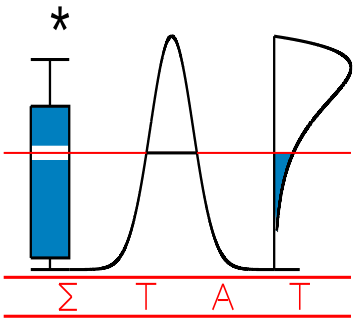
T E C H N I C A L

R E P O R T

0324

ON THE COMPARABILITY OF EFFICIENCY SCORES IN NONPARAMETRIC FRONTIER MODELS

STEINMANN, L. and L. SIMAR



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

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

On the comparability of efficiency scores in nonparametric frontier models

Lukas Steinmann*
Lukas.Steinmann@soi.unizh.ch
Socioeconomic Institute
University of Zurich
Hottingerstrasse 10
8032 Zurich, Switzerland

Léopold Simar†
simar@stat.ucl.ac.be
Institut de Statistique
Université Catholique de Louvain
20 voie du Roman Pays
1348 Louvain-La-Neuve, Belgium

July 11, 2003

Abstract

Data Envelopment Analysis (DEA) is widely used in the field of academic research, in business consulting and in a regulatory context. Usually it is the aim to estimate efficiency scores of decision making units (DMU). The attempt to infer from a sample on the true, but unknown production technology makes it a typical estimation procedure. Banker (1993) and Kneip, Park and Simar (1998) prove that the estimators obtained by DEA are biased, but under certain assumptions are consistent. Efficiency estimates obtained by DEA therefore seem to be suited for hypothesis testing, e.g. for comparison of mean efficiency between groups of observations. However, under certain circumstances—that will be analyzed in this paper—mean efficiency of groups of observations are biased to a different degree and thus differences in mean efficiency are also biased. Without bias correction, hypothesis tests of mean efficiencies between groups are then erroneous. In this paper an indicator is proposed to detect non-comparable mean efficiency scores. The procedure is illustrated in Monte Carlo simulations and applied to a real world data set.

Keywords: production frontier, nonparametric estimation, comparisons of efficiencies.
JEL Classification: D24, L60, 047.

*I would like to thank Peter Zweifel and Markus König (both Socioeconomic Institute, University of Zurich) for a vast number of helpful and valuable comments.

†Research support from "Projet d'Actions de Recherche Concertées" (No. 98/03–217) and from the "Interuniversity Attraction Pole", Phase V (No. P5/24) from the Belgian Government are also acknowledged.

1 Introduction

Data Envelopment Analysis (DEA) as a method to estimate efficiency measures of decision making units (DMU) is facing an increasing interest in the area of academic research, business consulting and regulation. DEA is usually referred to as a nonparametric, deterministic linear programming technique. Efficiency scores represent a potential for an increase in efficiency *relative* to a peer group. Therefore DEA efficiency scores are said to be *relative*, i.e. with regard to the most efficient observations. Not surprisingly, the statistical properties of DEA estimators—such as unbiasedness and consistency—have been neglected in most empirical research published up to date.

However, if one is willing to accept that there is an underlying reference technology, DEA turns out to be a typical estimation procedure: It is the attempt to infer from a sample to the true but unknown underlying technology. Therefore, statistical properties of the efficiency estimator become relevant. In view of a growing literature investigating the statistical properties of DEA efficiency estimators, there is obviously a change of mind. DEA estimators have been shown to be biased, but consistent under certain assumptions [see Banker (1993) and Kneip, Park and Simar (1998)]. Since efficiency is related to the true frontier, this represents the departure from the concept of relative efficiency towards a concept of *absolute* efficiency.

This paper first briefly summarizes the literature on the statistical properties of DEA, emphasizing the underlying assumptions for consistency (section 2). In section 3 the comparability of estimated inter-group mean efficiency will be analyzed. In the following section an indicator to assess the comparability of the relative efficiency estimates will be proposed (section 4), while in section 5 Monte Carlo simulations are presented. In section 6 the procedure is illustrated using a real world data set. Finally, concluding remarks and a summary are presented in section 7.

2 Biasedness and consistency

The DEA estimator of the true but unknown efficiency θ_i is obtained as $\hat{\theta}_i$[1] by solving the following linear problem [see Charnes, Cooper and Rhodes (1978)]:[2]

[1] A "hat" serves as an indicator for an estimated scalar. A single "bar" points to a mean of a vector and two "bars" are used for means of a vector of mean values.

[2] Here, the input oriented DEA version will be presented. Moreover, constant returns to scale (CRS) are assumed throughout this paper. However, the results hold for the output oriented and/or variable returns to scale (VRS) version (if the true technology exhibits VRS).

$$\begin{aligned} & \min & \theta_i & \\ \text{s.t.} \quad & & \theta_l X_i - \mathbf{X}\lambda_i \geq 0 & \\ & & Y_i - \mathbf{Y}\lambda_i \leq 0 & \\ & & \lambda \geq 0. & \end{aligned} \tag{2.1}$$

This linear program is solved for every decision making unit DMU_i ($i \in 1...i...n$) in the sample, where Y_i (X_i) is the output (input) vector of DMU_i of dimension $s \times 1$ ($m \times 1$, respectively). $\mathbf{Y}$ and $\mathbf{X}$ represent the output and input matrices of dimension $s \times n$ and $m \times n$. The scalar $\hat{\theta}_i$ is the efficiency score, indicating the maximally possible proportional reduction of all inputs with given output level.

The statistical properties of interest discussed here are biasedness and consistency of $\hat{\theta}_i$. An estimator is said to be unbiased if and only if the expected value $E(\hat{\theta}_i)$ of the estimator equals the true value θ_i at any given sample size n:

$$E(\hat{\theta}(n) - \theta) = 0 \ \forall n \in I\!N^+. \tag{2.2}$$

Weak consistency on the other hand is a large sample property. It requires that the estimated efficiency converges to the true value with increasing sample size n, that is

$$\lim_{n\to\infty} P(|\hat{\theta}(n) - \theta| < \epsilon) = 1. \tag{2.3}$$

Simar and Wilson (2000b) note:

> *Consistency is an essential property for any estimator. Indeed it would be rather meaningless to use an estimator that does not satisfy consistency, since even with an infinite amount of data, an inconsistent estimator cannot be expected to give an accurate estimate of the quantity of interest. ...we know that DEA (or FDH) efficiency estimators converge as the sample size increases (...), but by themselves, these results have little practical use other than to confirm that DEA and FDH estimators are* ***possibly*** *reasonable to use for efficiency estimation.*

The use of DEA implies the following production possibility set

$$\mathcal{P} = \{(X, Y) \in I\!R^{p+q} \,|\, X \text{ can produce } Y\}. \tag{2.4}$$

The data generating process (DGP), the frontier and the efficiency scores are defined as follows:

– The production process is characterized by the density $f(X, Y)$ of multivariate random variables (X, Y), which has support $\mathcal{P}$ and the frontier is a boundary of $\mathcal{P}$.

