Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP01/2020

Improving Finite Sample Approximation by Central Limit Theorems for Estimates from Data Envelopment Analysis

Léopold Simar, Valentin Zelenyuk

Date: January 2020

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

Improving Finite Sample Approximation by Central Limit Theorems for Estimates from Data Envelopment Analysis*

Léopold Simar† Valentin Zelenyuk¶

January 10, 2020

Abstract

We propose an improvement of the finite sample approximation of the central limit theorems (CLTs) that were recently derived for statistics involving production efficiency scores estimated via Data Envelopment Analysis (DEA) or Free Disposal Hull (FDH) approaches. The improvement is very easy to implement since it involves a simple correction of the variance estimator with an estimate of the bias of the already employed statistics without any additional computational burden and preserves the original asymptotic results such as consistency and asymptotic normality. The proposed approach persistently showed improvement in all the scenarios that we tried in various Monte-Carlo experiments, especially for relatively small samples or relatively large dimensions (measured by total number of inputs and outputs) of the underlying production model. This approach therefore is expected to produce more accurate estimates of confidence intervals of aggregates of individual efficiency scores in empirical research using DEA or FDH approaches and so must be valuable for practitioners. We also illustrate this method using a popular real data set to confirm that the difference in the estimated confidence intervals can be substantial. A step-by-step implementation algorithm of the proposed approach is included in the Appendix.

Key words: Data Envelopment Analysis, DEA; Free Disposal Hull, FDH; Statistical Inference; Production Efficiency; Productivity.

*This is a substantially revised version of CEPA WP 07/2018.

†Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Voie du Roman Pays 20, B1348 Louvain-la-Neuve, Belgium. leopold.simar@uclouvain.be

¶School of Economics and Centre for Efficiency and Productivity Analysis (CEPA), University of Queensland, Colin Clark Building (39), St Lucia, Brisbane, Qld 4072, Australia. v.zelenyuk@uq.edu.au

1 Introduction

Nonparametric envelopment estimators are frequently used by applied researchers in efficiency and productivity analysis to facilitate measurement of producers' performances. These estimators are obtained by enveloping the clouds of points for an observed sample of input and output levels of firms, and thus estimating the set of technologically feasible allocations of inputs and outputs, also referred to as the production possibility set or, simply, the attainable set. A distance from an input-output allocation to the boundary of the (estimated) attainable set gives (an estimate of) technical or production efficiency of that allocation.

The most popular estimators of this type are known as data envelopment analysis (DEA) if we assume the convexity of the set of attainable combinations of inputs and outputs (see Farrell, 1957 and Charnes, Cooper and Rhodes, 1978) and the free disposal hull (FDH) estimators that do not impose convexity, but only free disposability assumption on inputs and outputs (see Deprins, Simar and Tulkens, 1984).[1]

Both DEA and FDH estimators found their use and became popular in many areas of research from both theoretical and applied angles. They attracted significant attention in management and operations literature (e.g., see Boussofiane, Dyson and Thanassoulis (1991) and Cook and Seiford (2009) and references therein). Among the key conclusions from the statistical literature is that under certain regularity conditions, these estimators are consistent and have complicated limiting distributions, which are different depending on the dimension of the production model (number of inputs and outputs), type of returns to scale, and whether convexity is imposed or not. Moreover, inference for certain measures of efficiency of a particular fixed point (e.g., an observed point of an individual firm) in the attainable set is available by using appropriate bootstrap techniques; see Kneip, Simar and Wilson (2008, 2011) and Simar and Wilson (2011) for the DEA case and Jeong and Simar (2006) for the FDH case.[2]

On the other hand, in many situations, the applied researchers are interested to compare means of group(s) of firms, or they might be interested in testing the hypotheses about the attainable set (convexity, returns to scale, etc.) or in dynamic settings, they might be interested to analyze changes in productivity over time, etc. In all these cases, the inference will be based on statistics which are appropriate aggregating functions of the individual efficiency scores over the sample. A recent work of Kneip, Simar and Wilson (2015) (hereafter KSW) has shown that inference on simple sample means of estimated scores is problematic unless the dimension of the problem (the total number of inputs and outputs) is very small

[1] Also see Afriat (1972).

[2] Also see Dyson et al. (2001) for the discussions on the advantages, limitations and suggested protocols for DEA.

(e.g. as small as 1 for the FDH case and 2 for the most common DEA estimates under assumption of variable returns to scale). The basic problem comes from the fact that the DEA/FDH efficiency estimates have a bias that does not vanish fast enough when the sample size increases. However, KSW provide central limit theorems (CLT) based on a bias correction method. The important result is that, when using the appropriate statistics, regular quantiles of the standard normal can be used to derive confidence intervals or to compute p-values.

These basic results have been used in KSW for providing confidence intervals on the mean of efficiency scores in a group, and have been adapted in various testing situations in Kneip, Simar and Wilson (2016).[3] It was then extended to inference for aggregate efficiency in Simar and Zelenyuk (2018), and also to conditional measures of efficiency, with a test of the separability condition in Daraio, Simar and Wilson (2018). Recently similar CLT results have been derived for Malmquist indices (Kneip, Simar and Wilson, 2018) and for cost and allocative efficiency in Simar and Wilson (2018).

In all these papers, the authors show with extensive Monte-Carlo experiments that the CLT can be applied in many practical situations and that the results behave as expected: improving the accuracy of the normal approximations (size and power of the tests, or coverage of the resulting confidence intervals) when the sample size increases, but the accuracy may reduce when the dimension of the problem increases. In some situations, these results might be disappointing for the practitioner.

In this paper we suggest an "easy to compute" correction of the used statistics, that does not require any additional computational burden and that improves the accuracy of the normal approximation provided by the corresponding CLT. In a nutshell, our approach involves a simple correction of the variance estimator with an estimate of the bias of the already computed statistics so that it preserves the original asymptotic results such as consistency and asymptotic normality. We will describe in more details how the correction is motivated and derived for the simplest case of average of efficiency scores (which cover all the extensions mentioned above), and then for the case of aggregate efficiency, where the correction requires more care. We illustrate how this works in practice in these two cases, with several Monte-Carlo experiments (different dimensions and sample sizes) and with a real data set. In some cases, the improvements may be substantial.

The paper is organized as follows. The next section introduces the basic notions and notations and summarizes the background (i.e. the results from KSW). Section 3 gives the basic idea of our corrected statistics in the simplest case of a simple average of efficiency scores

[3]In this paper we will focus on the case of one group of individuals (e.g., firms) that share or are benchmarked with respect to the same technology.

then Section 4 shows how to adapt this basic idea to other situations, with some emphasis on the case of aggregate efficiency. Section 5 indicates through Monte-Carlo experiments how this works in practice. Section 6 illustrates the proposed method with a real data set. Section 7 summarizes the concluding remarks. Finally, the practical implementation algorithm of the proposed approach is summarized in the Appendix.

2 Theoretical Background

2.1 Key Definitions

Assume that each firm uses p inputs $x \in \mathbb{R}^p_+$ to produce a vector of q outputs, denoted by $y \in \mathbb{R}^q_+$ and that the technology set, Ψ is defined as[4]

$$\Psi = \{(x, y) \in \mathbb{R}^{p+q}_+ \mid x \text{ can produce } y\}, \tag{2.1}$$

which gives the set of combinations of inputs and outputs that are technologically feasible. The *technology*, or *efficient frontier* of Ψ, is given by

$$\Psi^\partial = \left\{(x, y) \in \Psi \mid \left(\gamma^{-1}x, \gamma y\right) \notin \Psi \text{ for all } \gamma > 1\right\}. \tag{2.2}$$

We assume the usual regularity conditions on technology in production theory (e.g., see Färe and Primont, 1995). Specifically, the key assumptions we need are:

Assumption 2.1 *Ψ is closed and Ψ^∂ exists.*

Assumption 2.2 *All outputs and all inputs are strongly disposable, i.e., $(x_o, y_o) \in \Psi \Rightarrow (x, y) \in \Psi \ \forall x \geqq x_o, y \leqq y_o$.*

For the sake of brevity we will focus on the case when efficiency is measured via the so-called Farrell–Debreu output oriented measure of technical (in)efficiency (similar results can be derived for other orientations). For a firm with an input-output allocation (x, y), this (in)efficiency is defined as

$$\lambda(x, y) = \sup\{\theta > 0 \mid (x, \theta y) \in \Psi\}. \tag{2.3}$$

Intuitively, (2.3) gives the maximum real number (equal or greater than unity) by which one can multiply the output vector y to expand or contract it (radially) to the boundary of

[4]We use $\mathbb{R}^p_+$ and $\mathbb{R}^p_{++}$ to denote the non-negative and strictly positive real spaces of dimension p, respectively.

Ψ, while keeping the same level of input x and the same technology. Thus, for any feasible input output allocation (i.e., $\forall (x,y) \in \Psi$) we have $\lambda(x,y) \geq 1$, while its reciprocal gives an efficiency level measured between 0 and 1. While this measure is defined for all points in $\mathbb{R}^p_+ \times \mathbb{R}^q_+$, for simplicity and practical considerations we also assume away the so-called singularity points, such as $(x, \mathbf{0}_q)$, $(\mathbf{0}_p, y)$ and $(\mathbf{0}_p, \mathbf{0}_q)$, where $\mathbf{0}_a$ is the origin of $\mathbb{R}^a_+$, which often lead to computational problems.

Many methods have been developed in the literature to estimate the technology frontier Ψ^∂ and the related values of efficiency scores, $\lambda(x,y)$ from a sample of n *iid* observations on inputs and outputs $\mathcal{X}_n = \{(X_i, Y_i) \mid i = 1, \ldots, n\}$. Here we will focus on the DEA and FDH nonparametric and consistent estimators that envelop the data under certain constraints on technology. Specifically, if one wishes to assume Constant Returns to Scale (CRS) for Ψ,[5] then the CRS-DEA estimator is given by

$$\widehat{\lambda}(x,y \mid \mathcal{X}_n) \equiv \max_{\xi_1,\ldots,\xi_n,\theta} \{\theta \mid \sum_{k=1}^n \xi_k Y_k \geqq \theta\, y,\ \sum_{k=1}^n \xi_k X_k \leqq x,\ \theta \geq 0, \xi_k \geq 0\}, \tag{2.4}$$

inspired by Farrell (1957) and popularized as a linear program (LP) by Charnes, Cooper and Rhodes (1978).[6] Alternatively, if one wishes to assume Variable Returns to Scale (VRS) for Ψ, then the VRS-DEA estimator is given by

$$\widehat{\lambda}(x,y \mid \mathcal{X}_n) \equiv \max_{\xi_1,\ldots,\xi_n,\theta} \{\theta \mid \sum_{k=1}^n \xi_k Y_k \geqq \theta\, y,\ \sum_{k=1}^n \xi_k X_k \leqq x, \\ \theta \geq 0,\ \xi_k \geq 0,\ \sum_{k=1}^n \xi_k = 1\}, \tag{2.5}$$

i.e., constraint $\sum_{k=1}^n \xi_k = 1$ is added to the CRS-DEA specification (2.4).

If convexity of Ψ is not assumed, but only strong disposability of inputs and outputs, then one can use the FDH estimator, introduced by Deprins, Simar and Tulkens (1984):

$$\widehat{\lambda}(x,y \mid \mathcal{X}_n) \equiv \max_{\theta} \{\theta \mid Y_k \geqq \theta\, y,\ X_k \leqq x,\ k = 1, \ldots, n,\ \theta \geq 0\}. \tag{2.6}$$

[5]To be precise, technology Ψ exhibits CRS if and only if $\delta\Psi = \Psi,\ \forall \delta > 0$, which sometimes is referred to as global CRS—the notion we will use here. This is different from the cases of the so-called local CRS, where this or an analogous property holds in some neighborhood and, for example measured by the scale elasticity (e.g., see Zelenyuk (2013) for the precise definitions in the DEA context and related references), which we do not consider here.

[6]To be precise, both papers focused on the input orientation, which here (due to CRS) is the reciprocal of the output orientation problem (2.4). See Sickles and Zelenyuk (2019) for more discussion and references on DEA and related methods.

