Tziogkidis, Panagiotis

Working Paper

Bootstrap DEA and hypothesis testing

Cardiff Economics Working Papers, No. E2012/18

Provided in Cooperation with:
Cardiff Business School, Cardiff University

Suggested Citation: Tziogkidis, Panagiotis (2012) : Bootstrap DEA and hypothesis testing, Cardiff Economics Working Papers, No. E2012/18, Cardiff University, Cardiff Business School, Cardiff

This Version is available at:
https://hdl.handle.net/10419/65746

Cardiff Economics Working Papers

Working Paper No. E2012/18

Bootstrap DEA and Hypothesis Testing

Panagiotis Tziogkidis

August 2012

Cardiff Business School
Aberconway Building
Colum Drive
Cardiff CF10 3EU
United Kingdom
t: +44 (0)29 2087 4000
f: +44 (0)29 2087 4419
business.cardiff.ac.uk

Bootstrap DEA and Hypothesis Testing

Panagiotis Tziogkidis
Economics Department, Cardiff Business School, CF10 3EU, email: tziogkidisp@cf.ac.uk

Abstract

Bootstrapping non-parametric models is a fairly complicated exercise which is associated with implicit assumptions or requirements that are not always obvious to the non-expert user. Bootstrap DEA is a significant development of the past decade; however, some of its assumptions and properties are still quite unclear, which may lead to mistakes in implementation and hypothesis testing. This paper clarifies these issues and proposes a hypothesis testing procedure, along with its limitations, which could be extended to test almost any hypothesis in bootstrap DEA. Moreover, it enhances the intuition behind bootstrap DEA and highlights logical and theoretical pitfalls that should be avoided.

Key words: *Data Envelopment Analysis, Efficiency, Bootstrap, Bootstrap DEA, Hypothesis Testing*

JEL Classification: *C12, C14, C15, C61, C67*

1 Introduction

Since the paper of Simar and Wilson (1998), which introduced a method to implement bootstrapping in data envelopment analysis (DEA), many empirical applications have used this technique and further theoretical advances have been proposed. One of the most frequent uses of bootstrap DEA is to test various hypotheses. The majority of DEA papers use the approach of Simar and Wilson (1998) or their more recently proposed method of confidence interval construction (Simar and Wilson, 2000) in order to test (usually) the hypothesis of whether: (i) two firms from the same sample differ significantly in efficiency, (ii) two firms from different samples differ significantly in efficiency and (iii) two samples have equal average efficiency. These hypothesis testing methods can be extended to the case of returns to scale and productivity change (not analyzed here).

Although bootstrap DEA is not a recent development, some assumptions or requirements related to its implementation have not been clarified yet, while quite a few authors fail to use these methods appropriately. The most common errors found in the literature relate to: (i) the use of theoretically inconsistent hypothesis testing procedures, (ii) the use of potentially inappropriate tests in hypothesis testing, and (iii) applying regression analysis using the bootstrapped efficiency scores. The method of Simar and Wilson (1998) for obtaining the bootstrapped DEA scores is technically consistent and comprises a valuable tool for implementing statistical inference on DEA. However, it requires a few further clarifications regarding its assumptions and its appropriate implementation on hypothesis testing.

In this paper we provide these clarifications and we introduce a universal approach for using the bootstrapped efficiency scores in a theoretically consistent way. We demonstrate how one could fall into theoretical pitfalls using bootstrap confidence intervals and how they can be avoided. We focus our analysis on using bootstrap DEA to test the hypothesis of significant efficiency differences between two firms and we propose a straightforward and theoretically consistent procedure which can be easily extended to test any hypothesis.

The remainder of the paper is structured as follows: section 2 provides an insight into bootstrapping, section 3 succinctly presents the DEA and bootstrap DEA techniques, section 4 explains the logic behind bootstrap DEA and clarifies the associated assumptions, section 5 outlines the important issues with the Simar and Wilson (1998) approach, section 6 proposes a theoretically consistent procedure to test hypotheses and highlights the theoretical pitfalls that one may face when using these methods, section 7 outlines how confidence intervals may be constructed, section 8 comments on other potential issues when using bootstrap DEA while section 9 concludes the paper.

2 The bootstrap

The bootstrap was first introduced by Efron (1979), while Efron and Tibshirani (1993) provide a nice exposition of various issues associated with bootstrapping. Although it is a well established approach, we need to “re-establish” it for the purposes of this study, emphasizing on certain issues which will help us understand the source of variation and the nature of bias in bootstrap DEA. We will expose our ideas mainly within the regression (OLS) framework, as the principles of bootstrapping within a model are relevant in DEA. A deep understanding on how

the bootstrap should be applied on DEA is required in order to design consistent hypotheses to be tested as well as to understand their limitations.

The bootstrap is a procedure of drawing with replacement from a sample, mimicking the data generating process of the underlying true model and producing multiple estimates which can be used for statistical inference. One of its most important uses is to test hypotheses, especially in cases where statistical inference is impossible otherwise. Resampling, within the framework of the bootstrap, relates to redistributing the assumed randomness of the model among observations. This randomness is reflected in the deviations of the model's variables from their expected values, as calculated (or estimated) by the model. The higher the variance of the residuals the, wider the constructed bootstrap confidence intervals will be in hypothesis testing.

In the regression framework (let us assume OLS) these deviations are the model's residuals and there are two methods to bootstrap: to bootstrap pairs (alternatively termed "case resampling") and to bootstrap residuals (or "fixed resampling", as the independent variable is the same in all iterations). In the first case we resample pairs of observations and apply OLS each time. In the second case we resample residuals, we add them up to the expected value of the dependent variable and we apply OLS each time on this new pseudo-variable and the initial independent variables. In each case we obtain a distribution for the estimated coefficients (beta's) of the model which, in the limit, should be equal under both procedures. Resampling residuals is more sensitive to model assumptions (Efron and Tibshirani, 1993), mainly due to the fact that it assumes that the distribution of residuals does not depend on the observed sample; it is the same no matter what the independent variable is. However, resampling residuals might be more intuitive and appropriate to be applied in some cases (Efrom and Tibshirani 1993).

The accuracy of the bootstrap estimates depends on two factors: the variance of the model residuals and the inherent bias of the bootstrap process, both of which vary with sample size. Residual variance is the source of variability for bootstrapping and the resulting bootstrap distributions should be similar to the residual distribution (at least the higher moments). In fact, the center of the bootstrap distribution of an estimator is expected to be equal to the value of the estimator computed by the model. Any deviation from that value is known as the bootstrap bias and it is due to the random resampling process of the bootstrap. Especially if the sample is small and the observations are scattered, the effect of this bias may propagate. Therefore, correcting for bootstrap bias centers the distribution of the estimator to its expected value.

