Tziogkidis, Panagiotis

Working Paper

Monte Carlo experiments on bootstrap DEA

Cardiff Economics Working Papers, No. E2012/19

Provided in Cooperation with:
Cardiff Business School, Cardiff University

Suggested Citation: Tziogkidis, Panagiotis (2012) : Monte Carlo experiments on bootstrap DEA, Cardiff Economics Working Papers, No. E2012/19, Cardiff University, Cardiff Business School, Cardiff

This Version is available at:
https://hdl.handle.net/10419/65806

Cardiff Economics Working Papers

Working Paper No. E2012/19

Monte Carlo Experiments on Bootstrap DEA

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

Monte Carlo Experiments on Bootstrap DEA

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

Abstract

Since the introduction of bootstrap DEA there is a growing literature on applications which use this method, mainly for hypothesis testing. It is therefore important to establish the consistency and evaluate the performance of bootstrap DEA. The few Monte Carlo experiments in the literature perform this exercise on the basis of coverage probabilities, using a certain population assumption and usually they analyze the simple case of 1 input and 1 output. However, it has been argued recently that coverage probabilities are not a good tool of assessment. In our study we evaluate the performance of bootstrap DEA using the standard approach of comparing moments. We use three different data generating processes over three different dimensions while for each case we compare results from both the smooth and "naïve" bootstrap. Our results are not in accordance with previous studies, as we find that the smooth bootstrap performs overall worse while we highlight the cases where the researcher should be cautious when using these techniques.

Key words: *Data Envelopment Analysis, Efficiency, Bootstrap, Bootstrap DEA, Monte Carlo*

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

1 Introduction

The implementation of the bootstrap in the non-parametric data envelopment analysis (DEA) models is a relatively recent and increasingly popular practice in DEA applications. It was introduced by Simar and Wilson (1998) and is now almost a requirement when statistical inference needs to be applied on DEA, especially hypothesis testing. Therefore, it is crucial to establish that the bootstrap provides consistent results and explore the conditions that might affect its performance.

However straightforward it seems to perform Monte Carlo simulations, great care needs to be taken in designing the experiment, especially the data generating process (DGP). And the most important point is to ensure that the DGP is theoretically consistent with the assumptions of bootstrap DEA and at the same time consistent with some economic interpretation. Some of these assumptions have been stated in Simar and Wilson (1998), however Tziogkidis (2012) has provided a deeper insight regarding their implications and has proposed refinements on the application of bootstrap DEA. Moreover, there is a clear motivation in Tziogkidis (2012) to explore certain aspects of bootstrap DEA with Monte Carlo experiments which have not been previously examined.

In this paper we perform the Monte Carlo experiments suggested by Tziogkidis (2012) to examine whether the performance of bootstrap DEA is affected by sample size and dimensions (number of inputs and outputs), by smoothing or not smoothing the empirical distribution of DEA scores as well as by introducing model biases such as specification and measurement biases. We also illustrate the different nature between the bootstrap bias and DEA bias which was thoroughly discussed in Tziogkidis (2012) while we provide empirical support to the author's theoretical arguments about the potential inconsistency of Simar and Wilson's (2000) method of confidence interval construction.

To perform our analysis we compare the moments of the efficiency score distribution of three hypothesized populations and various randomly drawn samples from them, where DEA and bootstrap DEA is applied. Previous studies on the performance of bootstrap DEA, assess the performance of the algorithm on the basis of "coverage probabilities" which, as explained in Tziogkidis (2012), might provide inconsistent results. However, we also compute coverage probabilities and we obtain different results for the particular cases examined in this paper, which is quite puzzling.

Our results suggest that bootstrap DEA is asymptotically consistent with satisfactory rates of convergence. However, as the dimensions of the linear program increase or at the presence of model biases, the rates of convergence reduce significantly. Another interesting result is that the "naïve" or non-smoothed bootstrap seems to perfom better compared to the smooth bootstrap, which contradicts the inconsistency arguments of Simar and Wilson (1998). Actually, this result also suggests that we could avoid using the complicated smoothing techniques which are also very sensitive to the choice of the smoothing parameter.

The remainder of the paper is structured as follows: section 2 reviews the relevant literature on the performance of bootstrap DEA, section 3 outlines the Monte Carlo experiments performed in this study, section 4 presents and discusses the simulation results, section 5 performs an extra exercise comparing coverage probabilities while section 6 concludes the paper.

2 Literature review

Data envelopment analysis (DEA) is a non-parametric technique which is used to assess the relative performance of decision making units (DMUs), introduced by Charnes, Cooper and Rhodes (1978). It uses linear programming to attach optimal weights to a set of inputs and outputs that DMUs use in their production process. The major advantage of DEA is that it does not require the specification of a production function, due to its non-parametric nature. On the other hand, its major disadvantage is that it is not possible to apply statistical inference due to the lack of stochastic elements.

To mitigate this disadvantage of DEA, Simar and Wilson (1998) introduced bootstrap DEA as a tool of extracting the sensitivity of DEA scores towards the randomness which is attributed to the distribution of (in)efficiency. This is a quite strong assumption; however, Tziogkidis (2012) has suggested that if the sample is homogeneous enough, then this assumption is quite reasonable. Moreover, the statistical properties of DEA scores have been explored [Korostelev et al. (1995), Kneip et al. (1998)], suggesting that they converge asymptotically with rates of convergence that depend on sample size and dimensions (number of inputs and outputs). This implies that bootstrap DEA should also perform asymptotically well; however, we should be clear about how good performance is defined.

Usually, to assess the performance of any bootstrap algorithm, Monte Carlo simulations are used where a true model and population is defined. Then, the assessed model is said to perform well if the moments of the bootstrapped models asymptotically converge towards the “true” or population moments of the population. Another aspect of good performance, usually in parametric models, is by modeling a null hypothesis and checking whether the hypothesized “true” value lies within the bootstrapped confidence intervals and whether the percentage of successes (known as coverage) converges towards the nominal level of confidence.

The performance evaluation of bootstrap DEA should be based on its uses, apart from the evaluation of moments or coverage. Using the distribution of bootstrapped efficiency scores and the observed bootstrap bias, Simar and Wilson (1998) suggest that it is possible to approximate the "true" efficiency scores of DMUs by correcting twice for bootstrap bias. Moreover, using the distribution of bootstrap bias, Simar and Wilson (2000) construct confidence intervals where the "true" efficiency scores of DMUs are supposed to lie. However, as Tziogkidis (2012) has argued, both of the aforementioned uses of bootstrap DEA require that the bootstrap bias is equal to the DEA bias, while he proves that using Simar and Wilson's (2000) confidence intervals for hypothesis testing might lead to inconsistent results and suggests that the assessment on the basis of coverage probabilities might be an invalid approach for bootstrap DEA.

The literature on bootstrap DEA performance is quite narrow and seems to be focusing only on coverage and confidence interval widths, while moments are not considered in Monte Carlo simulations. To our knowledge, Löthgren (1998) and Simar and Wilson (2000, 2004) are the only well known simulation studies which assess the performance of Simar and Wilson's (1998, 2000) approaches and in both studies this is done using coverage probabilities. Moreover, these simulations are only based on a single population specification under the assumption of one input and one output. Regarding moment comparisons, we are not aware of any simulation exercise; hence our study is further motivated.

