

IZA DP No. 5997

## **When, Where and How to Perform Efficiency Estimation**

Oleg Badunenko  
Daniel J. Henderson  
Subal C. Kumbhakar

September 2011

# When, Where and How to Perform Efficiency Estimation

**Oleg Badunenko**

*University of Cologne*

**Daniel J. Henderson**

*State University of New York at Binghamton  
and IZA*

**Subal C. Kumbhakar**

*State University of New York at Binghamton*

Discussion Paper No. 5997  
September 2011

IZA

P.O. Box 7240  
53072 Bonn  
Germany

Phone: +49-228-3894-0  
Fax: +49-228-3894-180  
E-mail: [iza@iza.org](mailto:iza@iza.org)

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

## **ABSTRACT**

### **When, Where and How to Perform Efficiency Estimation**

In this paper we compare two flexible estimators of technical efficiency in a cross-sectional setting: the nonparametric kernel SFA estimator of Fan, Li and Weersink (1996) to the nonparametric bias corrected DEA estimator of Kneip, Simar and Wilson (2008). We assess the finite sample performance of each estimator via Monte Carlo simulations and empirical examples. We find that the reliability of efficiency scores critically hinges upon the ratio of the variation in efficiency to the variation in noise. These results should be a valuable resource to both academic researchers and practitioners.

JEL Classification: C14

Keywords: bootstrap, nonparametric kernel, technical efficiency

Corresponding author:

Daniel J. Henderson  
Department of Economics  
State University of New York  
Binghamton, NY 13902-6000  
USA  
E-mail: [djhender@binghamton.edu](mailto:djhender@binghamton.edu)

# 1 Introduction

Performance measures via efficiency scores have become widespread for operators in business, government, infrastructure, public transportation, energy production and other sectors. This is especially true in regulated industries. The objective of the regulators is to ensure efficient use of resources as a competitive market would. In this setting, governmental regulators typically estimate the efficiency of firms and the resultant efficiency scores are used to compare firms against ‘a benchmark’ to set targets, incentives or similar criterion. In some scenarios, failure to improve towards benchmark firms or failure to achieve a given absolute efficiency level can lead to monetary or other penalties, and hence effect profits.

For example, on January 1, 2009, the German regulator (Federal Network Agency) initiated the “Incentive Regulation Program” which called to benchmark (in terms of efficiency) German electricity and gas network operators nationwide. Participation for the approximately 850 electricity and 730 gas companies was compulsory. The consequences of employing a given method to efficiency measurement here may have an enormous economic impact. If a particular method produces efficiency scores which are biased upwards, the regulator may not penalize some firms and hence the inefficiencies are allowed to remain. On the other hand, if the efficiency scores are biased downwards, firms may be unnecessarily penalized. This may lead to a reduction in earnings and hence a change in the corporate landscape and possible disruption of the supply of energy. The annual benchmark related network cost (of around 200 electricity and 190 gas companies) amounts to approximately 10 billion Euro in the electricity industry and 4 billion Euro in the gas sector. We note that even a one-percent efficiency improvement can lead to large increases in profit to the suppliers. Alternatively, if the regulator forces the suppliers to lower prices, this may lead to increases in consumer surplus.

To assess the efficiency score, the German regulator in this case employs both Stochastic Frontier Analysis (SFA) and Data Envelopment Analysis (DEA) methods and for each firm picks the one that reveals the largest efficiency level. In general, there is no obvious way to choose an estimator of efficiency. While we note that both methods provide an estimate of the efficiency score for a firm, unless the two are identical, one will be closer to the truth (although no one knows it) in practice. It is obvious that obtaining a reliable efficiency score here is important especially if the regulator (based on the incorrect efficiency score) sets a benchmark, viz., a target price that is below the “minimum” average cost. This could result in monetary losses to suppliers and such a policy will not be sustainable. Similar regulations are currently occurring in the UK as well as other nations (Smith and Street (2005)). Benchmarking and efficiency analysis is routinely performed in water utilities, electricity and natural gas, rail transportation and postal services (Thanassoulis (2000a,b), Cubbin (2005), Dassler et al. (2006)). Furthermore, if efficiency is not taken into account, some regulatory steps may lead to destructive consequences for some [inefficient] firms if they are enacted rapidly (Morrison Paul et al. (2000)).

Here we hope to uncover which measure performs best in a given situation. This study can also be used as a helpful guide when employing efficiency estimation. The two primary methods to estimate efficiency scores are SFA and DEA. There has been a long standing division between SFA and DEA. The historically perceived merit of SFA is that the estimator is stochastic. The historically perceived merit of DEA is that the estimator is nonparametric in nature. Recent research, however, has seen a relaxation of functional forms in SFA and the introduction of asymptotics in DEA. That being said, there has been no systematic comparison of these improved techniques. Most studies either focus on one or the other, or propose a naive alternative when comparing one model versus the other.

We attempt to compare two promising estimators of technical efficiency in the cross-sectional case. Specifically, we compare the nonparametric kernel estimator of Fan et al. (1996), hereafter FLW, to the nonparametric bias corrected DEA estimator of Kneip et al. (2008), hereafter KSW. Although the FLW model requires a parametric second stage (and hence it could be argued that it is semiparametric), it is more robust than the initial SFA model by Aigner et al. (1977) and Meeusen and van den Broeck (1977). Similarly, KSW (which introduces asymptotics via bootstrapping) has been shown to perform well as compared to the standard DEA model found in Charnes et al. (1978). To make the comparison between the two methods, first we briefly discuss the merits of each approach in different settings. Then, using a set of simulations which we believe are fair to each method, we assess the finite sample performance of each estimator via Monte Carlo methods. Finally, we examine three separate empirical examples.

Our results show that the estimation of efficiency scores critically hinges upon the estimated ratio of the variation in efficiency to the variation in noise. Specifically, when this ratio is high (high inefficiency variation relative to noise), estimation by both methods works well. Both the FLW and KSW estimators do a good job at estimating efficiency scores and efficiency rankings. However, estimation by KSW dominates in nearly each of our methods of comparison. This makes sense as this is closer to what KSW is best suited for.

When the ratio is near unity (similar efficiency to noise variation ratio), neither method dominates. While both methods underestimate efficiency in this scenario, FLW does so by less than KSW for all but the most efficient firms. In other words, FLW is better at measuring efficiency scores for most firms (relative to the KSW estimator), but KSW does a better job at estimating the efficiency scores of the most efficient firms (relative to the FLW estimator). In addition, we find that both methods are good at predicting the rankings of the top and bottom performing firms, but do a very poor job at identifying the median efficient firms ranking. Finally, when the estimated ratio is small (low inefficiency variation relative to noise), both methods do a poor job. Neither method can be trusted for estimating efficiency scores or rankings. Our empirical examples reflect the findings in the Monte Carlo section.

It is important to note that the DEA estimator assumes the absence of noise. While we firmly acknowledge this point, this estimator is being notoriously applied in the literature in the presence of noise (every data tend to have noise) and hence we look at potential consequences when this is the case. We further note that in the case where no noise exists (see Section 3.5.2), the KSW estimator does an excellent job at estimating efficiency scores and rankings.

It is worth emphasizing that in practice the regulator does not know from the outset whether the variation of the noise prevails over the variation of the inefficiency, or vice versa. Nor does the practitioner know what model is more appropriate for a particular dataset. In this paper we attempt to provide a comprehensive yet simple guide to choosing the estimator for technical efficiency measurement. We do not favor one model over the other, and will let the data speak for themselves. For a given dataset we first check to see if it even makes sense to perform the analysis with DEA or SFA. Second, depending on the purpose of the analysis, we seek to determine what method should be applied.