The bootstrap bias should not be confused with the model bias, which is defined as the difference of the model estimates from their true values. The latter occurs when other biases plague the model, which are not always observable, and the two most important ones are the measurement bias and the model specification bias, both of which violate the OLS assumptions in our example. These biases cause the model-estimated parameters to deviate significantly from their "true" value, even asymptotically. Therefore, the bootstrap estimates, which mimic the estimated model, will also fail to converge towards the true values (however, they will still converge towards the model estimates). In fact, considering that bootstrap estimators are also subject to bootstrap bias, it is possible that they will deviate from the true values even more than the model estimates. Since model biases are unobservable it is impossible to accurately compute the true value of an estimator using the bootstrap distribution; we could only approximate it under the assumption that there are no model biases.

3 DEA and bootstrapping

The concept of efficiency has been traditionally related to the ratio of outputs over inputs of a certain firm relative to others. However, in a multiple input-output setup it is necessary to attach weights to inputs and outputs, which reflect their relative rate of usage, in order to calculate the ratio of weighted outputs over weighted inputs. DEA is a non-parametric technique which is based on this logic and uses linear programming to determine optimal weights which minimize the distance between the frontier and the decision making unit (DMU) under consideration, subject to disposability and convexity constraints. The major advantage of DEA is that it does not require the specification of a production function: it just uses a set of inputs that DMUs want to minimize and a set of outputs that DMUs want to maximize.

DEA was first introduced by Charnes, Cooper and Rhodes (1978) with their CRS-consistent "CCR" model, while it was extended by Banker, Charnes and Cooper (1984) to account for VRS. We would like to avoid the exposition of the technical details involved since DEA is well established in the literature. Actually, the intended reader is expected to be already familiar with both DEA and bootstrap DEA methods.

Technical efficiency, as termed in DEA, is most commonly examined under the assumption of either input or output orientation. Under input orientation, DEA efficiency scores are interpreted as required input contractions to make a DMU efficient, keeping the level of outputs

fixed. Under output orientation efficiency scores correspond to required output expansions to make a DMU efficient, keeping input levels fixed[1]. Hence, in input orientation inputs behave as variables and outputs as model parameters, while in output orientation outputs are the variables and inputs the constants. In this paper we will be using the CRS technology assumption under input orientation, although the extension to the output oriented case or VRS should be straightforward.

One of the disadvantages of DEA is that statistical inference is very difficult to be applied on DEA scores. Therefore, bootstrap DEA was introduced by Simar and Wilson (1998), allowing to extract the sensitivity of efficiency scores which results from the distribution of (in)efficiency in the sample. Again, we would like to avoid demonstrating the technical details of the method since it is fairly established, while it would destruct the informed reader from the purpose of the paper. However, further details and analysis on related issues can be found in the papers of Simar and Wilson (1998, 2000) as well as their book chapters (Simar and Wilson, 2004, 2008). The outline of their proposed bootstrap procedure can be summarized in the following steps:

i. Use DEA to calculate efficiency scores.
ii. Draw with replacement from the empirical distribution (ED) of efficiency scores. Simar and Wilson (1998) suggest that smoothing the ED provides more consistent results.
iii. Divide the original efficient input levels by the pseudo-efficiency scores drawn from the (smoothed) empirical distribution to obtain a bootstrap set of pseudo-inputs.
iv. Apply DEA using the new set of pseudo-inputs and the same set of outputs and calculate the bootstrapped efficiency scores.
v. Repeat steps ii-iv B times and use bootstrapped scores for statistical inference and hypothesis testing.

4 The logic behind bootstrap DEA

The logic of bootstrapping within a model framework, as described in section 2, applies to a large extent in the case of DEA. The choice between bootstrapping “pairs” (case resampling) or

[1] For example, an efficiency score of 0.8 under input orientation (it varies between 0 and 1) implies that the DMU under assessment should use 80% of its inputs to become 100% efficient. On the other hand, an efficiency score of 1.2 under output orientation (it is greater than 1) implies that output should reach 120% (i.e. expand by 20) for a firm to be efficient.

"residuals" (fixed resampling) depends on the model of DEA we are using. In oriented models, where either inputs or outputs are fixed, it is more reasonable to use fixed resampling, while in non-oriented models such as the additive model, it is more reasonable to apply case resampling.

The source of variability in this special application of the bootstrap is the observed distribution of efficiency scores, which is treated as random. This variability can be translated into deviations of the efficient input levels from their efficient level. The analogies between bootstrapping regression models and bootstrap DEA are now clearer: (in)efficiency scores are assumed to be the model's "residuals" while the efficient input levels are the "dependent" variable or the model's expected value which is updated in each bootstrap loop. Thus, the distribution of (in)efficiency scores should be unaffected by (or uncorrelated with) the observed output levels (in input orientation).

It is therefore crucial to highlight the implications of the assumption of randomness for sampling DMUs which also justifies why the Simar & Wilson (1998) procedure is termed as the "homogeneous" bootstrap. In particular, this assumption is equivalent to the requirement that all DMUs in the sample should be similar in terms of technology and characteristics (that is, homogeneous). The assumption about technology is a well known one in DEA and refers to the fact that all DMUs in the sample should have access to similar processes in producing their outputs, as otherwise the model would be subject to model specification bias.

The assumption of homogeneity relates to the definition of a DMU and in particular to the fact that the DMUs under assessment should have similar attributes towards the concept of efficiency that is being examined (or efficiency approach as termed in many circumstances[2]). Otherwise, it is possible that inefficiency would be correlated with the output variables (in input orientation), which would violate the principles of fixed resampling (or "residual" bootstrapping). An intuitive example is required to clarify this statement: suppose that we want to compare the efficiency of basketball players towards their efficiency in terms of retrieving rebounds. If we use height as an input variable and number of rebounds as an output variable, obviously tall players would probably be very efficient and short ones would be very inefficient. Hence, inefficiency would be correlated with the number of rebounds (and height), while input

[2] An approach in efficiency literature is the set of inputs that a DMU is assumed to be using in order to produce a certain set of outputs, according to some theory. For example in banking there are several approaches: the intermediation approach, the production approach, the user-cost approach etc.

orientation would suggest desirable height contractions for players to become more efficient, which is logically inconsistent.

If, instead, we used training effort on rebounds as an input and number of rebounds as an output and included only players of the same height and ability, then the analysis would be more meaningful while inefficiency would be fairly random. That is, obtaining rebounds would be largely attributed to exogenous factors (such as timing, position on the field, injuries, opponent's strategy etc). The idea is that using a set of homogeneous DMUs[3] in a consistent DEA model, we should expected all DMUs to perform similarly (hence be input efficient) and therefore inefficiency should be random. The homogeneity assumption is very strong but careful sample selection would make the use of the "homogeneous" bootstrap DEA of Simar and Wilson (1998) more plausible. Respecting this assumption is the only way we could consider inefficiency of the DMUs in the sample to be randomly distributed.