In Löthgren (1998) the Monte Carlo experiment is used to compare the performance of three different bootstrap DEA procedures, including that of Simar and Wilson (1998). Coverage probabilities are calculated for the simple case of 1 input and 1 output (under both CRS and VRS) over a range of levels of significance and number of DMUs. Table 1 summarizes the results of Löthgren (1998) for the bootstrap DEA method of Simar and Wilson (1998). In many cases coverage probabilities fall as sample size increases, however Löthgren (1998) does not provide any explanation as to why this behavior is observed.

Table 1. Löthgren Monte Carlo results on Simar and Wilson (1998)

	Nominal Coverage Levels				
n	*0.8*	*0.9*	*0.95*	*0.975*	*0.99*
20	0.758	0.853	0.860	0.865	0.869
30	0.715	0.886	0.892	0.894	0.900
60	0.617	0.900	0.914	0.914	0.919
120	0.400	0.833	0.947	0.952	0.953
250	0.245	0.501	0.890	0.968	0.972
500	0.091	0.226	0.481	0.836	0.974

Source: Löthgren (1998), Table 5.2

The results in Simar and Wilson (2000) are quite different as the reported coverage probabilities, based on the "enhanced" confidence intervals introduced in the same paper[1], behave ideally as they converge towards the nominal probabilities surprisingly well. Moreover, confidence intervals narrow down which is expected since the bootstrap bias reduces with sample size. Their Monte Carlo exercise involves a one-input, one-output specification under the assumption of output orientation and under both CRS and VRS[2]. Their results for the CRS technology assumption are summarized in Table 2 below. Column 1 reports the sample sizes, columns 2 to 6 present the coverage probabilities for the 80%, 90%, 95%, 97.5% and 99% levels of confidence, respectively, while the last column reports the average 95% confidence interval widths. One interesting result is that smaller samples exhibit lower coverage although their width is quite large, giving the impression that the sample descriptives massively change as sample size increases. More information about the moments of the samples used and of the bootstrap results would have provided a deeper insight.

[1] In their previous paper (Simar and Wilson, 1998) the confidence interval construction was based on the percentiles of bootstrap distribution of bias-corrected efficiency scores. In Simar and Wilson (2000) it was suggested that this approach introduced unnecessary noise and it was proposed to use the percentiles of the distribution of bootstrap biases instead. The resulting confidence intervals have 4 times less variance and their efficiency is the same to the previously used ones.

[2] We would like to note that it is not clear to us whether they employ the homogeneous bootstrap of their 1998 paper or the heterogeneous one introduced in 2000. In either case, the results are worthwhile to present.

Table 2. Simar and Wilson (2000) Monte Carlo results: CRS case

	Nominal Coverage Levels					Av. CI width
n	*0.8*	*0.9*	*0.95*	*0.975*	*0.99*	(95%)
10	0.693	0.814	0.886	0.919	0.942	0.911
25	0.772	0.883	0.935	0.973	0.983	0.586
50	0.784	0.894	0.940	0.970	0.985	0.351
100	0.794	0.911	0.946	0.973	0.988	0.187
200	0.810	0.899	0.946	0.970	0.994	0.095
400	0.807	0.903	0.953	0.977	0.995	0.047

Source: Simar and Wilson (2000), Table 1

Similar evidence are found in Simar and Wilson (2004), who perform Monte Carlo experiments to compare their smooth bootstrap procedure against two variants of the "naïve" bootstrap, under a simple, one input and one output setup. The first of the two "naïve" procedures draws from the empirical distribution of inputs and outputs (known as case or pairs resampling) whereas the latter draws from the empirical distribution of efficiency scores (known as fixed or "residual" resampling). They find that the smooth bootstrap outperforms the other two, while the "naïve" bootstrap which draws from the input and output data provides better results than the other one. However, drawing directly from the data would be computationally more intensive in applied research as it would require applying the bootstrap procedure for each DMU separately (i.e. it can be only applied for one reference DMU at a time), while it would be more intuitive in the case of non-oriented models (Tziogkidis, 2012). Their results for the CRS case and for a 95% level of significance are summarized in Table 3. In particular, Table 3 reports the coverage probabilities and the average confidence interval widths for the three different cases; the Simar and Wilson (2000) enhanced method ("SW2000"), the "naïve" bootstrap with pair resampling ("Pairs") and the "naïve" bootstrap with fixed resampling ("Fixed"). Again, the results indicate that the proposed method of Simar and Wilson (1998, 2000) performs well, in terms of coverage, and that it is superior compared to the non-smooth or "naïve" bootstrap.

In this simulation study we assess the performance of bootstrap DEA, both smooth and "naïve", on the basis of moments although we also provide results for comparison purposes. Apart from evaluating moments, we examine the behavior of certain variables to assess the validity of certain assumptions or suggestions expressed in Simar and Wilson (1998, 2000) and were criticized by Tziogkidis (2012). The results of this study provide a better understanding of

bootstrap DEA and suggest another approach of performance evaluation which has been unexplored until now.

Table 3. Simar and Wilson (2004) Monte Carlo CRS results (95%)

	Coverage Probabilities (95%)			Av. CI Width (95%)		
n	*SW2000*	*Pairs*	*Fixed*	*SW2000*	*Pairs*	*Fixed*
10	0.916	0.899	0.899	0.1384	0.2018	0.2018
25	0.932	0.894	0.890	0.0551	0.0664	0.0693
50	0.920	0.896	0.891	0.0283	0.0320	0.0315
100	0.921	0.889	0.891	0.0146	0.0154	0.0157
200	0.937	0.879	0.888	0.0076	0.0078	0.0074
400	0.936	0.883	0.889	0.0039	0.0037	0.0038
800	0.950	0.886	0.871	0.0019	0.0019	0.0019
1600	0.957	0.876	0.868	0.0010	0.0009	0.0009
3200	0.951	0.897	0.864	0.0005	0.0005	0.0005
6400	0.960	0.878	0.868	0.0003	0.0002	0.0002

Source: Simar and Wilson (2004), Tables 10.1 and 10.3

3 The Monte Carlo experiments

3.1 The experiment outline

We perform our Monte Carlo experiments using three different population assumptions of size *N*=10,000, which are assumed to reflect one theoretically consistent case and two cases which are associated with model biases. The efficiency scores of the three populations and their distributions are unobservable and are expected to be approached by the samples only asymptotically. The simulations are run over sample sizes of 10, 15, 20, 25, 30, 60 and 120 and three different dimensions: 1 input and 1 output (1I/1O), 2 inputs and 1 output (2I/1O), and 2 inputs and 2 outputs (2I/2O). Each bootstrap DEA model involves $B = 2000$ replications while the Monte Carlo experiment is run $M = 1000$ times. Our experiments were applied once using the smooth bootstrap[3] and once the "naïve" bootstrap. All calculations were performed in

[3] In our simulation study we choose the least squares cross validation (LSCV) method to determine the smoothing parameter (after correcting for sample size). Compared to "plug-in" methods, LSCV has the advantage of performing well, even when the target distribution has a non-standard shape. A nice comparison is provided in Loader (1999).

