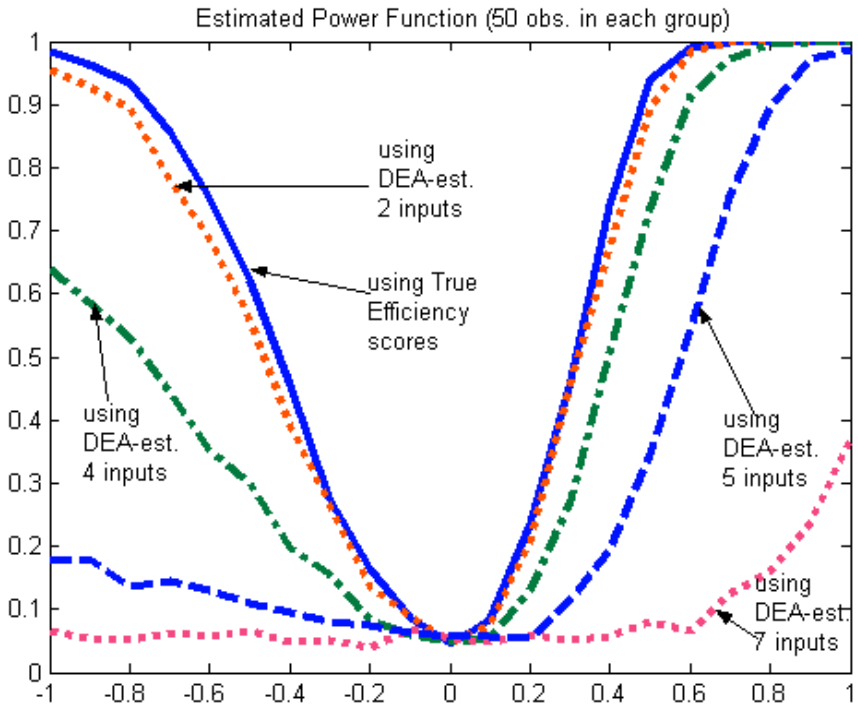

Figure 4. Estimated Power Function (50 observations in each group).

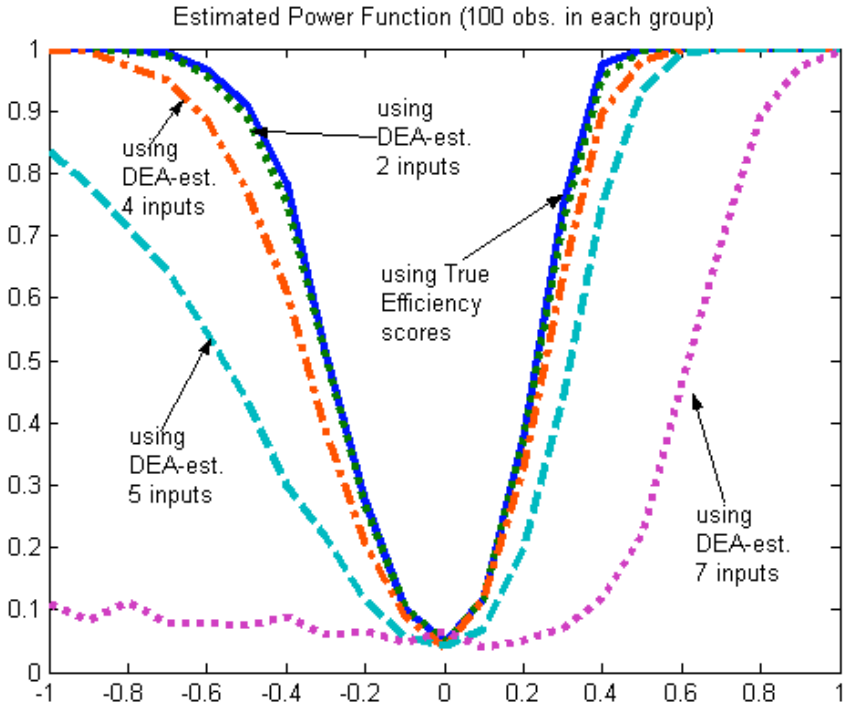

Figure 5. Estimated Power Function (100 observations in each group).