A common misunderstanding with bootstrap DEA is that the source of variability relates to the sampling bias inherent in DEA. The DEA sampling bias is associated with the fact that the observed sample is (randomly) drawn from an underlying, unobserved population and the efficiency scores of the DMUs in the sample depend on the DMUs that define the frontier. This causes DEA efficiency scores to be overestimated compared to the "true" frontier, with the only highly unlikely exception that the DMUs which define the population frontier are all included in the sample. Sampling bias, therefore, is one of the factors that cause deviations between the sample DEA scores and the population ones and it should be common to all DMUs as it relates to the sample size.

However, sampling bias is irrelevant to the bootstrap bias, as the sample size is the same in all bootstrap replications. To obtain an approximation of the sampling bias we could either subsample or generate a hypothesized population from the sample distribution moments and draw pseudo-samples of the same size as the sample under consideration. On the other hand, the bootstrap bias, defined as the observed difference between the DEA scores and the mean (or median) of their bootstrap distribution, is caused because of the bias inherent in the random

[3] We could therefore define DMU in this framework as a unit which process inputs to produce outputs, subject to barriers which are imposed by its characteristics. Depending on the analysis, these constraints could be physical (height, weight, age, etc), technical (capacity, temperature, surface, volume, etc), other measurable operational characteristics (e.g. commercial banks' loans to deposits rations, dividend policy, corporate income taxation, etc), or the environment of the DMU (location, industry, competition, etc).

resampling procedure of the bootstrap, as already explained. Hence, correcting for bootstrap bias removes the latter effect and centers the bootstrap distribution on the DEA score of the DMUs that is being examined. To move from the sample DEA efficiency scores to the population ones, we would need to have a good estimate of the unobservable sampling bias while other biases (such as model specification or measurement) should not exist.

5 Issues in Simar and Wilson (1988)

Simar and Wilson (1998) in their seminal paper, apart from introducing the methodology outlined in the previous section, they also use the resulting bootstrap distribution of efficiency scores to construct confidence intervals where the "true" or population efficiency score is expected to lie. They uncover this region using the assumption that subtracting twice for bootstrap bias should center the "true" efficiency score (θ_A) and not the "biased" DEA one $\left(\hat{\theta}_A\right)$. The bootstrap bias for DMU A is defined as:

$$\widehat{bias}_A = \overline{\hat{\theta}_A^b} - \hat{\theta}_A \quad (1)$$

where $\overline{\hat{\theta}_A^b}$ is the median (or mean) of the bootstrapped efficiency scores of DMU A. The success of this logic is based on the assumption that the distribution of the bootstrap bias is similar to that of the model or "DEA bias", that is $\left(\hat{\theta}_A^b - \hat{\theta}_A\right) \sim \left(\hat{\theta}_A - \theta_A\right)$. The (double) bias-corrected distribution of efficiency scores in Simar and Wilson (1998) is:

$$\hat{\hat{\theta}}_A^b = \hat{\theta}_A^b - 2\widehat{bias}_A \quad (2)$$

with a median which is assumed to center θ_A.

Indeed, the bootstrapped efficiency scores are subject to two biases: the bootstrap bias and the unobservable sampling bias. More precisely, the distance between the bootstrap DEA scores and the "true" ones is the sum of the sampling and bootstrap bias which, as already mentioned, they are quite different and they would only be equal by chance:

$$\left(\overline{\hat{\theta}_A^b} - \theta_A\right) = \left(\overline{\hat{\theta}_A^b} - \hat{\theta}_A\right) + \left(\hat{\theta}_A - \theta_A\right) \neq 2\widehat{bias}_A \quad (3)$$

The DEA bias $\left(\hat{\theta}_A - \theta_A\right)$ is due to sampling variations and the bootstrap bias $\left(\overline{\hat{\theta}_A^b} - \hat{\theta}_A\right)$ is due to the resampling process, thus they cannot be equal. Hence, the double-bias-corrected efficiency score of Simar and Wilson (1998) seems to be lacking theoretical support. Moreover, the DEA bias might also incorporate other errors, such as model specification or measurement biases, which would be very hard to detect and rule out.

Correcting for bootstrap bias once and under the assumption of no other biases we obtain the sensitivity of a DMU's DEA score towards inefficiency randomness, which was the purpose of bootstrap DEA when first introduced. Indeed, the median (or mean) of the single-bias-corrected bootstrap distribution is:

$$\overline{\hat{\theta}_A^{b*}} = \overline{\hat{\theta}_A^b - \widehat{bias}_A} = \overline{\hat{\theta}_A^b} - \widehat{bias}_A = \overline{\hat{\theta}_A^b} - \left(\overline{\hat{\theta}_A^b} - \hat{\theta}_A\right) = \hat{\theta}_A \quad (4)$$

whereas the median (or mean) of the twice-bias corrected bootstrap distribution is

$$\overline{\hat{\tilde{\theta}}_A^b} = \overline{\hat{\theta}_A^b - 2\widehat{bias}_A} = \hat{\theta}_A - \widehat{bias}_A = 2\hat{\theta}_A - \overline{\hat{\theta}_A^b} \quad (5)$$

Simar and Wilson (1998) claim that $\overline{\hat{\tilde{\theta}}_A^b} \simeq \theta_A$, however this also lacks theoretical support due to (3). To provide support to this result it would be necessary to perform a Monte Carlo exercise which compares the first four moments of the distribution of $\overline{\hat{\tilde{\theta}}_i^b}, i = 1,2, \dots n$ (the best estimates of the "population" efficiency score) with these of the population distribution $\theta_j, j = 1,2 \dots N$.

6 Hypothesis testing

The literature on testing hypotheses using bootstrap DEA is either not very clear or limited to specific examples. Simar and Wilson (2008) provide guidance on using their techniques and demonstrate an example of hypothesis testing for the case of mean efficiency score differences between two groups. Among their general rules they suggest that: the test statistic used has to be a function of the data, the critical value should result from the bootstrap distribution while the null hypothesis and the alternative should be clearly stated and be theoretically sensible. However, it is not straightforward how one could use their methods to test hypotheses and which should be the principles which should be respected when testing hypotheses.

In this section we will provide an outline for designing and implementing hypothesis testing using the bootstrap distribution of efficiency scores. Then we will explore how the testing procedures implied by Simar and Wilson (1998, 2000) should be appropriately used to test hypotheses and we will prove that they can be converted in our proposed one. This analysis is important as these methods have not always been used appropriately in the literature and there seems to be confusion as to what is being tested. We will use the example of testing the hypothesis of efficiency score differences between two DMUs and highlight its limitations.