2.2 Key results from KSW.

As pointed out above, under certain regularity conditions, the asymptotic statistical properties of DEA and FDH estimators for $\lambda(x,y)$ have been established when evaluated at fixed points (x,y), and bootstrap techniques have to be used for practical inference due to the complex nature of the asymptotic distribution of the estimators.[7]

When confidence intervals for means or testing issues about the shape of Ψ^{∂} are concerned, the practitioner needs the statistical properties of some statistics built from the estimators of efficiency scores at the random points. The simplest statistic one may consider is the mean of the efficiency scores to draw inference on μ_λ, the average efficiency score in the population

$$\mu_\lambda = \mathrm{E}\big(\lambda(X,Y)\big), \tag{2.7}$$

where the expectation is over the random variables (X,Y). Below we will also need to estimate the variance σ^2_λ of the efficiency scores in the population:

$$\sigma^2_\lambda = \mathrm{Var}\big(\lambda(X,Y)\big). \tag{2.8}$$

Inference on μ_λ could be obtained from the sample mean of the estimated efficiency scores:

$$\overline{\lambda}_n = n^{-1}\sum_{i=1}^{n}\widehat{\lambda}(X_i,Y_i \mid \mathcal{X}_n), \tag{2.9}$$

where one of the nonparametric estimators above would be evaluated at the sample points (X_i,Y_i).[8] However, the basic results in KSW indicate that this is not so easy and that a simple CLT is not directly available as we now briefly explain.

Specifically, even the inference on the simple sample means of estimated scores is problematic except for cases when the dimension of the problem (the number of inputs and outputs) is very small: e.g., 1 for the FDH case and 2 for the most common VRS-DEA case.

In a nutshell, the problem comes from the fact that the DEA/FDH efficiency estimates have a bias that does not disappear fast enough with the increase in the sample size for

[7] See Simar and Wilson (2015) for more details, a comprehensive discussion and more references. It is worth noting that among others, a key assumption is that $\Pr\{(X_i,Y_i)\in\Psi\}=1,\ \forall i=1,\ldots,n$, i.e., all observations are feasible with respect to the technology Ψ and that there is no noise. If there is noise in the data, then a version of Stochastic DEA or Stochastic FDH (e.g., see Simar and Zelenyuk (2011)) can be used to filter out the noise or outliers.

[8] Note a slight complication of notation here: $\lambda(X,Y)$ and $\lambda(X_i,Y_i)$ mean that the measure that was defined in (2.3) for any point (x,y) is evaluated at a random point (X,Y) and its i^{th} realization (X_i,Y_i), respectively. Meanwhile, $\widehat{\lambda}(X_i,Y_i \mid \mathcal{X}_n)$ denotes the estimate of $\lambda(X_i,Y_i)$, which was defined in (2.4)–(2.6) for any point (x,y), here evaluated at the i^{th} realization (X_i,Y_i), where the estimation involved all the sample $\mathcal{X}_n$, rather than a subsample of it, as will also be used later.

the estimation. To mitigate the problem, KSW developed new central limit theorems that involve the bias correction method (using a generalized jackknife) and, if that is not enough (depending on the chosen estimator and on the dimension of the problem), using means of the efficiency estimates on a subsample of points. As a result, when using the appropriate statistics, regular quantiles of the standard normal can be used to derive confidence intervals or to compute p-values. In particular, KSW used these results for providing confidence intervals for the mean of efficiency scores of a group, and provided a foundation for developing the related results in various testing situations.

To be more precise, the key results from KSW can be summarized as follows: for the FDH and DEA estimators, under some regularity conditions and as $n \to \infty$,

$$\mathrm{E}\left(\widehat{\lambda}(X_i, Y_i \mid \mathcal{X}_n) - \lambda(X_i, Y_i)\right) = Cn^{-\kappa} + R_{n,\kappa} \tag{2.10}$$

$$\mathrm{E}\left(\left(\widehat{\lambda}(X_i, Y_i \mid \mathcal{X}_n) - \lambda(X_i, Y_i)\right)^2\right) = o\big(n^{-\kappa}\big), \tag{2.11}$$

$$\left|\mathrm{Cov}\left(\widehat{\lambda}(X_i, Y_i \mid \mathcal{X}_n) - \lambda(X_i, Y_i),\ \widehat{\lambda}(X_j, Y_j \mid \mathcal{X}_n) - \lambda(X_j, Y_j)\right)\right| = o\big(n^{-1}\big) \tag{2.12}$$

for some finite constant C and for all $i, j \in \{1,\ \ldots,\ n\}$, $i \neq j$ and where $R_{n,\kappa}$ is the remainder term of the series approximation, which is of order $o\left(n^{-\kappa}\right)$ or smaller. It is worth noting that $R_{n,\kappa}$ does not affect the asymptotic results of interest here, although specific rates of it are available in KSW.[9]

Importantly, note that the values of the constant C, the rate κ, and the remainder term $R_{n,\kappa}$ depend on which estimator is used. Specifically:

- when we assume CRS, the CRS-DEA estimator (2.4) achieve the rate $\kappa = 2/(p+q)$.
- when we assume VRS, the VRS-DEA estimator (2.5) achieves the rate $\kappa = 2/(p+q+1)$.
- when we do not assume convexity and maintain only the strong disposability assumption, the appropriate estimator is the FDH estimator (2.6). It achieves the rate $\kappa = 1/(p+q)$.

Using these results KSW showed that under certain regularity assumptions, as $n \to \infty$,

$$\sqrt{n}\big(\overline{\lambda}_n - \mu_\lambda - Cn^{-\kappa} + o(n^{-\kappa})\big) \xrightarrow{\mathcal{L}} N(0, \sigma_\lambda^2). \tag{2.13}$$

[9]To avoid confusion we remind the reader of the basic notations for the rates of convergence (see e.g. Serfling, 1980, Sections 1.1.2 and 1.2.5). For a non-stochastic sequence of real numbers $\{c_n\}$, we write $c_n = O\big(n^{-\kappa}\big)$ as $n \to \infty$, when $|n^\kappa c_n|$ remains bounded as $n \to \infty$. We write $c_n = o\big(n^{-\kappa}\big)$ when $\lim_{n\to\infty} n^\kappa c_n = 0$. Similarly a sequence of random variables $\{\xi_n\}$ is said to be bounded in probability if, for every $\varepsilon > 0$, there exist M_ε and a n_ε such that for all $n > n_\varepsilon$, $\Pr(\xi_n > M_\varepsilon) < \varepsilon$. The notation $\xi_n = O_P(1)$ will be used. Then $\xi_n = O_P\big(n^{-\kappa}\big)$ means that $n^\kappa \xi_n = O_P(1)$. Further the notation $\xi_n = o_P\big(n^{-\kappa}\big)$ will be used when $n^\kappa \xi_n$ converges in probability to zero as $n \to \infty$.

This leads to a problem: when $\kappa \leq 1/2$ (e.g., for a CRS-DEA this happens already when $p+q>3$ for VRS-DEA when $p+q>2$ and for FDH when $p+q>1$!), the bias does not converge to zero fast enough with the increase in the sample size, thus distorting the standard CLT results, e.g., by centering the corresponding confidence intervals at a wrong place, making such confidence intervals unreliable.

The solution suggested by KSW is to correct for the leading term in the inherent bias of DEA and FDH, $Cn^{-\kappa}$, via its consistent estimator, defined as

$$\widehat{B}_{n,\kappa} = (2^{\kappa}-1)^{-1}(\overline{\lambda}^{*}_{n/2} - \overline{\lambda}_n), \tag{2.14}$$

where $\overline{\lambda}^{*}_{n/2}$ is a generalized jackknife analog of $\overline{\lambda}_n$, obtained as

$$\overline{\lambda}^{*}_{n/2} = \big(\overline{\lambda}^{(1)}_{n/2} + \overline{\lambda}^{(2)}_{n/2}\big)/2, \tag{2.15}$$

where $\overline{\lambda}^{(1)}_{n/2}$ is a version of $\overline{\lambda}_n$ based on a random subset of $\mathcal{X}_n$ of size $n/2$ and $\overline{\lambda}^{(2)}_{n/2}$ is the analog based on the remaining part of the sample (for simplicity assume n is even). That is, formally, for $\ell = 1, 2$, one computes

$$\overline{\lambda}^{(\ell)}_{n/2} = 2n^{-1} \sum_{\{i|(X_i,Y_i)\in\mathcal{X}^{(\ell)}_{n/2}\}} \widehat{\lambda}(X_i, Y_i \mid \mathcal{X}^{(\ell)}_{n/2}),$$

where $\mathcal{X}^{(1)}_{n/2} \cap \mathcal{X}^{(2)}_{n/2} = \emptyset$ and $\mathcal{X}^{(1)}_{n/2} \cup \mathcal{X}^{(2)}_{n/2} = \mathcal{X}_n$.

The intuitive idea behind the bias correction in (2.14) can be summarized as follows (for details and proofs, see KSW). Equation (2.13) indicates that, asymptotically, the leading term of the bias when using $\bar{\lambda}_n$ for estimating μ_λ is driven by $Cn^{-\kappa}$, where C is a constant. So if we use half of the full sample, i.e. of size $n/2$, equation (2.13) indicates that asymptotically, the leading term of the bias when using $\bar{\lambda}^{(\ell)}_{n/2}$ for estimating μ_λ is driven by $C2^{\kappa}n^{-\kappa}$ for the same constant C and for $\ell = 1, 2$. Then by simple algebra (see KSW for details), it is clear that $\widehat{B}_{n,\kappa}$ will approximate the bias term and what is important is that the error described in equation (2.16) has the order of magnitude smaller than or equal to the remainder $o(n^{-\kappa})$ in (2.14).[10]

KSW also show that

$$\widehat{B}_{n,\kappa} = Cn^{-\kappa} + R'_{n,\kappa} + o_P\big(n^{-1/2}\big), \tag{2.16}$$

where $R'_{n,\kappa}$ has the same order of magnitude as the original $R_{n,\kappa}$ introduced above, thus

[10] As suggested in Kneip et al. (2016), this operation of a random split in two parts is repeated a large number, L, of times by shuffling the observations in $\mathcal{X}_n$ before each split, providing $\widehat{B}^{(l)}_{n,\kappa}$ for $l = 1, \ldots, L$. Then to reduce the variance of the bias estimator we use $\widehat{B}_{n,\kappa} = L^{-1}\sum_{l=1}^{L} \widehat{B}^{(l)}_{n,\kappa}$.

making it a suitable correction of the bias.

With these results KSW obtain the following new CLTs.

Theorem 2.1 *Under assumptions for Theorem 3.1, 3.2 or 3.3 of Kneip et al. (2015), for $p+q \leq 5$ if a CRS-DEA estimator is used and Ψ exhibits CRS and is convex, for $p+q \leq 4$ if a VRS-DEA estimator is used and Ψ is convex, for $p+q \leq 3$ if an FDH estimator is used and Ψ satisfies free disposability of inputs and outputs, when $n \to \infty$ we have*

$$\sqrt{n}\left(\overline{\lambda}_n - \widehat{B}_{n,\kappa} - \mu_\lambda + R_{n,\kappa}\right) \xrightarrow{\mathcal{L}} N(0, \sigma_\lambda^2), \tag{2.17}$$

and when $\kappa < 1/2$ then, as $n \to \infty$ we have

$$\sqrt{n_\kappa}\left(\overline{\lambda}_{n_\kappa} - \widehat{B}_{n,\kappa} - \mu_\lambda + R_{n,\kappa}\right) \xrightarrow{\mathcal{L}} N(0, \sigma_\lambda^2), \tag{2.18}$$

with $\overline{\lambda}_{n_\kappa}$ being a subsample version of $\overline{\lambda}_n$, where the averaging is taken over a random subsample $\mathcal{X}^*_{n_\kappa} \subset \mathcal{X}_n$ of size $n_\kappa = \lfloor n^{2\kappa} \rfloor < n$. Formally

$$\overline{\lambda}_{n_\kappa} = n_\kappa^{-1} \sum_{\{j|(X_j,Y_j)\in\mathcal{X}^*_{n_\kappa}\}} \widehat{\lambda}(X_j, Y_j \mid \mathcal{X}_n). \tag{2.19}$$

Note that the bounds on $p+q$ given for being able to apply (2.17) are due to the particular definitions of $R_{n,\kappa}$ for each estimator (see KSW for details).[11] It is also worth noting that for $p+q=5$ in the CRS case, $p+q=4$ in the VRS case and $p+q=3$ for the FDH case, both versions of the CLT are applicable. As commented in KSW, looking to the specific form of the remainder term, the second version of the CLT in these limiting cases is preferable, because the neglected terms are of small order $O(\cdot)$. This is confirmed by the Monte-Carlo experiments (see KSW and our simulation results below for details).

