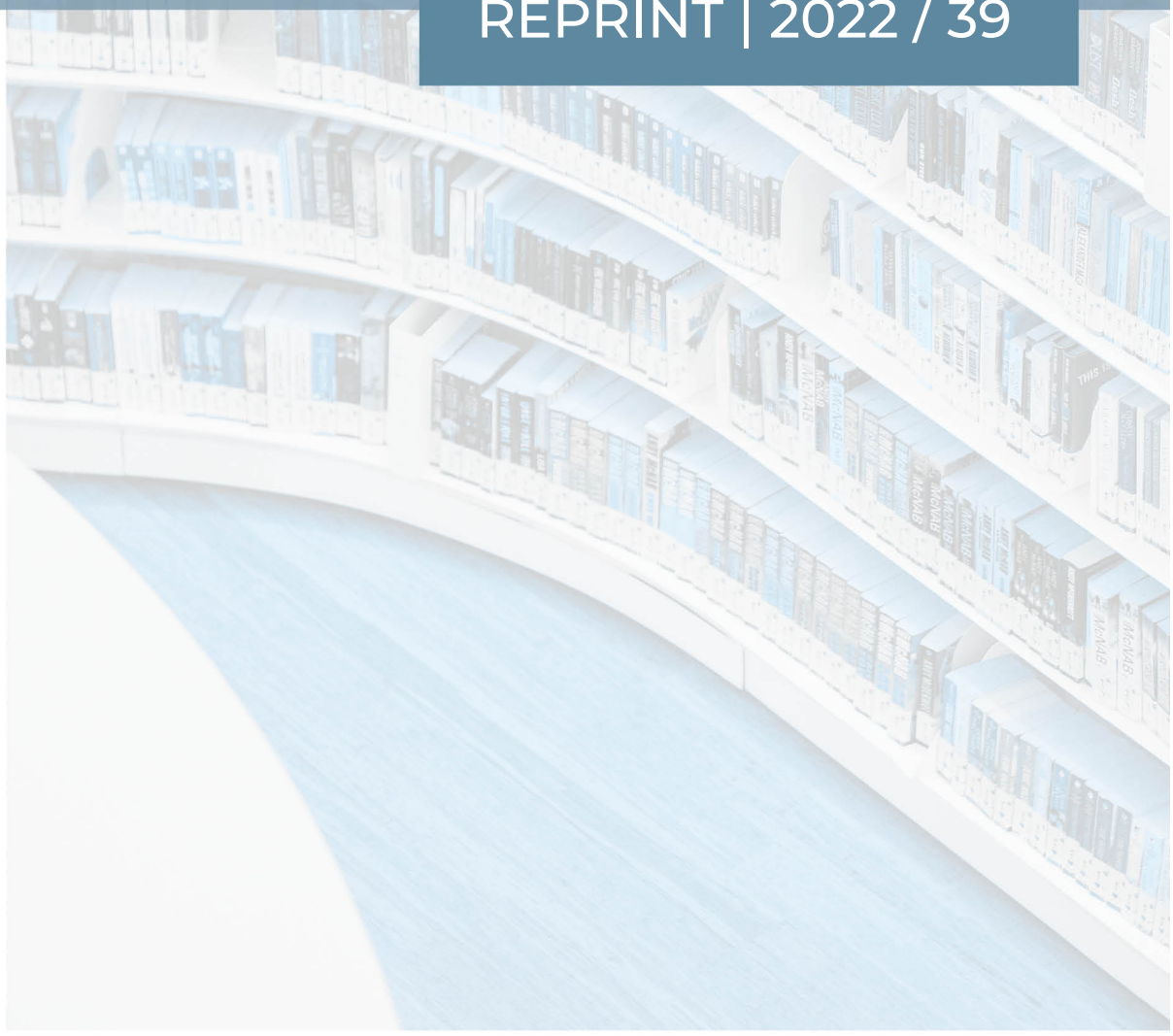
DATA SHARPENING FOR IMPROVING CENTRAL LIMIT THEOREM APPROXIMATIONS FOR DATA ENVELOPMENT ANALYSIS–TYPE EFFICIENCY ESTIMATORS

Bao Hoang Nguyen, Léopold Simar, Valentin Zelenyuk

REPRINT | 2022 / 39



UCLouvain

ISBA | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

European Journal of Operational Research 303 (2022) 1469–1480

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Data sharpening for improving central limit theorem approximations for data envelopment analysis–type efficiency estimators

Bao Hoang Nguyen[a], Léopold Simar[b], Valentin Zelenyuk[c,*]

[a] School of Economics, University of Queensland, Brisbane, Qld 4072, Australia
[b] Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Voie du Roman Pays 20, Louvain-la-Neuve B1348, Belgium
[c] School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, Qld 4072, Australia

ARTICLE INFO

Article history:
Received 21 September 2021
Accepted 20 March 2022
Available online 24 March 2022

Keywords:
Data envelopment analysis (DEA)
Free disposal hull (FDH)
Production efficiency
Statistical inference

ABSTRACT

Asymptotic statistical inference on productivity and production efficiency, using nonparametric envelopment estimators, is now available thanks to the basic central limit theorems (CLTs) developed in Kneip, Simar, and Wilson (2015). They provide asymptotic distributions of averages of Data Envelopment Analysis (DEA) and Free Disposal Hull (FDH) estimators of production efficiency. As shown in their Monte-Carlo experiments, due to the curse of dimensionality, the accuracy of the normal approximation is disappointing when the sample size is not large enough. Simar and Zelenyuk (2020) have suggested a simple way to improve the approximation by using a more appropriate estimator of the variances. In this paper we suggest a novel way to improve the approximation, by smoothing out the spurious values of efficiency estimates when they are in a neighborhood of 1. This results in sharpening the data for observations near the estimated efficient frontier. The method is very easy to implement and does not require more computations than the original method. We compare our approach using Monte-Carlo experiments, both with the basic method and with the improved method suggested in Simar & Zelenyuk (2020) and in both cases we observe significant improvements. We show also that the Simar & Zelenyuk (2020) idea of the variance correction can also be adapted to our sharpening method, bringing additional improvements. We illustrate the method with some real data sets.

1. Introduction

Envelopment estimators have been widely used in performance analysis as a powerful tool to estimate the efficiency of decision making units (DMUs). The application of envelopment estimators, such as Data Envelopment Analysis (DEA) along the lines of Farrell (1957), Charnes, Cooper, & Rhodes (1978) and Banker, Charnes, & Cooper (1984), and the Free Disposal Hull (FDH) originated by Deprins, Simar, & Tulkens (1984), is wide and deep, ranging from private sectors (e.g., manufacturing, banking, insurance, etc.) to public sectors (e.g., education, health care, etc.), from micro level (e.g., departments within firms, or firms themselves) to macro level (regions, or countries). Applied researchers in the field, besides (and in addition to) investigating the estimates of individual efficiency, are usually interested in analyzing the performance of some groups of DMUs that share common characteristics (e.g., domestic banks vs. foreign banks, public hospitals vs. private hospitals, etc.). These analyses are largely based on the statistical inference for sample statistics, such as the simple mean or weighted mean, of the estimated efficiency.

The literature on statistical properties of envelopment estimators of technical efficiency first emerged around the early 1990s and flourished over the last two decades with many breakthroughs. The statistical properties of envelopment estimators at a fixed point have been well-established by the seminal works of Kneip, Park, & Simar (1998), Park, Simar, & Weiner (2000), Kneip, Simar, & Wilson (2008), and Park, Jeong, & Simar (2010). Meanwhile, the statistical properties of the estimators at a random point have been recently brought to the literature by the seminal work of Kneip, Simar, & Wilson (2015). Based on these properties, Kneip et al. (2015) derive the Centre Limit Theorems (CLTs) for the envelopment estimators of technical efficiency. Many other important works also have been leveraged by the important results in Kneip et al. (2015), such as hypothesis testing in the context of DEA/FDH (Kneip, Simar, & Wilson, 2016), the CLTs for aggregate efficiency (Simar & Zelenyuk, 2018), the CLTs for conditional efficiency (Daraio, Simar, & Wilson, 2018), the CLTs for Malmquist indices (Kneip, Simar, & Wilson, 2020), and the CLTs for cost and allocative efficiency (Simar & Wilson, 2020b), to mention a few.

* Corresponding author.
E-mail address: v.zelenyuk@uq.edu.au (V. Zelenyuk).

https://doi.org/10.1016/j.ejor.2022.03.038
0377-2217/

Although the results in Kneip et al. (2015) provide a solid ground for many of the theoretical developments in the field, the performance of the CLTs is sometimes disappointing for practitioners, especially for small samples with a large dimension of inputs and outputs. To improve the finite sample approximation of the CLTs, Simar & Zelenyuk (2020) proposed a 'simple to compute' approach, which is based on a bias-corrected version of the variance estimator. Simar & Zelenyuk (2020) explore the performance of their proposed approach with various Monte Carlo simulation scenarios. The simulation results suggest that the improvement in finite samples is sometimes substantial, but there is still room for further improvement.

In this study, we propose another way to improve the CLT approximations. The proposed improvements are mirroring the work of Simar & Zelenyuk (2006) and Kneip, Simar, & Wilson (2011), who addressed the discontinuity issue to improve the statistical inference in the DEA/FDH context. Specifically, the idea is to smooth out, in an appropriate way, the spurious values of efficiency estimates when they are in a neighborhood of 1 (i.e., the 100%-efficiency bound). We know indeed from previous works that this "discretization" near the boundary creates problems when estimating the distribution of efficiency estimates for the purpose of bootstrap approximations, see e.g., Simar & Wilson (1998), Simar & Zelenyuk (2006), Kneip et al. (2008) and Kneip et al. (2011). In our context, this smoothing-out results in sharpening the data for observations located near the estimated efficient frontier before applying the CLTs. To the best of our knowledge, this theoretical justification is novel for the DEA (and productivity and efficiency in general). It also brings the early attempts to a new level that is now better justified from a theoretical perspective based on the recently developed asymptotic theory for this field and related (yet not at all the same) "data sharpening" ideas in statistics.

The method is very easy to implement and does not require more computations than the original method. We investigate the size of the improvements by using Monte-Carlo experiments, compared both to the basic method in Kneip et al. (2015) and to the improved method suggested in Simar & Zelenyuk (2020). We show also that the latter can also be adapted to our sharpening method, bringing additional improvements.

Our paper is organized as follows. The next section briefly discusses the production theory underlying the analysis of technical efficiency. Section 3 introduces the envelopment estimators of technical efficiency. Section 4 summarizes the statistical properties of these estimators. Section 5 presents our data sharpening idea to improve the accuracy of the CLTs. Section 6 provides Monte Carlo simulation evidence about the performance of the proposed approach for various sample sizes and various dimensions of inputs/outputs. Section 7 provides some illustrations with real data sets and Section 8 summarizes the concluding remarks.

2. Theoretical background

To facilitate our discussion, in this section we briefly discuss the production theory underlying the analysis of technical efficiency.