6.1 A simple approach

Suppose that we want to test whether the DEA score of DMU A $\left(\hat{\theta}_A\right)$ differs significantly from the DEA score of DMU B $\left(\hat{\theta}_B\right)$, due to their sensitivity imposed by the distribution of (in)efficiency. Using the distribution of bootstrapped efficiency scores we could construct an acceptance region for DMU A and calculate the probability of observing the efficiency score of DMU B within this region. Hence, the hypothesis to be tested is:

$$H_0: \hat{\theta}_A = \hat{\theta}_B, \qquad H_1: \hat{\theta}_A \neq \hat{\theta}_B \tag{6}$$

If the desired significance level is a then we could use the $\left({}^{a}/_{2}\right)$ and $\left(1 - {}^{a}/_{2}\right)$ percentiles of the bootstrap distribution $\left(\hat{\theta}_A^b\right)$ in our two-tailed test. If we denote these percentiles with $\hat{p}_{{}^{a}/_{2}}$ and $\hat{p}_{1-{}^{a}/_{2}}$, respectively, we have:

$$\Pr\left(\hat{p}_{{}^{a}/_{2}} < \hat{\theta}_A^b < \hat{p}_{1-{}^{a}/_{2}}\right) = 1 - a \tag{7}$$

Using straightforward manipulations (subtracting $\hat{\theta}_A^b$ and adding $\hat{\theta}_A$) we get:

$$\Pr\left(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-{}^{a}/_{2}} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{{}^{a}/_{2}}\right) = 1 - a \tag{8}$$

Hence, we have constructed from (7) a $(1 - a)$% region where the efficiency score of DMU A is expected to be observed, using the distribution of its bootstrapped efficiency scores. This implies that we could use a related p-value to calculate the probability of observing the DEA score of DMU B within the "region" of DMU A:

$$p = \frac{\#\left(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-{}^{a}/_{2}} < \hat{\theta}_B < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{{}^{a}/_{2}}\right)}{B} \tag{9}$$

This is a standard indicator function used in bootstrap applications where the hash sign stands for "number of times". As usual, if $p > a$ the null hypothesis of no difference cannot be rejected. This straightforward logic can be extended to test any hypothesis.

One of the most troublesome limitations of this approach, which is common to all statistics or tests on bootstrap DEA, relates to the fact that the distribution of efficiency scores of each DMU is not normal (in most cases skewed) and at the same time it is not identical to that of other DMUs even of the same sample. The importance of this result is that:

$$p_{AB} = \frac{\#\left(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{A,1-a/2} < \hat{\theta}_B < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{A,a/2}\right)}{B} \neq \frac{\#\left(\hat{\theta}_B + \hat{\theta}_B^b - \hat{p}_{B,1-a/2} < \hat{\theta}_A < \hat{\theta}_B + \hat{\theta}_B^b - \hat{p}_{B,a/2}\right)}{B} = p_{BA} \quad (10)$$

Hence, the calculated p-value of (9) will differ depending on the reference DMU; that is, it is possible that $p_{AB} > a$ and $p_{BA} < a$ and vice versa. In the case of skewed distributions it is preferable to use the median of the bootstrapped distribution while it is necessary to apply alternative methods to construct confidence intervals. Such methods have been proposed by Efron (1982, 1987) and will be discussed in Section 6. However, although these methods improve the endpoints of the confidence intervals, it is still possible that the aforementioned problem persists. We therefore suggest for these few cases to reject the null hypothesis, since this seems to be a more conservative decision compared to accepting it.

Another limitation of this approach is that the extension to different samples requires the two samples to have similar distribution of inefficiency (to ensure similar source of variability). To some extent we could mitigate this issue by applying our test on the standardized efficiency scores, although the higher moments (skewness and kurtosis) would still need to be similar. Hence, if the standardized efficiency score of any DMU $k = \{A, B\}$ is $\hat{\zeta}_k = \left(\hat{\theta}_k - \overline{\hat{\theta}^{(k)}}\right) / S^{(k)}\left(\hat{\theta}\right)$, where $\overline{\hat{\theta}^{(k)}}$ and $S^{(k)}\left(\hat{\theta}\right)$ are the mean and the standard deviation of the efficiency scores of the group where DMU k belongs to, then (8) becomes:

$$\Pr\left(\frac{\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-a/2} - \overline{\hat{\theta}^{(A)}}}{S^{(A)}\left(\hat{\theta}\right)} < \hat{\zeta}_A < \frac{\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{a/2} - \overline{\hat{\theta}^{A}}}{S^{(A)}\left(\hat{\theta}\right)}\right) = 1 - a \quad (11)$$

Then we could substitute for $\hat{\zeta}_B$ and calculate the associated p-value as in (9), which should be exactly the same if DMUs A and B were from the same sample. However, if they do not, the two different groups need to be comparable (ideally homogeneous) while any differences in their means should be assumed to be random. Hence, standardizing would ensure that both groups have the same mean (zero) and variance (one), while the resulting variables would be comparable as they reflect standardized deviations from the mean. Note, though, that the skewness and kurtosis of the standardized efficiency scores are identical to the non-standardized ones, which supports our previous argument that higher moments still need to be similar to get meaningful results.

Finally, there are two important limitations when comparing DMUs from different samples which, unfortunately, cannot be mitigated. These are the issues of different technologies (which cause model specification bias) and of different sample sizes (which are associated with different sampling errors which do not cancel out). Hence, comparisons should be made only between groups of similar technology and size. It is clear that the requirement of homogeneity applies to the case of different samples as well.

6.2 Simar and Wilson (1998) approach

We will show how one may move from the Simar and Wilson (1998) logic to (9) with a few simple steps. Simar and Wilson (1998) suggest using the percentiles of $\hat{\hat{\theta}}_A^b$ to construct a region where the "true" efficiency score θ_A lies. Hence, the implied hypothesis to be tested for the case of efficiency score differences between two DMUs of the same sample is:

$$H_0: \theta_A = \theta_B, \qquad H_1: \theta_A \neq \theta_B \tag{12}$$

However, since θ's are unobservable we need to use their estimates for this test in order to be theoretically consistent:

$$H_0: \overline{\hat{\hat{\theta}}_A^b} = \overline{\hat{\hat{\theta}}_B^b}, \qquad H_1: \overline{\hat{\hat{\theta}}_A^b} \neq \overline{\hat{\hat{\theta}}_B^b} \tag{13}$$