The remainder of the paper proceeds as follows: Section 2 gives a short description of each estimation method as well as a methodological comparison. The third section provides the details of the Monte Carlo as well as the results of the simulations and several robustness checks. Section 4 summarizes the results of the simulations and gives a suggestion of how and when to perform efficiency estimation in practice. In Section 5, we use three empirical examples to see how the estimators work in practice and how the results in each of these data sets compare to the simulations. The final section concludes.

## 2 The competitors

In this section we present two efficiency estimators in the cross-sectional setting. The first is the SFA estimator of FLW which uses nonparametric kernel regression in the first stage to estimate the technology, and then uses the approach due to Aigner et al. (1977) to estimate the efficiency scores. The second is the DEA estimator of KSW which takes the standard DEA model of Charnes et al. (1978) and uses bootstrapping techniques to construct bias corrected efficiency scores. Each of the methods has its own merits both theoretically and empirically. In this section we first discuss each estimator and then give a brief methodological and conceptual comparison of the two methods.

### 2.1 (Nonparametric) estimation in a stochastic frontier framework

Historically, one of the drawbacks of SFA was that it was parametric in nature. It is well known that if the functional form in a parametric regression is misspecified then the parameter estimates will likely be inconsistent. This inconsistency can pass onto the

second stage in SFA, and have a detrimental impact on the estimation of the efficiency scores. Gong and Sickles (1992) have shown in Monte Carlo experiments that the performance of SFA is improved when the underlying technology is closer to that of the true underlying model. To address this concern when the technology is unknown, FLW use nonparametric kernel methods in the first stage.

FLW consider that the production frontier is a (smooth) unknown function of the  $p \times 1$  vector of regressors  $x$ , and the structure of the composite error is known. Specifically,

$$y_i = g(x_i) + v_i - u_i, \quad i = 1, 2, \dots, n, \quad (1)$$

where  $y_i$  is a scalar output variable,  $g(\cdot)$  is the unknown smooth function,  $v_i$  is the standard error disturbance and  $u_i$  is the one-sided error term representing technical inefficiency. Estimation of  $u_i$  can be accomplished in two steps. First, the conditional expectation of  $E(y_i|x_i)$  can be consistently estimated (see FLW) by nonparametric regression, yielding an estimate of the conditional mean  $\widehat{E}(y_i|x_i)$ . Note that  $g(\cdot)$  is not the conditional mean as  $E(y_i|x_i) = g(\cdot) + E(u_i|x_i) \equiv g(x_i) + E(u_i) \neq g(x_i)$ . We cannot separate the conditional mean from  $g(\cdot)$ . When running a nonparametric regression of  $y$  on  $x$  and we obtain  $\widehat{E}(y_i|x_i)$ , not  $\widehat{g}(x_i)$ . We recommend using local-linear least-squares estimation with the  $AIC_c$  method for bandwidth selection (Hurvich et al. (1998)) to estimate the conditional mean. Nonparametric estimation can be influenced by the choice of the bandwidth and regression type and so we have also run our simulations using different bandwidth choices (Kullback-Leibler and least-squares cross-validation) and different regression types (local-constant and local-linear). Although these variations brought about some individual changes, qualitatively the conclusions remain unchanged.

Second, by making distributional assumptions on both  $v$  and  $u$  we can construct the log-likelihood function to be maximized in order to obtain the efficiency scores. For example, following the normal, half-normal assumptions of Aigner et al. (1977), the concentrated log-likelihood function  $\ln l(\lambda)$  is maximized with respect to a single parameter  $\lambda$ . Specifically,

$$\max_{\lambda} \ln l(\lambda) = \max_{\lambda} \left\{ -n \ln \widehat{\sigma} + \sum_{i=1}^n \ln \left[ 1 - \Phi \left( \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}} \lambda \right) \right] - \frac{1}{2\widehat{\sigma}^2} \sum_{i=1}^n \widehat{\varepsilon}_i^2 \right\}, \quad (2)$$

where  $\widehat{\varepsilon}_i = y_i - \widehat{E}(y_i|x_i) - \mu(\widehat{\sigma}^2, \lambda)$ ,  $\mu(\lambda) = (\sqrt{2}\lambda\widehat{\sigma}) / [\pi(1 + \lambda^2)]^{1/2}$ , and

$$\widehat{\sigma} = \left\{ \frac{1}{n} \sum_{i=1}^n \left[ y_i - \widehat{E}(y_i|x_i) \right]^2 / \left[ 1 - \frac{2\lambda^2}{\pi(1 + \lambda^2)} \right] \right\}^{1/2}, \quad (3)$$

where  $\lambda = \sigma_u/\sigma_v$  and  $\sigma^2 = \sigma_u^2 + \sigma_v^2$ .

The obtained pseudo likelihood estimate of the variation ratio  $\widehat{\lambda}$  can be plugged in Eq. (3) to retrieve  $\widehat{\sigma}$ . Given  $\widehat{\lambda}$ ,  $\widehat{\sigma}$ , and  $\widehat{\varepsilon}_i$ , it is relatively easy to obtain point estimates of the technical efficiencies as well as their confidence intervals (see Horrace and Schmidt

(1996)). The point estimator of inefficiency via Jondrow et al. (1982) can be obtained from the mean of the conditional distribution of the marginal density  $f(u|\varepsilon)$ ,  $N(\mu_*, \sigma_*^2)$  as

$$E(u_i|\varepsilon_i) = \mu_{*i} + \sigma_* \left[ \frac{\phi(-\mu_{*i}/\sigma_*)}{1 - \Phi(-\mu_{*i}/\sigma_*)} \right], \quad (4)$$

where  $\mu_* = -\varepsilon\sigma_u^2/(\sigma_u^2 + \sigma_v^2)$  and  $\sigma_*^2 = \sigma_u^2\sigma_v^2/(\sigma_u^2 + \sigma_v^2)$ .

We also have interest in constructing confidence bounds for the technical efficiency scores. The lower and upper bounds of a  $(1 - \alpha) 100\%$  confidence interval for technical efficiency, modeled as  $\exp(-u) | \varepsilon$ , are given by

$$L_i = \exp(-\mu_{*i} - z_L\sigma_*) \quad (5)$$

and

$$U_i = \exp(-\mu_{*i} - z_U\sigma_*), \quad (6)$$

where  $Pr(Z > z_{Li}) = \frac{\alpha}{2} \left[ 1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right) \right]$  and  $Pr(Z > z_{Ui}) = (1 - \frac{\alpha}{2}) \left[ 1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right) \right]$ ,  $Z$  is distributed as  $N(0, 1)$ , and  $\Phi$  is the standard normal CDF. Hence,

$$z_{Li} = \Phi^{-1} \left( 1 - \frac{\alpha}{2} \left[ 1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right) \right] \right) \quad (7)$$

and

$$z_{Ui} = \Phi^{-1} \left( 1 - \left( 1 - \frac{\alpha}{2} \right) \left[ 1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right) \right] \right). \quad (8)$$

As the underlying technology is generally unknown in practice, one of the major benefits of this estimator is that the first stage is nonparametric. No distributional or functional form assumptions are needed to obtain a consistent estimate of the conditional mean,  $\hat{E}(y_i|x_i)$ . However, in the second stage, we must use distributional assumptions on  $v$  and  $u$  in order to obtain observation specific estimates of technical efficiency. Panel data can lead to a relaxation of these assumptions as in Kneip and Simar (1996), but that is beyond the scope of this paper.