– Free disposability and convexity: inputs and outputs are freely disposable and $\mathcal{P}$ is convex.

– The Farrell radial scores are by definition:

$$\theta(X, Y) = \inf\{\theta > 0 \,|\, (\theta X, Y) \in \mathcal{P}\} \tag{2.5}$$

– Deterministic frontier: $\text{Prob}[(X_i, Y_i) \in \mathcal{P}] = 1$. There are no errors in measurement, all DMUs lie in $\mathcal{P}$.

Banker (1993) provided a formal statistical basis for the efficiency evaluation techniques of DEA and proved that if the DGP satisfies certain assumptions, DEA efficiency estimators are consistent. A more general formulation of the DGP, which allows for multiple inputs and outputs is presented in Kneip *et al.* (1998). The following assumptions are sufficient for consistency:

Assumption 1 *(X_i, Y_i) for $i = 1, \ldots, n$ are i.i.d. on the production possibility set $\mathcal{P}$.*

As quoted in Kneip *et al.* (1998), the probability law of (X, Y) induces a density on $f(\theta, \eta, Y)$, where η is the input-mix. This density can be factorized in $f(\theta|\eta, Y)f(\eta, Y)$, where $f(\theta|\eta, Y)$ expresses the fact that the density of the efficiency θ may depend on the value of the input-mix η (and this will be the case in some scenarios analyzed in the following Monte Carlo simulations) and the output level.

Assumption 2 *Likelihood of efficient production: The efficiency density function $f(\theta|\eta, Y)$ takes on a (strict) positive value in the neighborhood of the efficient frontier, i.e. at the fully efficient end.*

3 Comparability of efficiency scores

The question of interest is whether the estimated efficiency scores may or may not be compared between groups of DMUs. Usually, one compares groups of DMUs by analyzing differences in mean efficiency. Comparability requires that group-specific mean efficiencies are biased to the same degree, i.e. that the *difference* is unbiased. Thus, for comparison of mean efficiencies unbiased *differences* in mean efficiency are sufficient. As long as all subsamples exhibit the same bias, comparison and thus hypothesis tests are possible without

increased probability of errors. This section analyzes conditions under which estimated efficiency scores may not be compared between groups of observations by simply looking at the difference in their mean efficiencies. In the remainder of this paper, the following scenario will be analyzed: There are two groups, A (diamonds) and B (circles), of DMUs that form a sample, each DMU produces one single output using two inputs.

DEA implicitly assumes that all observations belong to the same production possibility set. Such a scenario is illustrated in figure 1, left panel. Both groups produce outputs with similar input-mix (technically speaking, they have an almost identical distribution of η). Each part of the frontier serves as a reference for DMUs of both subsamples, assuring that all DMUs are measured with the same scale. Although estimates are biased, biases are about equal for both groups.[3] Consequently, differences in mean group efficiencies are unbiased and thus comparable.

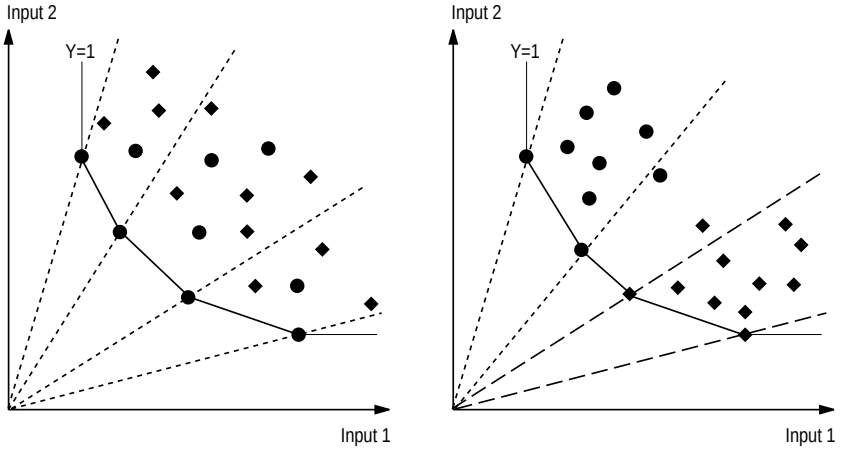


Figure 1: Comparable (left panel) and potentially non-comparable (right panel) mean efficiencies

The assumption of a common production possibility set may be appropriate or not. Supposed that the two groups at evaluation face different relative prices for inputs, it would be optimal to produce with a different input-mix (a different η) to be in accordance with microeconomic theory. Or, it might be that one group is restricted in its choice of inputs, which may result in a similar effect on the input-mix. Such a scenario is illustrated in figure 1, right panel, where the distribution of η is conditional on the membership to the subsample. Obviously, each group produces with different input-mix, the production cones[4] defined by

[3]However, Gijbles, Mammen, Park and Simar (1999) show in the bivariate case that DMU specific biases are not only conditional on the efficiency density but also on the curvature of the frontier. For the moment we consider this source of biases as negligible but we will return to this issue in the simulations.

[4]The term "production cone" is defined here as the intersection between the closed, convex and pointed

each group do not intersect. Both subsamples do not have overlapping distributions of their input-mix η. Thus, all DMUs are evaluated at a reference set defined by DMUs of the same group; they define a group-specific part of the frontier. This means that there is no common benchmark, no common frontier of the production possibility set that assures that the upper limit of the *estimated* efficiency distributions (which is always equal 1 in DEA) reflect the same level of true efficiency.

Obviously the comparability of mean efficiencies is related to assumption 1 which requires that all DMUs are identically and independently distributed in the production possibility set.[5] This assumption fosters groups to share a common frontier of the production possibility set, making them comparable as defined above. In figure 1, left panel, DMUs are likely to be identically distributed; their input-mix η does not depend on the membership to one of the two groups: $f(\eta_A) = f(\eta_B)$. On the other hand, in figure 1, right panel, the two groups differ with respect to input-mix, the distribution of η is conditional on membership of subsample, *i.e.* $f(\eta_A) \neq f(\eta_B)$. The two groups are thus not identically distributed when considering them as pooled sample.

But even with identical input-mix distribution, the two subsamples may also differ with respect to their efficiency density functions, which also contradicts the i.i.d. assumption in a pooled sample. This is the case when $f(\theta_A|\eta, Y) \neq f(\theta_B|\eta, Y)$. Thus, we consider two sources of differences in the distribution of DMUs in the input-output-space: first, differences in input-mix distributions and second, differences in conditional efficiency distributions. In any case, however, consistency is ensured as long as assumption 2 is satisfied—with increasing the two subsample sizes, the estimated frontier approaches the true frontier. No matter if they are comparable or not.