Of course, in practice the σ_λ^2 that appears in the theorem is not observed and for practical inference it can be replaced by a consistent estimator. KSW suggest the following one

$$\widehat{\sigma}^2_{\lambda,n} = \frac{1}{n}\sum_{j=1}^{n}(\widehat{\lambda}(X_j, Y_j|\mathcal{X}_n) - \overline{\lambda}_n)^2, \tag{2.20}$$

thus providing, when (2.17) can be applied, an asymptotically correct $(1-\alpha)$ confidence interval for μ_λ given by

$$\left[\overline{\lambda}_n - \widehat{B}_{n,\kappa} \pm z_{1-\alpha/2}\widehat{\sigma}_{\lambda,n}/\sqrt{n}\right], \tag{2.21}$$

where $z_{1-\alpha/2}$ is the corresponding quantile of the standard normal distribution. When the

[11] Again, we note that in all the cases, since $R_{n,\kappa} = o(n^{-\kappa})$, the remainder term can be neglected here.

dimension $p+q$ increases and $\kappa < 1/2$, an asymptotically correct $(1-\alpha)$ confidence interval for μ_λ is given by

$$\left[\overline{\lambda}_{n_\kappa} - \widehat{B}_{n,\kappa} \pm z_{1-\alpha/2}\widehat{\sigma}_{\lambda,n}/\sqrt{n_\kappa}\right]. \tag{2.22}$$

KSW also present some Monte-Carlo evidence, supporting the theory, confirming poor performance of the standard CLT in this case and substantially better performance by the CLTs they developed. Our goal in the next section is to try to improve the finite sample approximations of these new CLTs.

3 Improving Approximations

As pointed out above, for practical inference, one needs a consistent estimator of σ^2_λ, the variance of the efficiencies in the population. We know that

$$\sigma^2_\lambda = \mathrm{E}(\lambda(X,Y) - \mu_\lambda)^2$$

and KSW have shown that $\widehat{\sigma}^2_{\lambda,n}$ is such an estimator. Looking to this choice, $\widehat{\sigma}^2_{\lambda,n}$ appears to be an empirical version of $\mathrm{Var}\big(\widehat{\lambda}(X,Y \mid \mathcal{X}_n)\big)$. However, note that, by construction, we know that this estimator underestimates $\sigma^2_\lambda = \mathrm{Var}\big(\lambda(X,Y)\big)$. Indeed,

$$\lambda(X,Y) = \sup\{\theta \mid (X,\theta Y) \in \Psi\}$$

whereas

$$\widehat{\lambda}(X,Y \mid \mathcal{X}_n) = \sup\{\theta \mid (X,\theta Y) \in \widehat{\Psi}\},$$

where, according to the chosen estimator, $\widehat{\Psi}$ is the conical convex hull (a CRS-DEA case), or the convex hull (a VRS-DEA case) or the free disposal hull (an FDH case) of the clouds of points $\mathcal{X}_n$, implicitly defined in the expressions (2.4) to (2.6) above. However, note that in all the cases, we have

$$\widehat{\Psi} \subset \Psi$$

implying, with probability one, that

$$1 \leq \widehat{\lambda}(X,Y \mid \mathcal{X}_n) \leq \lambda(X,Y),$$

implying that the random variables $\widehat{\lambda}(X,Y \mid \mathcal{X}_n)$ are likely to have lower variance than the random variables they try to predict, $\lambda(X,Y)$.[12] Note also that, in practice, there are many

[12]Similar inequalities hold for the input efficiency scores, where $0 < \theta(X,Y) \leq \widehat{\theta}(X,Y \mid \mathcal{X}_n) \leq 1$.

instances when $\widehat{\lambda}(X_i, Y_i \mid \mathcal{X}_n) = 1$, which is likely to reduce the variability of the estimators even more, relative to the variability of the unobserved random variables they try to predict. This is a part of the source of the lower coverages than the nominal levels, that we observed in all the Monte-Carlo experiments.

Our goal therefore is to provide and explore another consistent estimator of σ_λ^2 that should at least partially correct for this disappointing property. The motivation for the new variance estimator goes along the following lines.

A natural estimator of σ_λ^2 should be its empirical version (if available), i.e.

$$\widetilde{\sigma}_\lambda^2 := \frac{1}{n}\sum_{i=1}^{n}(\lambda(X_i, Y_i) - \widetilde{\lambda}_n)^2$$

where

$$\widetilde{\lambda}_n := \frac{1}{n}\sum_{i=1}^{n}\lambda(X_i, Y_i)$$

but the problem of course is that $\lambda(X_i, Y_i)$ is unavailable and so $\widetilde{\lambda}_n$ and $\widetilde{\sigma}_\lambda^2$ are also unavailable–we only know from existing asymptotic theory (see KSW), that

$$\widetilde{\lambda}_n = \mu_\lambda + O_P\big(n^{-1/2}\big).$$

A natural approach (taken by KSW) is to replace the unavailable $\lambda(X_i, Y_i)$ by their consistent estimates $\widehat{\lambda}(X_i, Y_i \mid \mathcal{X}_n)$ and the sample mean $\widetilde{\lambda}_n$ by $\overline{\lambda}_n$, but we know (as a consequence of Lemma 4.1 in KSW) that

$$\overline{\lambda}_n = \mu_\lambda + B_{n,\kappa} + o(n^{-\kappa})$$

where the leading term for the bias is given by

$$B_{n,\kappa} = Cn^{-\kappa}$$

as described above. Now, by (2.16) we have a consistent estimator of this bias term which is already computed for deriving the CLT. So, we suggest replacing the unavailable $\widetilde{\lambda}_n$ by its bias corrected version, defined as

$$\widetilde{\lambda}_n^{\mathrm{bc}} := \overline{\lambda}_n - \widehat{B}_{n,\kappa}$$

(instead of $\overline{\lambda}_n$ as in KSW).

So, in the end, the new consistent estimator of σ^2_λ we propose is given by

$$\begin{aligned}\widehat{\widehat{\sigma}}^2_{\lambda,n} &= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\lambda}(X_i, Y_i \mid \mathcal{X}_n) - \widetilde{\lambda}^{\text{bc}}_n\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\lambda}(X_i, Y_i \mid \mathcal{X}_n) - \overline{\lambda}_n + \widehat{B}_{n,\kappa}\right)^2.\end{aligned} \tag{3.1}$$

It is easy to check, using (2.16), that

$$\widetilde{\lambda}^{\text{bc}}_n = \mu_\lambda + o(n^{-\kappa}) + o_P(n^{-1/2}), \tag{3.2}$$

meaning that this correction will preserve the existing asymptotic properties (consistency and asymptotic normality) derived by KSW, when $\widehat{\widehat{\sigma}}^2_{\lambda,n}$ is used in place of unknown σ^2_λ, and has a chance of improving its approximation in practice, i.e., for finite samples (simulated or real data), relative to cases when $\widehat{\sigma}^2_{\lambda,n}$ is used.

The resulting estimated $(1-\alpha)$ confidence intervals for μ_λ will therefore be

$$\left[\overline{\lambda}_n - \widehat{B}_{n,\kappa} \pm z_{1-\alpha/2}\widehat{\widehat{\sigma}}_{\lambda,n}/\sqrt{n}\right], \tag{3.3}$$

when (2.17) can be applied, and

$$\left[\overline{\lambda}_{n_\kappa} - \widehat{B}_{n,\kappa} \pm z_{1-\alpha/2}\widehat{\widehat{\sigma}}_{\lambda,n}/\sqrt{n_\kappa}\right] \tag{3.4}$$

when (2.18) can be applied, and where, as before, $z_{1-\alpha/2}$ is the corresponding quantile of the standard normal distribution.

Also note that after some elementary manipulations we can simplify the expression for the new variance estimator to be

$$\widehat{\widehat{\sigma}}^2_{\lambda,n} = \widehat{\sigma}^2_{\lambda,n} + \widehat{B}^2_{n,\kappa}, \tag{3.5}$$

and so one can immediately see that, as expected, we increase slightly the original KSW estimator ($\widehat{\sigma}^2_{\lambda,n}$) with a positive term, $\widehat{B}^2_{n,\kappa}$, which is of a small order, $O_P\big(n^{-2\kappa}\big)$. So, as $n \to \infty$, this correction does not play any role asymptotically, but in finite samples the bias may be so big that the difference is significant (in particular when κ decreases, i.e. when the dimension $p+q$ increases). We will investigate the practical consequences in some Monte-Carlo experiments, confirming the usefulness of this correction, even with sample sizes as large as $n = 1000$. Meanwhile, it might be worth noting that the theoretical developments outlined above, in (3.1)–(3.5) are, to the best of our knowledge, novel to the literature.

4 Implications to Other Results

4.1 Some Obvious Extensions

The basic results of KSW have been extended in various situations. Kneip et al. (2016) indicate how to use these CLTs for testing the equality of the means of two groups of firms, for testing the convexity of the attainable set and for testing returns to scale of the technology. All these tests are built on a statistic which is the difference of two sample means of efficiency scores computed on two independent random samples. Then the above CLTs can be used for both sample means, corrected for the bias term and rescaled by an estimator of the variance of the efficiency scores. It is clear that the accuracy of the procedures will be improved by using in each case the corrected estimator defined in (3.5), adapted to each particular case.

The same applies to the inference suggested in Kneip et al. (2018), where CLTs are derived for Malmquist indices, for the CLTs for conditional efficiency scores derived in Daraio et al. (2018) (and for their test of the separability condition), and all the CLTs derived in Simar and Wilson (2018) for various cost, revenue and allocative efficiencies. All the CLTs there have a similar shape to the ones derived in KSW (and summarized above) with a bias correction factor and a rescaling by an estimator of the variance of the efficiency scores.

Another natural application of our developments is to the context of the so-called 'two-stage approach', where at the first stage one obtains DEA estimates and at the second stage they are regressed on some variables that are believed to be influencing the inefficiency. Developing new CLTs that account for the bias of estimates of dependent variables in this approach has been analyzed in some details in KSW (see their section 5), where it was proven that under the so-called 'separability condition' (see their Assumption 5.1), the CLTs are also available for the second-stage regression (i.e., conditional mean estimation). For our context, it is worth noting that their Theorem 5.1 also provided a consistent estimator of the variance of the error term in a simple OLS framework. Clearly, the procedure suggested in our paper (and illustrated for the context of unconditional means) can be relatively easily adapted to improve their estimator (for the context of conditional means). The extension for a more general second-stage approach (involving MLE-based truncated regression rather than OLS) would be also similar and is one of natural future research directions.

To save space we do not give the details of all these extensions because they are rather obvious to derive. The derivations are a bit more tedious for aggregate efficiencies as explained in the next section.

4.2 Aggregate Efficiency

Simar and Zelenyuk (2018) (hereafter SZ) adapt and generalize the results from KSW to the case of CLT for aggregate or weighted efficiency (e.g., such as industry efficiency), which can be defined in terms of a ratio of two true means. They focus their presentation on the output aggregate technical efficiencies of a group of firms derived in Färe and Zelenyuk (2003), which rest on certain assumptions on the aggregate technology and the law of one price $w \in \mathbb{R}^q_{++}$ on output markets.

Here it is worth noting that the assumption of common prices is not needed for the developments in SZ or in this paper. This assumption is dictated from the aggregation theory, as a necessary condition, in order to establish the equivalence of the aggregate revenue (defined on aggregate technology) with the sum of individual revenues (defined on the individual technologies). The common price can be understood as a market equilibrium price, which in turn enables the theoretical solution to be obtained, while in practical applications it can be replaced by an average price in the industry.

The output oriented aggregate technical efficiency over a group of n firms, can be written as

$$\tau_n = \sum_{i=1}^{n} \lambda(X_i, Y_i) S_i, \tag{4.1}$$

where $S_i = \frac{w^T Y_i^{\partial}}{\sum_{i=1}^{n} w^T Y_i}$ is the weight of observation i (derived using economic theory reasoning by Färe and Zelenyuk (2003)). The advantage of this aggregate efficiency measure is that it accounts for the economic weight of each observation i in the sample, which can be very important when the group consists of observations of different sizes (typically a case in practice). Because the sample mean of efficiency scores does not account for an economic weight of an observation being aggregated, the aggregate efficiency can suggest values, conclusions and policy implications that are very different from those suggested by the sample mean.