## 2.2 (Bias-corrected) estimation in nonparametric frontier models

DEA has been nonparametric from its conception (Charnes et al. (1978)). However, the DEA estimator produces a biased estimate of the frontier. It can only measure the best practice technology. Specifically, the true frontier lies somewhere above the DEA estimated frontier. Hence, DEA estimates of efficiencies relative to the estimated frontier are too optimistic. The bootstrap procedures proposed by Simar and Wilson (1998), Simar and Wilson (2000) and KSW to correct this sampling bias use the idea



that the known distribution of the difference between estimated and bootstrapped efficiency scores mimics the unknown distribution of the difference between the true and the estimated efficiency scores. This relationship facilitates estimation of the bias and confidence intervals for the individual estimated efficiency scores.

In DEA, the functional form of the production process is not specified, it is solely observed that  $p$  input factors  $x$  are used to produce  $q$  outputs  $y$ , employing some technology  $\mathcal{T}$ , where

$$\mathcal{T} = \{(x, y) \mid x \text{ can produce } y\}. \quad (9)$$

Adopting the Shephard (1970) assumptions, we measure output oriented technical efficiency ( $\theta_j$ ) for a point  $(x_j, y_j)$  due to Farrell (1957) as

$$F_j^o(x_j, y_j) = \sup \{\theta_j \mid (x_j, y_j/\theta_j) \in \mathcal{T}\}, \quad (10)$$

which is the DEA estimator based on linear programming techniques.

In practice, the  $\mathcal{T}$ , and therefore  $F_j^o(x_j, y_j)$ , are unknown and must be estimated from a finite sample. Under arguably reasonable assumptions, KSW derive the asymptotic distribution of the DEA estimator and propose a way to perform statistical inference for the estimator given in Eq. (10) using bootstrapping techniques. If the true score ( $F^o$ ), comes from an unknown data generating process ( $\mathcal{P}$ ), and the estimator ( $\widehat{F}^o$ ), comes from a known data generating process  $\widehat{\mathcal{P}}(x, y)$ , then for a consistent bootstrap estimator ( $\widehat{F}^{o*}$ ), the following holds true

$$\left(\widehat{F}^{o*}/\widehat{F}^o - 1\right) \mid \widehat{\mathcal{P}}(x, y) \overset{\text{approximately}}{\sim} \left(\widehat{F}^o/F^o - 1\right) \mid \mathcal{P}. \quad (11)$$

This equivalence allows us to get closer to  $F^o$  and also suggests a way to derive the asymptotics of an estimator given by Eq. (10). KSW give a consistent bootstrap method (subsampling) which can be implemented in two steps. First, let (the integer)  $m = n^\kappa$  for some  $\kappa \in (0, 1)$ , where  $n$  is the sample size, and  $m$  is the size of the subsample. Second, perform the bootstrap as follows:

1. Generate a bootstrap subsample of size  $m$ ,  $S_m^* = \{(x_j^*, y_j^*)\}_{j=1}^m$ , by randomly drawing (independently, uniformly, and with replacement) from the original sample,  $S_n = \{(x_j, y_j)\}_{j=1}^n$ .
2. Apply the DEA estimator given by Eq. (10) where the technology  $\mathcal{T}^*$  is defined by the subsample drawn in step (1),  $S_m^* = \{(x_j^*, y_j^*)\}_{j=1}^m$ , to obtain the bootstrap estimate,  $\widehat{F}^{o*}$ .

Repeating steps (1) and (2) a large number ( $B$ ) of times (yielding  $\widehat{F}_b^{o*}$ ,  $b = 1, \dots, B$ ) allows us to approximate the conditional distribution of  $m^{2/(p+q+1)} \left(\widehat{F}^{o*}/\widehat{F}^o - 1\right)$ , the left hand side of Eq. (11), which in turn, allows us to approximate the unknown distribution

of  $n^{2/(p+q+1)} (\widehat{F}^o/F^o - 1)$ , the right hand side of Eq. (11). The bias-corrected DEA efficiency score is given by

$$\widehat{F}^o = \widehat{F}^o - \widehat{\text{bias}}_B, \quad (12)$$

where the bias is adjusted by employing the size  $m$  subsample

$$\widehat{\text{bias}}_B = \left(\frac{m}{n}\right)^{2/(p+q+1)} \left[ \frac{1}{B} \sum_{b=1}^B \widehat{F}_b^{o*} - \widehat{F}^o \right]. \quad (13)$$

Given the conditional distribution of  $m^{2/(p+q+1)} (\widehat{F}^{o*}/\widehat{F}^o - 1)$ , it is easy to find the quantiles  $\delta_{\alpha/2,m}$  and  $\delta_{1-\alpha/2,m}$  for an arbitrary  $\alpha \in (0, 1)$  such that

$$\text{Prob} \left[ m^{2/(p+q+1)} (\widehat{F}^{o*}/\widehat{F}^o - 1) \leq \delta_{\alpha/2,m} \right] = \frac{\alpha}{2} \quad (14)$$

and

$$\text{Prob} \left[ m^{2/(p+q+1)} (\widehat{F}^{o*}/\widehat{F}^o - 1) \leq \delta_{1-\alpha/2,m} \right] = 1 - \frac{\alpha}{2} \quad (15)$$

in order to compute a symmetric  $(1 - \alpha) 100\%$  confidence interval for  $F^o$  with  $\widehat{F}^o / (1 + n^{-2/(p+q+1)} \delta_{1-\alpha/2,m})$  and  $\widehat{F}^o / (1 + n^{-2/(p+q+1)} \delta_{\alpha/2,m})$  as lower and upper bounds.

One of the benefits of this estimator is that it allows for classical statistical inference. It improves upon the traditional DEA estimator by reducing the inherent bias and KSW have shown that it performs well in finite samples. However, it is left to the researcher to “decide” upon the value for  $\kappa$  (the size of the subsample). Although the bootstrap is consistent for all  $\kappa \in (0, 1)$ , there is *an* optimal value that makes the estimator more precise. The optimal  $\kappa$  depends on the dimensionality of the problem and therefore has to be set for each specific application (see KSW for details). We explore more of the benefits and shortcomings of each method in the next sub-section.

## 2.3 Methodological comparison

Prior to examining the finite sample properties of each method via simulations, it is useful to make conceptual comparisons. First, SFA was invented as an average practice analysis, while DEA is conceptually a best practice analysis. No observation point lies above the DEA frontier, while the constructed SFA frontier allows some data points to be above the frontier. The entire distance to the frontier in the DEA case is attributed to inefficiency, while SFA treats this distance as a combination of both inefficiency and statistical noise. Thus, it can be expected that one estimator will have an advantage depending upon the data generating process. We hope to view both scenarios. Furthermore, although these estimators are improvements over their predecessors: SFA using nonparametric methods and DEA using bootstrapping, the introduction of these “features” into DEA and SFA, does not change the “average” and “best practice” labels.

One benefit of SFA is the ability to handle the panel structure of data. For example, panel data can relax some of the functional form restrictions required of FLW (e.g., see

Kneip and Simar (1996)). Although attempts have been made to use DEA in a panel data setting (Henderson and Russell (2005); Ruggiero (2007)), we are unaware of a DEA estimator (with asymptotics proven) which can adequately handle this type of data. On the other hand, a benefit of DEA is that it can easily handle both multiple inputs and outputs. Under certain assumptions and data requirements, SFA can rely on distance functions, but these are not always met. In order to compare these estimators on a ‘relatively’ level playing field, we will focus on cross-sectional data with a single output. If the user is interested in using panel data or multiple outputs (in specific cases), the choice of estimator is obvious.

### 3 Simulations

In this section we examine the finite sample performance of each of our estimators. We are well aware that these have been examined for each estimator and that many comparisons have been made between traditional DEA and SFA estimators (for example, Gong and Sickles (1992), Banker et al. (1993), Thanassoulis (1993), Bojanic et al. (1998), Cubin and Tzanidakis (1998), Park and Lesourd (2000)). However, no study has compared these efficiency estimators against one another. Many comparisons in the literature often compare their particular model to simplistic counterparts. Given that we do not have an a priori preference for either estimator, we will try to present the most fair comparison possible. The goal here is to uncover useful information for the practitioner.