Now, let $\hat{k}_{a/2}$ and $\hat{k}_{1-a/2}$ be the lower and upper percentiles of $\hat{\hat{\theta}}_A^b$, so that:

$$\Pr\left(\hat{k}_{a/2} < \hat{\hat{\theta}}_A^b < \hat{k}_{1-a/2}\right) = 1 - a \tag{14}$$

These percentiles should be equal to the ones of $\hat{\theta}_A^b$ ($\hat{p}_{a/2}$ and $\hat{p}_{1-a/2}$), reduced by twice the scalar of bias, so using relatively relaxed notation:

$$k\left(\hat{\hat{\theta}}_A^b\right) = k\left(\hat{\theta}_A^b - 2\widehat{bias}_A\right) = percentile\left(\hat{\theta}_A^b\right) - 2\widehat{bias}_A = p\left(\hat{\theta}_A^b\right) - 2\widehat{bias}_A \tag{15}$$

This result is valid due to fact that all percentiles in (15) are *transformation respecting*[4] (Efron and Tibshirani, 1993). Substituting (2) in (14) and applying the percentile property in (15) on (14), we can easily move from (14) to (7):

[4] The *transformation respecting* property refers to the fact that the lower and upper boundaries of the confidence intervals transform correctly if, instead of the parameter we are considering ($\hat{\hat{\theta}}_A^b$) we use some function or transformation of it ($\hat{\theta}_A^b$). Practically, this property assures that the $\hat{p}$ percentiles are a valid transformation of the $\hat{k}$ percentiles in our case.

$$
\begin{aligned}
1-a &= \Pr\left(\hat{k}_{a/2} < \hat{\delta}_A^b < \hat{k}_{1-a/2}\right) \\
&= \Pr\left(\hat{p}_{a/2} - 2\widehat{bias}_A < \hat{\theta}_A^b - 2\widehat{bias}_A < \hat{p}_{1-a/2} - 2\widehat{bias}_A\right) \\
&= \Pr\left(\hat{p}_{a/2} < \hat{\theta}_A^b < \hat{p}_{1-a/2}\right) = 1-a
\end{aligned}
\tag{16}
$$

This proves that the resulting p-values from Simar and Wilson's (1998) approach would be identical to the ones in (9). This is due to the fact that $\hat{\theta}_A^b$ is the only source of variability in our hypothesis testing. Therefore, if one uses the bootstrap to test hypotheses it would be preferable to avoid the increased complexity introduced by double bias-correction.

6.3 Simar and Wilson (2000) approach

Simar and Wilson (2000) use the distribution of the deviations of the bootstrapped efficiency scores from the DEA scores to construct confidence intervals where the "true" efficiency score lies. They state that this method is better in that it reduces the unnecessary excess variation of the confidence intervals in their previous paper and that the new confidence intervals are more accurate for hypothesis testing.

In particular, the proposed confidence intervals are narrower and they are claimed to uncover the regions where the "true" efficiency score of a certain DMU lies, with the same precision as in Simar and Wilson (1998). Again, the accuracy of this approach depends on the validity of the assumption that $\left(\hat{\theta}_A^b - \hat{\theta}_A\right) \sim \left(\hat{\theta}_A - \theta_A\right)$ which might not be valid for the reasons explained previously. Assuming again a confidence level of $(1-a)\%$, the associated confidence region is defined by:

$$
1-a = \Pr\left(s_{a/2} < \hat{\theta}_A - \theta_A < s_{1-a/2}\right) = \Pr\left(\hat{s}_{a/2} < \hat{\theta}_A^b - \hat{\theta}_A < \hat{s}_{1-a/2}\right) \tag{17}
$$

where s and $\hat{s}$ represent the true and bootstrap percentiles of the relevant distributions in (17). Then they proceed by further assuming that $s_{a/2} \simeq \hat{s}_{a/2}$ and $s_{1-a/2} \simeq \hat{s}_{1-a/2}$, which is actually an extension of the that $\left(\hat{\theta}_A^b - \hat{\theta}_A\right) \sim \left(\hat{\theta}_A - \theta_A\right)$. Assuming that this is true, we have:

$$
\begin{aligned}
1-a &= \Pr\left(\hat{\theta}_A - s_{1-a/2} < \theta_A < \hat{\theta}_A - s_{a/2}\right) \\
&\simeq \Pr\left(\hat{\theta}_A - \hat{s}_{1-a/2} < \theta_A < \hat{\theta}_A - \hat{s}_{a/2}\right)
\end{aligned}
\tag{18}
$$

To test the hypothesis of significant efficiency differences, we could use (18) to check whether the "true" efficiency score of DMU B lies within that range defined for DMU A in order to test

whether they differ significantly in technical efficiency. However, θ_B is unobserved and its best estimate $\overline{\hat{\hat{\theta}}_B^b}$ should be used instead, so the hypothesis test becomes:

$$H_0: \theta_A = \theta_B \simeq \overline{\hat{\hat{\theta}}_B^b}, \qquad H_1: \theta_A \neq \theta_B \simeq \overline{\hat{\hat{\theta}}_B^b} \tag{19}$$

and the associated p-value is:

$$p = \frac{\#\left(\hat{\theta}_A - \hat{s}_{1-a/2} < \overline{\hat{\hat{\theta}}_B^b} < \hat{\theta}_A - \hat{s}_{1-a/2}\right)}{B} \tag{20}$$

which seems a bit problematic as it would wield either 0 or 1 as the source of variation $\left(\hat{\theta}_A^b\right)$ has been removed (has been averaged out). That is, there is actually no p-value to be calculated under this formulation. However, it is generally true that a desirable property of hypothesis testing (if not requirement), is that there should be a measurable p-value to show how strongly the null hypothesis is accepted or rejected, which is not the case here. If we want to relax the assumptions of this approach we would need to transform the null hypothesis as implied by their first paper, which was introduced in (13).