2.1. Characterization and axioms of production technology

A production technology in which a production unit utilizes p inputs, denoted as a $p-$dimensional column vector $x \in \Re_+^p$, to produce q outputs, denoted as a $q-$dimensional column vector $y \in \Re_+^q$, can be characterized generally by a technology set, defined as

$$\Psi = \left\{(x,y) \in \Re_+^p \times \Re_+^q : \ x \text{ can produce } y\right\}. \tag{2.1}$$

The production technology can be equivalently characterized by a portion of the technology set in the output space, which is referred to as the output set. The output set is defined as

$$P(x) = \left\{y \in \Re_+^q : \ x \text{ can produce } y\right\}. \tag{2.2}$$

When the efficiency measure is a concern, the boundary of the technology set is of interest. The boundary of the technology set is defined as

$$\Psi^{\partial} = \left\{(x,y) \in \Psi : \left(\delta^{-1}x, \delta y\right) \notin \Psi, \forall \delta > 1\right\}. \tag{2.3}$$

Usually, some standard axioms are imposed on the production technology, which can be summarized as follows (see Sickles & Zelenyuk (2019) and references therein for more detailed discussion).

A1. It is impossible to produce outputs without any inputs, i.e., $y \notin P(\mathbf{0}_p), \ \forall y \geq \mathbf{0}_q$ and $y \neq \mathbf{0}_q$.
A2. It is possible to produce nothing, i.e., $\mathbf{0}_q \in P(x), \ \forall x \in \Re_+^p$.
A3. $P(x)$ is a bounded set for all $x \in \Re_+^p$.
A4. Ψ is a closed set.
A5. All inputs and outputs are strongly disposable, i.e., $(x_0, y_0) \in \Psi \Rightarrow (x,y) \in \Psi, \forall x \geq x_0, y \leq y_0$.

2.2. Measure of technical efficiency

There are various measures of efficiency in the literature, yet the most popular efficiency measure appears to be the Farrell-Debreu technical efficiency. The Farrell-Debreu technical efficiency measures the radical distance from a point representing a production unit in output space (input space) to the boundary of the technology set in an output (input) orientation. For the sake of brevity, our discussion here focuses on the Farrell-Debreu output-oriented technical efficiency, where the radical distance is measured in an output orientation. For a production unit with input-output allocation (x,y), the Farrell-Debreu output-oriented technical efficiency measure is defined as

$$\theta(x,y) = \sup_{\theta} \{\theta > 0 : (x, \theta y) \in \Psi\}. \tag{2.4}$$

By construction, for all $(x,y) \in \Psi$, we have $\theta(x,y) \geq 1$, the value one characterizes a production plan (x,y) that belongs to the efficient boundary Ψ^{∂}. More generally, $\theta(x,y) > 1$ indicates the proportionate increase of the outputs the firm located at (x,y) should perform to reach the efficient frontier.

3. Envelopment estimators of technical efficiency

In practice, the technology set is not observable, and so neither is the efficiency measure. Researchers need to estimate the technology set and the related efficiency from a random sample of data, say, $\mathcal{X}_n = \left\{(X_i, Y_i) \ \middle| \ i = 1, \ldots, n\right\}$. Envelopment estimators appear to be one of the most popular methods to estimate the technology and efficiency. Among these, the estimator requires a minimum set of assumptions is the Free Disposal Hull (FDH) estimator. This estimator only requires the assumption on strong disposability of all inputs and outputs and is formulated as

$$\widehat{\theta}\left(X_i, Y_i \ \middle| \ \mathcal{X}_n\right) \equiv \max_{\zeta_1, \ldots, \zeta_n, \theta} \left\{\theta : \sum_{k=1}^n \zeta_k Y_k \geq \theta Y_i, \sum_{k=1}^n \zeta_k X_k \leq X_i, \right.$$
$$\left. \theta \geq 0, \zeta_k \in \{0,1\}, \sum_{k=1}^n \zeta_k = 1\right\}. \tag{3.1}$$

If in addition to strong disposability, one assumes that the technology set is convex and global variable returns to scale (VRS), one can then use the VRS-DEA estimator, which is formulated as

$$\widehat{\theta}\left(X_i, Y_i \mid \mathcal{X}_n\right) \equiv \max_{\zeta_1,\ldots,\zeta_n,\theta} \left\{ \theta : \sum_{k=1}^{n} \zeta_k Y_k \geq \theta Y_i, \sum_{k=1}^{n} \zeta_k X_k \leq X_i, \right.$$
$$\left. \theta \geq 0, \forall \zeta_k \geq 0, \sum_{k=1}^{n} \zeta_k = 1 \right\}. \tag{3.2}$$

Finally, if instead of global variable return to scale, one assumes that the technology is global constant return to scale (CRS), one can then utilize the CRS-DEA estimator, which is formulated as

$$\widehat{\theta}\left(X_i, Y_i \mid \mathcal{X}_n\right) \equiv \max_{\zeta_1,\ldots,\zeta_n,\theta} \left\{ \theta : \sum_{k=1}^{n} \zeta_k Y_k \geq \theta Y_i, \sum_{k=1}^{n} \zeta_k X_k \leq X_i, \right.$$
$$\left. \theta \geq 0, \forall \zeta_k \geq 0 \right\}. \tag{3.3}$$

The flexibility on assumptions about the reference technology does not come without cost, which is a slower rate of convergence of efficiency estimators, compared to the rate achieved in restrictive parametric models, when the dimension of the model, $p+q$, increases. This is known as the "curse of dimensionality". We will discuss it in more detail in the next section.

4. Statistical properties

4.1. Basic CLTs from Kneip et al. (2015)

Although statistical properties of envelopment estimators of technical efficiency at a fixed point are well-established in the literature, the statistical properties of these estimators at a random point, which is necessary to establish central limit theorems for mean efficiency, has recently been developed by Kneip et al. (2015). The three most important results in the work of Kneip et al. (2015) can be summarized as follows[1]

$$E\left[\widehat{\theta}\left(X_i, Y_i \mid \mathcal{X}_n\right) - \theta\left(X_i, Y_i\right)\right] = Cn^{-\kappa} + R_{n,\kappa}, \tag{4.1}$$

$$E\left[\left(\widehat{\theta}\left(X_i, Y_i \mid \mathcal{X}_n\right) - \theta\left(X_i, Y_i\right)\right)^2\right] = o\left(n^{-\kappa}\right), \tag{4.2}$$

$$\left|COV\left[\widehat{\theta}\left(X_i, Y_i \mid \mathcal{X}_n\right) - \theta\left(X_i, Y_i\right), \widehat{\theta}\left(X_j, Y_j \mid \mathcal{X}_n\right) - \theta\left(X_j, Y_j\right)\right]\right|$$
$$= o\left(n^{-1}\right), \tag{4.3}$$

where C is a constant, $R_{n,\kappa}$ is a remainder term of order $o(n^{-\kappa})$ or smaller, κ is the rate of convergence and depends on the types of estimators and the dimension of the production space. Specifically, κ equals to $1/(p+q)$, $2/(p+q+1)$ and $2/(p+q)$ for FDH, VRS-DEA, and CRS-DEA estimators, respectively. We see that the achieved rates of convergence for these non-parametric estimators, i.e., n^{κ}, may be much lower than the usual rate achieved in most parametric models, i.e., $\sqrt{n}$, when $p+q$ increases.

From the above results, one can see that the envelopment estimators are biased and more importantly the bias vanishes with a slower rate compared to the variance when the sample size increases. As a result, to derive limiting distribution for the mean efficiency, Kneip et al. (2015) suggest correcting for the bias and controlling for the scaling factor when necessary. To discuss the procedure in Kneip et al. (2015) in more detail, let us first define μ_θ as the population mean of the true efficiency, i.e.,

$$\mu_\theta = E(\theta(X,Y)), \tag{4.4}$$

σ_θ^2 as the population variance of the true efficiency, i.e.,

$$\sigma_\theta^2 = Var(\theta(X,Y)), \tag{4.5}$$

and $\overline{\theta}_n$ as the sample mean of the estimated efficiency, i.e.,

$$\overline{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}\left(X_i, Y_i \mid \mathcal{X}_n\right). \tag{4.6}$$

Clearly we see from (4.1) that

$$E\left[\overline{\theta}_n\right] = \mu_\theta + Cn^{-\kappa} + R_{n,\kappa}. \tag{4.7}$$

The bias correction is obtained through a generalized jackknife method. Specifically, the original sample $\mathcal{X}_n$ is randomly divided into two disjoint subsets $\mathcal{X}_{n/2}^{(1)}$ and $\mathcal{X}_{n/2}^{(2)}$ both of size $n/2$. Then, in each subset, efficiency scores are estimated and the averages are computed by using only the sample of size $n/2$. Formally, for $\ell = 1, 2$ we have

$$\overline{\theta}_{n/2}^{(\ell)} = 2n^{-1} \sum_{\left\{i \mid (X_i, Y_i) \in \mathcal{X}_{n/2}^{(\ell)}\right\}} \widehat{\theta}\left(X_i, Y_i \mid \mathcal{X}_{n/2}^{(\ell)}\right). \tag{4.8}$$

Kneip et al. (2015) show that the consistent estimator of the bias of $\overline{\theta}_n$ is then given by

$$\widehat{B}_{n,\kappa} = (2^{\kappa} - 1)^{-1}(\overline{\theta}_{n/2}^{*} - \overline{\theta}_n), \tag{4.9}$$

where

$$\overline{\theta}_{n/2}^{*} = \frac{\overline{\theta}_{n/2}^{(1)} + \overline{\theta}_{n/2}^{(2)}}{2}. \tag{4.10}$$

The main argument in Kneip et al. (2015) is that

$$\widehat{B}_{n,\kappa} = B_{n,\kappa} + R_{n,\kappa} + o_p(n^{-1/2}), \tag{4.11}$$

where $B_{n,\kappa} = Cn^{-\kappa}$ is the leading term in the bias of the envelopment estimators as shown in (4.1) and (4.7). As explained in Kneip et al. (2016), this jackknife operation can be repeated J times, providing at each step the bias correction $\widehat{B}_{n,\kappa}^{(j)}$, then by averaging over the J replications, we obtain an estimator of the bias with less variance

$$\widehat{B}_{n,\kappa} = J^{-1} \sum_{j=1}^{J} \widehat{B}_{n,\kappa}^{(j)}. \tag{4.12}$$

The CLTs in Kneip et al. (2015) can then be summarized in the following theorem.