In small samples however, the degree of the bias depends on the probability of observing DMUs arbitrary close to the boundary. This probability depends on subsample sizes and efficiency density functions. In non-comparable scenarios [as presented in figure 1, right panel], subsamples may define their own frontiers. In this case, if the two groups have different efficiency density functions and/or unequal subsample sizes, finite sample biases will be group specific and differences will be biased. Since the group specific bias is unknown the efficiency scores and their mean values may not be compared between the two subsamples unless more elaborate and demanding techniques are applied [for instance the general "heterogeneous"

cone defined by one facet of the frontier and the production possibility set. All DMUs inside the same production cone are evaluated at the same facet or, equivalently, peer group.

[5]It is worth noting that this is an assumption about the allocation of DMUs in the input-output space that concerns the data generating process (DGP), i.e. how DMUs are distributed in the true but unknown production possibility set. Estimated efficiency scores on the other hand are not independently distributed since the scores of some DMUs depend on other DMU's location, the DMUs that define the corresponding part of the frontier.

bootstrap algorithm, see Simar and Wilson (2000b)].

Conclusion 1 *In different input-mix scenarios, unbiased differences between groups of DMUs crucially depend on the efficiency density functions in conjunction with subsample sizes. Therefore scenarios with different input-mix will be called potentially non-comparable. Comparison of mean efficiencies requires a bias correction which in turn calls for more elaborate and demanding procedures such as the general bootstrapping algorithm.*

In small samples with the same input-mix for the two groups, estimated efficiency scores are biased at least by the ratio between the estimated efficiency score of the most efficient DMU (which is per definition one) and its true efficiency. Supposed *all* DMUs defining the efficient frontier have a true efficiency of e.g. 0.9, all efficiency scores are biased upwards by factor 1/0.9 [see figure 2 left panel]. For the most efficient DMUs, like DMU 1, the ratio between the true efficiency a/a' and the estimated efficiency $a'/a' = 1$ equals a/a'. The same holds for inefficient observations, like DMU 2, where the ratio between true efficiency and estimated efficiency ($\frac{a/c}{a'/c}$) is also equal to a/a'.

Two issues have been neglected up to now. First, it will rarely be the case that *all* DMUs defining the efficient frontier have the same true efficiency and thus relative biases will not be exactly equal for any DMU in the sample. In figure 2, right panel for example, DMU 1 has a higher relative bias than DMU 2. DMU 3 is projected on the reference set defined by DMU 1 and DMU 2, and may be seen as a linear combination of these two DMUs—obviously its relative bias is also a linear combination. Second, the curvature of the true frontier is responsible for an additional bias for DMUs that are projected on the "middle" of a reference set, cf. DMU 3 in figure 2 left panel, the bias is illustrated by the shaded area.

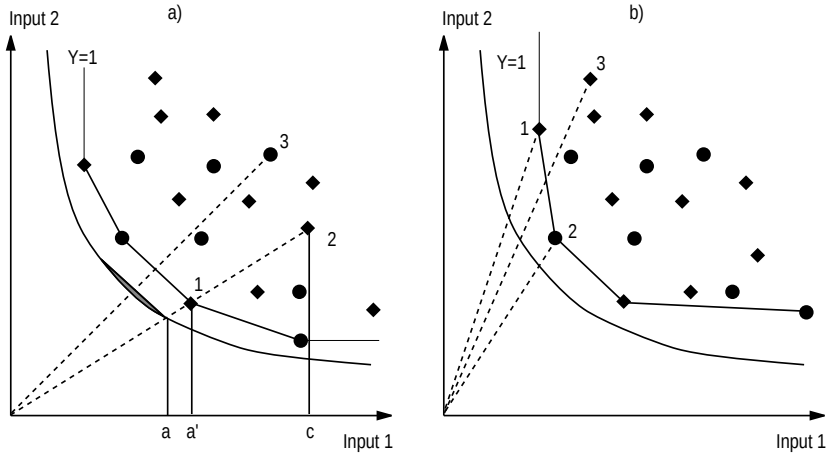


Figure 2: Differences in relative biases

However, as long all DMUs are i.i.d., both expected biases due to the curvature and unequal finite sample biases are on average equal for both subsamples and consequently, the difference in mean efficiency is expected to be unbiased.

Conclusion 2 *If the subsamples share a common distribution of input-mix (i.e. are i.i.d. w.r.t. input-mix), expected differences of mean efficiency between groups of DMUs are likely to be unbiased—regardless of assumptions about efficiency density functions. Therefore a scenario characterized by a common input-mix will be called comparable.*

4 A method to assess comparability

In this section a measure to assess the comparability of mean efficiencies will be proposed. We have seen above that this crucially depends on the fact that both subsamples share a common distribution of their input-mix. The underlying idea is based on the weak monotonicity property of DEA: a proportional increase of all inputs of a DMU must lead to an inverse decrease of its efficiency score. This monotonicity property is now applied on the whole subsample: if a change of the data of one subsample does not result in a corresponding change in the difference of the mean efficiencies, subsamples are likely to define their own production cones and comparability of group mean efficiencies between the groups may be flawed by differences in biases.

Coming back to figure 1, left panel, it appears that a proportional increase of all inputs of group A (diamonds) by a factor $\Psi = 1 + \varepsilon$, where ε is arbitrary small[6], would result in an inversely decreased efficiency of all DMUs belonging to subsample A. Normalizing the relative change in the mean efficiency (measured in percentage points, see the numerator in equation 4.1), by the expected relative change (again, in percentage points, see the denominator in 4.1) due to the arbitrary small increase of all inputs would result in a value of 1:

$$\Delta_A = \frac{\frac{\bar{\hat{\theta}}_A - \bar{\hat{\theta}}_A^{\Psi}}{\bar{\hat{\theta}}_A}}{1 - 1/\Psi} = 1 \tag{4.1}$$

Conversely, efficiency scores of all DMUs of group A (diamonds) will remain unchanged in the non-comparable case [see figure 1, right panel], and hence the numerator and thus Δ_A take on a value of zero:

$$\Delta_A = \frac{\frac{\bar{\hat{\theta}}_A - \bar{\hat{\theta}}_A^{\Psi}}{\bar{\hat{\theta}}_A}}{1 - 1/\Psi} = 0. \tag{4.2}$$

Suppose that DMUs of subsample A define the frontier [like in the left panel of figure 1 where DMUs of group A are now be represented by circles]. Here, an arbitrary small increase

[6]Sensitivity analysis with respect to the choice of ε will be discussed below.