To prove why the approach of Simar and Wilson (2000) might lead to inconsistent results note that from (18) and using the calculated $\overline{\hat{\hat{\theta}}_A^b}$ instead of the unobservable θ_A, we have:

$$\begin{aligned} 1 - a &= \Pr\left(\hat{\theta}_A - s_{1-a/2} < \theta_A < \hat{\theta}_A - sa_{/2}\right) \\ &\simeq \Pr\left(\hat{\theta}_A - \hat{s}_{1-a/2} < \overline{\hat{\hat{\theta}}_A^b} < \hat{\theta}_A - \hat{s}a_{/2}\right) \\ &\xrightarrow{\textbf{\textit{Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.}}} \Pr\left(\overline{\hat{\theta}_A^b} - \hat{s}_{1-a/2} < \hat{\theta}_A \right. \\ &\left. < \overline{\hat{\theta}_A^b} - \hat{s}a_{/2}\right) \end{aligned} \tag{21}$$

And, using ease of notation, note that the s-percentiles are equivalent to:

$$s\left(\hat{\theta}_A^b - \hat{\theta}_A\right) = p\left(\hat{\theta}_A^b\right) - \hat{\theta}_A \tag{22}$$

Substituting back in (21) we have:

$$1 - a = \Pr\left(\hat{\theta}_A + \overline{\hat{\theta}_A^b} - \hat{p}_{1-a/2} < \hat{\theta}_A < \hat{\theta}_A + \overline{\hat{\theta}_A^b} - \hat{p}a_{/2}\right) \tag{23}$$

However, this result is different compared to what we have already derived for the consistent probability statement about $\hat{\theta}_A$ in (9). That is:

$$\begin{aligned} 1 - a &= \Pr\left(\hat{\theta}_A + \overline{\hat{\theta}_A^b} - \hat{p}_{1-a/2} < \hat{\theta}_A < \hat{\theta}_A + \overline{\hat{\theta}_A^b} - \hat{p}a_{/2}\right) \\ &\neq \Pr\left(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-a/2} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}a_{/2}\right) \end{aligned} \tag{24}$$

The reason is very straightforward: evaluating the confidence region of $\overline{\hat{\hat{\theta}}_A^b}$, which is assumed to be a good approximation of θ_A, requires an appropriate percentile as in (14). Hence, when using

the Simar and Wilson's (2000) approach for our particular hypothesis testing example, the median of the bootstrapped efficiency scores will be used instead of the actual bootstrapped efficiency scores. This implies that the calculated p-values will be binary as the source of variability has been replaced by its central tendency and that there is a high probability of a Type I error. Especially in bigger samples where variability is usually reduced, the constructed confidence intervals will be tighter implying that Type I error becomes almost a fact. The only chance to avoid, to some extent, Type I errors is for the distribution of $\hat{\theta}_A^b$ to be highly leptokurtic and centered on $\overline{\hat{\theta}_A^b}$, while there should be no model selection or measurement biases. Furthermore, this suggests that assessment of bootstrap DEA with Monte Carlo simulations on the basis of "coverage probabilities" (that is the probability to observe the population efficiency score within the boundaries calculated in (18)) might not be a good idea. In fact, we would expect that our statement, if true, would be reflected in diminishing coverage probabilities as sample size increases.

We conclude, regarding the use of Simar and Wilson (2000) confidence intervals for testing our example hypothesis, that it may be used in two situations: if we have knowledge of the "true" (or population) efficiency score of the DMU under assessment or if the bootstrapped efficiency scores have a very leptokurtic distribution (and preferably non-skewed), which in turn requires a very small variance. Yet, even if all these requirements are met, this approach does not allow for p-values to be calculated.

Using the distribution of $\hat{\theta}_A^b - \hat{\theta}_A$ in hypothesis testing consistently should always be possible to produce the same p-values as in our approach in (9). In particular:

$$1 - a = \Pr\left(\hat{s}_{a/2} < \hat{\theta}_A^b - \hat{\theta}_A < \hat{s}_{1-a/2}\right) = \Pr\left(\hat{\theta}_A^b - \hat{s}_{1-a/2} < \hat{\theta}_A < \hat{\theta}_A^b - \hat{s}_{a/2}\right) \quad (25)$$

and using (22) we get:

$$\Pr\left(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-a/2} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{a/2}\right) = 1 - a \quad (26)$$

which is the same as (8), hence proving our statement. The difference here is that we avoid assuming that $s \simeq \hat{s}$ which has caused the aforementioned issues with this approach. Also it proves that the argument of Simar and Wilson (2000) regarding the excess variability in their 1998 paper has no effect on the computation of the p-values in hypothesis testing.

7 Confidence interval construction

In the previous sections we established that testing for efficiency score differences between two DMUs of the same sample is associated with the probability statement in (8):

$$\Pr\left(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-a/2} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{a/2}\right) = 1 - a \quad (27)$$

This information can be used to construct confidence intervals or acceptance regions about $\hat{\theta}_A$. Hence, if the efficiency score of another DMU falls within the region of DMU A we could state that the two DMUs do not differ significantly in efficiency and this will be due to the implied sensitivity of efficiency scores introduced by the distribution of (in)efficiency. To perform this task we will need to calculate two percentiles: one for the lower bound and one for the upper bound in (27). Denote the $\left(a/2\right)^{\text{th}}$ percentile of $\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-a/2}$ with $\hat{t}_L$ and the $\left(1 - a/2\right)^{\text{th}}$ percentile of $\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{a/2}$ with $\hat{t}_U$. These percentiles are associated with the following one-tailed probability statements which we will need to use to construct our central $(1-a)\%$ confidence interval (Efron, 1982):

$$\Pr\left(\hat{t}_L < \hat{\theta}_A\right) = \Pr\left(\hat{\theta}_A < \hat{t}_U\right) = 1 - \frac{a}{2} \quad (28)$$

It is straightforward to verify that:

$$\Pr\left(\hat{t}_L < \hat{\theta}_A < \hat{t}_U\right) = 1 - a \quad (29)$$

Therefore the lower and upper bounds of our confidence intervals are $\hat{t}_L$ and $\hat{t}_U$, respectively. The confidence intervals will be centered on $\hat{\theta}_A$ by construction, if the medians of the bootstrapped distributions are used in all relevant calculations (for example, for the bias).

However, the bootstrap distributions of efficiency scores are usually skewed and the calculated confidence intervals will be biased to some extent. Therefore appropriate techniques should be implemented which correct for skewness and provide more accurate endpoints for the constructed confidence intervals. Simar and Wilson (1998) suggest using the bias corrected (BC) intervals of Efron (1982), however it is not the best option when dealing with skewness. A more appropriate method is that of Efron (1987), where the "bias corrected and accelerated" (BC_a) confidence intervals account for skewness through the *acceleration parameter*. The first step to construct the central BC_a confidence intervals with coverage $1 - a$ is to calculate corrected percentiles of the bootstrap distribution endpoints. Without loss of generality, if we are using our definition of the bias-corrected bootstrap distribution from (4), that is $\hat{\theta}_A^{b*} = \hat{\theta}_A^b -$