Specifically, we first discuss the basic design of the experiment. We define the data generating processes used, the distributional assumptions on the data as well as the other parameters of the experiment. Next, we discuss the methods by which we compare the performance of the efficiency estimators. We take several standard measures as well as one that we propose just for this experiment. Third, we run our simulations and discuss the relative performance of the estimators under the various scenarios. Finally, we briefly mention a subset of the vast number of robustness checks we tried and summarize the results of those runs. The results of these additional experiments are available in Appendices A-F.

We use the programming language R (R Development Core Team (2008)) for all simulations and empirical examples in the paper. To obtain the results for the FLW estimator we employ the np (Hayfield and Racine (2008)) package for kernel estimation and the bbmle package for obtaining an estimate of the parameter  $\lambda$ . For the KSW estimator we use the FEAR (Wilson (2008)) package. All R code and data are available from the authors upon request.

### 3.1 Design of the experiment

We conduct simulations for a production process which employs two inputs ( $x_1$  and  $x_2$ ) to produce a single output ( $y$ ). We consider two simple production functions: (1) Cobb-Douglas (CD)  $y = x_1^\alpha x_2^{\gamma-\alpha}$  and (2) constant elasticity of substitution (CES)  $y = [\beta x_1^\rho + (1 - \beta)x_2^\rho]^{\gamma/\rho}$ . For the primary analysis we assume the output is generated via a constant returns to scale (CRS) technology ( $\gamma = 1$ ). We set  $\alpha = 1/3$ ,  $\beta = 2/3$  and  $\rho = 0.5$ .

We assume that the true error term is distributed normally with mean zero and variance  $\sigma_v^2$ ,  $v \sim N(0, \sigma_v^2)$ . We further assume that the true technical efficiency is  $TE = \exp(-u)$ , where  $u$  is half-normally distributed with variance  $\sigma_u^2$ ,  $u \sim N^+(0, \sigma_u^2)$  and drawn independently of  $v$ . We introduce the noise and inefficiency in the production process of unit  $i$  as

$$y_i = x_{1i}^\alpha \cdot x_{2i}^{\gamma-\alpha} \exp(v_i - u_i) \quad (16)$$

in the CD case. Similarly, for the CES production function, the analog is

$$y_i = [\beta x_{1i}^\rho + (1 - \beta)x_{2i}^\rho]^{\gamma/\rho} \exp(v_i - u_i). \quad (17)$$

In an attempt to model various real-life scenarios, we have simulated four different combinations for the variation of the error and inefficiency terms. Table 1 shows the matrix of chosen combinations when  $\sigma_u$  and  $\sigma_v$  take values 0.01 and 0.05.

The first row of the table is for the case where the error variation is relatively small. In this case, the model has relatively little variation in the error term. The first column of the table is for the case where the variation of the inefficiency term is relatively small. In this case, we say that the decision making units are relatively efficient. The second row and second column represent the cases where the error variation and inefficiency variation are relatively large, respectively. We are specifically interested in particular cells of this table. In scenario s1 ( $\sigma_v = \sigma_u = 0.01, \lambda = 1.0$ ), both terms are relatively small. In other words, the data are measured with relatively little error and the units are relatively efficient. In scenario s2 ( $\sigma_v = 0.01$  and  $\sigma_u = 0.05, \lambda = 5.0$ ), the data have relatively little noise, but the units under consideration are relatively inefficient. In scenario s3 ( $\sigma_v = 0.05$  and  $\sigma_u = 0.01, \lambda = 0.2$ ), the data are relatively noisy and the the firms are relatively efficient. The fourth scenario s4 ( $\sigma_v = \sigma_u = 0.05, \lambda = 1.0$ ) is redundant as  $\lambda = 1.0$  as in s1. However, we show this case to emphasize that the results of the experiment depend upon the ratio of  $\sigma_u$  to  $\sigma_v$  and not their absolute values.

All experiments consist of 2000 Monte Carlo trials. Within each set of experiments, we analyze three sample sizes,  $n = 50, 100$ , and 200. For each Monte Carlo trial, we simulate a DGP by drawing observations for inputs  $x_1$  and  $x_2$  distributed uniformly on  $[1, 2]$ . For each of the four combinations of  $\sigma_u$  and  $\sigma_v$ , we then compute the ‘‘observed’’ output observations as in Eq. (16) and (17) in the CD and CES cases, respectively. We thus have four base scenarios for each of the assumed production functions.

KSW emphasized that the choice of the parameter  $\kappa$  that determines the subsample of the bootstrap is important for accuracy of the estimator and the optimal value varies depending on the dimensionality of the input/output setup. We have followed their suggestion and set this value at 0.7.

We again want to emphasize that the estimator in KSW does not allow for noise. That being said, we want to see the potential consequences of such an exercise. We consider the case of no noise in our DGP for both estimators in Section 3.5.2.

It is important to note that an efficiency model only makes sense if the distribution of  $(v - u)$  has negative skewness (Kumbhakar and Lovell (2003)). Therefore, we checked whether the skewness of the sampled distribution of  $(v - u)$  was negative and rejected those Monte Carlo replications that yielded the wrong (positive) skewness. When the wrong skewness was detected we went back and sampled again. In other words, the number of simulations (2000) was preserved. We also ran the results where we did not throw away runs with the wrong skewness. The results of this experiment showed that the nonparametric kernel model would (essentially) predict perfect efficiency for nearly all observations when  $(v - u)$  yields the wrong skewness. Including these cases led to efficiency scores which were close to unity.

## 3.2 Comparative measures

To compare the finite sample performance of our estimators we consider the following median (over the 2000 simulations) measures of

1. Bias(TE) =  $\frac{1}{n} \sum_{i=1}^n \left( \widehat{\text{TE}}_i - \text{TE}_i \right)$
2. RMSE(TE) =  $\left[ \frac{1}{n} \sum_{i=1}^n \left( \widehat{\text{TE}}_i - \text{TE}_i \right)^2 \right]^{1/2}$
3. Upward Bias (TE) =  $\frac{1}{n} \sum_{i=1}^n 1 \left( \widehat{\text{TE}}_i > \text{TE}_i \right)$
4. Kendall's  $\tau$  (TE) =  $\frac{n_c - n_d}{0.5n(n - 1)}$

where  $n_c$  is the number of concordant pairs, and  $n_d$  is the number of discordant pairs in the data set (efficiency ranks of FLW or KSW).  $1(A)$  denotes the indicator function that the event  $A$  is true,  $\widehat{\text{TE}}_i$  is the estimate of technical efficiency of unit  $i$  in a given Monte Carlo replication (by a given estimation method) and  $\text{TE}_i$  is the true efficiency score (determined by the DGP). We should note that we choose to use the median (over the 2000 simulations) instead of the average because it eliminates the chance that a few large values can skew the results. The mean results are similar and are available from the authors upon request.

As a further comparison, we report the empirical coverage accuracy (ECA) of each estimator. ECA reports the proportion of the true efficiencies that are within the bounds of the (predicted) 95% confidence intervals for the estimated technical efficiencies. In the tables we report the median of these proportions across the Monte Carlo simulations.