Case	Group A	Δ_A	Δ_B	Δ_{AB}
Figure 1, left panel	diamonds	1	0	1
Figure 1, right panel	diamonds	0	0	0
Figure 1, left panel	circles	0	1	1

Table 1: Values of Δ for figure 1

of all inputs would leave group A's mean efficiency unaffected since the frontier is shifted to the north-east. Accordingly, Δ_A will take on a value of zero. However, mean efficiency of group B *proportionally* increases since all DMUs of group B now lie closer to the shifted frontier. The relative change of the mean efficiency of group B, normalized by the the relative initial shift results in:

$$\Delta_B = \frac{\frac{\bar{\hat{\theta}}_B - \bar{\hat{\theta}}_B^{\Psi}}{\bar{\hat{\theta}}_B}}{1 - \Psi} = 1. \tag{4.3}$$

Thus, the difference in mean efficiency of both groups is affected as expected, i.e. according to the shift of the frontier. With subsample specific input-mix, a change of the inputs of subsample A will not affect group B's mean efficiency, i.e. $\Delta_B = 0$.

A change in the difference of the mean efficiencies due to a change of the inputs of group A may materialize through a change of the mean efficiency of either group. As long as a change leads to a corresponding change in the difference in the mean efficiency, the monotonicity property is satisfied, indicating that both groups largely share the same frontier. Δ_A and Δ_B measure the degree of transmission of a change of the data on each groups' mean efficiency. Taken together, they indicate how a change in group A's data affects estimated difference in mean efficiency:

$$\Delta_{AB} = \Delta_A + \Delta_B. \tag{4.4}$$

Table 1 summarizes the different scenarios described above.

Thus, Δ_{AB} may serve as an indicator of difference of input-mix distribution: if Δ_{AB} equals zero[7], production cones of either group do not contain any DMUs of the other group. In such a situation, they do not have a common frontier and biases in mean efficiencies might be group specific, depending on efficiency density functions and subsample sizes. In fact, comparison will be valid only if, by chance, the two subgroups have the same conditional efficiency distribution and the same sample size. In practice, when $\Delta_{AB} \approx 0$, comparison of mean efficiencies is likely to be erroneous, unless more elaborate statistical techniques are used (appropriate group specific bias correction, ...).

If Δ_A takes on a value close to one, (almost) all DMUs belonging to group A lie inside the production cone(s) defined by group B. At the same time Δ_B must be close to zero, and

[7]This implies that both Δ_A and Δ_B are equal to zero since they are both defined on a domain [0,1].

both Δs are summing up to around one. Values of Δ_B close to one indicate that (almost) all DMUs of group B lie inside group A's production cone; this again implies that Δ_A is close to zero. However, they again sum up to one, indicating that the change of the data leads to a corresponding change of the estimated difference in mean efficiency. In these situations, the monotonicity property holds when applied to group means instead of single DMUs.

However, with real world data it will rarely be the case that one group defines the whole production possibility set or that both groups do not share any parts of their relevant production possibility sets. Some DMUs of one group may define parts of the efficient frontier and be a reference for DMUs of both groups while the other group also defines parts of the frontier which serves as a reference for DMUs of both groups. It will also be the case that DMUs of both groups define some parts of the frontier that constitutes a reference for DMUs of one or both subsamples. In these cases Δ_A and Δ_B will take on values between zero and one. However, their sum is again indicating to what extent the change of the data materializes in a change of the difference in the mean efficiencies between the groups.

5 Monte Carlo Simulations

The following DGP has been specified for the Monte Carlo simulations. In particular a Cobb-Douglas production function has been chosen:

$$\mathcal{P} = \{(X,Y) \in I\!R^{p+q} | F(Y,X) < 0)\}, \;\; F(Y,X) = y - x_1^{\beta} x_2^{1-\beta}. \tag{5.1}$$

The output quantities are realizations of a normal distribution with an expected value of 100 units and standard deviation of 20.[8]

The input-mix η is also a normally distributed random variable with an expected value of $\frac{1}{4}\pi$ and a standard deviation of $\frac{1}{32}\pi$. Since this parameter is crucial for the i.i.d. assumption its expected value will be altered throughout the trials. In the case with $f(\eta_A) = f(\eta_B)$ the i.i.d. assumption is satisfied, in all other trials where η is conditional on group membership, subsamples differ with regard to their expected input-mix $[f(\eta_A) \neq f(\eta_B)]$.

Once outputs and input-mix are obtained, efficient input quantities x_1^*, x_2^* may be computed. Since each set of x_1^*, x_2^*, y lies on the frontier, radial inefficiency is introduced by multiplying the efficient values with an inefficiency term (cf. equation 5.2). The factor of excess inputs is a one-sided random variable that is gamma distributed[9]:

$$x_m = x_m^*(1+\gamma) \;\; m = 1, 2. \tag{5.2}$$

[8] In the quite rare case of a negative output quantity the value is set equal to one.

[9] The gamma probability density function is $f(x|\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{(\alpha-1)} e^{-\frac{x}{\beta}}$. The expected value is $\alpha\beta$, the variance $\alpha\beta^2$.

$\eta_A = \eta_B = \frac{1}{4}\Pi$, $\alpha_A = \alpha_B = 2$, $\beta_A = \beta_B = 0.125$, # runs: 10,000			
	group A	group B	Difference/Ratio
Expected Efficiency $\overline{\overline{\theta}}$	79.9922	80.0093	-0.0171
Estimated Expected Efficiency $E(\overline{\overline{\widehat{\theta}}})$	84.3280	84.3472	-0.0192
Estimated absolute Bias $\mathcal{B}_a = E(\overline{\overline{\widehat{\theta}}}) - \overline{\overline{\theta}}$	4.3358	4.3378	-0.0020
Estimated relative Bias $\mathcal{B}_r = E(\overline{\overline{\widehat{\theta}}})/\overline{\overline{\theta}}$	1.054	1.054	1.0000
$\overline{\Delta}_A, \overline{\Delta}_B, \overline{\Delta}_{AB}$	0.5492	0.4085	0.9577

Table 2: Equal input-mix and efficiency densities

There will be two different density functions of excess inputs in the Monte Carlo simulations. The first one has an expected value of 0.25 ($\alpha = 2, \beta = 0.125$), which results in an expected efficiency of 0.8 (according to equation 5.2). The central 95 percent of all DMUs exhibiting this density function have a true efficiency between 0.59 and 0.97. The second one has an expected value of about 0.42 ($\alpha = 2, \beta \approx 0.21$), which leads to an expected efficiency of 0.7. Here, 95 percent of all DMUs will have a true efficiency between 0.46 and 0.95.[10]

In the following sections four different trials will be presented, the combinations of equal/nonequal input-mix and the two density functions. The main goal is to verify if Δ_{AB} serves to detect different input-mix distributions and thus to identify potentially non-comparable mean efficiency scores. Second, it will be seen that, as pointed out above, with the i.i.d. assumption violated, unbiased differences in inter-group mean efficiency crucially depend on the conditional efficiency density functions of the two groups[11]. In each section the expected mean efficiencies, the estimated expected mean efficiencies, and their differences (i.e. the estimated absolute bias) and their ratios (i.e. the estimated relative bias) are presented along with Δ_{AB}.