Matlab using a Monte Carlo code written by the author, which repeatedly calls an appropriately modified bootstrap DEA MatLab code written by L. Simar (last updated in Nov. 2002). The computational costs in seconds, using a standard PC Intel i3 2.8MHz processor, are presented for each case in Table 4 below, while the cumulative runtime is 56.6 days.

Table 4. Computational costs (seconds) of Monte Carlo simulations

	Standard			**Alternative A**			**Alternative B**		
	1I/1O	*2I/1O*	*2I/2O*	*1I/1O*	*2I/1O*	*2I/2O*	*1I/1O*	*2I/1O*	*2I/2O*
Smooth	145717	277203	420790	153163	257373	358620	150952	298104	448240
Naïve	142347	265474	410790	146450	240212	356740	146802	247749	424690

3.2 The data generating processes

The data generating processes (DGPs) are similar to the ones used in previous Monte Carlo studies; however, we apply small variations in order to examine more cases and to attach an economic interpretation to the cases examined. We generate three different types of populations, the intuition of which is explained in the next subsection, which we call as "Standard", "Alternative A", and Alternative "B". All experiments are performed over three different dimensions: 1 input and 1 output, 2 inputs and 1 output and 2 inputs and 2 outputs.

For the 1 input and 1 output case, the true production function is a simple case of CRS technology where the efficient levels of input $\left(x^{eff}\right)$ are uniformly distributed on the [10,20] interval: $y = x^{eff} \sim U(10{,}20)$. Regarding the case of 2 inputs and 1 output, we use again a standard Cobb Douglas CRS production function to generate output: $y = \left(x_1^{eff}\right)^{0.5}\left(x_2^{eff}\right)^{0.5}$. It is assumed that the efficient levels of inputs are uniformly distributed with $x_1^{eff} \sim U(10, 20)$ and $x_2^{eff} \sim U(20, 30)$. Finally, for the case of 2 inputs and 2 outputs, the CRS production function is Cobb Douglas for both outputs: $y_1 = \left(x_1^{eff}\right)^{0.5}\left(x_2^{eff}\right)^{0.5}$ and $y_2 = \left(x_1^{eff}\right)^{0.3}\left(x_2^{eff}\right)^{0.7}$. It is assumed that efficient levels of inputs are normally distributed with $x_1^{eff} \sim U(10{,}20)$ and $x_2^{eff} \sim U(20{,}30)$.

In all previous cases actual inputs (x_i) are assumed to deviate from their efficient levels in three different ways which reflect three different assumptions on the assumed error (the three DGP assumptions):

Standard: $x_i = x_i^{eff} e^{0.2|u|}$ where $u \sim N(0,1)$

Alternative A: $x_i = x_i^{eff} e^{0.2u}$ where $u \sim N(0,1)$

Alternative B: $x_i = x_i^{eff} e^{0.8u}$ where $u \sim U(0,1)$

A necessary clarification for the cases where 2 inputs are used is that the error term (u) is a common vector used by both inputs. The logic for this specification lies in the definition of DEA efficiency scores in input orientation which are actually contraction factors applied to all inputs simultaneously. For example, if a DMU has an efficiency score of 0.8, then it will need to use only 80% of *all* its inputs in order to become technically efficient. Therefore our experiment design is consistent with DEA principles.

Once we have obtained the pairs of inputs and outputs we apply DEA on the assumed populations and the resulting efficiency scores are treated as the "true" ones which, however, we call them "population scores" to avoid confusion. After constructing the hypothesized populations we produce the 1000 samples where both DEA and bootstrap DEA will be applied on. In order to avoid discriminating against the smooth or the "naïve" bootstrap we randomly draw these samples in advance and use exactly the same ones in the smooth and "naïve" bootstrap evaluation. Hence, any error imposed by the randomness of the Monte Carlo exercise is mitigated since both procedures have the same base of comparison.

3.3 The intuition behind the DGPs

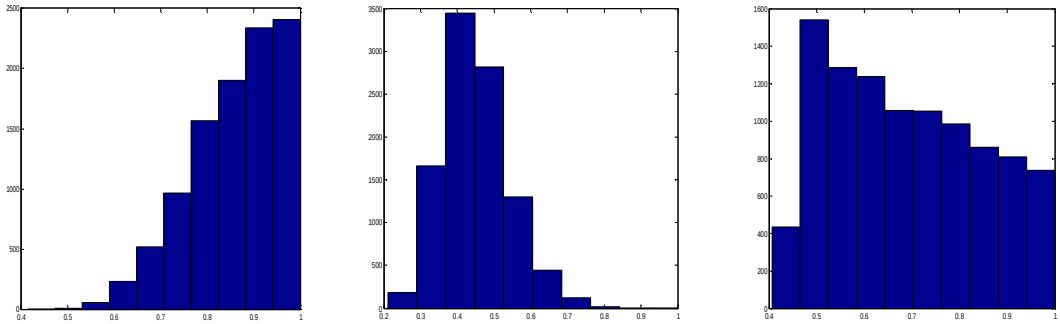
The population distribution of efficiency scores is presented in Figure 1 for the case of 1 input and 1 output[4]. The first histogram corresponds to the "Standard" case, while the second and the third histograms correspond to "Alternative A" and "Alternative B", respectively. It is obvious that the standard case is associated with a half-normal distribution, alternative A with a slightly skewed normal distribution and alternative B mostly with a uniform distribution.

The choice of the three different populations is intentional and aims in assessing bootstrap DEA under different conditions. In particular, the standard case uses the absolute normal deviations of inputs from their efficient levels. Indeed, a theoretically consistent exercise should use only positive deviations of efficient levels, which respect the DEA assumption in input orientation that $x > x_{eff}$, hence we named it "Standard". Also, it is consistent with a perfectly

[4] The histograms for the higher dimensions look almost exactly the same and are not presented here to conserve space.

competitive market as it consists of quite homogeneous firms which exhibit constant returns to scale and have access to the same technology in order to produce the same output. All firms are expected to perform efficiently, however, due to random exogenous factors some firms use proportionately more outputs than the efficient level.

Figure 1. Population distribution of efficiency scores for each of the three alternatives



Alternative A, exhibits an inconsistency with DEA: the input deviations may also be negative which implies that the input levels might be lower than the efficient input levels. We deem this inconsistency as a possible source of DEA bias which is associated with technology. In particular, in the DEA world it would be a model specification bias to include in a sample DMUs which have access to unique technology, while the reference technology is believed to be homogeneous for all firms. That would result in an unfair comparison while the few efficient DMUs would be the technologically privileged ones. Hence, Alternative A aims in demonstrating the effects of specification bias in DEA and bootstrap DEA and examines the consequences of not respecting the requirement of homogeneity in technology.