We are also interested in comparing the density estimates of technical efficiency across the Monte Carlo draws. Solely looking at the median (over the simulations) may hide interesting results. Specifically, for each draw, we sort the data by the relative value of the true efficiency. We are interested in comparing the true distribution of technical efficiency for a percentile across all Monte Carlo simulations. We therefore plot these (kernel) densities for the  $\alpha$ -percentile ( $\alpha n$ ), the median ( $n/2$ ) and the  $(1-\alpha)$ -percentile ( $(1-\alpha)n$ ) of the efficiency scores. These show us how well each of the estimators perform at estimating particular portions of our sample. For example, if interest lies in estimating the benchmark firms, then we would be interested in knowing which estimator does the best job at estimating the  $(1-\alpha)$ -percentile of the efficiency distribution.

We should note that the results likely include sampling variation apart from what we would find for individual, fixed points that do not change from one trial to the next. To state things differently, the  $\alpha$ -percentile firm in trial 1 is likely different from the  $\alpha$ -percentile firm in trial 2, and so on. An alternative approach could focus on one or perhaps a small set of fixed points in the input-output space that are held constant over Monte Carlo trials. We have conducted limited simulations which show that these two approaches lead to qualitatively similar results and these results are available upon request.

Finally, we also plot the distribution of the ranks for each of these percentiles. This will show us how well each estimator does at predicting the correct rankings of the firms. These densities should be centered around  $\alpha n$ ,  $n/2$ , and  $(1-\alpha)n$ . In this paper we set  $\alpha = 0.05$ . We note that most studies surrounding efficiency scores essentially look at average behavior. It will be shown that we uncover different findings depending upon which percentile is being estimated. In other words, the relative performance of our estimators are heterogenous across the sample.

### 3.3 Baseline case

In this sub-section we provide our baseline results. The results are given in Tables 2 and 3 and Figures 1-4. Table 2 as well as Figures 1 and 3 correspond to the CD technology. Similarly, Table 3 as well as Figures 2 and 4 are for the CES production function. Recall that KSW represents the DEA estimator and FLW represents the SFA estimator. Here we consider the case where the distributional assumptions on the error terms in the FLW model are correctly specified. Many of these assumptions will be relaxed later.

### 3.3.1 Estimation of technical efficiency

Tables 2 and 3 show the performance measures of the technical efficiency estimates for the CD and CES technologies, respectively. In Table 2, FLW generally tends to outperform KSW in terms of ECA. In Table 3, this dominance by FLW is uniform across the scenarios and sample sizes. Moreover, in Table 3, ECA improves with the sample size for FLW, but gets worse with KSW. A potential answer for this phenomenon will become apparent later.

Turning to the bias of the technical efficiency scores, we note that the median bias is negative in nearly all cases. This implies that both of our estimators (at the median) tend to underestimate the true technical efficiency scores. In terms of RMSE, the RMSE of technical efficiency is relatively small when  $\sigma_v$  is small. The opposite result holds when  $\sigma_v$  is relatively large. A peculiar result emerges for the KSW estimator. Here we see that the RMSE of the technical efficiency estimates increases with the sample size in each scenario (recall that KSW assume the absence of noise).

In addition to the median bias, we also consider an alternative approach to compare across estimation methods. Specifically, we give an estimate of upward bias. The upward bias measure shows the percentage of points for which technical efficiency has been overestimated. The measure checks whether an estimate is larger than the true value. If so, then the indicator function returns a value of one. If not, then a zero value is given. The number of estimates above the truth are then divided by the sample size to give the percentage of estimates which were over estimated. Values near 0.5 are ideal for this measure. Anything in excess of one-half suggest an upward bias and anything less than a half suggests a downward bias. For both estimators we see that as the sample size goes up, the percentage of over-estimated results goes down. In other words, given that the desired value of upward bias is 0.5, both FLW and KSW systematically underestimate the true technical efficiency.

It is difficult to conclude from the tables whether or not one method underestimates more than the other. Therefore, we look at other percentiles of the efficiency distribution as well. Figures 1 and 2 show the distributions of the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles of the true and estimated technical efficiencies. Each row contains three panels corresponding to the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles of efficiency for a particular scenario. For visual simplicity, we have chosen to show only the case when  $n = 200$ . The other figures do not differ much and are available upon request.

Figures 1 and 2 confirm that KSW performs best in scenario s2. Here we see that the KSW densities are very close to the true distributions of efficiency. This is expected as this estimator is more closely designed for this particular scenario (again, KSW assumes no noise). FLW systematically underestimates the efficiency of the chosen percentiles. The mode of the FLW estimator is anywhere from 3 to 13 percentage points lower than the true mode.

Scenario 2 is the only obvious case between estimators here. In contrast, when  $\lambda$  is unity (s1 and s4), FLW and KSW consistently underestimate (each of the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles) the efficiency scores. If we examine the mode of the densities, FLW appears to be closer to the truth when looking at the 5<sup>th</sup> or 50<sup>th</sup> percentile, but the opposite is true at the 95<sup>th</sup> percentile. What this says (for s1 and s4) is that KSW is better at determining the benchmark firms, but FLW is better at estimating a majority of the efficiency scores. It is up to the user to determine which method is needed for their particular problem. That being said, we should note that the spread of the efficiency estimates is much larger for FLW than KSW. In other words, FLW sometimes confounds noise with inefficiency.

Perhaps the most discouraging result is for scenario s3. Here we have the case where the error variation is relatively large and the inefficiency variation is relatively small. We see that when  $\lambda$  is relatively small that both estimators do a poor job at predicting efficiency scores. It can be argued that FLW is slightly better when looking at the 5<sup>th</sup> or 50<sup>th</sup> percentile and KSW is relatively better when looking at the 95<sup>th</sup> percentile, but neither does a good job. It appears that both estimators break down when trying to predict inefficiency here. Neither estimator is able to correctly draw efficiency scores when the noise variation is large relative to that of the inefficiency variation. We tried other combinations which produce the same value for  $\lambda$  and as expected we find that the results depend only upon parameter  $\lambda$  and not the absolute values of  $\sigma_v$  and  $\sigma_u$ .

Finally, we would like to note that plotting the densities also gives us an explanation to why ECA decreases in KSW and increases in FLW with the sample size. As  $n$  increases (not shown) we get more probability mass around the true means of the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles. In other words, as the variation of the true efficiencies decreases with the sample size, given that the KSW performs poorly with a small sample, when the sample size is increased, it does even worse. On the other side, given that FLW is better in predicting certain percentiles in small samples, when  $n$  increases, FLW gets better. This helps validate our decision to examine specific percentiles of the efficiency distribution.

### 3.3.2 Efficiency rankings

The last element of Tables 2 and 3 give Kendall's  $\tau$  for the efficiency ranks between the true and estimated efficiency scores. In each scenario, for each sample size, FLW has a large Kendall's  $\tau$ . This measure suggests that FLW does a slightly better job at identifying the ranks of efficiency scores.

We again note that the values in Tables 2 and 3 correspond to median behavior. Figures 3 and 4 look for differences at particular percentiles of the efficiency (ranking) distributions. These figures show the empirical distribution of the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles of the true (vertical line) and estimated ranks. Each row contains three panels corresponding to the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles of ranks for a particular



scenario. Ideally we would like to see the densities lie on top of the vertical (true) line with minimal variation. Again we show only the case of  $n = 200$  for visual simplicity.

The first thing to note in the figures is that the FLW and KSW densities are difficult to distinguish from one another. The best case scenario for both estimators is s2. Here we see that regardless of the percentile, both estimators do a good job at identifying the ranks. In other words, in the s2 world, we can be somewhat confident that the rankings provided by both estimators are correct. However, the results for s3 are alarming. Recall that the efficiency scores were poorly estimated in s3. Here we see that these densities are more or less uniform over the observations. What this means is that both estimators have a non-trivial probability of ranking the 95<sup>th</sup> percentile as the 5<sup>th</sup> percentile, and vice versa. In other words, the practitioner should place relatively little confidence in efficiency rankings from either estimator when  $\lambda$  is relatively small.