Theorem 4.1. *Under the appropriate set of assumptions of Theorem 3.1, 3.2, or 3.3 specified in 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 a VRS-DEA estimator is used and Ψ is convex, for $p+q \leq 3$ if a FDH estimator is used and satisfies free disposability of inputs and outputs, as $n \to \infty$ we have*

$$\sqrt{n}\left(\overline{\theta}_n - \widehat{B}_{n,\kappa} - \mu_\theta + R_{n,\kappa}\right) \xrightarrow{d} N\left(0, \sigma_\theta^2\right), \tag{4.13}$$

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

$$\sqrt{n_\kappa}\left(\overline{\theta}_{n_\kappa} - \widehat{B}_{n,\kappa} - \mu_\theta + R_{n,\kappa}\right) \xrightarrow{d} N\left(0, \sigma_\theta^2\right), \tag{4.14}$$

where $\overline{\theta}_{n_\kappa}$ is a subsample version of $\overline{\theta}_n$, in the sense that the average is taken over a random subsample $\mathcal{X}_{n_\kappa}^{*} \subseteq \mathcal{X}_n$ of size $n_\kappa = \lfloor n^{2\kappa} \rfloor < n$. Formally

$$\overline{\theta}_{n_\kappa} = \frac{1}{n_\kappa} \sum_{\left\{i : (X_i, Y_i) \in \mathcal{X}_{n_\kappa}^{*}\right\}} \widehat{\theta}\left(X_i, Y_i \middle| \mathcal{X}_n\right). \tag{4.15}$$

It should be noticed that in (4.15), the notation is explicit: we compute an average over a random subsample of points in $\mathcal{X}_{n_\kappa}^{*}$, but the reference set for each efficiency estimator is the full sample $\mathcal{X}_n$.

[1] All the results here are based on a set of mild regularity assumptions about the data generating process specified in Kneip et al. (2015). Typically they concern the smoothness of the frontier, the continuity of the density of (X,Y) on Ψ and the strict positivity of the latter on the frontier.

Remark. As discussed in Kneip et al. (2015), both results in (4.13) and (4.14) are applicable for the CRS-DEA estimator with $p+q=5$, for the VRS-DEA estimator with $p+q=4$, and for the FDH estimator with $p+q=3$, yet the result in (4.14) is more preferable for these cases because the neglected term in (4.14) (i.e., $\sqrt{n_\kappa}R_{n,\kappa}$) converges to zero faster than the neglected term in (4.13) (i.e., $\sqrt{n}R_{n,\kappa}$).

In order to apply the CLTs in practice, one needs to obtain a consistent estimate of population variance of efficiency. Kneip et al. (2015) suggest using the following estimator

$$\widehat{\sigma}^2_{\theta,n} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right) - \overline{\theta}_n\right)^2. \tag{4.16}$$

With these results, one can estimate a confidence interval for the population mean of efficiency, μ_θ. Specifically, when (4.13) is applied, the $(1-\alpha)100\%$ confidence interval for μ_θ is given by

$$\left[\overline{\theta}_n - \widehat{B}_{n,\kappa} - z_{\alpha/2}\frac{\widehat{\sigma}_{\theta,n}}{\sqrt{n}}, \overline{\theta}_n - \widehat{B}_{n,\kappa} + z_{\alpha/2}\frac{\widehat{\sigma}_{\theta,n}}{\sqrt{n}}\right], \tag{4.17}$$

where $z_{\alpha/2}$ is the critical value corresponding to the area of $\alpha/2$ on the right tail of the standard normal distribution. Meanwhile, when (4.14) is applied, the $(1-\alpha)100\%$ confidence interval for μ_θ is given by

$$\left[\overline{\theta}_{n,\kappa} - \widehat{B}_{n,\kappa} - z_{\alpha/2}\frac{\widehat{\sigma}_{\theta,n}}{\sqrt{n_\kappa}}, \overline{\theta}_{n,\kappa} - \widehat{B}_{n,\kappa} + z_{\alpha/2}\frac{\widehat{\sigma}_{\theta,n}}{\sqrt{n_\kappa}}\right]. \tag{4.18}$$

Kneip et al. (2015) investigate the quality of the CLT approximations by intensive Monte-Carlo experiments. Kneip et al. (2016) do the same for evaluating the quality of the approximations in various testing situations exploiting Theorem 4.1 to build appropriate test statistics. The approximations work reasonably well in most of the cases but become, as expected, disappointing when $p+q$ increases with small n. This will be illustrated and confirmed in our Monte-Carlo experiments below.

4.2. Improvements suggested by Simar & Zelenyuk (2020)

Simar & Zelenyuk (2020) point out that $\widehat{\sigma}^2_{\theta,n}$ in (4.16) is based on $\overline{\theta}_n$, a biased estimator of the population mean, and thus it might be a source of error in estimating σ^2_θ, with implications for finite sample accuracy of the CLT approximations. To correct for this, Simar & Zelenyuk (2020) suggest using the available first-order bias-corrected estimator of the mean, i.e.,

$$\overline{\theta}_{n,\mathrm{bc}} = \overline{\theta}_n - \widehat{B}_{n,\kappa}. \tag{4.19}$$

It turns out that this provides the following estimator for the population variance

$$\widehat{\widehat{\sigma}}^2_{\theta,n} = \widehat{\sigma}^2_{\theta,n} + \widehat{B}^2_{n,\kappa}. \tag{4.20}$$

Simar & Zelenyuk (2020) show that the confidence intervals in (4.17) and (4.18) are still valid when replacing $\widehat{\sigma}^2_{\theta,n}$ by $\widehat{\widehat{\sigma}}^2_{\theta,n}$ and show in Monte Carlo experiments that this provides significant improvements on the accuracy of the coverages of the resulting confidence intervals.

But, even with this enhancement, the CLT approximations are still disappointing, especially for the cases of small sample sizes and for the large number of inputs and outputs. So there is room for additional improvements, as explained and confirmed below.

5. Data sharpening

It is known that the estimation of the distribution of the efficiency scores from the empirical distribution of their DEA/FDH estimators is jeopardized by the discretization of the latter near its boundary, i.e., values at data points where the efficiency is not far from one. In particular, a lot of the resulting estimates will have the value of one, where the DGP, by continuity, assumes the probability of zero of these values. This is referred in the literature to as "spurious" ones. So far this issue has been raised when using bootstrap approximations where the "naive" bootstrap is seen as being inconsistent. In this case, some smoothing of the efficiency scores is requested, see e.g., Simar & Wilson (1998) and Kneip et al. (2008). A simplified smoothing approach, smoothing-out only the values for data points near the efficient boundary, has been used in the related bootstrap setups by Simar & Zelenyuk (2006) and Kneip et al. (2011). We will use the latter idea in our setup here, i.e., for improving the approximations in Theorem 4.1. We first indicate that this discretization issue may impact the quality of the CLT approximations from two sources.

First, evidently, we know that $\mathrm{Prob}(\theta(X_i,Y_i)=1)=0$, but by construction, the envelopment estimators will provide many values $\widehat{\theta}(X_i,Y_i \mid \mathcal{X}_n)=1$, whereas the true $\theta(X_i,Y_i)>1$ with probability one, especially when the dimension of the problem $p+q$ increases. Thus, these spurious ones provide additional bias in the estimation of μ_θ when using $\overline{\theta}_n$ or $\overline{\theta}_{n,\kappa}$. In addition, it can be seen that this discretization also impacts the estimator of the bias $\widehat{B}_{n,\kappa}$ defined in (4.9). To see this, we can look at a different way to compute $\widehat{B}_{n,\kappa}$. Let us define for $i=1,\ldots,n$,

$$\widehat{\theta}^*_i = \begin{cases} \widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}^{(1)}_{n/2}\right) & \text{if } (X_i,Y_i)\in\mathcal{X}^{(1)}_{n/2} \\ \widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}^{(2)}_{n/2}\right) & \text{if } (X_i,Y_i)\in\mathcal{X}^{(2)}_{n/2}, \end{cases} \tag{5.1}$$

and define a bias correction for $\widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right)$ as

$$\widehat{B}_{i,n,\kappa} = (2^\kappa-1)^{-1}\left(\widehat{\theta}^*_i - \widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right)\right). \tag{5.2}$$

Then, it is easy to verify that we can compute $\widehat{B}_{n,\kappa}$ in (4.9) as

$$\widehat{B}_{n,\kappa} = n^{-1}\sum_{i=1}^{n}\widehat{B}_{i,n,\kappa}. \tag{5.3}$$

Note that if we repeat the jackknife operation J times, we can define for each $j=1,\ldots,J$, the individual measure $\widehat{B}^{(j)}_{i,n,\kappa}$ and then define $\widehat{B}_{i,n,\kappa}=J^{-1}\sum_{j=1}^{J}\widehat{B}^{(j)}_{i,n,\kappa}$.[2] By looking into the decomposition in (5.1) and (5.2), we see that if $\widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right)=1$, we will also have $\widehat{\theta}^*_i=1$, and thus the individual observation i will not contribute to the bias correction in (5.3). Moreover, there is a non-negligible probability that the same will happen for observations i such that $\widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right)=1+\tau$, for some small value of τ.

So we hope to improve the accuracy of the CLT approximation by smoothing-out the estimated efficiency scores that are equal or near one. There are several asymptotically equivalent ways to achieve this. Given our purpose, we choose a way which gives, with probability one, smoothed values $\widetilde{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right) \geq \widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right)$ when $\widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right) \leq 1+\tau$. Specifically, the smoothed values are defined as

$$\widetilde{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right) = \begin{cases} \widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right) & \text{if } \widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right) > 1+\tau \\ \widehat{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right)\times\tilde{\xi}_i & \text{otherwise}, \end{cases} \tag{5.4}$$

where $\tilde{\xi}_i$ is a random independent draw from a uniform distribution on the interval $[1,1+\tau]$. The underlying idea is to approximate the density of $\theta(X,Y)$, locally, in the τ-neighborhood of the frontier, by a uniform density.[3] Clearly, not every τ is suitable for

[2] This equivalent way to define $\widehat{B}_{n,\kappa}$ was already mentioned in Kneip et al. (2016).

[3] This smoothing is asymptotically equivalent to the one suggested in Kneip et al. (2011) and Simar & Zelenyuk (2006), where in the second line of (5.4), they define $\widetilde{\theta}\left(X_i,Y_i \mid \mathcal{X}_n\right)=\tilde{\xi}_i$. Note that this latter way of smoothing has also been evaluated in

this purpose: while it should be 'big enough' to do the smoothing-out job, it should also be 'small enough' to assure the central limit theorems that we leverage on here are still applicable. Below we derive the lower and upper bounds for τ.

To assure the consistency of the density estimator after smoothing, the neighborhood of 1, tuned by τ, should be large enough: we must have $\tau \to 0$ and $n^{-\kappa}/\tau \to 0$ as $n \to \infty$, see Theorem 4.1 in Kneip et al. (2011). So if we define $\tau = n^{-\gamma}$, this requires $0 < \gamma < \kappa$. Intuitively, we must smooth out the efficiency estimates that are in a neighborhood of the frontier of the bigger order than the statistical precision of the estimator.