Finally, Alternative B exhibits no such inconsistency as the assumed deviations are by definition positive (they are positively uniformly distributed). Although it looks similar to the standard case, the assumption of uniform deviations imposes a more random structure in the market, which is no longer associated with perfect competition. This is due to the fact that mean deviation is no longer zero, as in the “Standard” case. Hence, this could imply a market with no particular or unclear competitive conditions, but it could also be due measurement errors which result in such structures. Indeed, even in a perfectly competitive market, an error in measurement of input, either due to misreporting or due to methodological mistakes in measuring inputs, would largely affect the distribution of efficiency scores. However, such an

error would not affect results on average as much as in the previous case of the model specification bias.

4 Simulation results

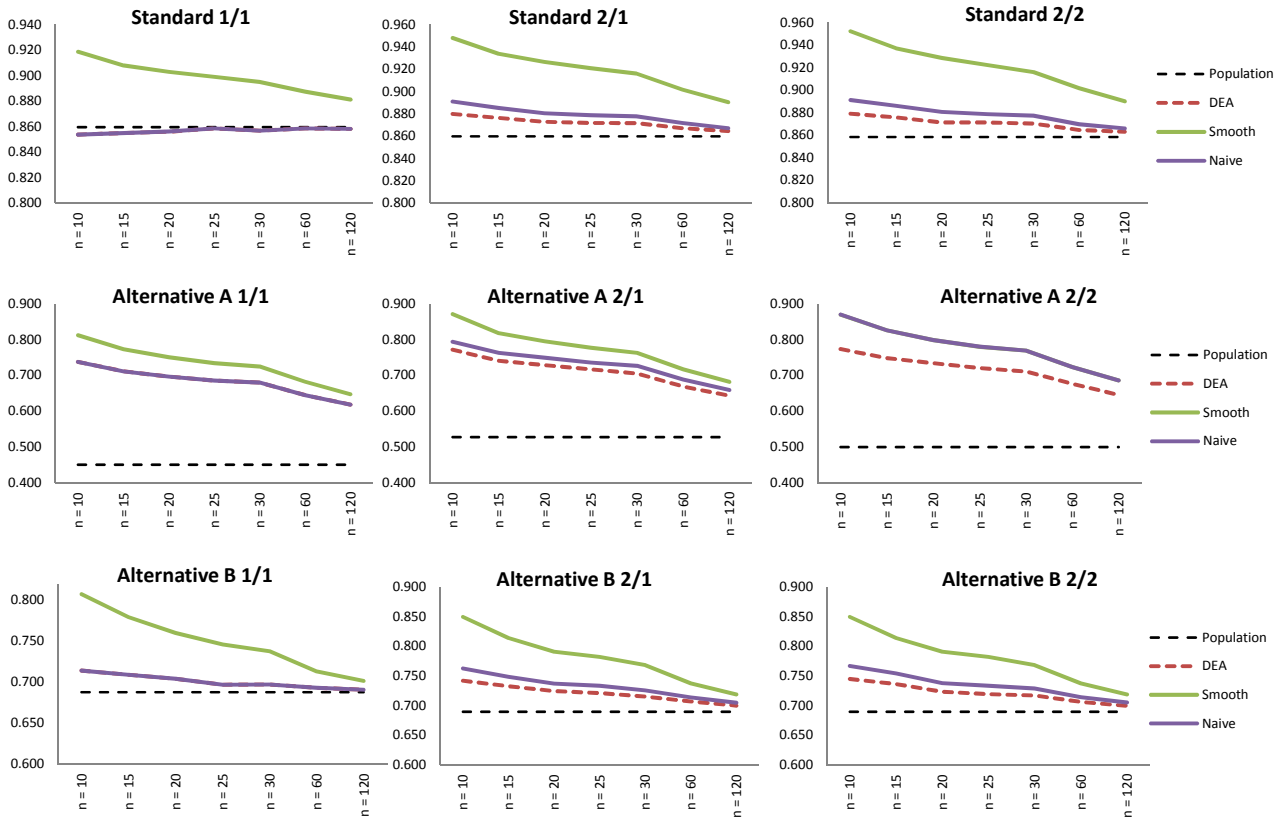
As with every Monte Carlo exercise, the most interesting moment is the mean. We therefore first present a graph which summarizes our Monte Carlo results on mean convergence. In particular, Figure 2 presents the 9 different combinations of DGPs and model dimensions examined, plotted against sample size. The black dotted line represents the population mean efficiency score, the red dotted line corresponds to the center of the distribution of sample DEA means, while the green and the blue lines represent the centers of the distributions of the smooth and "naïve" bootstrap DEA means, respectively[5].

The results in Figure 2 are quite intuitive. First and foremost, they suggest that the DEA means converge towards the population mean quite fast, with the exception of Alternative A where the rates of convergence are very low which is attributed to the hypothesized model specification bias. The model (or DEA) bias, reflected by the distance between the black dotted line and the red dotted line, is therefore affected significantly by the inclusion of DMUs with access to different technology of production. Especially in this case it is obvious that the bootstrap bias is very different compared to the DEA bias for both bootstrap DEA procedures, which provides further support to the arguments in Tziogkidis (2012) against the practice of correcting twice for bootstrap bias to obtain an approximation of the population score.

Regarding bootstrap DEA results, the bootstrap means converge towards the DEA means, which is the expected behavior. The rate of convergence is not affected by model biases but is only affected by the dimensions of the program. This is also expected as the sample DEA scores are assumed to be free of any external biases when bootstrap DEA is applied and the bootstrap mimics the behavior of the target distribution of DEA scores. The evaluation of performance of bootstrap DEA should therefore focus on bootstrap bias, which is represented by the distance between the green or the blue line and the red dotted one. Indeed, bootstrap bias decreases

[5] To be precise, by center we mean the median. Also for bootstrap DEA means we calculate the mean of every bootstrapped sample and then its median reflects the bootstrap mean. From the Monte Carlo experiment a distribution of bootstrap means is computed and the values reported here are the medians of these distributions.

Figure 2. Monte Carlo results on mean convergence



with sample size, however it increases with the number of inputs and outputs. It is obvious from our results that the bootstrap bias of the "naïve" bootstrap is smaller compared to the bias of the smooth bootstrap, which contradicts the claims of Simar and Wilson (1998) that the "naïve" bootstrap is inconsistent; in fact it performs quite better than the smooth, while in some cases the bootstrap bias is almost zero.

The latter is not clear in Figure 2 where there are 3 cases (the 1 input and 1 output cases) where the naïve bootstrap means coincide with the DEA ones. We have therefore summarized the results on moments (mean, standard deviation, skewness and kurtosis) in Table 5, where the names in each column and row are self explanative. Regarding the standard case, all moments of the bootstrap DEA procedures converge towards the DEA ones, which, in turn, converge towards the population moments. In fact, even for small samples the results for both DEA and bootstrap DEA are quite satisfactory. However, for $n = 10$, where the rule of thumb for the minimum number of DMUs is violated, we observe excess skewness and kurtosis, although the first two moments behave well. Comparing the smooth and the "naïve" bootstrap we see that their higher moments (skewness and kurtosis) perform quite closely, while the "naïve" is closer to the DEA mean and variance, implying that it performs better.

Exactly the same conclusions are reached for Alternative B. Moments converge to their "true" values while very small samples might be problematic in terms of higher moments. The "naïve" bootstrap still performs better regarding mean and variance.