On the other hand, when  $\lambda$  is relatively large (s1 and s4), we have both good and bad outcomes. Both estimators appear to do a good job at identifying the best and worst performing firms. However, neither estimator can accurately predict the median efficient firm. Therefore, if interest lies in detecting the best or worst performing firms, then either estimator may be sufficient in this scenario. However, if the goal is to do a ranking of all firms across the sample, those ranks in the middle are questionable.

### 3.4 Finite sample properties of lambda

The conclusions of our simulations depend heavily upon the value of  $\lambda$ . Although FLW give simulation results and claim that the estimator of  $\lambda$  is root- $n$  consistent, we feel the need to show the finite sample properties of our estimates of  $\lambda$  across our simulations. The first three columns of Tables 4 and 5 present the median value of the difference between the estimated and true values of  $\lambda$ , its interquartile range (IQR), and median of the squared difference between the estimated and true values of  $\lambda$  across our 2000 simulations for each of our scenarios for each sample size. It can be seen that as the sample size rises, all three measures fall. Additionally, three last columns of the Tables 4 and 5 give the sizes of the bootstrapped tests for testing the null that estimated value of  $\lambda$  equals the true value of  $\lambda$  at 1%, 5%, and 10% levels. We use 500 bootstrap replications in each of our 2000 Monte Carlo simulations. The numbers in the tables are close to the nominal values suggesting neither systematic under- nor overrejection of the null. The sizes of the tests in a battery of robustness checks that we describe below are generally in line with the baseline cases. As expected, there some distortions when the true distribution of  $u$  is Gamma. However, when the true noise is heteroskedastic, the estimated sizes are close to nominal sizes.

We note here that we choose to use median (over the 2000 simulations) estimates here and throughout the paper because kernel methods occasionally give large (in absolute terms) values in the tails. For the case of  $\lambda$ , for each scenario, one or two of the 2000

simulations would give an estimated value of  $\lambda$  which was extremely large (sometimes in excess of 50) and this skews mean square error values.

### 3.5 Robustness checks

The baseline results above are informative, but we can often wonder whether they are truly representative of the estimators. Therefore, in this sub-section, we consider several robustness checks to make sure that our simulations accurately depict the performance of our estimators. The results for each of these robustness checks are given in Appendices A-F.

#### 3.5.1 Changes in true distribution of error term

To check the responsiveness of both estimators to changes in the distribution of the true technical efficiencies, we alternatively draw the inefficiency term ( $u$ ) from the gamma distribution,  $u \sim \Gamma(k, \theta)$ . It is expected that this will not qualitatively change the results for KSW as the estimator does not require an assumption on the distribution of the error term. However, this is not the case for FLW. Here we consider the case where FLW is misspecified in the second stage. In other words, while we change the distribution of  $u$  from half-normal to gamma, we still estimate the FLW model assuming the distribution of  $u$  is half-normal. Within our experiments, we examine three shapes of the gamma distribution:  $k \in \{\exp(-1), 1, \exp(1)\}$ . To make this Monte Carlo extension comparable to our baseline results, we set  $\theta$  equal to (36.79, 100, 271.83) so that  $\sigma_u = 0.01$  and equal to (7.36, 20, 54.37) so that  $\sigma_u = 0.05$  (recall that the variance of the gamma distribution is  $k\theta^2$ ).

Not surprisingly, the results (Appendix A) for the exponential case where  $k = 1$  are nearly identical to the half-normal case discussed in the previous sub-section. The true gamma distribution of the inefficiency for  $k = \exp(-1)$  looks more “squeezed” than the half-normal and therefore exposes much less true inefficiency. Conceptually, even if the standard deviation remains equal to that of the half-normal case, all scenarios get “closer” to the s3 scenario for the half-normal case, which proved to be difficult for both estimators. We thus expect deterioration in each estimators’ performance.

The bias of technical efficiency is still negative, implying aggregate underestimation of true scores. As expected, now that the distribution of the true inefficiency is “squeezed”, both estimators perform poorly at predicting the 95<sup>th</sup> percentile of the true efficiencies. Further, neither estimator does a good job at identifying the true ranks. In short, with some minor quantitative changes, the results from the baseline case seem to generally hold.

When considering the case where  $k = \exp(1)$ , the median inefficiency gets larger and improves the performance of both estimators. The main change from the baseline case is that the bias of KSW is now positive in scenario s2. Thus, KSW tends to overestimate

when there is little noise especially at higher true levels of efficiency. The upward bias measure is now in excess of 0.5 for KSW in scenario s2 but falls as sample size increases.

In short, we find that changing the true distribution of  $u$  can have an effect on the results of the exercise. However, the results still depend upon the estimated value of  $\lambda$ . Larger values lead to better estimation for both models, smaller values lead to worse estimation for both models and values near unity come with mixed results.

### 3.5.2 No noise

We have stressed in the paper that the KSW estimator assumes the absence of noise. However, our simulations include noise in each scenario. Further, we saw that when the efficiency to noise ratio was high, the KSW estimator performed well.

Here we consider the case where there is no noise in the DGP. The results in Appendix B show that KSW clearly dominates the FLW estimator. This is expected as the KSW estimator assumes no noise and the FLW estimator assumes noise exists. Given that the FLW estimator assumes noise exists, it appears to treat some of the inefficiency as noise and hence further underestimates technical efficiency. That being said, the FLW efficiency estimates are similar to (but smaller than) those in scenario s2. In fact, the upward bias measure is close to zero. On the other hand, for the KSW estimator, this measure is close to the desired value of 1/2. In short, when no noise exists, we recommend using KSW.

### 3.5.3 Returns to scale

As noted before, all of our DGP's assume a CRS technology. We also consider the case where the technology exhibits decreasing returns to scale ( $\gamma < 1$ ). Specifically, we consider the case where  $\gamma = 0.8$ . We ran this specification for various technologies and report the results most closely related to the baseline case in Appendix C. We expected the performance of the KSW estimator to deteriorate as we employed the KSW efficiency measure which assumes CRS. However, the results only differed slightly. For both this modified baseline case and others, we do not see any changes to the qualitative results of the experiment.

### 3.5.4 Changes in technology

We also performed simulations using the more complicated technology in Guilkey et al. (1983):

$$ye^{\theta y} = [\delta_1 x_{1i}^{\rho_1} + \delta_2 x_{2i}^{\rho_2}]^{\gamma/\rho}. \quad (18)$$

This relatively flexible (as compared to CD and CES) technology allows for different elasticities of substitution between inputs by tweaking the parameters  $\rho_1$ ,  $\rho_2$ , and  $\rho$  (see Guilkey et al. (1983)). We tried various combinations of the parameters  $\rho_1$ ,  $\rho_2$ , and  $\rho$ .

The results of this experiment appear in Appendix D. Even though this is a more complicated technology, we do not see any major differences from the baseline results. This is likely the case because both estimators are nonparametric and thus make no assumptions on the functional form. Hence, both are equally good at estimating a more complicated technology as compared to their respective performances in the baseline results.

### 3.5.5 Heterogenous parameters

The choice of homogenous parameters has been questioned in empirical work (see Hastie and Tibshirani (1993)). It can be argued that firms differ both in their intercepts and their slope parameters. To see whether or not this has an effect on the results of our study we considered the Cobb-Douglas DGP in Eq. (16) where the coefficient  $\alpha$  was unique to each  $i$ . Specifically, we generated  $\alpha = 1/3$  plus a mean zero normally distributed term with standard deviation 0.01. We report the results where we allow  $\alpha_i$  to vary across Monte Carlo replications and note that the results do not differ significantly if we fix  $\alpha_i$  across replications. CRS was maintained. The results of this experiment can be found in Appendix E. The qualitative results do not change. This is expected as nonparametric methods do not make assumptions of homogenous parameters.