On the other hand, we do not want to lose the basic property of our estimator of the mean characterized by (4.7). The CLT will now be centered on the new estimator of μ_θ given by

$$\overline{\widetilde{\theta}}_n = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\theta}\big(X_i, Y_i \mid \mathcal{X}_n\big), \tag{5.5}$$

and we need τ to be small enough to keep the asymptotic properties of $\overline{\theta}_n$. So we need $\overline{\widetilde{\theta}}_n = \overline{\theta}_n + o_p(n^{-\kappa})$. It is easy to see that

$$\overline{\widetilde{\theta}}_n = \overline{\theta}_n + O_p(\tau^2). \tag{5.6}$$

The deviation $O_p(\tau^2)$ comes from the fact that $\tilde{\xi}_i = 1 + U_i$ where U_i is a uniform on $[0, \tau]$, so the deviation, when it applies, is of the order $O_p(\tau)$ and this happens with a probability of order $O(\tau) = \text{Prob}\big(\widehat{\theta}(X_i, Y_i \mid \mathcal{X}_n) < 1 + \tau\big)$. So we need $O_p(\tau^2) = o_p(n^{-\kappa})$, which implies $\gamma > \kappa/2$, thus providing the lower bound for γ to keep the desired asymptotic properties of $\overline{\theta}_n$. Therefore, the smoothing-out keeps the asymptotic properties of $\overline{\theta}_n$ for all values of $\tau = n^{-\gamma}$ with $\kappa/2 < \gamma < \kappa$. Thus, we have exact (and fixed) values for lower and upper bounds for γ, which in turn imply the lower and upper bounds for τ, which will vary with n and the dimension of the production model.

Interestingly, the smoothing technique in (5.4) is equivalent to "data sharpening" of the original observations in $\mathcal{X}_n$, which are close to the frontier.[4] Indeed, we can see the smoothed estimator of the efficiencies as being the DEA/FDH estimators at sharpened data points $(X_i, \widetilde{Y}_i)$, $i = 1, \ldots, n$, but computed with respect to the same reference sample $\mathcal{X}_n$. Indeed let us define

$$\widetilde{Y}_i = \begin{cases} Y_i & \text{if } \widehat{\theta}\big(X_i, Y_i \mid \mathcal{X}_n\big) > 1 + \tau \\ Y_i/\tilde{\xi}_i & \text{otherwise .} \end{cases} \tag{5.7}$$

It is easy to check that $\widetilde{\theta}\big(X_i, Y_i \mid \mathcal{X}_n\big) = \widehat{\theta}\big(X_i, \widetilde{Y}_i \mid \mathcal{X}_n\big)$, i.e., the regular FDH/DEA estimator for the sharpened sample point $(X_i, \widetilde{Y}_i)$, but with the reference set being the original sample $\mathcal{X}_n$.

Of course, for coherence, the jackknife estimator of the bias in $\overline{\widetilde{\theta}}_n$ has to be adapted. Specifically, in (5.1) and (5.2), the efficiency scores are now evaluated for points in the sharpened sample with the reference sets remaining the same (i.e., $\mathcal{X}_n$ and its two subsets $\mathcal{X}_{n/2}^{(\ell)}$, $\ell = 1, 2$). In other words, in the jackknife procedure, $\widehat{\theta}\big(X_i, \widetilde{Y}_i \mid \mathcal{X}_n\big)$, $\widehat{\theta}\big(X_i, \widetilde{Y}_i \mid \mathcal{X}_{n/2}^{(1)}\big)$, and $\widehat{\theta}\big(X_i, \widetilde{Y}_i \mid \mathcal{X}_{n/2}^{(2)}\big)$ are now used in the places of the original $\widehat{\theta}\big(X_i, Y_i \mid \mathcal{X}_n\big)$, $\widehat{\theta}\big(X_i, Y_i \mid \mathcal{X}_{n/2}^{(1)}\big)$, and $\widehat{\theta}\big(X_i, Y_i \mid \mathcal{X}_{n/2}^{(2)}\big)$, respectively. We denote the resulting bias estimate $\widetilde{B}_{n,\kappa}$. Due to (5.6), it shares the same property in (4.11) as $\widehat{B}_{n,\kappa}$.

Finally, the more appropriate consistent estimator of σ_θ^2 may be computed as

$$\widetilde{\sigma}^2_{\theta,n} = \frac{1}{n}\sum_{i=1}^{n}\Big(\widetilde{\theta}\big(X_i, Y_i \mid \mathcal{X}_n\big) - \overline{\widetilde{\theta}}_n\Big)^2. \tag{5.8}$$

So the confidence intervals described above in (4.17) and (4.18) for μ_θ are now given by

$$\left[\overline{\widetilde{\theta}}_n - \widetilde{B}_{n,\kappa} - z_{\alpha/2}\frac{\widetilde{\sigma}_{\theta,n}}{\sqrt{n}}, \overline{\widetilde{\theta}}_n - \widetilde{B}_{n,\kappa} + z_{\alpha/2}\frac{\widetilde{\sigma}_{\theta,n}}{\sqrt{n}}\right], \tag{5.9}$$

and when (4.14) is applied,

$$\left[\overline{\widetilde{\theta}}_{n,\kappa} - \widetilde{B}_{n,\kappa} - z_{\alpha/2}\frac{\widetilde{\sigma}_{\theta,n}}{\sqrt{n_\kappa}}, \overline{\widetilde{\theta}}_{n,\kappa} - \widetilde{B}_{n,\kappa} + z_{\alpha/2}\frac{\widetilde{\sigma}_{\theta,n}}{\sqrt{n_\kappa}}\right]. \tag{5.10}$$

Finally the improvement suggested by Simar & Zelenyuk (2020) can also be applied. It can be shown that this provides the following estimator of the variance

$$\widetilde{\widetilde{\sigma}}^2_{\theta,n} = \widetilde{\sigma}^2_{\theta,n} + \widetilde{B}^2_{n,\kappa}. \tag{5.11}$$

So the alternative is to use $\widetilde{\widetilde{\sigma}}_{\theta,n}$ in place of $\widetilde{\sigma}_{\theta,n}$ in the intervals (5.9) and (5.10).

We will investigate through our Monte-Carlo experiments, if these changes introduced by our data sharpening described in (5.7) provide, as expected, significant improvements on the coverages of the resulting confidence intervals.

6. Monte carlo experiments

6.1. Data generating process

To investigate the performance of the proposed improvements, we perform Monte Carlo simulations using the same data generating process as in Simar & Zelenyuk (2020) (for the multiple outputs scenario), who in turn adapted the data generating process from Zelenyuk (2020). Specifically, the technology set characterizing the production technology is given by

$$\Psi = \left\{(x, y) : \left(\sum_{\ell=1}^{q}\beta_\ell (y_\ell)^2\right)^{1/2} \le \prod_{s=1}^{p}(x_s)^{\alpha_s}\right\}, \tag{6.1}$$

where $\alpha_s \ge 0$, $\beta_\ell \ge 0$. The technology specified in (6.1) is characterized by the weighted Euclidian norm on the output side, and without loss of generality the weights β_ℓ are chosen such that $\sum_{\ell=1}^{q}\beta_\ell = 1$. This specification ensures that the output set is convex and satisfies the inactivity axiom of the production theory (i.e., producing nothing is possible). Meanwhile, the specification on the right hand side of the inequality in (6.1) follows the well-known Cobb-Douglas functional form, which is widely used in theoretical and empirical studies and can be viewed as the first order approximation (in logs) of the unknown production frontier. Moreover, this specification is more flexible than other extreme functional forms, such as Leontief or linear. We also select α_s such that the technology exhibits non-increasing returns to scale, i.e., $\sum_{s=1}^{q}\alpha_s \le 1$.[5]

The values of α_s and β_ℓ for different scenarios are summarized in Table 1.[6]

To generate the data, for each observation $i \in \{1, \ldots, n\}$, we first generate efficient output, $Y_{\ell i}^{\partial}$, such that $Y_{\ell i}^{\partial} \overset{iid}{\sim} Uniform(0.1, 1)$ for

our Monte-Carlo experiments, giving qualitatively similar results, although in some cases, they are slightly inferior in terms of the achieved coverages of the resulting confidence intervals.

[4] The terminology "data sharpening" was used, e.g., by Choi, Hall, & Rousson (2000) in the context of non-parametric regression and more recently by Doosti & Hall (2016) in the context of improving the accuracy of density estimators by perturbation of the data. The perturbation can be additive or multiplicative, as in our case.

[5] The results from the Monte Carlo simulations with different choices of α_s and β_ℓ are qualitatively the same and available from the authors upon request.

[6] The Monte Carlo simulation for each scenario involves 3000 MC simulations, 20 reshuffles (to estimate bias), and sample sizes of up to 1000.

Table 1
Parameters for Data Generating Process.

p	q	β_ℓ	α_s
1	1	$\beta_1 = 1$	$\alpha_1 = 0.4$
2	1	$\beta_1 = 1$	$\alpha_1 = 0.4, \alpha_2 = 0.2$
2	2	$\beta_1 = 0.5, \beta_2 = 0.5$	$\alpha_1 = 0.4, \alpha_2 = 0.2$
3	2	$\beta_1 = 0.5, \beta_2 = 0.5$	$\alpha_1 = 0.4, \alpha_2 = 0.2, \alpha_3 = 0.1$
3	3	$\beta_1 = 0.5, \beta_2 = 0.3, \beta_3 = 0.2$	$\alpha_1 = 0.4, \alpha_2 = 0.2, \alpha_3 = 0.1$

each $\ell \in \{1, \ldots, q\}$. Then for the case $p = 1$, we let

$$X_{1i} = \left(\sum_{\ell=1}^{q} \beta_\ell \left(Y_{\ell i}^{\partial} \right)^2 \right)^{1/2\alpha_1}, \tag{6.2}$$

for each $i \in \{1, \ldots, n\}$. For the case $p \geq 2$, for each $i \in \{1, \ldots, n\}$, we generate $p-1$ inputs, X_{si}, for each $s \in \{2, \ldots, p\}$, such that $X_{si} \overset{iid}{\sim} Uniform(0, 1)$ and let

$$X_{1i} = \left(\frac{\left(\sum_{\ell=1}^{q} \beta_\ell \left(Y_{\ell i}^{\partial} \right)^2 \right)^{1/2}}{\prod_{s=2}^{p} (X_s i)^{\alpha_s}} \right)^{1/\alpha_1}. \tag{6.3}$$

We then generate for each observation $i \in \{1, \ldots, n\}$ the true efficiency, θ_i, such that[7]

$$\theta_i \sim |N(0, \sigma_\theta^2)| + 1, \tag{6.4}$$

and the observed output is determined by the efficient output and the true efficiency, given as

$$Y_i = \frac{Y_i^{\partial}}{\theta_i}. \tag{6.5}$$

We investigate the performance of the CLTs under different approaches by examining the empirical coverage (corresponding to the nominal coverage of 0.95) of the estimated confidence interval. The empirical coverage is the percentage of times (out of total number of Monte Carlo (MC) replications) that the estimated confidence interval includes the true value of the population mean. Moreover, we also report the averages of the bias-corrected point estimates of mean and standard deviation over all MC replications.

In our results below we use the middle of the range for the possible values of $\tau = n^{-\gamma}$, with $\kappa/2 < \gamma < \kappa$, i.e., $\tau = n^{-0.75\kappa}$. We did the same simulations with γ in a grid of values ranging from 0.55κ to 0.95κ and the results are summarized in the Appendix A. While in many cases the results are similar, the level of γ near 0.75κ appears to provide the best performance. Developing a rule for choosing an optimal level of γ might be a fruitful path forward.

6.2. Results and discussion

Before discussing the results, it is important to note that in our Monte Carlo experiments, we investigate the sizes of the improvements of our sharpening method both compared to the basic method in Kneip et al. (2015) and to the improved method suggested in Simar & Zelenyuk (2020). Moreover, we also examine the additional improvements by adapting the latter to our sharpening method. For further reference, let us denote these methods by the following solutions.