5.1 Trial 1: Equal input-mix densities and efficiency densities

In the first trial both groups share the same parameters, which should result in comparable efficiency scores. This should be characterized by a Δ_{AB} close to one. It is expected that biases are of the same magnitude for both groups and therefore the differences in mean efficiencies are unbiased. The results with $\Psi = 1.01$ are displayed in table 5.1. In the header of the table, all parameters are presented. For example, both groups have an expected input

[10] These two efficiency density functions and their cumulative functions are presented in figure 6 in the Appendix.

[11] To save place, the dependence on subsample sizes will not be explored explicitly, since this is straightforward. So we will always use $n_A = n_B = 50$.

mix of $E(\eta_{A,B}) = \frac{1}{4}\pi$, which is equal to 45 degrees. In total, 10,000 runs have been performed with this parameter setting.

The first row of the table contains the expected efficiency, the mean of the true mean efficiencies of every run (all efficiencies are presented in percentage points). The last column shows the difference in the expected mean efficiency between the two subsamples. In the next row the mean of estimated mean efficiencies of every run is reported, the difference between the first and the second row are the estimated (absolute) biases in mean efficiency (row 3). The fourth row reports the ratios of the estimated expected and the expected mean efficiencies, which indicates to what degree efficiency scores were upward biased throughout this trial. Since DEA efficiency scores are always upwards biased, theses ratios, which are a measure for relative biasedness, must be equal or larger than one. In the last column of the fourth row, the ratio of the relative biases are reported. If this ratio takes on a value of one, mean efficiencies of both groups are equally biased, values below one indicate that subsample B's mean efficiencies were more biased and vice versa.

As expected, biases are about the same for both groups, and differences in mean efficiency take on a value of 0.002 percentage points, a bias that certainly may be ignored. $\overline{\Delta}_{AB}$ takes on a value close to one—as expected. It is interesting to see that $\overline{\Delta}_A$ takes on a larger value than $\overline{\Delta}_B$. In this specific setup with identical DGPs for both groups one would expect that both elements of the indicator equal 0.5. On average each subsample should be equally frequent involved in the definition of the production possibility set, at least for marginal changes of the excess inputs (i.e. a Ψ close to one). Once group A's DMUs are shifted away from the frontier, the rest of the excess inputs only affects $\overline{\Delta}_A$. The rather small $\Psi = 1.01...$ in the simulation proved to be large enough to give $\overline{\Delta}_A$ (=0.5492) more weight compared to $\overline{\Delta}_B$ (=0.4085).

An open issue is how to choose ε or equivalently, Ψ. On principle, ε should be as small as possible. However, one must keep in mind that computers have limited precision and thus a ε too close to zero may result in a change of the data that may not be recognized by the DEA software. The simulations indicated that different values of ε resulted in almost equal Δ_{AB}. However, this may be due to the specific DGP of the simulation and may not be generalized. Moreover, in an empirical context the DGP is unknown and consequently rules based on a DGP are useless.

Therefore, we propose the following procedure, illustrated with one randomly chosen data set produced in trial 1. We computed Δ_A, Δ_B and Δ_{AB} for different Ψ. The results of this procedure are plotted in figure 3. Δ_{AB} is almost constant over a wide range of Ψ. With increasing values of Ψ, group A's DMUs will be moved further away from the true frontier. This is reflected by an increasing value of Δ_A and the decreasing value of Δ_B, which indicates

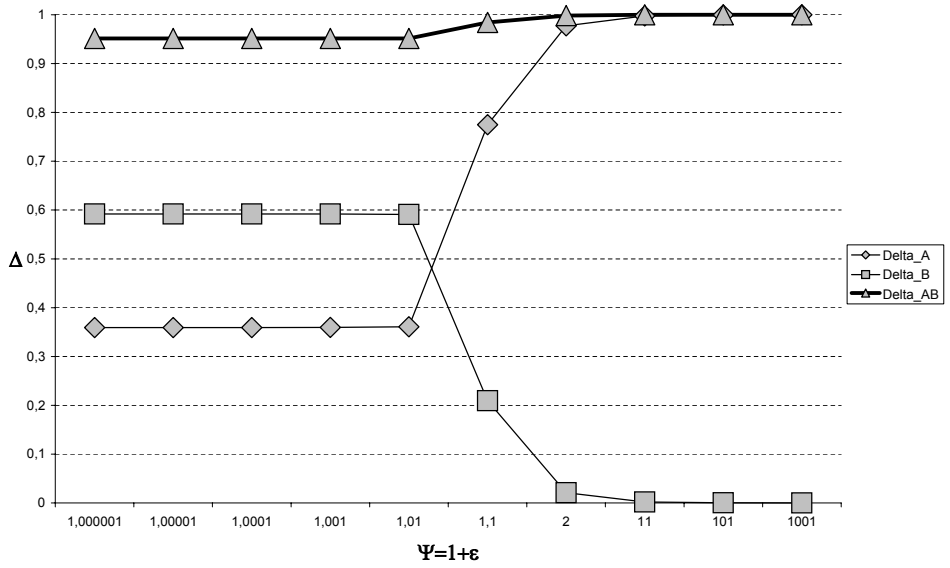


Figure 3: Dependence of Δ w.r.t. Ψ in trial 1. The scale for Ψ is logarithmic.

that more and more DMUs of group B define the frontier. Finally the production possibility set is defined by DMUs of group B only, Δ_A and Δ_{AB} both take on a value of one, while Δ_B equals zero.

In this case the choice of Ψ has no effect on the value of Δ_{AB}. However, the two components reveal some information about the structure of the data. Δ_A and Δ_B may be interpreted as the share of observations of the corresponding group that lie inside the other group's production possibility set. With increasing Ψ, both Δ_A and the share of DMUs belonging to group A that lie inside the frontier defined by group B, increases.