### 3.5.6 Heteroskedasticity

Our final robustness check is to introduce heteroskedasticity in the error term. This is potentially problematic for the FLW estimator as their second stage assumes the variance components are constant. Specifically, we have generated our error term ( $v_i$ ) as before except allow the variance to be a function of  $x_1$  and specific to each observation  $i$  (with varying degrees of heteroskedasticity). The results of this experiment can be found in Appendix F. We find a slight decrease in the upward bias measure for each estimator relative to the baseline results. We do not find any other qualitative differences in the results.

## 4 Which estimator and when

Throughout our simulations and robustness checks we have seen that our results depend upon the relative sizes of  $\sigma_u$  and  $\sigma_v$ . Assuming this holds true for other data generating processes not considered, we want to offer a simple outline of when to use efficiency analysis and what estimator to be employed in various scenarios.

The practitioner should of course start by checking whether there is evidence of technical inefficiency in the data. One way to do this is to perform the test due to Schmidt and Lin (1984). Next, recall that we essentially have three different possible

outcomes: (1) The value of  $\lambda$  can be near unity. This occurs when the variation in the error ( $\sigma_v$ ) and efficiency ( $\sigma_u$ ) are of similar magnitudes (scenarios s1 and s4). (2) The value of  $\lambda$  is relatively large. This occurs when the variation from the efficiency term is large relative to that of the noise (scenario s2). (3) Finally, the value of  $\lambda$  can be relatively small. Here we have scenario s3, where the error variation is large relative to the variation of the efficiency term.

We argue that the first step in any efficiency analysis should be to check the estimated value of  $\lambda$  (as estimated by FLW). If the estimated value of  $\lambda$  is relatively small, then, as we have seen, the efficiency scores and ranks will likely be poorly estimated. In this case, we do not recommend efficiency estimation with these particular methods. This does not imply that all firms are perfectly efficient. It is just difficult to extract the inefficiency from the (relatively large) noise component. In essence, studies which conduct efficiency in this scenario may be unreliable.

The second scenario is the case where  $\lambda$  is relatively large. Here both estimators do a good job at estimating efficiency and estimating the ranks. Here, KSW appears to be dominant in terms of estimating efficiency. On the other hand, even though there appears to be a slight performance improvement for FLW in terms of the rankings, KSW is not much different and should be considered reliable in this situation as well. In short, the KSW estimator should get the nod in general for this scenario, but FLW will not necessarily produce poor estimates. As we will see in the empirical section, for the scenario in which  $\lambda$  is large, both estimators produce similar efficiency scores and we find a high correlation between the (efficiency) rankings of the estimators.

The more difficult scenario to detect is when the ratio is near unity. Here we do not see a well defined winner between the two estimators, nor can we be confident of our rankings for a significant portion of our samples. Figures 1 and 2 show that when examining the 5<sup>th</sup> or 50<sup>th</sup> percentile of the efficiency scores that the FLW estimator does a better job at estimating the true efficiency score. That being said, the estimator generally gives efficiency scores which are too small. On the other hand, when examining the 95<sup>th</sup> percentile of the efficiency scores, the KSW estimator does a better job at estimating the true efficiency scores. Again, KSW underestimates the true efficiency score at each percentile shown. When moving to the rankings, it is obvious that both estimators do a decent job at ranking the top and bottom performers, but both do a poor job at identifying the median rank. In short, FLW appears to be better at identifying the worst performers and KSW appears to be better at identifying the best performers. However, a close examination of the plots of the rankings show very minor differences and we do not want to place too much emphasis on the differences between the estimators here. What we do want to point out is that trying to use this to pick rankings at the median is difficult at best.

To summarize this discussion, we have reported these results into two tables: Table 6 for efficiency scores and Table 7 for rankings. Table 6 gives a summary of the results for each scenario and a suggestion for which estimator to use. The first row for each

combination of  $\sigma_u$  and  $\sigma_v$  shows the relative magnitude of the estimates. For example, when  $\sigma_u = 0.01$  and  $\sigma_v = 0.01$ , FLW and KSW are relatively similar in terms of efficiency scores. Specifically, sometimes FLW is larger and sometimes KSW is larger. However, both are smaller than the true efficiency scores. Similarly, for scenario s2 ( $\sigma_u = 0.05$  and  $\sigma_v = 0.01$ ), FLW typically produces smaller efficiency scores than KSW which is relatively close to the truth. The second through fourth rows corresponding to each scenario are our suggestion for which estimator to use depending on which part of the population the researcher is concerned with estimating. For example, when  $\lambda$  is near unity, we suggest that authors use FLW for all but the best decision making units. For those near the top, KSW is preferable. Finally, we want to note that even though we have suggestions for s3, we do not recommend performing efficiency analysis with these procedures for this scenario.

Table 7 is not divided between FLW and KSW. We do not give a recommendation between these two estimators for ranks as they are quite similar. We conclude that they are (roughly) equally good or equally bad in estimating rankings under each scenario. Here we just provide the situations in which a practitioner can trust the rankings of the efficiency scores. In the table, for each scenario, we report the performance of the estimators in terms of determining the true ranking at different percentiles of the efficiency distribution. For example, when  $\sigma_u$  and  $\sigma_v$  are of similar magnitude (s1 and s4), both estimators do a fair job at identifying the best and the worst performing decision making units. However, each does a relatively poor job at identifying the median performer. At the same time, under scenario s2, both estimators do a good job at identifying the rankings. As expected, both estimators do a poor job when  $\lambda$  is relatively small.

To further simplify the results, we would like to offer the following four step approach to choosing an estimator

1. Estimate the model via FLW and obtain an estimate of  $\lambda$ .
2. Refer to Table 1 to determine what scenario your data lies in.
3. Check to see if efficiency estimation is recommended.
4. If your data pass step 3, use Tables 6 and/or 7 to help choose the appropriate estimator for your particular needs.

Here we argue that a one size fits all approach is inappropriate for efficiency analysis. When using these results for policy or in court, the user must be aware of which observations are of particular interest and whether or not interest lies in the estimate or the ranking. We have shown that for most (but not all) cases the estimators underestimate the efficiency. This should be considered if setting a threshold for firms. We have also shown when rankings are reliable and when they are not. Therefore we stress that these four steps be taken before discussing the results from an efficiency exercise.

For a practical example, we note that for the case mentioned in the Introduction, identifying a benchmark firm(s) is often important to regulators. A benchmark firm(s) (say top 5%) is often used to decide the penalty (or carrot) for the bottom firms in yardstick competition. In such a case, it is important to accurately estimate the top 5% and bottom 5% of firms. We note that identification of both the top 5% and bottom 5% of firms is feasible in scenarios s1, s2 and s4. In general, FLW would produce more reliable results for the bottom 5% (except for s2) and KSW would generally produce more reliable results for the upper 5%.

## 5 Empirical applications

We have shown the performance of the estimators in Monte Carlo simulations, but it is also worthwhile considering how the estimators work in practice. Here we consider three separate datasets used in empirical papers which correspond to estimated values of  $\lambda$  which are less than, greater than and close to unity. We report the parameters from the FLW nonparametric regression of each model in Table 8. In this table we also report a measure for  $R^2$  of the first stage regression in FLW, defined as the Kendall correlation coefficient between the predicted and true output. Table 9 provides the summary statistics of the technical efficiencies whereas the panels in Figure 5 give the kernel densities of the estimated technical efficiencies. Finally, Table 10 presents the rank correlation coefficients between the technical efficiency estimates of the two estimators.