- Solution 1: The basic method in Kneip et al. (2015), i.e., estimating confidence intervals using (4.17) or (4.18).
- Solution 2: The improved method suggested in Simar & Zelenyuk (2020), i.e., estimating confidence intervals using (4.17) or (4.18) with $\widehat{\sigma}_{\theta,n}$ being replaced by $\widehat{\widetilde{\sigma}}_{\theta,n}$.
- Solution 3: Our sharpening method, i.e., estimating confidence intervals using (5.9) or (5.10).
- Solution 4: The improved method suggested in Simar & Zelenyuk (2020) adapted to our sharpening method, i.e., estimating confidence intervals using (5.9) or (5.10) with $\widetilde{\sigma}_{\theta,n}$ being replaced by $\widetilde{\widetilde{\sigma}}_{\theta,n}$.

From Fig. 1 and Table 2, some observations are in order. First, as in Kneip et al. (2015) and Simar & Zelenyuk (2020), the current methods in the literature already work well for large samples (e.g., $n = 1000$ or above) with the estimated means and standard deviations close to their true values as well as the empirical coverages close to the nominal coverage.[8] Thus, in our discussion we focus on the performance in relatively small samples where these methods perform relatively poorly. Second, in virtually all the scenarios with relatively small samples (e.g., $n \leq 500$), our sharpening method provides the bias-corrected estimates of mean that are closest to the true value, on average, especially for relatively large dimension of inputs and outputs. Third, the improved method suggested in Simar & Zelenyuk (2020) does actually help to correct the downward bias in the estimation of standard deviation.

Moreover, our Monte Carlo experiments also confirm the result in Simar & Zelenyuk (2020) that Solution 2 provides persistent improvements compared to Solution 1, and the improvements are substantial for relatively small samples and large dimensions of inputs and outputs. For example, for $n = 100$, the improvements provided by Solution 2 (relative to Solution 1) are around 0.15 with $p = 3, q = 2$ and around 0.20 with $p = 3, q = 3$. However, even with the enhancement, the empirical coverages from Solution 2 under these scenarios (i.e., 0.646 and 0.710) are still far from the nominal coverages of 0.95.

Regarding our sharpening method, Solution 3 provides significant improvements compared to both Solution 1 and Solution 2.[9] The improvements provided by Solution 3 range from 0.02 to 0.34 (relative to Solution 1) and from 0.02 to 0.14 (relative to Solution 2). For instance, with $p = 3, q = 2$ and $n = 100$, the empirical coverage under Solution 3 is around 0.29 higher than Solution 1 and 0.14 higher than Solution 2. When the sample size increases, the magnitudes of improvements diminish (as expected, because of the consistency of these approaches). For example, with the same numbers of inputs and outputs, the improvements decrease to around 0.11 and 0.03, respectively, for $n = 300$. Moreover, it is worth noting here that the improvements provided by Solution 3 are, to some extent, not persistent. For instance, when the sample size increases to $n = 500$, with $p = 3, q = 3$, the empirical coverage under Solution 3 becomes 0.05 less than the empirical coverage under Solution 2.

When adapting the improved method suggested in Simar & Zelenyuk (2020) to our sharpening method, the improvements are more substantial, and more importantly, are persistent. As can be seen from Table 2, Solution 4 provides improvements ranging from 0.02 to 0.46 compared to Solution 1 and ranging from 0.02 to 0.27 compared to Solution 2. For example, with $n = 100$, the improvement provided by Solution 4 relative to Solution 2 is around 0.15

[7] The tables in the next section report the results where $\sigma_\theta = 1$.

[8] In this paper, we consider the sample sizes up to $n = 1000$. It is important to note that going beyond such sample size, although doable (at much greater computational cost), is outside of the focus of the current paper. Indeed, the explicit focus of our paper is to improve the performance of the CLTs in relatively small samples, where the current methods in the literature perform relatively poorly.

[9] It is worth clarifying here that we apply the rule-of-thumb that the magnitude of difference between an empirical coverage and the nominal coverage is significant if it is greater than or equal to 0.01 (i.e., the "error bound" corresponding to the 95% level of confidence given by $2 \times \sqrt{\frac{0.95(1-0.95)}{M}} \approx 0.01$, where $M = 3000$ is the number of MC replications) and the magnitude of difference between two empirical coverages is significant if it is greater than or equal to 0.02 (i.e., two times the "error bound" corresponding to the 95% level of confidence).

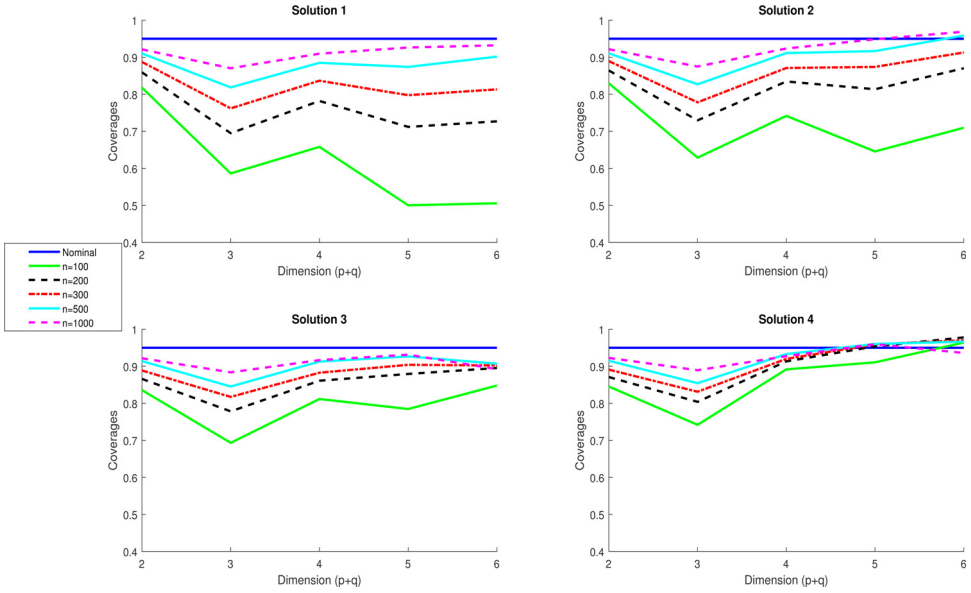


Fig. 1. Empirical Coverages for the Mean Efficiency with $\tau = n^{-0.75\kappa}$.

Table 2
Empirical Coverages for the Mean Efficiency with $\tau = n^{-0.75\kappa}$.

n	p	q	$\overline{n-m}$	Solution 1			Solution 2			Solution 3			Solution 4		
				Mean	Std	Cvg	Mean	Std	Cvg	Mean	Std	Cvg	Mean	Std	Cvg
100	1	1	17.07	1.769	0.554	0.818	1.769	0.568	0.830	1.778	0.545	0.836	1.778	0.559	0.845
	2	1	35.800	1.720	0.510	0.586	1.720	0.556	0.629	1.754	0.483	0.693	1.754	0.531	0.742
	2	2	49.055	1.703	0.466	0.658	1.703	0.549	0.742	1.773	0.423	0.812	1.773	0.516	0.892
	3	2	64.970	1.631	0.408	0.500	1.631	0.525	0.646	1.750	0.358	0.785	1.750	0.494	0.911
	3	3	73.107	1.610	0.371	0.506	1.610	0.515	0.710	1.775	0.322	0.848	1.775	0.496	0.964
200	1	1	22.33	1.785	0.571	0.859	1.785	0.578	0.864	1.790	0.566	0.866	1.790	0.573	0.871
	2	1	52.043	1.757	0.539	0.695	1.757	0.569	0.730	1.776	0.521	0.778	1.776	0.552	0.804
	2	2	75.642	1.747	0.501	0.782	1.747	0.562	0.835	1.790	0.469	0.861	1.790	0.535	0.913
	3	2	106.481	1.703	0.457	0.712	1.703	0.557	0.814	1.784	0.411	0.879	1.784	0.524	0.954
	3	3	124.397	1.688	0.423	0.727	1.688	0.554	0.870	1.807	0.370	0.895	1.807	0.524	0.978
300	1	1	26.16	1.791	0.579	0.887	1.791	0.583	0.890	1.793	0.576	0.889	1.793	0.580	0.891
	2	1	64.756	1.770	0.551	0.762	1.770	0.573	0.778	1.783	0.537	0.817	1.783	0.560	0.831
	2	2	97.189	1.763	0.519	0.837	1.763	0.568	0.871	1.795	0.492	0.883	1.795	0.544	0.920
	3	2	140.871	1.728	0.477	0.798	1.728	0.564	0.874	1.793	0.436	0.904	1.793	0.533	0.959
	3	3	167.806	1.723	0.448	0.813	1.723	0.567	0.913	1.821	0.397	0.902	1.821	0.535	0.970
500	1	1	32.25	1.795	0.585	0.910	1.795	0.588	0.911	1.796	0.584	0.914	1.796	0.586	0.915
	2	1	85.269	1.782	0.563	0.819	1.782	0.578	0.827	1.790	0.554	0.845	1.790	0.569	0.854
	2	2	132.419	1.780	0.535	0.885	1.780	0.572	0.911	1.803	0.515	0.913	1.803	0.553	0.933
	3	2	199.225	1.758	0.501	0.874	1.758	0.572	0.917	1.806	0.467	0.927	1.806	0.544	0.960
	3	3	242.857	1.758	0.474	0.902	1.758	0.577	0.959	1.833	0.428	0.907	1.833	0.545	0.967
1000	1	1	42.74	1.797	0.592	0.921	1.797	0.593	0.922	1.798	0.591	0.922	1.798	0.592	0.923
	2	1	122.861	1.791	0.574	0.870	1.791	0.582	0.875	1.796	0.569	0.884	1.796	0.577	0.889
	2	2	202.333	1.792	0.552	0.910	1.792	0.575	0.924	1.805	0.539	0.917	1.805	0.562	0.927
	3	2	316.829	1.779	0.524	0.926	1.779	0.576	0.949	1.810	0.498	0.931	1.810	0.553	0.960
	3	3	397.859	1.786	0.502	0.932	1.786	0.581	0.969	1.838	0.465	0.892	1.838	0.552	0.936

Note: (i) (4.13) applies for $p+q<4$, (4.14) applies for $p+q\geq 4$. (ii) $\overline{n-m}$: is the average number of "smoothed-out" observations over 3000 MC replications. (iii) For each solution, "Mean" column reports the averages of bias-corrected point estimates of simple mean over 3000 MC replications, "Std" column reports the averages of point estimates of standard deviation over 3000 MC replications, and "Cvg" column reports the empirical coverages over 3000 MC replications. (iv) For all scenarios, the true mean is 1.798 and the true standard deviation is 0.603

for $p=2, q=2$ (i.e., 0.892 versus 0.742) and around 0.27 for $p=3, q=2$ (i.e., 0.911 versus 0.646). Even with moderate sample sizes, such as $n=500$, Solution 4 still provides significant improvements. For instance, for $n=500$ and with $p=3, q=2$, the empirical coverage under Solution 4 is 0.960, which is around 0.04 higher than the empirical coverage under Solution 2 (i.e., 0.917). Meanwhile, it is worth noting here that for relatively large dimensions of inputs and outputs, Solution 4 provides a confidence level slightly greater than 95% (i.e., more conservative). For example, with $p=3, q=3$, the empirical coverages under Solution 4 are around 0.02 higher than the nominal coverage of 0.95 for $n\leq 500$. In this regard we note that while ideally one may wish an approach to always attain exactly the perfect (or nominal) levels, in practice there is often some under– or over–estimation for finite samples and so the key questions are: how often?, by how much?, and how substantial it is for practical implications (e.g., in terms of the length of a confidence interval)? In this respect, the slightly more conservative confidence intervals that we observe for Solution 4 at moderate sample sizes, like $n=500$, can be viewed as the positive feature of an approach, relative to the substantial under–estimation of confidence intervals (which may lead to over–rejection of related hypotheses).