5.2 Trial 2: Different input-mix densities, but equal efficiency densities

In the second trial both groups again share the same density function of excess inputs, but they differ with respect to the expected input-mix η. This should result in a situation roughly described in figure 1, right panel. Consequently, it is expected that the disjunct production cones will result in a Δ_{AB} close to zero. Moreover, it is also expected that estimated mean efficiency is slightly larger than in the previous trial since both groups produce with different

$\eta_A = \frac{1}{16}\Pi$, $\eta_B = \frac{7}{16}\Pi$, $\alpha_A = \alpha_B = 2$, $\beta_A = \beta_B = 0.125$, # runs: 10,000			
	group A	group B	Difference/Ratio
Expected Efficiency $\overline{\overline{\theta}}$	80.0177	80.0606	-0.0429
Estimated Expected Efficiency $E(\overline{\overline{\widehat{\theta}}})$	86.5551	86.5636	-0.0085
Estimated absolute Bias $\mathcal{B}_a = E(\overline{\overline{\widehat{\theta}}}) - \overline{\overline{\theta}}$	6.5374	6.5029	0.0345
Estimated relative Bias $\mathcal{B}_r = E(\overline{\overline{\widehat{\theta}}})/\overline{\overline{\theta}}$	1.0817	1.0812	1.0004
$\overline{\Delta}_A, \overline{\Delta}_B, \overline{\Delta}_{AB}$	0.0041	0.0038	0.0079

Table 3: Different input-mix, but equal efficiency densities

input-mix, virtually dividing the sample into two parts [of half the size]. However, since both subsamples have the same size and density functions, the expected biases in mean efficiency should be equal and differences in mean efficiency should therefore be unbiased. The results are in table 5.2.

These results indicate that Δ_{AB}, which takes on a value close to zero, serves as a measure to reflect disjunct production cones. As expected, biases are larger than in the previous trial. This is the consequence of a virtually partitioned sample, in which the probability of an almost efficient DMU is reduced, but still being equal for both subsamples. As expected, and commented above, estimated mean efficiencies are indeed equally biased and differences in mean efficiency are unbiased. So here, in this trial, it turns out that the group means can be compared. But this is only the case since both subsamples have identical efficiency distributions and subsample sizes in this particular Monte-Carlo scenario. However, the value of Δ_{AB} warns us that there might be a problem because the input-mix are different.

For illustration, we again did the sensitivity analysis and computed Δ_A, Δ_B and Δ_{AB} for different values of Ψ, as displayed in figure 4. In trial 2, Δ_A, Δ_B and Δ_{AB} are again almost constant over a wide range of Ψ. However, with increasing values of Ψ group A's DMUs will be moved further away from the true frontier. But since group B has another input-mix they define their own frontier. Only for very large values of Ψ, DMUs belonging to group A will be shifted into the production possibility set defined by group B. As this happens, Δ_A starts to increase. Finally, all DMUs of group A lie inside the production possibility set solely defined by DMUs of group B and Δ_A and Δ_{AB} both take on a value of one.

This example shows that the information of interest is extracted for small values of Ψ where Δ_A, Δ_B and Δ_{AB} are almost constant. Again, Δ_A and Δ_B may be interpreted as the share of DMUs of the corresponding group that lie inside the other group's production possibility set.

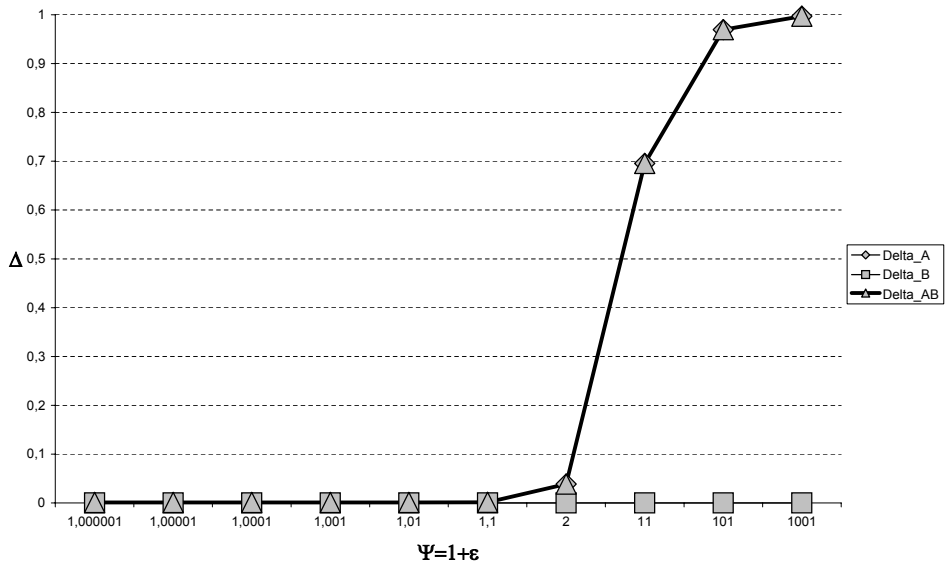


Figure 4: Dependence of Δ w.r.t. Ψ in trial 2. The scale for Ψ is logarithmic.

5.3 Trial 3: Equal input-mix densities, but different efficiency densities

In the third trial both groups have their own efficiency density function. Group A has an expected mean efficiency of 0.8, while group B should be 10 percentage points less efficient. However, they again have an equal expected input-mix. $\overline{\Delta}_{AB}$ is thus expected to take on a value of one, which should be due to a rather high $\overline{\Delta}_B$, indicating that most of group B's DMUs indeed lie inside group A's production possibility set. Furthermore, it is expected that biases lie between those in the first and the second trial. The reason for this is that group B may in principal contribute in the definition of the production possibility set but will not do so with the same frequency as in trial 1 since the probability of observing a almost fully efficient DMU is smaller. As shown in table 5.3, again, $\overline{\Delta}_{AB}$ takes on a value close to one, and as expected, $\overline{\Delta}_B$ is larger, confirming that group B's DMUs lie inside group A's production cones in most runs. It is surprising that the estimated expected bias is larger in absolute value for group B, even though only by about 1 percentage point. Thus 90 percent of the true difference in mean efficiency is identified by DEA in this specific setup. However, this in contrast to the prediction which states that differences should be unbiased

$\eta_A = \eta_B = \frac{1}{4}\Pi$, $\alpha_A = \alpha_B = 2$, $\beta_A = 0.125, \beta_B \approx 0.214$, # runs: 10,000			
	group A	group B	Difference/Ratio
Expected Efficiency $\overline{\overline{\theta}}$	80.0191	70.0843	9.9348
Estimated Expected Efficiency $E(\overline{\overline{\hat{\theta}}})$	84.8852	75.9387	8.9465
Estimated absolute Bias $\mathcal{B}_a = E(\overline{\overline{\hat{\theta}}}) - \overline{\overline{\theta}}$	4.8662	5.8544	-0.9882
Estimated relative Bias $\mathcal{B}_r = E(\overline{\overline{\hat{\theta}}})/\overline{\overline{\theta}}$	1.0608	1.0835	0.9790
$\overline{\Delta}_A, \overline{\Delta}_B, \overline{\Delta}_{AB}$	0.3070	0.6575	0.9645

Table 4: Equal input-mix, but different efficiency densities

for $\overline{\Delta}_{AB} \approx 1$. The explanation is that in this setup DMUs of group B are much more likely to be projected on a weak efficient part of the frontier defined by DMUs of group A since their expected efficiency is lower. By changing the data by factor Ψ these DMUs change their efficiency scores as expected and contribute to a $\overline{\Delta}_{AB} \approx 1$.