For example, consider a hypothetical case of an industry with 100 firms, where 99 firms are 100% efficient yet are very small, say representing together only 1% of the industry in terms of the market share, while suppose the remaining firm that controls the rest of the market is only 50% efficient. Here, the sample mean of the efficiency scores will suggest that the industry is nearly 100% efficient and this will grossly misrepresent the situation in this industry, almost completely dominated by a very inefficient firm. The weighted mean will show the efficiency of approximately 50% for this industry.

To develop their theory, SZ first show another (and equivalent) form of the aggregate efficiency, namely

$$\tau_n = \frac{n^{-1}\sum_{i=1}^n w^T Y_i^\partial}{n^{-1}\sum_{i=1}^n w^T Y_i}, \tag{4.2}$$

where Y_i^∂ is the counterfactual output vector for the firm i obtained by its radial projection on the efficient frontier Ψ^∂, i.e. $Y_i^\partial = \lambda(X_i, Y_i)Y_i$. That is, it is the ratio of the average (or total) industry revenue evaluated at technically efficient levels of output to the average (or total) industry revenue that was actually observed. To simplify further notation, we will denote $Z_i = w^T Y_i$ and $Z_i^\partial = \lambda(X_i, Y_i)Z_i$.

SZ consider the ratio of the true means, $\tau = \mu_1/\mu_2$, as the parameter of interest, where $\mu_1 = \mathrm{E}\big(Z_i^\partial\big)$ and $\mu_2 = \mathrm{E}\big(Z_i\big)$, and then by standard arguments SZ show that

$$\sqrt{n}\big(\tau_n - \tau\big) \xrightarrow{\mathcal{L}} N(0, \sigma_\tau^2), \tag{4.3}$$

where

$$\sigma_\tau^2 = \tau^2\left(\frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} - 2\frac{\sigma_{12}}{\mu_1\mu_2}\right), \tag{4.4}$$

and where $\sigma_1^2 = \mathrm{Var}(Z_i^\partial)$, $\sigma_2^2 = \mathrm{Var}(Z_i)$ and $\sigma_{12} = \mathrm{Cov}(Z_i^\partial, Z_i)$. Since τ_n is not available, we will estimate τ by using our nonparametric frontier estimators to obtain

$$\widehat{\tau}_n = \frac{\widehat{\mu}_{1,n}}{\widehat{\mu}_{2,n}} = \frac{(1/n)\sum_{i=1}^n \widehat{Z}_i^\partial}{(1/n)\sum_{i=1}^n Z_i}, \tag{4.5}$$

where $\widehat{Z}_i^\partial = \widehat{\lambda}(X_i, Y_i|\mathcal{X}_n)Z_i$. Extending the theory of KSW, they derive an estimator of the bias of $\widehat{\tau}_n$ and a consistent estimator of its variance providing the CLT which is summarized below.

The bias estimator deserves some details: the leading part of the bias is governed by the inherent bias of the numerator due to the nonparametric estimators of $\lambda(X_i, Y_i)$. As in KSW, the generalized jackknife is used for the estimation of the bias of the numerator only, i.e. $B_{\mu_{1,n},\kappa}$. By repeating the random splits in two parts L times (see above (2.14) and Footnote 10) we end up with the bias correction for $\widehat{\tau}_n$ given by

$$\widehat{B}_{\tau_n,\kappa} = \frac{\widehat{B}_{\mu_{1,n},\kappa}}{\widehat{\mu}_{2,n}} = \frac{L^{-1}\sum_{\ell=1}^L (2^\kappa - 1)^{-1}\big(\widehat{\mu}_{1,n,\ell}^* - \widehat{\mu}_{1,n}\big)}{\widehat{\mu}_{2,n}}, \tag{4.6}$$

where $\widehat{\mu}_{1,n,\ell}^*$ is the analog of $\overline{\lambda}_{n/2}^*$, as defined in (2.14) and (2.15), and where the use of $\widehat{\mu}_{2,n}$ at the denominator does not change the order of the approximation (see SZ for details). They obtain the following CLT.

Theorem 4.1 *Under the appropriate set of Assumptions described in Theorem 3.1, 3.2 or*

3.3 of Kneip et al. (2015), for $p+q \leq 5$ if a CRS-DEA estimator is used and Ψ is CRS and is convex, for $p+q \leq 4$ if VRS-DEA estimator is used and Ψ is convex, for $p+q \leq 3$ if FDH estimator is used and Ψ satisfies free disposability of inputs and outputs, when $n \to \infty$ we have

$$\sqrt{n}\ \left(\widehat{\tau}_n - \widehat{B}_{\tau_n,\kappa} - \tau + R_{n,\kappa}\right) \xrightarrow{\mathcal{L}} N(0, \sigma_\tau^2), \tag{4.7}$$

where $R_{n,\kappa} = o\big(n^{-\kappa}\big)$.

For $\kappa < 1/2$, then as $n \to \infty$ we have

$$\sqrt{n_\kappa}\ \left(\widehat{\tau}_{n_\kappa} - \widehat{B}_{\tau_n,\kappa} - \tau + R_{n,\kappa}\right) \xrightarrow{\mathcal{L}} N(0, \sigma_\tau^2), \tag{4.8}$$

with $\widehat{\tau}_{n_\kappa}$ being a subsample version of $\widehat{\tau}_n$, in the sense that the averages are taken over a random subsample $\mathcal{X}^*_{n_\kappa} \subset \mathcal{X}_n$ of size $n_\kappa = \lfloor n^{2\kappa} \rfloor < n$. Formally

$$\widehat{\tau}_{n_\kappa} = \frac{n_\kappa^{-1} \sum_{\{j|(X_j,Y_j)\in\mathcal{X}^*_{n_\kappa}\}} \widehat{Z}_j^\partial}{n_\kappa^{-1} \sum_{\{j|(X_j,Y_j)\in\mathcal{X}^*_{n_\kappa}\}} Z_j}, \tag{4.9}$$

where $\widehat{Z}_j^\partial = \widehat{\lambda}(X_j, Y_j|\mathcal{X}_n) Z_j$.

The remark after Theorem 2.1 above is still valid here. For practical inference, SZ propose to estimate σ_τ^2 by plugging estimators of its components in equation (4.4), i.e.

$$\widehat{\sigma}^2_{\tau,n} = \widehat{\tau}_n^2 \left[\frac{\widehat{\sigma}^2_{1,n}}{\widehat{\mu}^2_{1,n}} + \frac{\widehat{\sigma}^2_{2,n}}{\widehat{\mu}^2_{2,n}} - 2\,\frac{\widehat{\sigma}_{12,n}}{\widehat{\mu}_{1,n}\widehat{\mu}_{2,n}}\right] \tag{4.10}$$

where $\widehat{\sigma}^2_{1,n}$ and $\widehat{\sigma}^2_{2,n}$ are the empirical variances of $\widehat{Z}_i^\partial$ and Z_i, respectively, and $\widehat{\sigma}_{12,n}$ is the empirical covariance between $\widehat{Z}_i^\partial$ and Z_i. While SZ showed that this estimator is consistent, i.e., $\widehat{\sigma}^2_\tau \xrightarrow{p} \sigma^2_\tau$, they also showed that in the Monte-Carlo experiments, the estimated confidence intervals based on the CLT with this estimator of σ^2_τ were significantly under-covering the true values (similarly as in KSW), even for samples like $n = 1000$.

The resulting estimated $(1-\alpha)$ confidence intervals for τ can thus be obtained as

$$\left[\widehat{\tau}_n - \widehat{B}_{\tau_n,\kappa} \pm z_{1-\alpha/2}\widehat{\sigma}_{\tau,n}/\sqrt{n}\right], \tag{4.11}$$

when (4.7) can be applied, and

$$\left[\widehat{\tau}_{n_\kappa} - \widehat{B}_{\tau_n,\kappa} \pm z_{1-\alpha/2}\widehat{\sigma}_{\tau,n}/\sqrt{n_\kappa}\right] \tag{4.12}$$

when (4.8) can be applied.

Using similar logic as in the section above, an alternative estimator for σ_τ^2 can be suggested by correcting for the bias of $\widehat{\mu}_{1,n}$ in the estimator of the variance of the nonparametric estimators, i.e. $\widehat{\sigma}_{1,n}^2$ (note that this correction has no influence on $\widehat{\sigma}_{12,n}$). Following the same arguments as above we propose to use

$$\widehat{\widehat{\sigma}}_{1,n}^2 = \widehat{\sigma}_{1,n}^2 + \widehat{B}_{\mu_{1,n},\kappa}^2, \tag{4.13}$$

where $\widehat{B}_{\mu_{1,n},\kappa}^2$ is defined above as the numerator of (4.6). So we have

$$\widehat{\widehat{\sigma}}_{\tau,n}^2 = \widehat{\tau}_n^2 \left[\frac{\widehat{\widehat{\sigma}}_{1,n}^2}{\widehat{\mu}_{1,n}^2} + \frac{\widehat{\sigma}_{2,n}^2}{\widehat{\mu}_{2,n}^2} - 2\,\frac{\widehat{\sigma}_{12,n}}{\widehat{\mu}_{1,n}\widehat{\mu}_{2,n}} \right], \tag{4.14}$$

giving an alternative consistent estimator of the variance σ_τ^2 with larger values as expected.

Thus, the corrected estimates of $(1-\alpha)$ confidence intervals for τ will therefore be

$$\left[\widehat{\tau}_n - \widehat{B}_{\tau_n,\kappa} \pm z_{1-\alpha/2} \widehat{\widehat{\sigma}}_{\tau,n} / \sqrt{n} \right], \tag{4.15}$$

when (4.7) can be applied, and

$$\left[\widehat{\tau}_{n_\kappa} - \widehat{B}_{\tau_n,\kappa} \pm z_{1-\alpha/2} \widehat{\widehat{\sigma}}_{\tau,n} / \sqrt{n_\kappa} \right] \tag{4.16}$$

when (4.8) can be applied.

To the best of our knowledge, the theoretical developments outlined above, in (4.13)–(4.16) are novel to the literature. Again, the proposed correction is of a small order, it vanishes asymptotically fast enough to preserve the asymptotic results from SZ, yet it might give significant improvements in practice in terms of more accurate confidence intervals, especially for moderate sample sizes and large dimensions, as illustrated in our Monte-Carlo experiments below. To help a practitioner, a step-by-step implementation algorithm of the proposed approach is summarized in the Appendix.

5 Monte-Carlo Evidence

In this section we summarize results from many Monte-Carlo (MC) experiments we conducted to analyze the performance of the newly proposed estimators for the variances described above, which aim to improve the approximation of the CLTs in finite samples.

As in the related works, to evaluate the performance we will estimate and present the

mean coverage of confidence intervals for different approaches for various types of the data generating process (DGP).[13] Intuitively, a better coverage by a particular method suggests that such method is likely to provide more accurate (and thus more reliable) confidence intervals than a method with relatively poorer coverage. We focus on the usual choices for the nominal confidence levels: $0.90, 0.95, 0.99$ and for various sample sizes, for 1000 of MC replications. Meanwhile, for the procedure of the bias estimation described above, we choose $L = 20$ reshuffles.[14]

5.1 Simplified Scenarios

For all of the scenarios in this sub-section, the technology set is assumed to be characterized by

$$\Psi = \{(x, y) \in \mathbb{R}^p_+ \times \mathbb{R}^1_+ \mid y \leq \psi(x)\},$$

where for $\psi(x)$ we use the Cobb–Douglas production function, as it appears to be the most popular in empirical works and in past simulations. Specifically,

$$\psi(x) = y^{\partial}(x) = a_0 \Pi_{r=1}^{p} x_r^{\beta_r}, \tag{5.1}$$

where β_r $(r = 1, ..., p)$ are the partial scale-elasticity coefficients, which are the parameters that are constant in a given scenario, whose sum reveals the returns to scale in the production. We have varied these parameters across scenarios to check the sensitivity of results.[15]

Following common practice and without much loss of generality, each input is generated from the standard uniform distribution, while the inefficiency scores are generated as $\lambda_i \sim |N(0, \sigma^2_\lambda)| + 1$ for various choices of σ^2_λ.[16] The 'observed outputs' are then generated by projecting the frontier points $y^{\partial}(x)$ inside the technology set Ψ, i.e., $Y_i = \psi(X_i)/\lambda_i$.[17]

[13] Recall that this mean coverage is defined as the percentage of times an estimated confidence interval with selected nominal confidence level covers the true value, out of all (here 1000) MC replications.