$\widehat{bias}_A$, we would be replacing the percentiles $\hat{s}_{a/2}$ and $\hat{s}_{1-a/2}$ with the BC_a ones $\hat{s}^{(a_1)}$ and $\hat{s}^{(a_2)}$, where

$$s^{(a_1)} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(a/2)}}{1 - \hat{\alpha}\left(\hat{z}_0 + z^{(a/2)}\right)}\right) \tag{30}$$

and:

$$s^{(a_2)} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-a/2)}}{1 - \hat{\alpha}\left(\hat{z}_0 + z^{(1-a/2)}\right)}\right) \tag{31}$$

where Φ is the standard normal cumulative density function and $z^{(a/2)}$ is the normalized value that corresponds to the $^{a}/_{2}{}^{\text{th}}$ percentile of the standard normal distribution, so that $\Phi\left(z^{(a/2)}\right) = {}^{a}/_{2}$. The parameter $\hat{z}_0$ is called the bias correction parameter and depends on the proportion of bootstrap estimates that are lower than the model estimates: $G\left(\hat{\theta}_A\right) = \Pr\left(\hat{\theta}_A^{b*} < \hat{\theta}_A\right)$ and $\hat{z}_0 = \Phi^{-1}\left[G\left(\hat{\theta}_A\right)\right]$ is the standard normal value that corresponds to that probability. In our particular example $\hat{\theta}_A^{b*}$ is already bias-corrected by the median of its distribution, hence $G\left(\hat{\theta}_A^{b*}\right) = 0.5$ and therefore $\hat{z}_0 = 0$. We have chosen this example on purpose since this way we can isolate the effect of skewness bias, calculated by the acceleration parameter $\hat{\alpha}$, while the resulting intervals are less variant compared to using the mean as the center of the bootstrap distribution.

Finally, the acceleration parameter for the non-parametric case can be calculated in various ways (see Efron and Tibshirani, 1993 for more information) one of which involves using the jackknife of the bias-corrected bootstrap efficiency score of DMU A in each bootstrap loop, $\hat{\theta}_{(-A),i}^{b*}$, so that:

$$\hat{\alpha} = \frac{\sum_{i=1}^{B}\left(\hat{\theta}_{(-A),i}^{b*} - \hat{\theta}_{A,i}^{b*}\right)^3}{6\left[\sum_{i=1}^{B}\left(\hat{\theta}_{(-A),i}^{b*} - \hat{\theta}_{A,i}^{b*}\right)^2\right]^{3/2}} \tag{32}$$

where $\hat{\theta}_{A,i}^{b*}$ is the bias-corrected bootstrap efficiency score of DMU A on the i^{th} bootstrap repetition. The obvious problem is that (32) is inconsistent since the jackknife requires deleting the DMU under consideration in each bootstrap repetition[5]. One possible alternative would be to approximate the acceleration parameter through the jackknife estimate of bias for DMU A:

[5] It is quite tempting to think that a more reasonable way to approximate the marginal contribution of DMU A, as required by the "acceleration" parameter, would be to apply a form of cross-validation on each

$$\hat{a} = \frac{\sum_{i=1}^{B}\left(\hat{\theta}_{A,(\cdot)}^{b*} - \hat{\theta}_{A,i}^{b*}\right)^3}{6\left[\sum_{i=1}^{B}\left(\hat{\theta}_{A,(\cdot)}^{b*} - \hat{\theta}_{A,i}^{b*}\right)^2\right]^{3/2}} \simeq \frac{\sum_{i=1}^{B}\left(mean\left(\hat{\theta}_{A}^{b*}\right) - \hat{\theta}_{A,i}^{b*}\right)^3}{6\left[\sum_{i=1}^{B}\left(mean\left(\hat{\theta}_{A}^{b*}\right) - \hat{\theta}_{A,i}^{b*}\right)^2\right]^{3/2}} \tag{33}$$

where $\hat{\theta}_{A,(\cdot)}^{b*} = \sum_{i=1}^{B}\hat{\theta}_{A,(-i)}^{b*}/B \simeq mean\left(\hat{\theta}_{A}^{b*}\right)$ and $\hat{\theta}_{A,(-i)}^{b*}$ is the mean of $\hat{\theta}_{A}^{b*}$ distribution leaving out the i^{th} element. Although (33) is not exactly what we wish to calculate in (32), it is a fair approximation for the case of testing hypotheses between different DMUs while future research should focus on finding efficient ways of properly calculating the acceleration parameter in bootstrap DEA. Obviously, if we were testing the hypothesis of differences in mean efficiency scores between two groups of DMUs, then the jackknife of the mean would be more straightforward to be calculated.

To summarize, in order to obtain appropriate confidence intervals for the DEA score of DMU A, we suggest using the BC_a method of Efron (1987) to appropriately compute the endpoints a_1 and a_2 of the distribution of $\hat{\theta}_{A}^{b*}$, that is $\left(\hat{\theta}_{A}^{b*,(a_1)}, \hat{\theta}_{A}^{b*,(a_2)}\right)$. Denote, then, the percentiles of this distribution as $\hat{s}^{(a_1)}$ and $\hat{s}^{(a_2)}$ and by applying appropriate transformations we have:

$$\begin{aligned}&\Pr\left(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}^{(a_2)} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}^{(a_1)}\right)\\ &\qquad = \Pr\left(\hat{\theta}_A + \hat{\theta}_A^{b*} - \hat{s}^{(a_2)} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^{b*} - \hat{s}^{(a_1)}\right) = 1 - a\end{aligned} \tag{34}$$

Like previously, we will need to use the ${a_1}^{\text{th}}$ and ${a_2}^{\text{th}}$ percentiles of the two endpoints in (34), which we denote as $\hat{t}^{(a_1)}$ and $\hat{t}^{(a_2)}$, respectively. Finally, the central BC_a percentiles with coverage $1 - a$ are calculated by $\left(\hat{t}^{(a_1)}, \hat{t}^{(a_2)}\right)$.

To maximize intuition we have graphically represented in Figure 1 what bootstrap DEA does and how hypothesis testing is performed. In our simple one input (x), one output (y) case, we consider a sample of 30 DMUs which is randomly drawn from an underlying population. The sample CRS frontier is defined by DMU A, while the hypothesized (unobserved) population frontier is also drawn for comparison. Sampling bias in this case is considered to be the distance between the population frontier and the sample frontier. The bias is common to all DMUs and it