However, the results are different for Alternative A. In the presence of model specification bias, we observe that the DEA higher moments are far from the population ones. This also affects the higher moments of the bootstrap DEA procedures, which, however, still perform quite close to the true ones. This implies that the performance of the bootstrap is not affected by the DEA bias, hence applying hypothesis testing to compare two DMUs of the same sample would give consistent results. On the other hand, comparing a DMU from the "biased" sample to a DMU from another sample would be inappropriate, which confirms the relevant argument in Tziogkidis (2012).

The simulation results indicate that bootstrap DEA is a consistent procedure, which performs well even under the existence of model biases. However, care needs to be taken when comparing different samples as the results would be inconsistent. Moreover, comparing the smooth bootstrap of Simar and Wilson (1998) and the non-smooth or "naïve" bootstrap, we find

that the latter performs better; hence, the term "naïve" seems to be quite an unfair characterization.

Table 5. Moments of the Monte Carlo simulations

	Standard 1/1				Standard 2/1				Standard 2/2			
Population	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
N = 10,000	0.860	0.097	-0.682	2.919	0.860	0.097	-0.688	2.927	0.858	0.098	-0.656	2.836
DEA	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.854	0.125	-0.875	2.970	0.880	0.118	-0.862	2.762	0.879	0.117	-0.825	2.660
n = 15	0.855	0.116	-0.866	3.086	0.876	0.111	-0.843	2.891	0.876	0.112	-0.817	2.807
n = 20	0.856	0.111	-0.837	3.141	0.873	0.109	-0.781	2.844	0.871	0.109	-0.745	2.729
n = 25	0.859	0.108	-0.858	3.239	0.872	0.106	-0.791	2.891	0.871	0.107	-0.738	2.793
n = 30	0.857	0.107	-0.802	3.155	0.872	0.105	-0.760	2.883	0.870	0.106	-0.719	2.805
n = 60	0.858	0.102	-0.754	3.030	0.867	0.101	-0.716	2.854	0.865	0.102	-0.670	2.743
n = 120	0.858	0.100	-0.712	2.937	0.864	0.099	-0.699	2.877	0.863	0.100	-0.666	2.765
Smooth	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.919	0.134	-1.038	3.912	0.948	0.128	-0.919	3.436	0.952	0.130	-0.871	3.315
n = 15	0.908	0.123	-0.966	3.663	0.934	0.120	-0.875	3.336	0.937	0.121	-0.820	3.214
n = 20	0.903	0.117	-0.906	3.555	0.927	0.116	-0.800	3.127	0.929	0.117	-0.743	2.989
n = 25	0.899	0.114	-0.914	3.578	0.921	0.113	-0.796	3.118	0.922	0.114	-0.729	3.010
n = 30	0.895	0.111	-0.844	3.415	0.916	0.111	-0.765	3.078	0.916	0.112	-0.713	2.987
n = 60	0.887	0.106	-0.773	3.140	0.902	0.105	-0.715	2.938	0.902	0.107	-0.668	2.822
n = 120	0.881	0.102	-0.721	2.986	0.890	0.102	-0.700	2.922	0.890	0.103	-0.665	2.803
Naïve	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.854	0.125	-1.038	3.912	0.891	0.121	-0.931	3.475	0.891	0.122	-0.880	3.305
n = 15	0.855	0.116	-0.966	3.663	0.885	0.114	-0.885	3.346	0.886	0.115	-0.823	3.212
n = 20	0.856	0.111	-0.906	3.555	0.881	0.111	-0.801	3.132	0.881	0.112	-0.741	2.987
n = 25	0.859	0.108	-0.914	3.578	0.879	0.108	-0.800	3.121	0.879	0.109	-0.729	3.015
n = 30	0.857	0.107	-0.844	3.415	0.878	0.107	-0.768	3.080	0.877	0.108	-0.720	2.981
n = 60	0.859	0.102	-0.773	3.140	0.872	0.102	-0.718	2.938	0.870	0.103	-0.672	2.820
n = 120	0.858	0.100	-0.721	2.986	0.867	0.100	-0.699	2.923	0.866	0.100	-0.667	2.801
	Alternative A 1/1				**Alternative A 2/1**				**Alternative A 2/2**			
Population	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
N = 10,000	0.450	0.091	0.629	3.651	0.527	0.107	0.562	3.435	0.500	0.101	0.608	3.640
DEA	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.738	0.160	-0.012	2.555	0.772	0.157	-0.113	2.320	0.774	0.157	-0.119	2.365
n = 15	0.711	0.151	0.168	2.690	0.741	0.151	0.058	2.398	0.748	0.150	0.026	2.472
n = 20	0.697	0.145	0.265	2.778	0.729	0.146	0.124	2.530	0.734	0.147	0.119	2.503
n = 25	0.686	0.142	0.311	2.808	0.717	0.145	0.214	2.521	0.721	0.143	0.187	2.568
n = 30	0.680	0.140	0.327	2.838	0.705	0.141	0.239	2.616	0.711	0.141	0.259	2.566
n = 60	0.645	0.132	0.464	3.061	0.669	0.135	0.377	2.816	0.676	0.136	0.371	2.775
n = 120	0.619	0.125	0.524	3.214	0.643	0.129	0.442	2.970	0.646	0.129	0.445	2.988
Smooth	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.813	0.177	-0.015	3.177	0.872	0.183	-0.034	2.875	0.870	0.184	-0.016	3.015
n = 15	0.773	0.166	0.188	3.094	0.818	0.172	0.157	2.810	0.826	0.172	0.135	2.941
n = 20	0.751	0.157	0.287	3.083	0.795	0.164	0.220	2.907	0.799	0.165	0.202	2.895
n = 25	0.734	0.153	0.331	3.047	0.778	0.161	0.297	2.831	0.780	0.160	0.275	2.925
n = 30	0.725	0.149	0.345	3.038	0.763	0.156	0.323	2.902	0.769	0.156	0.347	2.921
n = 60	0.683	0.139	0.476	3.174	0.717	0.146	0.427	3.027	0.723	0.147	0.435	3.002
n = 120	0.647	0.132	0.531	3.275	0.682	0.137	0.480	3.112	0.686	0.138	0.483	3.162
Naïve	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.738	0.160	-0.015	3.177	0.794	0.168	-0.046	2.887	0.870	0.184	-0.017	2.961
n = 15	0.712	0.151	0.188	3.094	0.763	0.161	0.154	2.813	0.826	0.172	0.133	2.906
n = 20	0.697	0.145	0.287	3.083	0.750	0.154	0.220	2.904	0.799	0.165	0.203	2.872
n = 25	0.686	0.142	0.331	3.047	0.736	0.152	0.298	2.830	0.780	0.160	0.275	2.894
n = 30	0.680	0.140	0.345	3.038	0.727	0.149	0.326	2.911	0.769	0.156	0.347	2.883
n = 60	0.645	0.132	0.476	3.174	0.689	0.140	0.429	3.022	0.723	0.147	0.434	2.971
n = 120	0.619	0.125	0.531	3.275	0.659	0.133	0.478	3.107	0.686	0.138	0.483	3.142