[14] Here, it is worth reminding that, as shown in KSW, the value of $L = 1$ is sufficient to apply the relevant CLTs. In Kneip et al. (2016), they suggest repeating the shuffling L times in order to reduce the variance of the bias estimator by a factor $1/L^2$. Every value of L is theoretically fine, only the numerical burden is a limit for this choice. Other choices of L we tried (e.g., $L = 10$, 100, 1000) in pilot MC trials gave similar results and so we decided to limit L to 20 to reduce the numerical burden.

[15] Because what matters the most from a theoretical perspective (e.g., for the rate of convergence) is the total number of inputs and outputs $(p + q)$, here we focus on a single-output case, i.e., $q = 1$, and consider different numbers of inputs, changing p.

[16] We tried many choices of σ^2_λ and the results were similar, with generally the same conclusion. The tables present results for $\sigma^2_\lambda = 1$.

[17] Recall that the data generating process that justifies asymptotic properties of DEA and FDH assumes no noise and so we follow this assumption here as well.

As in SZ, the true values of $\tau = \mu_1/\mu_2$ were computed via a prior simulation of $n = 2,000,000$ realizations of $(Y_i, Y_i^{\partial} = \psi(X_i))$ and then using (4.5). While we will limit our focus to the VRS-DEA estimator, which appears to be the most popular in practice, analogous conclusions apply to the CRS-DEA and FDH estimators.

In the tables below, note that the shortcut 'CLT1' refers to the relevant part of CLT when $\kappa \geq 1/2$ (i.e., (2.17) for the sample mean and (4.7) for the aggregate efficiency) and 'CLT2' refers to the relevant part of CLT when $\kappa < 1/2$ (i.e., (2.18) for the sample mean and (4.8) for the aggregate efficiency) and the asterisk indicates that the improved estimator for the variance was used when applying these theoretical results.

We first consider the case where $p = 1$, $q = 1$, i.e., where the bias is very small and, in fact, converges to zero fast enough so that even the standard CLT applies, although KSW and SZ showed that their approaches (based on CLT1, i.e., (2.17) for the sample mean and (4.7) for the aggregate efficiency) perform substantially better in the finite samples. Table 1 presents the results for such a case, when $\beta_1 = 0.4$.[18] As expected, because the bias is very small and converges to zero very fast in this 1-by-1 case, the improvement due to the proposed variance estimator is relatively small, yet persistent at virtually all sample sizes and nominal levels we considered, whether for the sample mean or the aggregate efficiency (i.e., weighted sample mean).

Similar results are obtained for $p = 2$, $q = 1$, and Table 2 presents the results for such a case, where $\beta_1 = 0.4$, $\beta_2 = 0.2$. From this table, one can see that there is a slightly more pronounced improvement from the proposed variance estimator, as expected, since the estimation bias generally increases with an increase in dimension, *ceteris paribus*, yet it is still not so large here, for the 2-input-1-output case.

The next case we consider is when $p = 3$, $q = 1$ and so both CLT1 and CLT2 apply (in the case of DEA-VRS). Table 3 presents the results for $\beta_1 = 0.4$, $\beta_2 = 0.2$, $\beta_3 = 0.1$ (other choices show similar results). One can see that the improvement from using the proposed variance estimator is persistent and larger than that observed for smaller dimensions, and both for CLT1 and CLT2. It is especially substantial at relatively small samples. As expected, because the methods are asymptotically equivalent, the larger the sample the smaller the improvement (both in relative and in absolute terms), yet note that it is still observed even for samples like 1000, e.g., for nominal coverage set at 0.95, the original approach of KSW (when CLT2 is used) provides coverage of 0.918, while the improved version provides coverage of 0.934. A similar conclusion is also obtained when observing results for the aggregate

[18] Here and in the other scenarios, many other choices of β's were tried (including the case when the returns to scale is close to constant) and they produced similar results, confirming robustness of the qualitative conclusions.

Table 1: Coverage for the Sample Mean and Aggregate Efficiency, for $p = 1$, $q = 1$

	CI level = 0.90		CI level = 0.95		CI level = 0.99	
	CLT1	CLT1*	CLT1	CLT1*	CLT1	CLT1*
	CLT for the Sample Mean					
10	0.500	0.557	0.573	0.635	0.685	0.736
20	0.602	0.653	0.676	0.711	0.795	0.824
50	0.691	0.717	0.775	0.790	0.892	0.898
100	0.771	0.784	0.834	0.842	0.918	0.921
200	0.826	0.830	0.880	0.883	0.967	0.968
300	0.846	0.846	0.912	0.913	0.960	0.962
500	0.847	0.848	0.907	0.907	0.979	0.979
1000	0.867	0.867	0.914	0.914	0.979	0.979
	CLT for the Aggregate Efficiency					
10	0.529	0.580	0.587	0.657	0.706	0.757
20	0.611	0.646	0.675	0.716	0.825	0.845
50	0.717	0.738	0.797	0.812	0.903	0.918
100	0.769	0.775	0.846	0.856	0.933	0.935
200	0.816	0.821	0.892	0.896	0.973	0.975
300	0.844	0.844	0.914	0.916	0.976	0.976
500	0.843	0.843	0.912	0.912	0.981	0.981
1000	0.881	0.882	0.922	0.923	0.973	0.973

efficiency.

The MC results also suggest that all the estimated coverages, even with the improved variance estimator, are still underestimating the nominal levels, especially for relatively small samples. It is worth contrasting, however, that they are much better than the coverages based on the standard CLT (which were around 0 for these nominal levels and for most of the samples). This phenomenon is pertinent not just to our approach, but rather it is an 'Achilles heel' of all the nonparametric envelopment estimators of this type and is sometimes referred to as the 'curse of dimensionality' problem, which is, intuitively, a price to pay for the flexibility of being nonparametric. In particular, for DEA-VRS the order of the errors is $n^{-2/(p+q+1)}$, which makes it problematic to try to make an inference with $n = 50$ and $p + q \geq 6$. See Section 2.5.1, Table 2.1 in Simar and Wilson (2013), e.g., with $n = 50$ and $p + q = 6$, the achieved precision corresponds to the inference that one would expect with $n = 9$ observations in a fully parametric model.

Nevertheless, we acknowledge that more work is needed to find further improvements in the finite-sample approximations of the new CLTs. For example, looking into the higher order approximations of the bias or/and the variance might be a fruitful way forward in future research, e.g., similar to the approach taken in the literature on the fixed effects in

Table 2: Coverage for the Sample Mean and Aggregate Efficiency, for $p = 2$, $q = 1$

	CI level = 0.90		CI level = 0.95		CI level = 0.99	
	CLT1	CLT1*	CLT1	CLT1*	CLT1	CLT1*
	CLT for the Sample Mean					
10	0.282	0.352	0.321	0.405	0.407	0.484
20	0.395	0.445	0.437	0.505	0.549	0.636
50	0.503	0.563	0.591	0.638	0.705	0.752
100	0.611	0.650	0.708	0.735	0.836	0.859
200	0.693	0.712	0.770	0.784	0.880	0.892
300	0.733	0.748	0.815	0.829	0.923	0.926
500	0.789	0.793	0.864	0.868	0.945	0.948
1000	0.825	0.826	0.887	0.893	0.979	0.980
	CLT for the Aggregate Efficiency					
10	0.260	0.348	0.320	0.409	0.434	0.524
20	0.431	0.490	0.486	0.556	0.610	0.681
50	0.567	0.617	0.653	0.703	0.785	0.830
100	0.679	0.712	0.769	0.797	0.878	0.896
200	0.730	0.742	0.804	0.819	0.915	0.924
300	0.766	0.775	0.848	0.859	0.939	0.946
500	0.807	0.816	0.876	0.886	0.957	0.957
1000	0.839	0.842	0.897	0.901	0.970	0.973

the panel data estimation framework by Dhaene and Jochmans (2015).[19]

Next we consider the case when $p = 4$, $q = 1$ and so $\kappa < 1/2$, meaning that only the CLT2 case and its improvement (denoted as CLT2*) applies here. Table 4 presents the results for $\beta_1 = 0.4$, $\beta_2 = 0.2$, $\beta_3 = 0.1$, $\beta_4 = 0.15$ (other choices show similar results). The same conclusions as reached in the previous table apply here as well, except that the improvement is even more pronounced (and applicable only for CLT2). Specifically, one can see from Table 4 that the improvement from using the proposed variance estimator is also persistent here and more substantial (as expected, due to larger bias). The coverage is especially improved at relatively small samples, and there is even an improvement when $n = 1000$, where, for example, for 0.95 nominal coverage, the improved version provides coverage of 0.937, while the original approach of KSW provides coverage of 0.912.

[19]We thank Koen Jochmans for the fruitful discussions on this idea.

Table 3: Coverage for the Sample Mean and Aggregate Efficiency, for $p = 3$, $q = 1$.

	CI level = 0.90				CI level = 0.95				CI level = 0.99			
	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*
	CLT for the Sample Mean											
10	0.136	0.189	0.172	0.251	0.162	0.227	0.216	0.290	0.211	0.299	0.267	0.380
20	0.197	0.263	0.283	0.368	0.237	0.308	0.338	0.429	0.314	0.405	0.426	0.538
50	0.338	0.397	0.487	0.576	0.389	0.477	0.569	0.647	0.511	0.588	0.680	0.778
100	0.454	0.509	0.608	0.674	0.521	0.573	0.703	0.771	0.642	0.710	0.842	0.891
200	0.475	0.524	0.705	0.748	0.568	0.623	0.784	0.835	0.732	0.771	0.906	0.937
300	0.595	0.634	0.770	0.809	0.671	0.721	0.857	0.882	0.805	0.845	0.937	0.959
500	0.636	0.661	0.793	0.827	0.728	0.755	0.880	0.894	0.852	0.876	0.956	0.967
1000	0.762	0.776	0.859	0.870	0.830	0.841	0.918	0.934	0.934	0.945	0.975	0.983
	CLT for the Aggregate Efficiency											
10	0.146	0.212	0.191	0.289	0.172	0.249	0.243	0.323	0.230	0.330	0.305	0.417
20	0.210	0.283	0.310	0.407	0.251	0.346	0.365	0.474	0.342	0.450	0.481	0.609
50	0.413	0.497	0.567	0.656	0.486	0.576	0.654	0.736	0.615	0.703	0.778	0.847
100	0.534	0.593	0.695	0.752	0.612	0.674	0.781	0.842	0.750	0.799	0.896	0.933
200	0.604	0.644	0.766	0.804	0.684	0.722	0.844	0.877	0.821	0.862	0.942	0.963
300	0.667	0.703	0.815	0.846	0.756	0.789	0.876	0.892	0.876	0.898	0.961	0.967
500	0.709	0.732	0.842	0.861	0.785	0.813	0.902	0.917	0.907	0.927	0.967	0.971
1000	0.786	0.796	0.846	0.866	0.862	0.869	0.923	0.931	0.941	0.948	0.979	0.981

Table 4: Coverage for the Sample Mean and Aggregate Efficiency, for $p = 4$, $q = 1$.

	CI level = 0.90		CI level = 0.95		CI level = 0.99	
	CLT2	CLT2*	CLT2	CLT2*	CLT2	CLT2*
	CLT for the Sample Mean					
10	0.101	0.145	0.113	0.179	0.153	0.246
20	0.140	0.228	0.173	0.271	0.255	0.370
50	0.280	0.404	0.353	0.499	0.503	0.675
100	0.453	0.580	0.549	0.685	0.723	0.835
200	0.608	0.712	0.717	0.813	0.864	0.923
300	0.727	0.803	0.820	0.889	0.928	0.961
500	0.802	0.866	0.882	0.923	0.972	0.989
1000	0.849	0.872	0.912	0.937	0.976	0.984
	CLT for the Aggregate Efficiency					
10	0.108	0.172	0.129	0.206	0.180	0.276
20	0.174	0.271	0.212	0.334	0.302	0.443
50	0.386	0.524	0.475	0.597	0.619	0.760
100	0.562	0.684	0.662	0.785	0.834	0.914
200	0.725	0.806	0.820	0.889	0.926	0.964
300	0.811	0.867	0.888	0.929	0.957	0.979
500	0.856	0.891	0.921	0.944	0.983	0.992

Similar conclusions are obtained for the cases with larger dimensions. For the sake of space we only present two more cases: $p = 5$, $q = 1$, in Table 5 where $\beta_1 = 0.4$, $\beta_2 = 0.2$, $\beta_3 = 0.1$, $\beta_4 = 0.15$ and $\beta_5 = 0.05$ and $p = 7$, $q = 1$ in Table 6, where $\beta_1 = 0.05$, $\beta_2 = 0.1$, $\beta_3 = 0.15$, $\beta_4 = 0.2$, $\beta_5 = 0.125$, $\beta_6 = 0.075$, $\beta_7 = 0.025$ (other values of parameters gave similar conclusions).