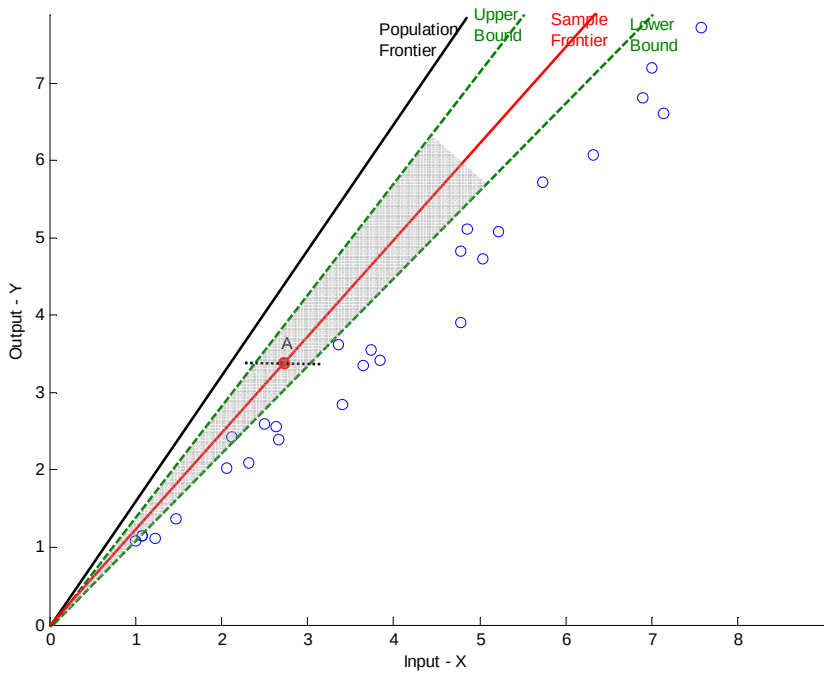
---

bootstrap replication. That is, instead of using $\hat{\theta}_{(-A)}^{b*}$, to use the average score for DMU A that results from deleting each time one of the other DMUs in the sample in each bootstrap repetition $\left(\hat{\theta}_{(A-j)}^{b*}\right)$. However, this approach would rocket the computational costs while the resulting estimate would be counter-intuitive since the values in $\hat{\theta}_{(A-j),i}^{b*}$ would be different from $\hat{\theta}_{A}^{b*}$ only for the cases where the j^{th} DMU excluded is one of the DMUs which defines the sample frontier (and only if A is inefficient), especially using the CRS assumption.

is reflected on the fixed angle between the frontiers, the tangent of which reflects technical efficiency. The widening gap reflects the fact that the fixed efficiency score differential is translated into bigger input contractions as input levels increase, which is very reasonable.

If we focus our analysis on DMU A, which is assumed to be the most efficient DMU in the sample, bootstrapping its efficiency scores can be translated into varying its input levels. We can center this variation about $\hat{\theta}_A$ by correcting for bias $\left(\hat{\theta}_A^{b*}\right)$ and the resulting input variation is represented here by the horizontal dotted line. Furthermore, following the aforementioned procedure we may construct confidence intervals, to see which firms do not differ significantly in performance. This is represented by the shaded area that is defined between the lower and upper bounds of the confidence interval. Note that we have taken care to intersect the horizontal dotted line close to the edges, leaving out some information at the tails. In our example we observe that about 3 DMUs fall within the confidence interval region, hence these DMUs do not differ significantly in efficiency from DMU A.

Figure 1. Graphical representation of hypothesis testing in bootstrap DEA



8 Other important issues

One of the most important issues in implementing hypothesis testing in bootstrap DEA is to use an appropriate testing procedure, if not following our proposed approach. A popular alternative is the use of non-parametric tests such as the Wilcoxon signed rank test and the Kolmogorov-Smirnov test. However, if the bootstrap distributions of DMUs are correlated with each other then these tests cannot be used. In particular, if high correlation is observed, the signs in the Wilcoxon test would almost always be the same, while the empirical cumulative distributions in the KS test would not intersect each other (perhaps just at the endpoints), hence these tests would rarely accept the null hypothesis. It is therefore better to use the proposed approach in this paper as it is distribution-free and independent of the correlation structure of bootstrapped scores.

We should clarify that correlation among bootstrap distributions for each DMU is different from the correlation between efficiency scores of bootstrapped samples. The latter appears always in both empirical work and simulation exercises. This correlation does not affect the validity of the aforementioned tests; however, it implies that applying second stage regressions on the bootstrapped samples of efficiency scores would yield inconsistent results. Regarding this case, it is preferable (if not necessary) to follow Simar and Wilson (2007).

9 Conclusion

In this paper we provided a deep insight in the workings of bootstrap DEA and we addressed the important issue of implementing hypothesis testing using bootstrapped efficiency scores. We emphasized on the implicit assumptions of the method and we explored the logical and theoretical pitfalls that should be avoided when using these methods to test hypotheses. We introduced a procedure for hypothesis testing which may be applied universally and we explained its associated limitations, while we proposed ways to deal with them. Finally, we used our theoretically consistent procedure to construct confidence intervals which serve as acceptance regions of the null hypothesis of no significant difference in efficiency scores.

The method of Simar and Wilson (1998) is a valid process for bootstrapping DEA scores and a valuable tool for statistical inference and hypothesis testing. However, the user has to respect the associated assumptions and be clear when using hypothesis testing about what is actually

being tested, otherwise inconsistent conclusions might be reached. We have illustrated how such an inconsistency might result from implementing the Simar and Wilson (2000) approach in the common case in the literature of comparing the performance of two DMUs. The paper serves as a guide for the users of bootstrap DEA and as a complement of the Simar and Wilson's (1998) paper, especially when hypothesis tests need to be carried out.

References

Banker R.D., Charnes A., Cooper W.W., (1984). "Some models for estimating technical and scale inefficiencies in data envelopment analysis", *Management Science*, Vol. 3, No. 9, pp. 1078-1092

Charnes A., Cooper W.W., Rhodes E., (1978). "Measuring the Inefficiency of Decision Making Units", *European Journal of Operational Research*, Vol. 2, pp. 429-444

Efron B., (1979). "Bootstrap methods: another look at the jackknife", *Annals of Statistics*, Vol. 9, pp. 1-26

Efron B., (1982)," The jackknife, the bootstrap and other resampling plans", *CBMS*, Vol. 38, SIAM-NSF

Efron B., (1987). "Better bootstrap confidence intervals", *Journal of the American Statistical Association*, Vol. 82, No. 397, pp. 171-185

Efron B., Tibshirani R.J., (1993). "An introduction to the bootstrap", *Chapman and Hall*, London

Simar L., Wilson W.P., (1998). "Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models", *Management Science*, Vol. 44, No. 1, pp. 49-61

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

Simar L., Wilson W.P., (2004). "Performance of the bootstrap for DEA estimators and iterating the principle", *ed. by* Cooper W.W., Seiford M.L., Zhu J., *in* Handbook on Data Envelopment Analysis, *Kluwer Academic Publishers*, pp. 265-298

Simar L., Wilson W.P., (2007). "Estimation and inference in two-stage, semi-parametric models of production processes", *Journal of Econometrics*, Vol. 136, pp. 31-64

Simar L., Wilson W.P., (2008). "Statistical inference in non-parametric frontier models" *ed. by* Fried O.H., Lovell C.A.K., Schmidt S.S.: "The measurement of productive efficiency and productivity growth", Oxford *University Press,* Oxford: New York, pp. 421-521