Table 3
Average Length of CI for the Mean Efficiency with $\tau = n^{-0.75\kappa}$.

n	p	q	$\overline{n-m}$	CI level = 0.95			
				Solution 1	Solution 2	Solution 3	Solution 4
100	1	1	17.07	0.22 (0.02)	0.22 (0.02)	0.21 (0.02)	0.22 (0.02)
	2	1	35.80	0.20 (0.02)	0.22 (0.02)	0.19 (0.02)	0.21 (0.02)
	2	2	49.05	0.29 (0.03)	0.34 (0.03)	0.27 (0.03)	0.32 (0.03)
	3	2	64.97	0.35 (0.04)	0.45 (0.05)	0.31 (0.04)	0.42 (0.05)
	3	3	73.11	0.40 (0.05)	0.56 (0.07)	0.35 (0.04)	0.54 (0.06)
200	1	1	22.33	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)
	2	1	52.04	0.15 (0.01)	0.16 (0.01)	0.14 (0.01)	0.15 (0.01)
	2	2	75.64	0.24 (0.02)	0.27 (0.02)	0.22 (0.02)	0.25 (0.02)
	3	2	106.48	0.31 (0.02)	0.37 (0.03)	0.28 (0.02)	0.35 (0.03)
	3	3	124.40	0.37 (0.03)	0.49 (0.04)	0.32 (0.03)	0.46 (0.04)
300	1	1	26.16	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)
	2	1	64.76	0.12 (0.01)	0.13 (0.01)	0.12 (0.01)	0.13 (0.01)
	2	2	97.19	0.21 (0.01)	0.23 (0.01)	0.20 (0.01)	0.22 (0.01)
	3	2	140.87	0.28 (0.02)	0.33 (0.02)	0.26 (0.02)	0.31 (0.02)
	3	3	167.81	0.34 (0.02)	0.44 (0.02)	0.31 (0.02)	0.41 (0.02)
500	1	1	32.25	0.10 (0.00)	0.10 (0.00)	0.10 (0.00)	0.10 (0.00)
	2	1	85.27	0.10 (0.00)	0.10 (0.00)	0.10 (0.00)	0.10 (0.00)
	2	2	132.42	0.17 (0.01)	0.19 (0.01)	0.17 (0.01)	0.18 (0.01)
	3	2	199.22	0.25 (0.01)	0.28 (0.01)	0.23 (0.01)	0.27 (0.01)
	3	3	242.86	0.32 (0.01)	0.39 (0.02)	0.29 (0.01)	0.37 (0.02)
1000	1	1	42.74	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)
	2	1	122.86	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)
	2	2	202.33	0.14 (0.00)	0.14 (0.00)	0.13 (0.00)	0.14 (0.00)
	3	2	316.83	0.21 (0.01)	0.23 (0.01)	0.20 (0.01)	0.22 (0.01)
	3	3	397.86	0.28 (0.01)	0.32 (0.01)	0.26 (0.01)	0.30 (0.01)

Note: (i) (4.13) applies for $p+q<4$, (4.14) applies for $p+q\geq 4$. (ii) $\overline{n-m}$: is the average number of "smoothed-out" observations over 3000 MC replications. (iii) Standard deviations are reported in parentheses.

For relatively large samples, e.g., $n = 1000$, as discussed earlier in this section, the current methods in the literature perform relatively well, and further improvements are not necessary. In some instances, Solution 4 provides a slightly better empirical coverages compared to Solution 2, yet the improvements are not significant. In the other instance, e.g., $p = 3, q = 3$, the empirical coverage from Solution 2 (i.e., 0.969) is slightly higher than the nominal coverage of 0.95, meanwhile the empirical coverage from Solution 4 (i.e., 0.936) is slightly lower than the nominal level.

With regard to the average lengths of confidence intervals, we can see from Table 3 that Solution 2 and Solution 4 provide, on average, relatively wider confidence intervals compared to Solution 1 and Solution 3 because they use bias-corrected versions of sample variances.[10] Moreover, in virtually all scenarios, the confidence intervals from Solution 4 appear to be narrower than those from Solution 2 because the estimates of standard deviation for Solution 4 are smaller (as shown in Table 2).[11] The smaller lengths of confidence interval indicate that the confidence interval estimation by Solution 4 is more accurate than the confidence interval estimation by Solution 2, ceteris paribus. The differences in the average lengths of confidence intervals are, however, not significant for relatively small dimensions of inputs and outputs, and diminish when sample sizes increase.

In summary, Solution 4 appears to be a winner by most counts here: in all our simulations it showed the most substantial and persistent improvement upon the current approaches in the literature. Moreover, it is interesting to note that the improvement is more substantial with higher dimensions of inputs and outputs, where it appears that the data sharpening becomes even more useful since the proportion of envelopment estimates in a neighborhood of 1 increases substantially with the increase in the dimension.

[10] We also examine the median lengths of confidence intervals and the conclusion is qualitatively similar.

[11] We thank the anonymous referee for this insight.

7. Illustration with real data

In this section, we provide empirical illustrations of the proposed improvements with some real data sets.[12]

7.1. Philippine rice data

Our first empirical illustration utilizes the same data set as in Simar & Zelenyuk (2020), which is a data set including the information about 43 rice producers in Tarlac, Philippines from 1990 to 1997.[13] From the data set, we obtain the information about three inputs and one output. Specifically, the three inputs include the area planted, labour used, and fertiliser used, which are measured in hectares, man-days of family and hired labour, and kilograms of active ingredients, respectively. Meanwhile, the output is measured in tonnes of freshly threshed rice.

Following Simar & Zelenyuk (2020), we apply the VRS-DEA estimator to estimate the Farrell-Debreu output oriented efficiency scores and the 99% confidence intervals of their simple mean for each year separately and also for the pooled data across the 8 years. The results are presented in Table 4.

From Table 4 we can see that the differences in the bias-corrected estimates of simple means and in the estimated confidence intervals are more substantial for the annual frontiers where the sample sizes are relatively small (i.e., $n = 43$). Meanwhile, for the pooled frontier where the sample size is relatively large (i.e., $n = 344$), the estimated results based on Solution 1 and Solution 2 are very similar to those based on Solution 3 and Solution 4, respectively.

[12] The primary purpose of this section is to illustrate for practitioners the proposed improvement, explaining how our method improves the quality of the inference. A real application using our method to investigate empirical questions of one's interest would require a separate paper on its own.

[13] The data set was popularized in the literature by Coelli, Rao, O'Donnell, & Battese (2005) and can be downloaded from the website of their book: http://www.uq.edu.au/economics/cepa/crob2005/software/CROB2005.zip.

Table 4
VRS-DEA Estimates of Simple Means of Efficiency and their 0.99% Confidence Intervals for the Philippine Rice Data with $\tau = n^{-0.75\kappa}$.

Year	Solution 1			Solution 2			Solution 3			Solution 4		
	Mean	Std	C.I.	Mean	Std	C.I.	Mean	Std	C.I.	Mean	Std	C.I.
1990	2.11	0.61	[1.76, 2.47]	2.11	0.77	[1.67, 2.56]	2.20	0.55	[1.89, 2.52]	2.20	0.73	[1.78, 2.62]
1991	1.84	0.55	[1.53, 2.15]	1.84	0.62	[1.48, 2.20]	1.94	0.49	[1.66, 2.22]	1.94	0.58	[1.61, 2.27]
1992	1.56	0.25	[1.41, 1.70]	1.56	0.34	[1.36, 1.75]	1.69	0.21	[1.57, 1.81]	1.69	0.32	[1.50, 1.88]
1993	1.50	0.34	[1.31, 1.70]	1.50	0.42	[1.27, 1.74]	1.65	0.29	[1.48, 1.82]	1.65	0.38	[1.43, 1.87]
1994	1.69	0.45	[1.43, 1.95]	1.69	0.57	[1.37, 2.02]	1.79	0.40	[1.56, 2.02]	1.79	0.54	[1.48, 2.10]
1995	1.62	0.39	[1.40, 1.85]	1.62	0.46	[1.36, 1.89]	1.73	0.33	[1.54, 1.92]	1.73	0.42	[1.49, 1.97]
1996	1.65	0.41	[1.42, 1.89]	1.65	0.53	[1.35, 1.96]	1.80	0.35	[1.60, 2.00]	1.80	0.50	[1.51, 2.09]
1997	1.91	0.38	[1.69, 2.13]	1.91	0.56	[1.59, 2.23]	1.97	0.34	[1.78, 2.17]	1.97	0.54	[1.66, 2.28]
Pooled	2.28	0.67	[2.12, 2.45]	2.28	0.81	[2.08, 2.48]	2.29	0.66	[2.13, 2.46]	2.29	0.80	[2.09, 2.49]

Note: For each solution, "Mean" column reports the corresponding bias-corrected estimate of simple mean, "Std" column reports the corresponding estimate of standard deviation, and "C.I." column reports the corresponding 99% confidence interval.

Table 5
CRS-DEA Estimates of Simple Means of Efficiency and their 0.95% Confidence Intervals for the Queensland Hospital Data with $\tau = n^{-0.75\kappa}$.

Hospital groups	Solution 1			Solution 2			Solution 3			Solution 4		
	Mean	Std	C.I.	Mean	Std	C.I.	Mean	Std	C.I.	Mean	Std	C.I.
Teaching	1.35	0.18	[1.29, 1.40]	1.35	0.23	[1.28, 1.42]	1.45	0.16	[1.40, 1.49]	1.45	0.22	[1.38, 1.51]
Non-teaching	1.92	0.43	[1.84, 2.00]	1.92	0.51	[1.83, 2.01]	1.93	0.42	[1.86, 2.01]	1.93	0.50	[1.84, 2.02]
All	1.97	0.40	[1.90, 2.03]	1.97	0.53	[1.88, 2.06]	1.98	0.39	[1.91, 2.04]	1.98	0.52	[1.89, 2.06]

Note: For each solution, "Mean" column reports the corresponding bias-corrected estimate of simple mean, "Std" column reports the corresponding estimate of standard deviation, and "C.I." column reports the corresponding 95% confidence interval.