However, efficiency scores of DMUs that are projected on weak efficient parts are by definition even more upwards biased than their reference DMUs. Thus, group B's mean efficiency is slightly more upwards biased, resulting in biased difference in mean efficiency. If computation of Δ_{AB} would be restricted on the subsample of DMUs that are on or projected on the Pareto-Koopmans efficient frontier, differences of mean efficiencies should be unbiased. In that case, biases in the difference in mean efficiency are due to unequal shares of DMUs that are projected on weak efficient parts of the frontier and not due to a lack of comparability. Given comparability, this is indeed likely to be the case since group B's expected efficiency is lower than the one of group A.

In order to save space, we do not present the results of the sensitivity analysis for trial 3 and trial 4 since, as above, the curves are shaped as predicted.

5.4 Trial 4: Different input-mix densities and efficiency densities

In this trial both groups differ with respect to group specific input-mix η and efficiency densities. This parameter setting should result in non-comparable situations. Therefore $\overline{\Delta}_{AB}$ is expected to take on a value close to zero. Since the probability mass at the fully efficient end of the efficiency density functions differ between the two subsamples, expected biases should not be equal, leading to biased differences in mean efficiency. This not only makes comparison impossible without bias correction but also renders statistical hypothesis tests erroneous. The results are in table 5.4. As expected, $\overline{\Delta}_{AB}$ takes on a value of around zero, indicating that comparability is not assured. All expectations are verified, particularly,

$\eta_A = \frac{1}{16}\Pi$, $\eta_B = \frac{7}{16}\Pi$, $\alpha_A = \alpha_B = 2$, $\beta_A = 0.125, \beta_B \approx 0.214$, # runs: 10,000			
	group A	group B	Difference/Ratio
Expected Efficiency $\overline{\overline{\theta}}$	80.0303	70.0731	9.9569
Estimated Expected Efficiency $E(\overline{\overline{\widehat{\theta}}})$	86.5889	79.8692	6.7197
Estimated absolute Bias $\mathcal{B}_a = E(\overline{\overline{\widehat{\theta}}}) - \overline{\overline{\theta}}$	6.5586	9.7961	-3.2375
Estimated relative Bias $\mathcal{B}_r = E(\overline{\overline{\widehat{\theta}}})/\overline{\overline{\theta}}$	1.0819	1.1396	0.9489
$\overline{\Delta}_A, \overline{\Delta}_B, \overline{\Delta}_{AB}$	0.0037	0.0075	0.0112

Table 5: Different input-mix and efficiency densities

the one that the estimated expected bias of the group with less likely fully efficient DMUs is larger resulting in biased differences in mean efficiency. While the true difference should be around 10 percentage points, DEA indicates an estimated expected difference of only 6.7 percentage points. One third of the difference is obscured by biases.

5.5 Conclusions of the Monte-Carlo experiments

These simulations indicate that Δ_{AB} is suited for identifying non-comparable production technologies. For $\Delta_{AB} \approx 0$ both subsamples define their own reference sets, having different input-mix distributions, which makes comparison of mean efficiency hazardous. Differences in mean efficiency are then only unbiased under very restrictive assumptions [same conditional density for the efficiencies and same subsample sizes] that are likely to be violated. To assess a correction for the group specific biases, time consuming and demanding bootstrap procedures must be applied. Values close to one on the other hand indicate comparable efficiency scores. However, it is not generally possible to infer directly and only from $\Delta_{AB} \approx 1$ on unbiased differences in estimated mean efficiency, but this is likely to be the case when the sample sizes increase.

The proposed indicator serves to decide if mean efficiencies may be compared or not and whether the general bootstrap procedure must be used to compute bias corrections.

6 An illustration with data on Program Follow Through schools

The proposed indicator Δ_{AB} and procedure will be illustrated with data on program follow through (PFT) of an experimental education program administered in US schools [for a description of the data see Charnes, Cooper and Rhodes (1981)]. First, the DEA efficiency

scores of the 70 schools in the data were computed. Mean efficiency of the group of schools with PFT (group A) is around 95.5% while the schools without PFT (group B) performs 93% on average.[12] Then Δ_A, Δ_B and Δ_{AB} are computed

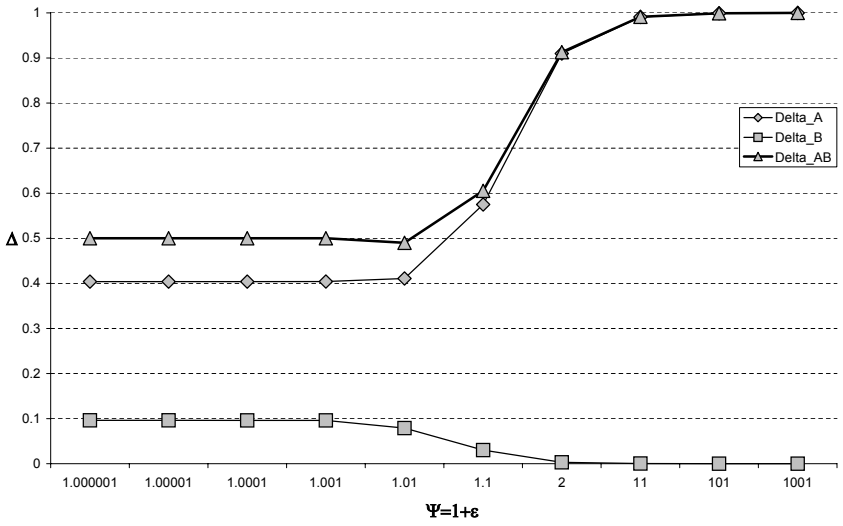


Figure 5: Dependence of Δ and Ψ in the PFT. The scale for Ψ is logarithmic.