Table 5. Moments of the Monte Carlo simulations (continued)

	Alternative B 1/1				Alternative B 2/1				Alternative B 2/2			
Population	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
N = 10,000	0.688	0.158	0.270	1.886	0.689	0.157	0.271	1.901	0.689	0.157	0.271	1.901
DEA	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.714	0.179	0.150	2.162	0.742	0.175	0.046	2.050	0.742	0.175	0.046	2.050
n = 15	0.709	0.172	0.176	2.093	0.732	0.172	0.117	2.005	0.732	0.172	0.117	2.005
n = 20	0.704	0.168	0.175	2.081	0.724	0.170	0.177	2.008	0.724	0.170	0.177	2.008
n = 25	0.697	0.166	0.228	2.071	0.721	0.167	0.178	1.983	0.721	0.167	0.178	1.983
n = 30	0.697	0.165	0.237	2.025	0.715	0.166	0.222	1.985	0.715	0.166	0.222	1.985
n = 60	0.693	0.162	0.247	1.952	0.707	0.163	0.239	1.930	0.707	0.163	0.239	1.930
n = 120	0.691	0.160	0.260	1.930	0.700	0.161	0.261	1.922	0.700	0.161	0.261	1.922
Smooth	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.807	0.202	0.178	2.482	0.849	0.206	0.133	2.377	0.849	0.206	0.133	2.377
n = 15	0.780	0.190	0.196	2.236	0.814	0.195	0.184	2.207	0.814	0.195	0.184	2.207
n = 20	0.760	0.182	0.190	2.175	0.791	0.189	0.242	2.153	0.791	0.189	0.242	2.153
n = 25	0.746	0.178	0.242	2.139	0.782	0.184	0.227	2.093	0.782	0.184	0.227	2.093
n = 30	0.738	0.175	0.250	2.071	0.768	0.180	0.264	2.069	0.768	0.180	0.264	2.069
n = 60	0.713	0.167	0.253	1.966	0.737	0.172	0.260	1.972	0.737	0.172	0.260	1.972
n = 120	0.701	0.162	0.264	1.936	0.718	0.165	0.270	1.943	0.718	0.165	0.270	1.943
Naïve	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
n = 10	0.714	0.179	0.178	2.482	0.762	0.186	0.131	2.372	0.767	0.191	0.124	2.314
n = 15	0.709	0.172	0.196	2.236	0.748	0.178	0.182	2.202	0.754	0.181	0.173	2.125
n = 20	0.704	0.168	0.190	2.175	0.737	0.175	0.240	2.148	0.738	0.178	0.243	2.071
n = 25	0.697	0.166	0.242	2.139	0.733	0.172	0.228	2.091	0.733	0.174	0.241	2.043
n = 30	0.697	0.165	0.250	2.071	0.725	0.169	0.262	2.064	0.728	0.173	0.248	1.985
n = 60	0.693	0.162	0.253	1.966	0.713	0.165	0.259	1.972	0.714	0.168	0.266	1.937
n = 120	0.691	0.160	0.264	1.936	0.705	0.162	0.270	1.941	0.705	0.164	0.276	1.905

5 Coverage results

Evaluating the bootstrap DEA procedures on the basis of coverage has been argued to be counter-intuitive, while results may not be reliable (Tziogkidis, 2012). However, since the existing literature has been using coverage probabilities thus far, we have also provided relevant results. Coverage probabilities evaluate the probability that the true efficiency score of a DMU lies within confidence intervals which have been determined using one of the available methodologies.

Simar and Wilson (2000), as already mentioned, construct these intervals using the percentiles of the distribution of bootstrap bias. Then the standard method to calculate coverage in Monte Carlo experiments is to fix a DMU to appear in every replication. We follow the literature and we fix the DMU to have as inputs and outputs the average values of the relevant input and output variable. Thus, for each bootstrap DEA run (or each Monte Carlo repetition) we construct the Simar and Wilson (2000) confidence intervals and we observe whether the population score of the "fixed" DMU lies within it. Applying the same procedure for

all $M = 1000$ Monte Carlo replications we calculate the proportion of times where our condition is satisfied, which is the coverage probability.

The results on coverage are presented in Table 6, for all cases examined, both for smooth and "naïve" bootstrap, for sample sizes of 10, 15, 20, 25, 30, 60 and 120 and for significance levels of 0.20, 0.10, 0.05 and 0.01. One of the findings is that in all cases, as the level of significance decreases the coverage probabilities increase which is reasonable as the confidence intervals widen. Moreover, coverage probabilities do not necessarily increase with sample size which is due to the fact that Simar and Wilson's (2000) confidence intervals narrow down towards a different point than the population efficiency score. The decreasing coverage is also evident in the results of Löthgren (1998) and to a less extent in Simar and Wilson (2004) while the decreasing confidence interval width is due to the decreasing bootstrap bias and is also demonstrated in Simar and Wilson (2004).

We also obtain different coverage probabilities across the three alternatives, with the significantly lower ones, overall, being those in Alternative A, which is subject to model specification biases. This highlights the fact that in the presence of such biases the bootstrap bias becomes very different compared to the model bias. The other two specifications exhibit similar results and the "performance" in terms of coverage is analogous to the performance in terms of moments, however the latter provides a more valid and accurate assessment of the methods.

The most astonishing result, however, is the significantly worse performance of the smooth bootstrap compared to the "naïve", which contradicts previous findings in the literature with the only exception of Löthgren (1998) where the coverage probabilities have similar behavior but quite different values. In fact, as sample size increases the smooth bootstrap always exhibits lower coverage while the lower values are reported for the standard case, which is an interesting result that requires further exploration[6].

The higher coverage for the naïve bootstrap can be easily justified by the convergence of the bootstrap and model biases (towards zero) which is a requirement for the consistency of the method of Simar and Wilson (2000). Still, the coverage probabilities do not converge towards the nominal levels. However, we would expect that both the smooth and the naïve bootstrap would exhibit the desirable results for very large sample sizes, where both the bootstrap and model biases would be negligible.

[6] This is also true in the moments comparison and can be easily verified by inspecting Figure 2.