Table 5: Coverage for the Sample Mean and Aggregate Efficiency, for $p = 5$, $q = 1$

	CI level = 0.90		CI level = 0.95		CI level = 0.99	
	CLT2	CLT2*	CLT2	CLT2*	CLT2	CLT2*
	CLT for the Sample Mean					
50	0.166	0.293	0.221	0.381	0.344	0.534
100	0.330	0.491	0.425	0.609	0.597	0.804
200	0.569	0.715	0.681	0.812	0.832	0.935
300	0.648	0.770	0.750	0.855	0.892	0.969
500	0.775	0.850	0.846	0.920	0.957	0.988
1000	0.855	0.905	0.920	0.954	0.976	0.991
	CLT for the Aggregate Efficiency					
50	0.235	0.385	0.304	0.488	0.452	0.664
100	0.457	0.612	0.563	0.726	0.730	0.883
200	0.689	0.797	0.781	0.880	0.901	0.966
300	0.750	0.841	0.836	0.914	0.948	0.990
500	0.835	0.898	0.904	0.954	0.979	0.992
1000	0.867	0.915	0.935	0.966	0.986	0.996

As expected, the improvement is much more substantial here, and almost everywhere. For example, for $p = 5$, $q = 1$, for the nominal coverage set at 0.95: for $n = 500$ the improved approach provides coverage of 0.92, while the original approach of KSW provides coverage of 0.846 and even for $n = 1000$, the original approach of KSW provides coverage of 0.92, while the improved approach provides coverage of 0.954, i.e., almost exactly recovering the nominal coverage.

Meanwhile, for $p = 7$, $q = 1$ and the nominal coverage is 0.95, one can see that for $n = 500$ the improved approach provides coverage of 0.941, while the original approach of KSW provides coverage of only 0.778, and when $n = 1000$ the improved approach provides coverage of 0.973 (i.e., slightly overshooting, by 0.023), while the original approach of KSW provides coverage of 0.895 (undershooting by 0.055). The results for the aggregate efficiency are similar, with a bit more overshooting observed.

Table 6: Coverage for the Sample Mean and Aggregate Efficiency, for $p = 7$, $q = 1$

	CI level = 0.90		CI level = 0.95		CI level = 0.99	
	CLT2	CLT2*	CLT2	CLT2*	CLT2	CLT2*
	CLT for the Sample Mean					
50	0.087	0.198	0.116	0.249	0.197	0.397
100	0.155	0.323	0.219	0.433	0.355	0.652
200	0.379	0.589	0.466	0.708	0.643	0.901
300	0.481	0.708	0.589	0.830	0.796	0.961
500	0.690	0.849	0.778	0.941	0.928	0.992
1000	0.801	0.923	0.895	0.973	0.976	1.000
	CLT for the Aggregate Efficiency					
50	0.109	0.226	0.141	0.299	0.220	0.461
100	0.219	0.391	0.276	0.523	0.419	0.741
200	0.429	0.682	0.547	0.793	0.733	0.952
300	0.573	0.801	0.685	0.890	0.860	0.983
500	0.750	0.903	0.840	0.966	0.960	0.996
1000	0.858	0.945	0.924	0.984	0.987	1.000

5.2 Scenarios with Correlated Inputs

All the DGPs considered so far had no correlation between inputs, i.e., the so-called 'orthogonal design' (as was the case in many other studies). As pointed out by one of the anonymous referees, the data observed in practice often has fairly high correlations among some inputs (e.g., think of production of concrete, cakes and many other products which require certain proportions of some or all inputs) and so it is indeed worth considering such scenarios in MC, especially because so far there appears to be no theory on this matter.

To explore how our approach performs and compares in such situations, we considered various DGPs with different levels of correlation and the results for a couple of typical scenarios are presented here. Below we present the results for the scenario defined exactly as in Table 3 except that the third input is defined as the sum of the standard uniform (i.e., as in the previous examples) plus a proportion of the 1^{st} input, while the second input is kept uncorrelated with the other two, i.e.,

$$X_{3i} = \gamma_{13} X_{1i} + U_i, \;\; X_{1i}, U_i \sim \text{uniform}(0,1), \tag{5.2}$$

Thus, the correlation between the first input and the third inputs is controlled via the coefficient of proportionality, γ_{13}. When this coefficient is zero then we are back in the orthogonal design and exactly as in Table 3, which makes it convenient for comparisons. Meanwhile, setting this coefficient to 1.4 makes the standard correlation coefficient between

the first and the third input being around 0.8, i.e., about the level often observed in real data on production. The MC results for this level are presented in Table 7 and one can see that the general conclusion that the correction of the variance leads to improvements in the accuracy of the confidence intervals estimation is maintained here as well, in this non-orthogonal design.

It is also interesting to compare Table 7 to Table 3. Doing so one can see evidence that for all the methods the coverage is usually more accurate (sometimes substantially) relative to the case of zero correlation among the variables, especially for relatively low samples (e.g., 500 and below). For example, for $n = 100$ with zero correlation in the DGP the coverage of CI for the sample mean with nominal level 0.95 when using CLT2 and CLT2*, respectively, is about 0.70 and 0.77 when there is no correlation and about 0.82 and 0.86 for the same DGP but with correlation of about 0.8 between two inputs. It is a similar story for the coverage of the aggregate efficiency, as well as for other levels and for other sample sizes, except for $n = 1000$, where there is a slight drop (relative to $n = 500$), which was sometimes observed in the coverage for the correlated case relative to no correlation, although note that the improved approach (CLT2*)is still doing better than the exiting approach (CLT2) at all levels of confidence.

Table 7: Coverage for the Sample Mean and Aggregate Efficiency, with Correlated Inputs with $\gamma_{13} = 1.4$

	CI level = 0.90				CI level = 0.95				CI level = 0.99			
	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*
	Coverage for the Sample Mean											
10	0.211	0.288	0.257	0.351	0.248	0.342	0.294	0.403	0.314	0.411	0.389	0.487
20	0.320	0.403	0.415	0.513	0.368	0.460	0.476	0.599	0.461	0.560	0.590	0.719
50	0.431	0.513	0.597	0.661	0.510	0.593	0.673	0.758	0.637	0.736	0.800	0.869
100	0.601	0.653	0.746	0.801	0.677	0.729	0.822	0.861	0.802	0.851	0.921	0.953
200	0.668	0.713	0.802	0.837	0.751	0.787	0.877	0.908	0.863	0.889	0.956	0.971
300	0.715	0.751	0.834	0.859	0.803	0.822	0.894	0.915	0.904	0.916	0.964	0.971
500	0.742	0.770	0.856	0.878	0.822	0.840	0.922	0.931	0.916	0.926	0.978	0.989
1000	0.778	0.785	0.874	0.880	0.858	0.865	0.933	0.939	0.950	0.961	0.980	0.984
	Coverage for the Aggregate Efficiency											
10	0.231	0.334	0.297	0.399	0.266	0.379	0.334	0.446	0.356	0.444	0.423	0.534
20	0.348	0.446	0.474	0.593	0.411	0.526	0.548	0.665	0.531	0.643	0.669	0.768
50	0.520	0.594	0.657	0.737	0.595	0.661	0.752	0.822	0.719	0.796	0.859	0.903
100	0.666	0.724	0.775	0.825	0.758	0.798	0.861	0.897	0.861	0.897	0.944	0.966
200	0.714	0.747	0.816	0.847	0.806	0.839	0.891	0.912	0.910	0.924	0.975	0.979
300	0.740	0.770	0.866	0.883	0.819	0.849	0.925	0.938	0.912	0.922	0.975	0.979
500	0.764	0.779	0.862	0.874	0.831	0.844	0.930	0.943	0.925	0.933	0.982	0.988
1000	0.753	0.757	0.862	0.870	0.839	0.858	0.927	0.930	0.934	0.940	0.979	0.983

Also note that for the improved approach in the DGP with correlated inputs, the coverage almost reaches the nominal level at $n = 500$, and is reasonably close to it already at $n = 200$. This is an encouraging evidence for our method, since indeed many production processes and real data often involve highly correlated inputs. We explored if this evidence is confirmed in other scenarios and with different DGPs and it appears to be so, although we believe a separate study is needed (both theoretical and MC) to explore this matter in more detail to reach a more robust conclusion.

We also tried the case where all inputs are correlated but at different levels. Specifically, we considered a scenario defined exactly as the previous scenario, except that the second input now also has a correlation with the first input (and therefore also with the third input) in a similar fashion, at the same or different levels. While we tried many different levels of correlation (and all confirmed the general conclusions), an interesting case to present here seems to be one where one input has a positive correlation while another input has a negative correlation, i.e., some inputs are complements while others are substitutes. Results of one of such scenarios is presented in Table 8, where in addition to the previous scenario, we also have

$$X_{2i} = -\gamma_{12}(X_{1i} - 1) + U_i, \;\; X_{1i}, U_i \sim \text{uniform}(0,1), \tag{5.3}$$

where $\gamma_{12} = 0.5$. Clearly, if $\gamma_{12} = 0$ then we are in exactly the same design as in Table 7. Comparing Table 8 to Table 7 suggests that the results are very similar as in the case of zero correlation.

Finally, in all such scenarios we observed that the general conclusions that we have arrived in previous sub-section (that correcting the variance usually lead to improved coverage) are maintained here when we have different levels of correlation between various inputs.

5.3 Scenarios with Multiple Outputs

In all of the simulations so far, we have considered the single output cases. This is because, theoretically, the convergence rate depends on the total number of inputs and outputs $(p+q)$, not on p alone or q alone. It is however, important (as pointed out by one of the anonymous referees) to check that this is confirmed in practice, especially for small samples.

While there are different ways to generate multi-output scenarios, here we try the following approach (adapted from Zelenyuk (2019)): the production technology is characterized by the following set

Table 8: Coverage for the Sample Mean and Aggregate Efficiency, with Correlated Inputs with $\gamma_{13} = 1.4$, $\gamma_{12} = 0.5$

	CI level = 0.90				CI level = 0.95				CI level = 0.99			
	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*
	Coverage for the Sample Mean											
10	0.212	0.293	0.243	0.351	0.250	0.344	0.297	0.419	0.311	0.419	0.387	0.494
20	0.329	0.417	0.432	0.526	0.377	0.470	0.493	0.607	0.474	0.590	0.613	0.724
50	0.450	0.525	0.597	0.678	0.533	0.612	0.682	0.768	0.664	0.763	0.826	0.874
100	0.631	0.680	0.769	0.815	0.714	0.771	0.836	0.887	0.818	0.861	0.934	0.958
200	0.692	0.723	0.816	0.846	0.764	0.799	0.891	0.920	0.882	0.914	0.966	0.977
300	0.723	0.760	0.834	0.863	0.808	0.828	0.907	0.926	0.912	0.929	0.967	0.979
500	0.746	0.774	0.862	0.875	0.829	0.841	0.919	0.926	0.932	0.947	0.986	0.987
1000	0.729	0.741	0.842	0.857	0.812	0.822	0.919	0.926	0.920	0.929	0.978	0.982
	Coverage for the Aggregate Efficiency											
10	0.214	0.314	0.275	0.368	0.257	0.364	0.324	0.426	0.338	0.440	0.406	0.525
20	0.333	0.436	0.451	0.565	0.402	0.508	0.525	0.651	0.515	0.621	0.659	0.760
50	0.514	0.583	0.652	0.725	0.581	0.659	0.729	0.800	0.717	0.795	0.855	0.909
100	0.665	0.723	0.779	0.828	0.747	0.794	0.859	0.908	0.858	0.899	0.942	0.969
200	0.710	0.748	0.819	0.855	0.795	0.823	0.898	0.924	0.906	0.929	0.975	0.983
300	0.735	0.756	0.856	0.882	0.814	0.839	0.917	0.930	0.904	0.920	0.974	0.980
500	0.763	0.774	0.863	0.878	0.833	0.849	0.930	0.939	0.920	0.931	0.985	0.989
1000	0.729	0.736	0.848	0.854	0.810	0.823	0.915	0.921	0.907	0.921	0.973	0.977

$$\Psi \;=\; \left\{ (x,y) \;:\; \left(\sum_{m=1}^{q} \alpha_m (y_m)^2 \right)^{1/2} \leq \prod_{r=1}^{p} (x_r)^{\beta_r} \right\} \tag{5.4}$$

where $\beta_r \geq 0$ and $\alpha_m \geq 0$. That is, the right side of the inequality in (5.4) is the same as in (5.1), while for the output-side of the technology characterization we use the weighted Euclidian norm where, without loss of generality, we normalize $\sum_{m=1}^{q} \alpha_m = 1$. To generate data coherent with this technology, we first generate efficient outputs for each $i \in \{1, ..., n\}$, e.g., as $\widetilde{Y}_{mi} \overset{iid}{\sim}$ Uniform$(0.1, 1)$ for each $m \in \{1, ..., q\}$, and then generate $p - 1$ of inputs, e.g., as $X_{ri} \overset{iid}{\sim}$ Uniform$(0, 1)$ for each $r \in \{2, ..., p\}$,[20] and then define the remaining input, X_{1i}, in terms of the other generated inputs and all the outputs according to (5.4), i.e., for each $i \in \{1, ..., n\}$, we let