For the annual frontiers, some remarks are in order. First, among other Solutions, Solution 3 appears to provide the narrowest confidence intervals. Meanwhile, the confidence intervals based on Solution 2 and Solution 4 are among the widest. For example, for the year 1995, the confidence interval based on Solution 3 is from 1.54 to 1.92, which is about 15% narrower than the confidence interval based on Solution 1 (from 1.40 to 1.85). On the other hand, the confidence intervals based on Solution 2 (i.e., from 1.36 to 1.89) and Solution 4 (i.e., from 1.49 to 1.97) are, respectively, around 17% and 7% wider than those based on Solution 1. Second, the lower bounds of the confidence intervals based on Solution 4 are around 7% to 13% higher than those based on Solution 2, and the corresponding upper bounds are around 2% to 7% higher. For instance, for the year 1992, the confidence interval based on Solution 4 is shifted to the right compared to the confidence interval based on Solution 2 by a distance of around 35% of their lengths.

In summary, our main take-away from the discussions above is that the differences in estimated results based on different methods are practically significant, especially for the relatively small sample sizes. And thus, despite the desire to have narrower confidence intervals, the Monte Carlo evidence from the previous section suggests the wider confidence intervals of Solution 4 (jackknife with sharpening for correction of bias and variance) are likely to be more accurate.

7.2. Queensland hospital data

Our second empirical illustration looks at the case of multiple outputs. We utilize the data set on Queensland public hospitals from Nguyen & Zelenyuk (2021). The data set includes annual data on 104 public acute hospitals in Queensland, Australia in the five financial years (FYs) from FY 2012/13 to FY 2016/17.[14] Following Nguyen & Zelenyuk (2021) and the common practice in the literature (e.g., see Berta, Callea, Martini, & Vittadini, 2010; Burgess & Wilson, 1996; Chowdhury & Zelenyuk, 2016; Ferrier & Trivitt, 2013; Grosskopf, Margaritis, & Valdmanis, 2001; Hao & Pegels, 1994; Harris, Ozgen, & Ozcan, 2000; Magnussen, 1996; Nayar, Ozcan, Yu, & Nguyen, 2013), we use three inputs (i.e., labor, capital, and consumable inputs) and two outputs (i.e., outpatient and inpatient outputs) to model the production process of hospitals. Labor input is an aggregation (based on the principle component analysis) of the full-time equivalent staff of six main categories of hospital personnel. Capital input is proxied by the number of beds, and consumable input is measured by expenditures on drug, surgical and medical supplies in 2013/14 constant prices. On the output side, outpatient output is measured by the number of non-admitted occasions of service. Meanwhile, inpatient output is measured by casemix weighted inpatient episodes.[15]

As in Nguyen & Zelenyuk (2021), we estimate the simple means of efficiency and their confidence intervals for teaching hospitals, non-teaching hospitals and for the whole sample. We also trim 5% of outliers in the right tail of the estimated efficiency distributions, resulting in a trimmed sample of 494 observations, in which 118 observations have teaching status. We apply the CRS-DEA estimator to estimate the Farrell-Debreu output oriented efficiency scores and the 95% confidence interval of their simple mean for each group of hospital separately as well as for the whole sample.[16] The results are presented in Table 5.

As can be seen from Table 5, the estimated (bias-corrected) simple means and confidence intervals are not significantly different among the solutions for non-teaching hospitals and for the whole sample, where the sample sizes are relatively large (i.e., $n = 376$ for non-teaching hospitals and $n = 494$ for the whole sample). However, for teaching hospitals (where the sample size equals to 118), our sharpening method provides significantly different results compared to the current approaches in the literature. Note that for teaching hospitals, Solution 2 provides a very similar confidence interval compared to Solution 1 (i.e., $[1.29, 1.40]$ versus $[1.28, 1.42]$). Meanwhile, the confidence interval based on Solution 3 (i.e., $[1.40, 1.49]$) is to the right and does not overlap with the

[14] In Australia, a financial year starts on 1 July and ends on 30 June of the next calendar year.

[15] See more detail discussion in Nguyen & Zelenyuk (2021).

[16] It is worth noting here that in Nguyen & Zelenyuk (2021), they use the whole sample data to estimate efficiency scores for both teaching and non-teaching hospitals.

confidence interval based on Solution 1. The confidence interval based on Solution 4 is similar to Solution 3, but is wider due to the use of biased-corrected variance.

Again, this empirical illustration shows that the differences in estimated results based on different methods are practically significant, especially for the relatively small sample sizes. As the Monte Carlo evidence from the previous section suggests the wider confidence intervals of Solution 4 (jackknife with sharpening for correction of bias and variance) are likely to be more accurate, we would recommend this approach for practical use.

8. Concluding remarks

In this paper, we propose a novel 'simple to compute' approach endeavoring to improve the finite sample approximation of the CLTs for the envelopment estimators of production efficiency. The method is based on data sharpening of observations that fall in a neighborhood of the efficient frontier. This is in the spirit of ways suggested by Simar & Zelenyuk (2006) and Kneip et al. (2011) to solve the discretization issue of the empirical distribution of DEA/FDH efficiency estimates near the efficient value (1), providing what is known as "spurious ones". The problem may become severe when the dimension of the problem increases. The method we developed does not involve additional computational burden compared to the up-to-date existing methods for CLTs derived by Kneip et al. (2015) and Simar & Zelenyuk (2020). Our Monte-Carlo experiments compare our method with the existing ones. The results suggest that the data sharpening we propose, applied together with the variance correction proposed by Simar & Zelenyuk (2020), provides persistently as good or better accuracy in the coverage of the resulting confidence intervals, than the current approaches in the literature. In many cases, the improvement is substantial and significant.

It is also worth mentioning here that even with the proposed improvements, the empirical coverages of the CLTs for small samples, with a relatively large dimension of inputs and outputs are far from the nominal coverages. There is no miracle: we have to face the well-known "curse of dimensionality" of the nonparametric approaches.

It is straightforward to apply our method to other CLTs based on Kneip et al. (2015). Among others, these include the CLTs for aggregate efficiency (Simar & Zelenyuk, 2018), the CLTs for conditional efficiency (Daraio et al., 2018), the CLTs for Malmquist indices (Kneip et al., 2020), and the CLTs for cost and allocative efficiency (Simar & Wilson, 2020b). The same is true for the problems of hypothesis testing in the context of DEA/FDH, such as testing for the equality of the efficiency of different groups of DMUs, testing for assumptions on technology sets (e.g., convexity or returns to scale), see Kneip et al. (2016) and Simar & Wilson (2020a), or testing the separability condition as in Daraio et al. (2018), among other possibilities. Due to the encouraging results of our Monte-Carlo experiments, we can hope that the improvements will be also observed in these extensions.

Acknowledgement

We thank the Editor and two anonymous referees for fruitful comments. Bao Hoang Nguyen and Valentin Zelenyuk acknowledge the support from The University of Queensland and the financial support from the Australian Research Council (FT170100401). We also thank Evelyn Smart and Zhichao Wang for their valuable feedback. We acknowledge and thank Queensland Health for providing part of the data that we used in this study. These individuals and organizations are not responsible for the views expressed in this paper.

Appendix A

In this Appendix, we examine the performance of the sharpening method with different choices of $\tau = n^{-\gamma}$ by varying γ in a grid of values ranging from 0.55κ to 0.95κ, i.e.,$\{0.55\kappa, 0.65\kappa, 0.75\kappa, 0.85\kappa, 0.95\kappa\}$.[17] From Fig. 2, we can see that while in many cases the results are similar, the level of γ near 0.75κ appears to provide the best performance. Specifically, with the values of γ close to its lower bound (i.e.,$\gamma = 0.55\kappa, 0.65\kappa$), Solution 3 and Solution 4 are not well-performing, i.e., the empirical coverages under these Solutions are further from the nominal coverage when the sample size increases. Meanwhile, with the values of γ close to its upper bound (i.e.,$\gamma = 0.85\kappa, 0.95\kappa$), Solution 3 and Solution 4 provide improvements relative to Solution 1 and Solution 2, but the improvements are not as substantial as if γ is in the middle of the range of its possible values. As a result, developing a rule for choosing an optimal level of γ might be a fruitful path forward.

Appendix B

In this Appendix, as an evidence for the sufficiency of the number of MC replications utilized in this study (MC = 3000), we re-

[17] We have also obtained the results for a finer grid, which, in addition to the values reported here, includes $\{0.60\kappa, 0.70\kappa, 0.80\kappa, 0.90\kappa\}$, but these figures do not provide any additional useful information, and therefore are omitted to save space.

Table B1
Coverage for the Mean Efficiency under Variant 2 with $\tau = n^{-0.75\kappa}$ and MC = 1000.

n	p	q	$\overline{n-m}$	Solution 1			Solution 2			Solution 3			Solution 4		
				Mean	Std	Cvg	Mean	Std	Cvg	Mean	Std	Cvg	Mean	Std	Cvg
100	1	1	17.14	1.767	0.553	0.824	1.767	0.567	0.834	1.776	0.544	0.844	1.776	0.558	0.852
	2	1	35.659	1.721	0.510	0.580	1.721	0.557	0.628	1.755	0.483	0.700	1.755	0.532	0.747
	2	2	48.910	1.705	0.465	0.654	1.705	0.549	0.734	1.775	0.422	0.806	1.775	0.516	0.884
	3	2	64.884	1.629	0.408	0.507	1.629	0.524	0.666	1.748	0.357	0.780	1.748	0.494	0.909
	3	3	73.114	1.609	0.374	0.505	1.609	0.518	0.711	1.774	0.324	0.853	1.774	0.499	0.962
200	1	1	22.23	1.787	0.572	0.864	1.787	0.579	0.871	1.791	0.568	0.871	1.791	0.575	0.874
	2	1	52.112	1.758	0.539	0.713	1.758	0.569	0.746	1.777	0.522	0.794	1.777	0.553	0.816
	2	2	75.416	1.747	0.502	0.783	1.747	0.563	0.833	1.790	0.470	0.852	1.790	0.536	0.909
	3	2	106.642	1.701	0.456	0.708	1.701	0.555	0.806	1.783	0.410	0.880	1.783	0.522	0.955
	3	3	124.509	1.688	0.424	0.736	1.688	0.554	0.877	1.808	0.371	0.893	1.808	0.524	0.976
300	1	1	26.11	1.789	0.578	0.893	1.789	0.582	0.894	1.791	0.575	0.894	1.791	0.579	0.899
	2	1	64.942	1.769	0.550	0.754	1.769	0.573	0.780	1.782	0.537	0.816	1.782	0.560	0.827
	2	2	97.278	1.762	0.519	0.848	1.762	0.568	0.870	1.795	0.492	0.871	1.795	0.544	0.911
	3	2	141.051	1.729	0.477	0.800	1.729	0.564	0.875	1.793	0.436	0.903	1.793	0.532	0.957
	3	3	167.841	1.718	0.448	0.807	1.718	0.567	0.903	1.817	0.397	0.908	1.817	0.535	0.979

(*continued on next page*)