Table 6. Monte Carlo results on coverage

	Standard 1/1				Standard 2/1				Standard 2/2			
Cov. Smooth	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**
n = 10	0.284	0.354	0.399	0.487	0.397	0.493	0.564	0.685	0.414	0.498	0.581	0.710
n = 15	0.171	0.209	0.242	0.307	0.327	0.403	0.464	0.563	0.275	0.358	0.419	0.529
n = 20	0.133	0.170	0.201	0.246	0.257	0.316	0.369	0.450	0.251	0.312	0.350	0.432
n = 25	0.124	0.151	0.171	0.202	0.225	0.277	0.310	0.379	0.206	0.257	0.286	0.346
n = 30	0.118	0.137	0.147	0.172	0.181	0.228	0.264	0.322	0.186	0.238	0.280	0.333
n = 60	0.069	0.082	0.086	0.095	0.111	0.130	0.146	0.171	0.112	0.140	0.152	0.168
n = 120	0.032	0.035	0.040	0.047	0.068	0.083	0.086	0.099	0.060	0.067	0.075	0.088
Cov. Naïve	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**
n = 10	0.722	0.756	0.871	0.942	0.527	0.642	0.742	0.862	0.530	0.654	0.749	0.871
n = 15	0.740	0.744	0.864	0.943	0.525	0.674	0.777	0.894	0.575	0.722	0.810	0.913
n = 20	0.758	0.759	0.889	0.948	0.580	0.715	0.809	0.912	0.544	0.695	0.781	0.908
n = 25	0.755	0.755	0.870	0.928	0.563	0.688	0.783	0.907	0.559	0.714	0.833	0.937
n = 30	0.779	0.786	0.884	0.942	0.569	0.704	0.805	0.915	0.576	0.700	0.810	0.925
n = 60	0.768	0.804	0.893	0.951	0.587	0.709	0.818	0.923	0.582	0.723	0.812	0.935
n = 120	0.780	0.825	0.904	0.955	0.527	0.686	0.808	0.939	0.500	0.675	0.795	0.926
	Alternative A 1/1				**Alternative A 2/1**				**Alternative A 2/2**			
Cov. Smooth	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**
n = 10	0.107	0.152	0.215	0.367	0.184	0.258	0.338	0.473	0.152	0.208	0.266	0.396
n = 15	0.092	0.137	0.183	0.304	0.167	0.243	0.308	0.441	0.145	0.184	0.231	0.345
n = 20	0.090	0.122	0.167	0.270	0.176	0.238	0.281	0.419	0.121	0.158	0.193	0.305
n = 25	0.078	0.120	0.159	0.256	0.161	0.232	0.274	0.387	0.121	0.171	0.213	0.298
n = 30	0.080	0.100	0.150	0.240	0.195	0.245	0.296	0.385	0.104	0.145	0.180	0.281
n = 60	0.106	0.131	0.158	0.240	0.208	0.255	0.310	0.413	0.105	0.133	0.171	0.260
n = 120	0.098	0.125	0.155	0.199	0.234	0.287	0.338	0.422	0.131	0.164	0.187	0.254
Cov. Naïve	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**
n = 10	0.143	0.162	0.234	0.333	0.159	0.199	0.267	0.370	0.152	0.209	0.268	0.398
n = 15	0.125	0.125	0.218	0.304	0.167	0.213	0.275	0.385	0.145	0.184	0.230	0.347
n = 20	0.124	0.126	0.187	0.271	0.159	0.221	0.267	0.382	0.122	0.157	0.194	0.309
n = 25	0.110	0.111	0.178	0.250	0.142	0.195	0.256	0.356	0.120	0.170	0.209	0.296
n = 30	0.098	0.100	0.167	0.245	0.188	0.234	0.276	0.373	0.104	0.143	0.181	0.284
n = 60	0.114	0.125	0.162	0.243	0.200	0.261	0.309	0.406	0.105	0.135	0.171	0.262
n = 120	0.108	0.122	0.150	0.212	0.241	0.297	0.337	0.419	0.132	0.164	0.187	0.254
	Alternative B 1/1				**Alternative B 2/1**				**Alternative B 2/2**			
Cov. Smooth	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**
n = 10	0.348	0.420	0.487	0.573	0.429	0.524	0.601	0.711	0.429	0.524	0.601	0.711
n = 15	0.277	0.333	0.379	0.446	0.372	0.455	0.527	0.632	0.372	0.455	0.527	0.632
n = 20	0.266	0.317	0.349	0.400	0.383	0.467	0.517	0.596	0.383	0.467	0.517	0.596
n = 25	0.228	0.267	0.298	0.346	0.350	0.418	0.463	0.546	0.350	0.418	0.463	0.546
n = 30	0.238	0.284	0.318	0.366	0.349	0.422	0.460	0.526	0.349	0.422	0.460	0.526
n = 60	0.230	0.273	0.298	0.327	0.307	0.387	0.447	0.520	0.307	0.387	0.447	0.520
n = 120	0.275	0.317	0.347	0.374	0.343	0.393	0.425	0.478	0.343	0.393	0.425	0.478
Cov. Naïve	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**	**p = 0.20**	**p = 0.10**	**p = 0.05**	**p = 0.01**
n = 10	0.748	0.788	0.866	0.937	0.635	0.720	0.806	0.899	0.607	0.718	0.820	0.903
n = 15	0.738	0.741	0.877	0.944	0.610	0.729	0.818	0.914	0.546	0.681	0.793	0.913
n = 20	0.758	0.761	0.889	0.946	0.574	0.715	0.807	0.919	0.587	0.729	0.836	0.939
n = 25	0.757	0.764	0.883	0.937	0.584	0.731	0.827	0.937	0.588	0.738	0.823	0.940
n = 30	0.755	0.764	0.890	0.952	0.575	0.721	0.825	0.934	0.581	0.719	0.821	0.934
n = 60	0.739	0.760	0.871	0.960	0.560	0.707	0.805	0.925	0.592	0.732	0.815	0.928
n = 120	0.714	0.761	0.845	0.950	0.608	0.734	0.821	0.942	0.611	0.747	0.837	0.944

To acquire a better understanding of the behavior of Simar and Wilson's (2000) smooth confidence intervals, we produced the graphs in Figure 3. Figure 3 the median lower and upper boundaries of the 95% confidence intervals (green and blue solid lines, respectively), along with

the population efficiency score of the DMU under evaluation (or fixed DMU represented by the black dotted line) as well as the median of the distribution of Simar and Wilson's (1998) bootstrap bias (twice) corrected efficiency scores of the "fixed DMU" (represented by the thin red dotted line). The latter is used only to demonstrate that the confidence intervals in Simar and Wilson (2000) do cover the bootstrap bias corrected scores of Simar and Wilson (1998), since they have the same theoretical foundations and since they are constructed on the assumption that the bootstrap bias has to be approximately equal to the model bias. Therefore, if this assumption is not respected, they both provide inconsistent results.

The plots in Figure 3 suggest that confidence intervals do narrow down with sample size, which is an expected property since the bootstrap bias decreases. However, they fail, on average to include the "fixed" DMU. However, inspecting the trend of the confidence intervals towards the "fixed" DMU, we observe that our previous point is correct: they will asymptotically return high coverage as the bootstrap and model biases converge to zero.

It is always possible that our results are restricted to the specific cases examined, which is implied by the different results in the literature. However, a solid theoretical explanation should be given which justifies the behavior of the smooth bootstrap towards these DGPs. In order to mitigate the chance that our results are due to programming mistakes, we performed a Monte Carlo experiment using the R-package FEAR, which has been developed and is continuously updated by P. Wilson. In particular, we examined the case of $n = 10$ under our "Standard" 1 input and 1 output specification and we obtained a coverage probability of 0.38 which is very close to our reported one (0.399).