[20]We also tried $X_{ri} \overset{iid}{\sim}$ Uniform$(0.1, 1)$ and results were similar (with slightly better coverage).

$$X_{1i} := \left(\frac{\left(\sum_{m=1}^{q} \alpha_m (\widetilde{Y}_{mi})^2 \right)^{1/2}}{\prod_{r=2}^{p} (X_{ri})^{\beta_r}} \right)^{1/\beta_1} . \tag{5.5}$$

The observed output vector is then defined in the same fashion as in the previous scenarios by correcting the efficient outputs with the corresponding efficiency scores, i.e.,

$$Y_i = \widetilde{Y}_i / \lambda_i, \tag{5.6}$$

where λ_i is generated as in previous scenarios, i.e., $\lambda_i \sim |N(0, \sigma_\lambda^2)| + 1$.

While we tried many scenarios and found similar results, for the sake of saving space here we only present two scenarios. One of these scenarios is presented in Table 9, where $p = 2$ and $q = 2$ and $\alpha_1 = 1/2$ and $\alpha_2 = 1/2$ while the prices of outputs are $w^T = (1, 1.5)$, and the rest of the parameters are the same as for Table 2. Also note that the scenario summarized in Table 3 has the same dimension, 4, as this new scenario summarize in Table 9 and so to some extent they are comparable on the grounds of equal dimensions and similar (though not the same) DGPs. Comparing these tables, one can see that the coverage is slightly lower, yet still very similar, for all approaches for samples of 100 and below and then much more similar for the larger samples, confirming the theory that it is the total dimension that matters. More importantly, note that the general conclusion that correcting the variance improves the coverage is also confirmed here.

Another scenario that has $p = 3$ and $q = 3$ is presented in Table 10, where now we have $\alpha_1 = 0.5$ and $\alpha_2 = 0.3$ and $\alpha_3 = 0.2$, while the prices of outputs are $w^T = (1, 1.5, 2)$, and the rest of the parameters are the same as for Table 5, which has the same dimension, 6. Observing these results and comparing them to a somewhat similar (though not the same) scenario with $p = 5$ and $q = 1$ in Table 5 that has the same dimension $p + q$ shows a slightly different tendency than in the previous scenario. Indeed, now the coverage is slightly better for most of the sample sizes and especially for the samples of 200 and below, though note that the difference becomes negligible for larger samples. Thus, while comparing the case with more outputs and fewer inputs to that with more inputs and fewer outputs, keeping the dimension $(p + q)$ the same, we see evidence that somewhat better coverage can be observed in some scenarios (as in this one), yet also the opposite performance can be observed for other scenarios (as in the previous example). This further confirms the theory that it is the total dimension that matters. Again, the key conclusion that correcting the variance usually improves the coverage is generally confirmed for all these and other scenarios we tried with multiple outputs.

Table 9: Coverage for the Sample Mean and Aggregate Efficiency for $p = 2$ and $q = 2$.

	CI level = 0.90				CI level = 0.95				CI level = 0.99			
	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*	CLT1	CLT1*	CLT2	CLT2*
	Coverage for the Sample Mean											
10	0.094	0.137	0.125	0.186	0.115	0.162	0.149	0.220	0.148	0.216	0.205	0.283
20	0.140	0.209	0.224	0.309	0.173	0.258	0.270	0.377	0.242	0.340	0.368	0.500
50	0.243	0.314	0.404	0.504	0.305	0.405	0.492	0.596	0.427	0.536	0.637	0.752
100	0.388	0.451	0.582	0.656	0.452	0.537	0.662	0.732	0.596	0.671	0.794	0.871
200	0.503	0.574	0.712	0.776	0.607	0.657	0.804	0.852	0.744	0.801	0.913	0.943
300	0.598	0.637	0.765	0.806	0.694	0.737	0.843	0.876	0.809	0.843	0.939	0.960
500	0.662	0.690	0.807	0.837	0.752	0.781	0.876	0.903	0.873	0.895	0.957	0.967
1000	0.755	0.774	0.865	0.877	0.830	0.857	0.917	0.936	0.934	0.942	0.986	0.987
	Coverage for the Aggregate Efficiency											
10	0.102	0.143	0.162	0.217	0.142	0.186	0.216	0.274	0.200	0.272	0.343	0.419
20	0.150	0.215	0.285	0.364	0.203	0.278	0.371	0.455	0.319	0.414	0.550	0.659
50	0.286	0.344	0.481	0.566	0.353	0.425	0.587	0.684	0.522	0.603	0.771	0.844
100	0.435	0.480	0.639	0.697	0.506	0.577	0.728	0.796	0.673	0.733	0.873	0.911
200	0.567	0.617	0.788	0.819	0.662	0.713	0.847	0.873	0.808	0.840	0.936	0.960
300	0.606	0.645	0.793	0.826	0.707	0.743	0.878	0.907	0.838	0.873	0.972	0.981
500	0.674	0.707	0.845	0.862	0.757	0.782	0.895	0.909	0.889	0.908	0.969	0.981
1000	0.759	0.775	0.882	0.893	0.856	0.875	0.934	0.946	0.942	0.951	0.986	0.988

5.4 Handling Higher Dimensions

Similar conclusions appear to also hold for larger dimensions than those we have presented here, typically with decreasing accuracy (for all methods) while increasing the dimensions, confirming the theory and illustrating the so-called 'curse of dimensionality' phenomenon, pertinent to DEA and FDH (as for many other nonparametric methods).

We also acknowledge that we have not explored the performance of higher dimensions as thoroughly as the ones we presented here for two reasons. First, this task would indeed be endless as one may always want to add yet another input or an output or both and in different DGPs. Second, the evidence we see from our Monte Carlo experiments, as well as from many other related studies we cited above, is that for DEA and FDH estimators, it is indeed advised to keep the dimensions as low as possible, e.g., as those we presented here, unless critical for the analysis.

Dimension reduction techniques are available for practitioners to achieve such a task. For example, one may use methods that are the variations of the principal component analysis adapted to DEA (e.g., see Daraio and Simar (2007) and Wilson (2018) and references therein) or use the economic aggregation approach adapted to DEA (e.g., see Zelenyuk (2019) and references therein), or a mixture of these two or other suitable approaches.

Table 10: Coverage for the Sample Mean and Aggregate Efficiency for $p = 3$ and $q = 3$.

	CI level = 0.90		CI level = 0.95		CI level = 0.99	
	CLT2	CLT2*	CLT2	CLT2*	CLT2	CLT2*
	Coverage for the Sample Mean					
50	0.203	0.346	0.253	0.427	0.372	0.614
100	0.420	0.599	0.514	0.708	0.678	0.872
200	0.627	0.783	0.727	0.872	0.876	0.962
300	0.741	0.853	0.827	0.921	0.939	0.988
500	0.813	0.895	0.887	0.942	0.967	0.995
1000	0.873	0.920	0.931	0.969	0.988	0.995
	Coverage for the Aggregate Efficiency					
	CLT2	CLT2*	CLT2	CLT2*	CLT2	CLT2*
50	0.278	0.401	0.373	0.532	0.632	0.791
100	0.486	0.650	0.622	0.768	0.824	0.926
200	0.659	0.801	0.782	0.885	0.929	0.977
300	0.747	0.853	0.845	0.920	0.952	0.989
500	0.814	0.886	0.891	0.941	0.965	0.993
1000	0.854	0.912	0.921	0.959	0.976	0.988

5.5 Summary of Monte Carlo Evidence

Overall, by comparing all the tables (those presented and those we omitted to conserve space), an expected general conclusion is confirmed: in the overwhelming majority of cases, the larger the dimension the larger the improvement tends to be due to the new variance estimator we proposed in this paper.

All in all, since a better coverage by a method typically implies better accuracy (and therefore greater reliability) of the estimated confidence intervals, the general conclusion from all the MC experiments we performed is that the proposed method should be very valuable for practitioners, at least as a robustness check.

6 Empirical Illustration

In this section we illustrate the methods that we discussed above for a real data set. For this purpose we have chosen a popular in the literature data set about rice producers in the Philippines, which was originally sourced from the International Rice Research Institute (IRRI) and popularized by Coelli et al. (2005) who used it for illustrating various examples of frontier and efficiency estimation and also made it available online.[21] This makes it a convenient resource for illustrations and comparisons and facilitates the possibility of repli-

[21] http://www.uq.edu.au/economics/cepa/crob2005/software/CROB2005.zip

cations by others. MATLAB codes handling the application are available upon request from the corresponding author.

The data set contains annual information on 43 rice producers in the Tarlac region of the Philippines, for 8 years (from 1990 to 1997). The particular information we use here is: one output (measured in tones of freshly threshed rice) and three inputs: (i) area planted (measured in hectares), (ii) labour used (measured in man-days of family and hired labor) and (iii) Fertilizers (measured in kilograms of active ingredients). For more details on the data, see Coelli et al. (2005, p. 325–326). It is worth noting that the correlation between all the inputs is fairly high in the data set (ranging from 0.83–0.92).

For this data we apply the VRS-DEA approach to estimate the Farrell–Debreu output oriented inefficiency scores (2.3). Table 11 presents the summary of the results. Specifically, columns 2 through 9 present the results for each year separately (i.e., using 43 observations to estimate an annual frontier) and column 10 presents results for the data pooled across the 8 years (i.e., using 344 observations to estimate the so-called 'grand' or 'pooled' or unconditional frontier for the 8 years). The reason for pooling the data (and thus ignoring possible technological change) is to illustrate the methods for a relatively large sample. We also tried pooling over a smaller number of years, e.g., 2, 3, 4 years, and the results are qualitatively the same.

The results include the VRS-DEA estimates of the simple and weighted sample means and their bias corrected versions and of the corresponding lower and upper bounds of CIs using the existing approaches (KSW for the simple mean and SZ for the weighted mean) and the lower and upper bounds of CIs from the improved approach proposed in this paper. (Results for other choices of significance α are qualitatively the same, yielding slightly narrower CIs, yet preserving the general conclusions.)

Before going into discussion of the estimates of confidence intervals, a few remarks are in order. First, the simple sample means of the (in)efficiency scores range from 1.20 to 1.51 (i.e., efficiency level of 0.83 to 0.66) with respect to an annual frontier and 1.8 (i.e., efficiency level of 0.55) with respect to the pooled frontier.[22] Second, note that, in all the periods, the aggregate efficiency (weighted sample means of the inefficiency scores) are substantially lower (i.e., showing higher average efficiency levels) than the simple sample means, e.g., for 1990 it is 1.30 vs. 1.51 and for the pooled frontier it is 1.57 vs. 1.8. Third, note on the substantially large bias correction for both the weighted means and especially for the simple means: e.g., it corrects the sample means of inefficiency for 1990 from 1.51 to 2.03, or for

[22]This difference between the averages of efficiency scores with respect to the annual frontiers and the pooled frontier suggests about a potential technological change over the whole period, which for simplicity of illustration and for staying within the scope of the paper we will ignore here.

the pooled frontier from 1.80 to 2.27, while for the aggregate efficiency the correction is from 1.3 to 1.57 in 1990 and from 1.57 to 1.9 for the pooled frontier.