The results are presented in figure 5. The value of Δ_{AB} is almost constant for $\Psi < 1.01$ and takes on a value of 0.5. For higher Ψ, Δ_{AB} increases, which indicates that schools without PFT more and more begin to define the efficient frontier as schools with PFT are moved further away from the true frontier. As expected, the bulk of schools with PFT are assigned inefficient when their inputs are multiplied by a large Ψ. This may also be seen when looking at the plot of Δ_B, which declines with increasing Ψ, indicating that the estimated frontier consists of non PFT schools only. In the end, mean efficiency of the schools without PFT is not affected by a change of the inputs of schools with PFT and $\Delta_B \approx 0$.

In the illustrated case the indicator Δ_{AB} takes on a value of about 0.5 for Ψ up to 1.01. This indicates that the data on PFT schools may not be compared to the four extreme

[12]In contrast to Charnes *et al.* (1981), we use DEA with constant returns to scale. This has been done to be in accordance with the previous parts of this paper. However, the restriction of constant returns to scale does not seem to be severe with this data set since the differences in mean efficiency are marginal (1% and 1.8% for group A and B, respectively).

scenarios in the simulations. The two groups do not have an entirely common frontier but obviously they are not characterized by non-overlapping production cones. The subsamples probably define group-specific parts of the frontier. It seems that about 10% of all schools without PFT are projected on segments of the frontier that are defined by schools with PFT ($\Delta_B = 0.1$). On the other hand, according to Δ_A, about 40% of all schools with PFT lie in areas with a boundary defined by schools without PFT.

This introduces the risk of biased difference of mean efficiencies. Interpretation without bias correction would probably lead to erroneous conclusions. Indeed, results from Simar and Wilson (2000b) indicate that the (heterogenous) bias correction is larger for schools without PFT. The proposed indicator Δ_{AB} suggests the following explanation: The fact that the two groups define their own frontier to a considerable extent—as indicated by Δ_{AB}—and since there are less schools without PFT, their bias in mean efficiency is expected to be larger. This is an *ex-post* explanation for the difference in the bias correction. However, it would be more efficient to compute the proposed indicator first and then compute the appropriate bias correction.

7 Conclusions

This paper analyzes under which conditions mean efficiency between groups of observations may be compared, which is the case if differences in mean efficiency are unbiased. Estimates of differences in mean efficiency may be either unbiased due to (1) a favorable data generating process (DGP) or (2) an appropriate bias correction.

In this paper it is argued that as long as both groups share the same reference set, the expected *difference* in mean efficiencies is unbiased although individual efficiency scores are biased. Such a scenario with comparable mean efficiencies turns out to be ensured as long as the i.i.d. assumption of the DGP is satisfied. In other scenarios with non-comparable, i.e. biased differences in mean efficiencies, a bias correction is necessary. Since these scenarios occur with group specific GDPs, the general bootstrap procedure must be performed, a demanding and tedious task. Before facing the costs of the bootstrapping procedure, one would like to know if this effort is necessary or not. We propose a simple, easily computable measure to assess the degree of comparability.

One could argue that it would be more straightforward to check the i.i.d. assumption using traditional statistical procedures such as correlation analysis of the input-mix between subsamples. Indeed, in the two-input one-output case this would be quite easy. The open question would, however, still be the threshold value of the correlation coefficient for rejection of the i.i.d. assumption. Moreover, with an increasing number of inputs and outputs,

comparability depends on an exponentially increasing number of input-mix, output-mix and also input-output-mix. While some input-mix (and output-mix and also input-output-mix) might be uncorrelated, others may be highly correlated—it would not be obvious whether DMUs are i.i.d. or not in the production possibility set.

In Monte Carlo simulations the proposed measure has been tested and proved to be a reliable indicator. High values of Δ indeed go along with comparable efficiency scores, i.e. with largely unbiased estimated mean differences. Low values of Δ on the other hand go along with non-comparable efficiency scores, and, depending on the true efficiency density functions, considerably biased differences in mean efficiencies.

There are three advantages of the proposed measure: (1) The matter of interest is tested, i.e. the consequences of a (violated) assumption on the estimated efficiency rather than testing the assumption itself, (2) it is simple in terms of interpretation, there is one single index instead of a correlation matrix (between inputs and outputs), (3) computation is simple, almost costless, and may be easily implemented in a DEA software package.

The advantages of the proposed indicator Δ are likely to become more favorable with more complex technologies, i.e. with multiple inputs and outputs. Since the presented simulations are based on a rather simple one output and two inputs technology, validation of the properties of Δ with a more general technology remains an open issue.

Appendix

Efficiency densities used in the simulations

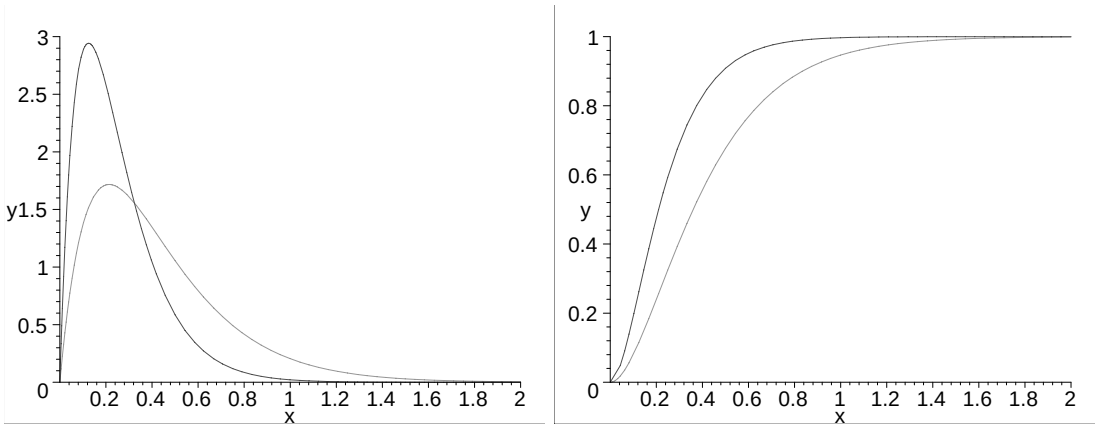


Figure 6: γ-distributed efficiency (probability and cumulative) density functions

References

[1] Banker, R.D. (1993), Maximum Likelihood, consistency and data envelopment analysis: a statistical foundation, *Management Science*, 39,10,1265-1273.

[2] Charnes, A., Cooper W.W. and E. Rhodes (1978), Measuring the inefficiency of decision making units, *European Journal of Operational Research* 2 (6), 429-444.

[3] Charnes, A., W. W. Cooper, and E. Rhodes (1981), Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through, *Management Science* 27, 668–697.

[4] 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*, vol 94, 445, 220-228.

[5] 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.

[6] Simar L. and P. Wilson (2000a), A General Methodology for Bootstrapping in Nonparametric Frontier Models, *Journal of Applied Statistics*, Vol 27, 6, 779-802.

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