Comparing our coverage results with Simar and Wilson's (2004) is not a straightforward task, since the DGP used in their paper is different. We did not try to replicate their results since we trust the reported ones, but we instead explored different DGPs and tried to attach an economic interpretation for each. In their experiments they use the following CRS 1 input 1 output specification: $y = xe^{-|v|}$ where $v \sim N(0,1)$ and $x \sim U(1,9)$ which returns the population of efficiency scores in Figure 4. The large differences in our results highlight the fact that the approach of Simar and Wilson (1998, 2000) might be more appropriate to be used in case where the population distribution of efficiency scores is believed to be similar to the one in Figure 4.

Figure 3. Simar and Wilson's (2000) simulated confidence intervals

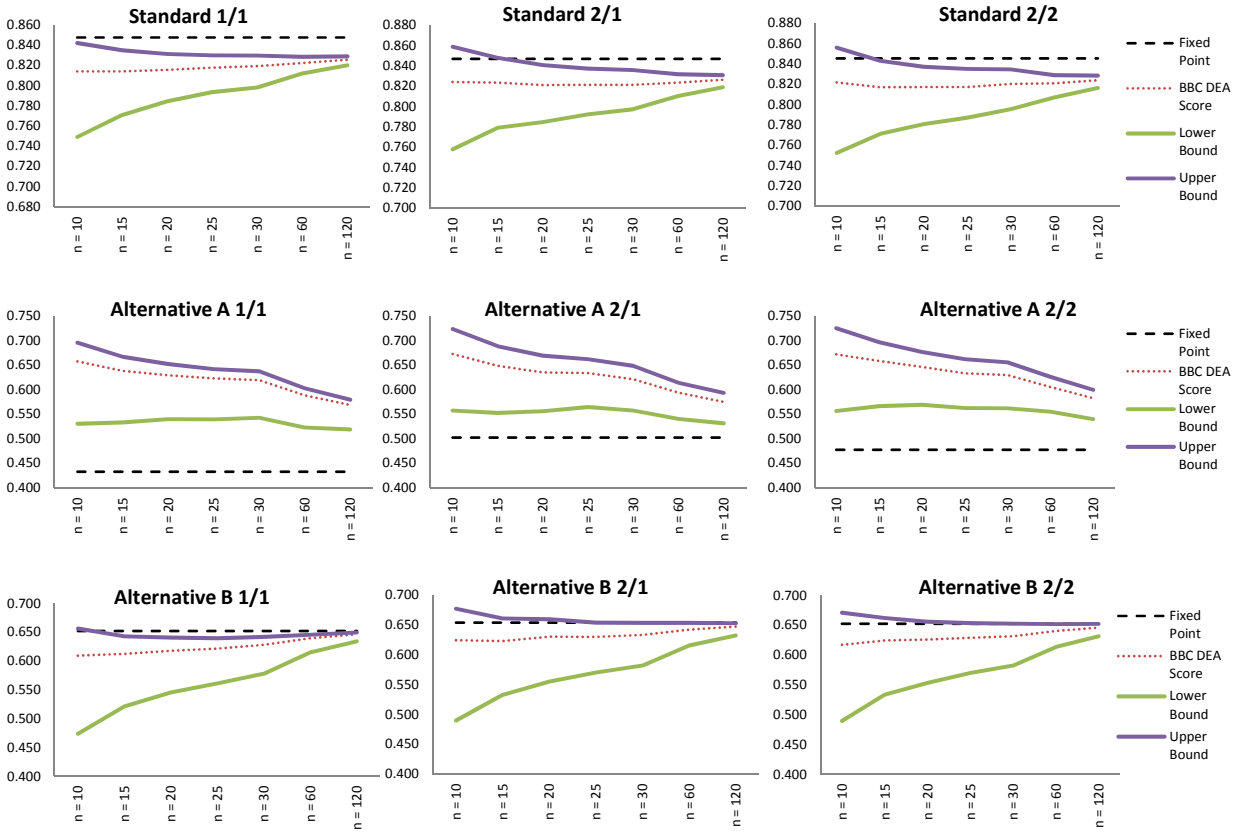
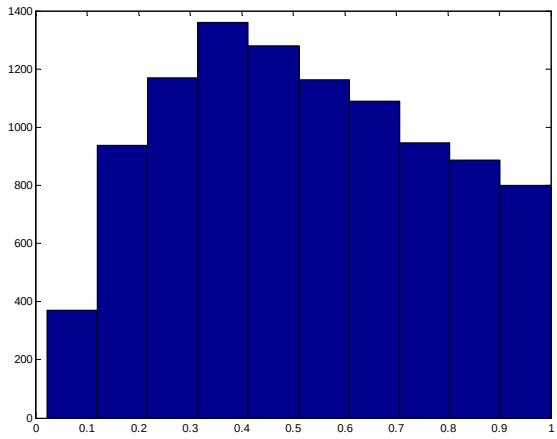


Figure 4. Simar and Wilson's hypothesized population



The results of this section indicate that for the cases examined the coverage probabilities do not necessarily converge towards their nominal probabilities. However, this does not necessarily imply that the bootstrap DEA is inconsistent, since the results from moment comparison are quite satisfactory. This is in accordance with the suggestion of Tziogkidis (2012) that using coverage probabilities to evaluate the performance of bootstrap DEA might not be appropriate. This point is supported by the different results we obtained compared to previous studies, which could be due to the different DGPs used but there could also be a theoretical justification which needs to be further explored.

6 Conclusions

The performance of bootstrap DEA is evaluated in this study using a different approach compared to the few previous studies. In particular, we compare moments to assess the performance of both the smooth and the "naïve" bootstrap instead of using coverage probabilities, which have been argued to be an invalid means of assessment (Tziogkidis, 2012).

Our results suggest that both smooth and naïve bootstrap procedures are always consistent. However, in the presence of model specification biases hypothesis testing should not be applied, especially if it involves different samples. Overall, the naïve bootstrap seems to be associated with smaller bootstrap biases and it is therefore preferred to the smooth one which requires a thorough evaluation to appropriately choose the smoothing parameter. Moreover,

our results support the arguments of Tziogkidis (2012) that the bootstrap bias can be quite different, in which cases the bootstrap bias corrected scores of Simar and Wilson (1998) as well as the confidence intervals in Simar and Wilson (2000) will only be consistent asymptotically.

As a further exercise we calculated coverage probabilities and we compared our results with these of previous studies. The findings of our research are not in accordance with previous simulation exercises, which might be attributed to differences in the assumed DGPs or to other theoretically oriented reasons which need to be explored in the future.

References

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

Kneip A., Park U.B Simar L., (1998). "A note on the convergence of nonparametric DEA estimators for production efficiency scores", *Econometric Theory*, Vol. 14, pp. 783–793

Korostelev A., Simar L., Tsybakov A.B, (1995). "Efficient estimation of monotone boundaries", *The Annals of Statistics*, Vol. 23, pp. 476–489

Loader R. C., (1999). "Bandwidth selection: Classical or plug-in?", *The Annals of Statistics*, Vol. 27, No. 2, pp. 415-438

Löthgren M., (1998). "How to bootstrap DEA estimators: a Monte Carlo comparison", *Working Paper Series in Finance and Economics*, No. 223, Stockholm School of Economics

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

Tziogkidis P., (2012). "Bootstrap DEA and hypothesis testing", *Cardiff Economics Working Paper Series*, Paper E2012/18