Regarding the confidence intervals, in general, from virtually all the results, one can see that they are larger (and sometimes substantially so) when the improved approach proposed in this paper is used—this is consistent with the MC results, where the estimated CI were also large, which enabled the provision of a more accurate coverage of the true values by the confidence intervals.

For example, for the simple sample mean of efficiency scores for 1990 (column 2) the CI based on the KSW approach is from 1.77 to 2.48, while the CI based on the improved approach is from 1.66 to 2.59, i.e., the difference in length of CI for the improved approach is about 31%. Similarly, when using the pooled data for 1990–1997 (column 10) the CI based on the KSW approach is from 2.13 to 2.47, while the CI based on the proposed approach is from 2.10 to 2.50, i.e., the difference in length of CI for the improved approach here is about 18%.

Meanwhile, for the aggregate efficiency (weighted sample mean), which gave fairly different numbers than the simple means, the estimated confidence intervals are also wider when using the improved methods relative to the SZ method, yet for this data the difference is not as large as for the simple means, although it can also be viewed as substantial. For example, for 1990 the CI based on the SZ approach is from 1.4 to 1.77, while the CI based on the improved approach is from 1.34 to 1.83, i.e., the difference in length of CI for the improved approach is about 32%. Similarly, in column 10 (pooled data for 1990–1997) the CIs based on the SZ approach are from 1.8 to 2.07, while the CI based on the proposed approach is from 1.78 to 2.09, i.e., the difference in length of CI for the improved approach is about 15%.

To conclude this section, it is worth reminding readers that CIs from both approaches are theoretically correct from an asymptotic point of view, since they are equivalent as $n \to \infty$. However, as can be seen from our simulations, for practitioners, who typically work with finite samples (and often sample below 1000 observations), the corrected method suggested in our paper is expected to provide more accurate estimates of confidence intervals (and typically wider) than those from the existing methods. In practical terms, this means that a too narrow confidence interval has a higher likelihood to mislead decision makers and stakeholders when, for example, comparing the efficiency of an industry or sub-groups of firms within it, etc.

All in all, since the proposed methods are expected to be more accurate and involve the same level of computational complexity as the existing methods, they therefore seem very appealing to be used regularly in practice, at least for the sake of reaching robust conclusions.

Table 11: Real Data Illustration: VRS-DEA Estimates of Simple and Weighted Means of Efficiency and their 0.99% Confidence Intervals for Rice Producers in the Philippines.

	1990	1991	1992	1993	1994	1995	1996	1997	pooled
	For the Sample Mean								
DEA estimate	1.51	1.41	1.20	1.27	1.40	1.33	1.38	1.40	1.80
Bias corrected	2.03	1.75	1.42	1.49	1.75	1.56	1.70	1.81	2.27
Lower bound of est. CI	1.77	1.42	1.23	1.32	1.51	1.34	1.44	1.49	2.13
Lower bound of est. CI-Improved	1.66	1.37	1.18	1.29	1.44	1.30	1.38	1.39	2.10
Upper bound of est. CI	2.48	2.06	1.53	1.72	2.03	1.79	1.92	1.93	2.47
Upper bound of est. CI-Improved	2.59	2.12	1.57	1.76	2.10	1.83	1.98	2.04	2.50
	For the Aggregate Efficiency								
DEA estimate	1.30	1.26	1.11	1.22	1.37	1.25	1.29	1.29	1.57
Bias corrected	1.57	1.48	1.24	1.38	1.69	1.44	1.54	1.58	1.90
Lower bound of est. CI	1.40	1.23	1.10	1.16	1.35	1.23	1.31	1.32	1.80
Lower bound of est. CI-Improved	1.34	1.19	1.08	1.14	1.30	1.20	1.27	1.25	1.78
Upper bound of est. CI	1.77	1.65	1.30	1.58	1.93	1.61	1.78	1.70	2.07
Upper bound of est. CI-Improved	1.83	1.68	1.32	1.60	1.98	1.63	1.82	1.76	2.09

7 Concluding Remarks

In this paper we propose a simple yet practically valuable improvement of the finite sample approximation of the recently derived central limit theorems for the statistics involving production efficiency estimates from DEA or FDH. This improvement is indeed very easy to implement because it amounts to a simple correction of the already developed statistics and involves no extra computational burden, while also preserving the existing asymptotic theory, including consistency and asymptotic normality.

The evidence from our Monte-Carlo experiments suggest that the proposed approach persistently gave improvements in most of the cases that we tried, especially for relatively small samples or relatively large dimensions of the assumed production model. The empirical illustration for a popular real data set we used confirms that the difference in the estimated confidence intervals can be substantial.

Given the importance of the central limit theorems in practical applications of statistical testing and constructions of confidence intervals, this approach is expected to be very useful (and at almost no additional computational costs) for practitioners analyzing production efficiency using DEA or FDH approaches—by providing more accurate and thus more reliable estimates of confidence intervals of aggregate functions of individual efficiencies. This, in turn, should help decision makers in making better decisions.

Some fruitful directions for future research would include a thorough investigation of how similar corrections can improve statistical testing based on the new CLTs. Among others,

this includes testing between groups within a population (e.g., industry), tests of returns to scale or convexity assumptions, as well as testing in the framework of the two-stage truncated regression context and the related separability condition. Finally, we note again that there is still fairly large scope for improving the finite sample approximations, especially for very small samples. In particular, exploring the higher order approximations of the bias or/and the variance might be a fruitful way forward for such improvements. It also seems fruitful to explore in greater details how the high correlation among inputs (as seems typical in many real data on production processes) can improve the accuracy of estimation and inference with these and other related methods.

Acknowledgements

We thank the Editor and three anonymous referees for the fruitful feedback that helped improve our paper substantially. We are also thankful for the feedback from participants of EWEPA2019 and seminars at the University of Cambridge, Monash University and University of Sydney and from our colleagues. The authors also acknowledge the support from their institutions. V. Zelenyuk acknowledges the support from the Australian Research Council (FT170100401). We also thank Bao Hoang Nguyen, Duc Manh Pham and Evelyn Smart for their feedback from proofreading. These individuals and organizations are not responsible for the views expressed in this paper.

Appendix: Practical Implementation Algorithm

Here, we summarize the algorithm for estimating confidence intervals, which was used in the simulations and for the real data illustration in this paper and can be adapted to other cases.

Step 1. Compute the efficiency scores using a chosen estimator (DEA or FDH).

Step 2. Compute the relevant statistic for the parameter of interest (e.g., simple sample mean using (2.9) or weighted sample mean using (4.1)).

Step 3. Estimate the bias of the statistic (e.g., using (2.14) for the simple sample mean or (4.6) for the weighted sample mean).

Step 4. Estimate the relevant variance:

- **(a)** for the simple sample mean, use (2.20) for the KSW approach and use (3.1) for the improved approach;
- **(b)** for the weighted sample mean, use (4.10) for the SZ approach and use (4.14) for the improved approach.

Step 5. Select the confidence interval α and estimate the lower and upper bounds for the asymptotic CIs of the parameter of interest, e.g.:

- **(i)** when $\kappa \geq 1/2$:
  - **(a)** for the simple sample mean, use (2.21) for the KSW approach and use (3.3) for the improved approach;
  - **(b)** for the weighted sample mean, use (4.11) for the SZ approach and use (4.15) for the improved approach;
- **(ii)** when $\kappa < 1/2$
  - **(a)** for the simple sample mean, use (2.22) for the KSW approach and use (3.4) for the improved approach;
  - **(b)** for the weighted sample mean, use (4.12) for the SZ approach and use (4.16) for the improved approach.

References

[1] Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review*, 568–598.

[2] Boussofiane, A., Dyson, R. G., and Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, 52(1),1–15.

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

[4] Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., and Battese, G. E. (2005) An Introduction to Efficiency and Productivity Analysis, Springer, New York.

[5] Cook, W. and Seiford, L. (2009). Data envelopment analysis (DEA)—thirty years on. *European Journal of Operational Research*, 192(1):1–17.

[6] Daraio, C. and Simar, L. (2007). Advanced robust and nonparametric methods in efficiency analysis: methodology and applications. New York, NY: Springer.

[7] Daraio, C., Simar, L. and P.W. Wilson (2018), Central limit theorems for conditional efficiency measures and tests of the 'separability' condition in nonparametric, two-stage models of production, Discussion paper 2016/27, ISBA, UCL, in press *The Econometrics Journal*, doi: 10.1111/ectj.12103

[8] Deprins, D., Simar, L. and H. Tulkens (1984), Measuring labor inefficiency in post offices. In *The Performance of Public Enterprises: Concepts and measurements*. M. Marchand, P. Pestieau and H. Tulkens (eds.), Amsterdam, North-Holland, 243–267.

[9] Dhaene, G. and K. Jochmans (2015), Split-panel jackknife estimation of fixed-effect models, *The Review of Economic Studies*, 82(3), 991–1030.

[10] Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., and Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, 132, 245–259.

[11] Emrouznejad, A., Parker, B., and Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Journal of Socio-Economics Planning Science*, 42(3), 151–157.

[12] Farrell, M.J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society*, A(120), 253–281.

[13] Färe, R, Primont D (1995) *Multi-Output Production and Duality: Theory and Applications*, Boston: Kluwer Academic Publishers.

[14] Färe, R. and V. Zelenyuk (2003), On Aggregate Farrell Efficiencies, *European Journal of Operational Research,* 146(3), 615–621.

[15] Jeong, S.O. and L. Simar (2006), Linearly interpolated FDH efficiency score for nonconvex frontiers, *Journal of Multivariate Analysis*, 97, 2141–2161.

[16] Kneip, A, L. Simar and P.W. Wilson (2008), Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models, *Econometric Theory*, 24, 1663–1697.

[17] Kneip, A., Simar, L. and P.W. Wilson (2011), A computational efficient, consistent bootstrap for inference with non-parametric DEA estimators. *Computational Economics,* 38, 483–515.

[18] Kneip, A., Simar, L. and P.W. Wilson (2015), When bias kills the variance: Central Limit Theorems for DEA and FDH efficiency scores, *Econometric Theory*, 31, 394–422.

[19] Kneip, A., Simar, L. and P.W. Wilson (2016), Testing Hypothesis in Nonparametric Models of Production, *Journal of Business and Economic Statistics*, 34(3), 435–456.

[20] Kneip, A., Simar, L. and P.W. Wilson (2018), Inference in Dynamic, Nonparametric Models of Production: Central Limit Theorems for Malmquist Indices, Discussion paper 2018/10, ISBA, UCL.

[21] Serfling, R. J. (1980), Approximation Theorems of Mathematical Statistics, New York: John Wiley & Sons, Inc.

[22] Sickles, R. C. and Zelenyuk, V., (2019). Measurement of Productivity and Efficiency, New York: Cambridge University Press.

[23] Simar, L. and P.W. Wilson (2011), Inference by the m out of n bootstrap in nonparametric frontier models. *Journal of Productivity Analysis*, 36, 33–53.

[24] Simar, L. and P.W. Wilson (2013), Estimation and inference in nonparametric frontier models: Recent developments and perspectives, *Foundations and Trends® in Econometrics*, 5(3-4), 183–337.

[25] Simar, L. and P.W. Wilson (2015). Statistical approaches for nonparametric frontier models: A guided tour. *International Statistical Review*, 83:1, 77–110.

[26] Simar, L. and P.W. Wilson (2018), Technical, Allocative and Overall Efficiency: Inference and Hypothesis Testing. Discussion paper 2018/18, ISBA, UCL.

[27] Simar, L. and Zelenyuk, V. (2011). Stochastic FDH/DEA estimators for frontier analysis. *Journal of Productivity Analysis*, 36(1), 1–20.

[28] Simar, L. and V. Zelenyuk (2018), Central Limit Theorems for Aggregate Efficiency. *Operations Research*, 66(1), 137–149.

[29] Wilson, P. W. (2018). Dimension reduction in nonparametric models of production. *European Journal of Operational Research*, 267(1), 349 – 367.

[30] Zelenyuk, V. (2013), A scale elasticity measure for directional distance function and its dual: Theory and DEA estimation. *European Journal of Operational Research*, 228(3), 592–600.

[31] Zelenyuk, V. (2019), Aggregation of inputs and outputs prior to Data Envelopment Analysis under big data, *European Journal of Operational Research*, forthcoming.