Table B1 (*continued*)

n	p	q	$\overline{n-m}$	Solution 1			Solution 2			Solution 3			Solution 4		
				Mean	Std	Cvg	Mean	Std	Cvg	Mean	Std	Cvg	Mean	Std	Cvg
500	1	1	32.39	1.795	0.586	0.912	1.795	0.588	0.912	1.797	0.584	0.916	1.797	0.586	0.918
	2	1	85.205	1.783	0.564	0.832	1.783	0.579	0.840	1.791	0.555	0.861	1.791	0.570	0.868
	2	2	132.781	1.782	0.536	0.884	1.782	0.572	0.916	1.804	0.515	0.905	1.804	0.553	0.930
	3	2	199.890	1.755	0.501	0.844	1.755	0.572	0.906	1.803	0.467	0.925	1.803	0.544	0.954
	3	3	242.833	1.757	0.474	0.902	1.757	0.577	0.954	1.832	0.428	0.916	1.832	0.545	0.970
1000	1	1	42.78	1.797	0.592	0.923	1.797	0.593	0.924	1.798	0.591	0.926	1.798	0.592	0.926
	2	1	122.869	1.792	0.574	0.885	1.792	0.583	0.893	1.797	0.569	0.895	1.797	0.578	0.904
	2	2	202.209	1.793	0.553	0.912	1.793	0.576	0.926	1.807	0.539	0.914	1.807	0.563	0.925
	3	2	316.290	1.780	0.525	0.922	1.780	0.576	0.944	1.811	0.499	0.921	1.811	0.554	0.949
	3	3	397.592	1.784	0.502	0.942	1.784	0.582	0.970	1.837	0.465	0.896	1.837	0.552	0.943

Note: (i) (4.13) applies for $p+q<4$, (4.14) applies for $p+q\geq 4$. (ii) $\overline{n-m}$: is the average number of "smoothed-out" observations over 1000 MC replications. (iii) For each solution, "Mean" column reports the averages of bias-corrected point estimates of simple mean over 1000 MC replications, "Std" column reports the averages of point estimates of standard deviation over 1000 MC replications, and "Cvg" column reports the empirical coverages over 1000 MC replications. (iv) For all scenarios, the true mean is 1.798 and the true standard deviation is 0.603

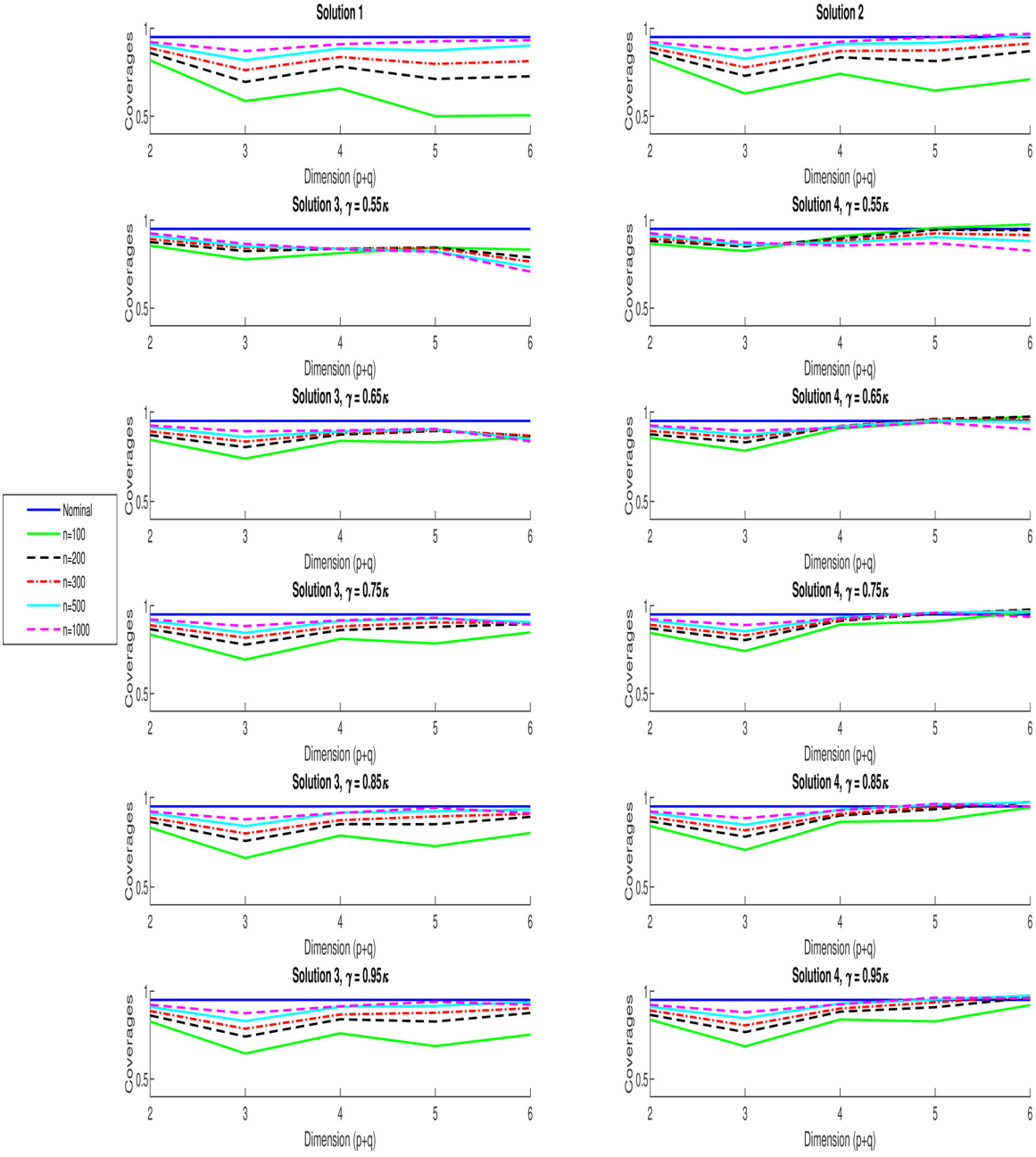


Fig. 2. Empirical Coverages for the Mean Efficiency with different choices of γ.

port here the simulation results with MC = 1000 and compare them with the results reported in Section 6. We can see that the results in Table B.6 are almost identical to those reported in Table 2 in Section 6. It indicates that further increasing the number of MC replications above 3000 will hardly bring any practical gain.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ejor.2022.03.038

References

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*(9), 1078–1092.

Berta, P., Callea, G., Martini, G., & Vittadini, G. (2010). The effects of upcoding, cream skimming and readmissions on the italian hospitals efficiency: A population-based investigation. *Economic Modelling, 27*(4), 812–821.

Burgess, J. F., & Wilson, P. W. (1996). Hospital ownership and technical inefficiency. *Management Science, 42*(1), 110–123.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444.

Choi, E., Hall, P., & Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression. *The Annals of Statistics, 28*(5), 1339–1355.

Chowdhury, H., & Zelenyuk, V. (2016). Performance of hospital services in ontario: DEA with truncated regression approach. *Omega, 63*, 111–122.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis*. New York: Springer.

Daraio, C., Simar, L., & Wilson, P. W. (2018). Central limit theorems for conditional efficiency measures and tests of the 'separability'condition in non-parametric, two-stage models of production. *The Econometrics Journal, 21*(2), 170–191.

Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor efficiency in post offices. In M. G. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The performance of public enterprises: Concepts and measurements, amsterdam, north-holland* (pp. 243–267).

Doosti, H., & Hall, P. (2016). Making a non-parametric density estimator more attractive, and more accurate, by data perturbation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78*(2), 445–462.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General), 120*(3), 253–290.

Ferrier, G. D., & Trivitt, J. S. (2013). Incorporating quality into the measurement of hospital efficiency: A double DEA approach. *Journal of Productivity Analysis, 40*(3), 337–355.

Grosskopf, S., Margaritis, D., & Valdmanis, V. (2001). Comparing teaching and non-teaching hospitals: A frontier approach (teaching vs. non-teaching hospitals). *Health Care Management Science, 4*(2), 83–90.

Hao, S. S., & Pegels, C. C. (1994). Evaluating relative efficiencies of veterans affairs medical centers using data envelopment, ratio, and multiple regression analysis. *Journal of Medical Systems, 18*(2), 55–67.

Harris, J., Ozgen, H., & Ozcan, Y. (2000). Do mergers enhance the performance of hospital efficiency? *The Journal of the Operational Research Society, 51*(7), 801–811.

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

Kneip, A., Simar, L., & Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory, 24*(6), 1663–1697.

Kneip, A., Simar, L., & Wilson, P. W. (2011). A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators. *Computational Economics, 38*(4), 483–515.

Kneip, A., Simar, L., & Wilson, P. W. (2015). When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores. *Econometric Theory, 31*(2), 394–422.

Kneip, A., Simar, L., & Wilson, P. W. (2016). Testing hypotheses in nonparametric models of production. *Journal of Business & Economic Statistics, 34*(3), 435–456.

Kneip, A., Simar, L., & Wilson, P. W. (2020). Inference in dynamic, nonparametric models of production: Central limit theorems for malmquist indices. *Econometric Theory*, 1–36.

Magnussen, J. (1996). Efficiency measurement and the operationalization of hospital production. *Health Services Research, 31*(1), 21–37.

Nayar, P., Ozcan, Y. A., Yu, F., & Nguyen, A. T. (2013). Benchmarking urban acute care hospitals: Efficiency and quality perspectives. *Health Care Management Review, 38*(2), 137–145.

Nguyen, B. H., & Zelenyuk, V. (2021). Aggregate efficiency of industry and its groups: The case of queensland public hospitals. *Empirical Economics, 60*(6), 2795–2836.

Park, B. U., Jeong, S.-O., & Simar, L. (2010). Asymptotic distribution of conical-hull estimators of directional edges. *The Annals of Statistics, 38*(3), 1320–1340.

Park, B. U., Simar, L., & Weiner, C. (2000). FDH Efficiency scores from a stochastic point of view. *Econometric Theory, 16*, 855–877.

Sickles, R. C., & Zelenyuk, V. (2019). *Measurement of productivity and efficiency*. New York: Cambridge University Press.

Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science, 44*(1), 49–61.

Simar, L., & Wilson, P. W. (2020a). Hypothesis testing in nonparametric models of production using multiple sample splits. *Journal of Productivity Analysis, 53*(3), 287–303.

Simar, L., & Wilson, P. W. (2020b). Technical, allocative and overall efficiency: estimation and inference. *European Journal of Operational Research, 282*(3), 1164–1176.

Simar, L., & Zelenyuk, V. (2006). On testing equality of distributions of technical efficiency scores. *Econometric Reviews, 25*(4), 497–522.

Simar, L., & Zelenyuk, V. (2018). Central limit theorems for aggregate efficiency. *Operations Research, 66*(1), 137–149.

Simar, L., & Zelenyuk, V. (2020). Improving finite sample approximation by central limit theorems for estimates from data envelopment analysis. *European Journal of Operational Research, 284*(3), 1002–1015.

Zelenyuk, V. (2020). Aggregation of inputs and outputs prior to data envelopment analysis under big data. *European Journal of Operational Research, 282*(1), 172–187.