### 5.1 Utilities data

The first data set comes from Kumbhakar and Tsionas (2011), hereafter KT. KT estimate a production function for 72 fossil fuel-fired steam electric power generating plants in the United States over the period 1986-1999. Specifically, they model output (net steam electric power generation in megawatt-hours) as a function of labor, fuel and capital. Quantities of labor are calculated by dividing the aggregate costs of labor by a cost-share weighted price for labor. The fuel quantities are calculated by dividing the fuel expenses by the Tornqvist price of fuel aggregate. The values of capital stocks are calculated by the valuation of base and peak load capacity at replacement cost to estimate capital stocks in a base year and then updating it in the subsequent years based upon the value of additions and retirements to steam power plants.

Here we consider the year 1998 (the year with the most observations), where the sample size  $n$  is 81 firms. Note that using other years does not change the qualitative results of the experiment. The first thing to notice in Table 8 is the relatively large estimate of  $\lambda = 4.0961$  with bootstrapped standard error 1.3206 (we used 1000 bootstrap replications). This puts us in scenario s2. Recall that this scenario leads to the most reliable results for technical efficiency estimates. The summary statistics for the efficiency estimates can be found in Table 9. Notice that the estimates for all points considered

are relatively similar between the estimation methods. This is further shown in the first panel of Figure 5. Other than the small dip in the upper tail of the distribution, these densities are quite similar.

The correlation rankings can be found in Table 10. Here we consider three different correlation coefficients: Kendall, Pearson and Spearman. Each of them show a relatively high correlation between the rank estimates of the two estimators. Even with this relatively small sample ( $n = 81$ ), both estimators appear to be very similar. We expect that these estimated efficiencies are similar to the true efficiencies.

## 5.2 Manufacturing industry database

The second data comes from the NBER-CES Manufacturing Industry Database (May, 2009), which is compiled by Randy A. Becker, and Wayne B. Gray. This database is a joint effort between the National Bureau of Economic Research and U.S. Census Bureau's Center for Economic Studies, containing annual industry-level data on output, employment, payroll and other input costs, investment, capital stocks, total factor productivity, and various industry-specific price indexes (<http://www.nber.org/data/nbprod2005.html>). This data has been used in many studies (e.g. Berndt and Morrison (1995), Amato and Amato (2000)). We consider a production function,  $Y = f(K, L, M)$ , where  $Y$  is defined as the total value of shipments divided by a respective deflator,  $K$  is the total real capital stock,  $L$  is the total employment, and  $M$  is the total cost of materials, electric and fuels divided by a respective deflator. We consider the year 2004 for our analysis, but note that other years produce similar results.

For the year 2004 (the most recent year available), we have  $n = 473$  firms. Our estimate of  $\lambda$  that is significantly less than unity (0.0016) and indistinguishable from zero when considering the bootstrapped standard error (0.7271). This does not imply that the regression is poor. In fact, the  $R^2$  of this regression is in excess of 0.90. However, we caution that the results of this exercise may be suspect given what we found in our simulations. Table 9 gives the descriptive statistics of the efficiency estimates for each estimation. The results are striking. Here we see that the minimum and maximum efficiency score for FLW are the same to four decimal places. Each of the values here show the efficiency score to be equal to 0.9998 for each observation. It is difficult to believe that all firms are (nearly) perfectly efficient. At the same time, KSW show that the efficiency scores go from 0.0774 to 0.7621. It is obvious that these results conflict one another. Our belief, based on our Monte Carlo simulations, is that neither is likely correct. This is a prime example of why it is problematic to conduct efficiency analysis when the estimated value of  $\lambda$  is significantly less than unity.



### 5.3 Forestry data

Our final example concerns a production function for the timber industry. Specifically, we take the data from Lien et al. (2007), hereafter LSB, and estimate output as a function of labor, forest area cut and capital. These data come from the Sample Survey of Agriculture and Forestry, compiled by Statistics Norway in 2004. All data are for the year 2003 and the sample consists of  $n = 3249$  active forest owners. Output is measured as the annual timber sales from the forest. Labor is obtained as the hours worked by the owner, his/her family or hired labor plus the hours worked by contractors. Forest area cut is measured in hectares. This variable only includes the total hectares cut, not the total number of hectares the owner possesses. Finally, capital is an estimate of the value of the increment from the forest.

The results from the nonparametric regression of output on labor, forest area cut and capital can be found in Table 8. Here we see that this model produces the highest goodness-of-fit measure of our three empirical data sets ( $R^2 = 0.9339$ ). The next point of interest from the table is  $\hat{\lambda}$ , which is close to unity (1.2222), with bootstrapped standard error 0.5157. This roughly corresponds to scenarios s1 and s4 from above. Recall that in these scenarios, FLW did a better job at estimating the lower end of the distribution and KSW did a better job at estimating the higher end of the distribution of efficiency scores. Table 9 gives us a description of the efficiency scores. First, note that the minimum efficiency score shown by KSW (0.1524) is much smaller than that reported by FLW (0.5518). Given what we have seen from above, we believe that KSW severely underestimates the true efficiency of this particular farmer. At the same time, both the median and the upper quartile ( $Q3$ ) are very different between the two estimators ( $\approx 0.5$  each). It is also likely that KSW underestimates these efficiency scores as well. However, at the high end, both estimators give similar efficiency scores for the top performer and we believe that KSW provides more reliable scores for this particular observation.

## 6 Conclusion

In this paper, we have attempted to provide a careful comparison of two efficiency estimators. The first estimator, by FLW improved upon past SFA estimators, by estimating the conditional mean nonparametrically. The estimator by KSW, improved upon past DEA estimators, by introducing asymptotics via bootstrapping. Both approaches are well studied in the literature, but little is known about their relative performance against one another.

We first gave a brief description of each method and discussed their methodological differences. We then used an extensive set of Monte Carlo simulations to determine when each estimator performed best. We found that when the variation ratio of efficiency to noise was large, the KSW estimator performed very well with all measures (noting that KSW assumes the absence of noise). Alternatively, when that ratio was small, neither

estimator did a very good job at estimating efficiency or efficiency ranks. When the ratio was near unity, we saw benefits of both the KSW and FLW approaches for particular observations of interest. Further, we provided a four step procedure to suggest where, when and how to perform efficiency estimation in a cross-sectional setting with a single output. We hope that this proves to be valuable to academic researchers as well as practitioners. Finally, we used the methods in three separate empirical exercises. Each of the exercises could be compared back to our simulations and we were able to provide some information of the “potential” reliability of each of the estimates.

We hope that the information provided in this paper can pinpoint some of the perceived problems associated with efficiency analysis as pointed out by Stone (2002) and others. Efficiency analysis is a large and growing field and authors need to be aware of where and when to use particular estimators. Failing to go through these checks could lead to estimates which are essentially meaningless. Policies based on these estimates could lead to detrimental outcomes.

## Acknowledgements

The authors would like to thank two anonymous referees, an anonymous associate editor, the editor Simon Day, Vasyl Golosnoy, Helmut Herwartz, Bill Horrace, Uwe Jensen, Kristiaan Kerstens, Roman Liesenfeld, Chris Parmeter, Subhash Ray and John Ruggiero as well as participants of the Seminar on Statistics and Econometrics at the Institute for Statistics and Econometrics (Christian Albrechts University at Kiel), the 2010 North American Productivity Workshop (Houston, TX) and the 2010 Annual Meeting of the Southern Economic Association (Atlanta, GA) for useful comments which led to a substantially improved version of this paper. All data and R code for this paper are available from the authors upon request.

## References

- Aigner, D., Lovell, C. K. and Schmidt, P.: 1977, Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics* **6**(1), 21–37.
- Amato, L. H. and Amato, C. H.: 2000, The impact of high tech production techniques on productivity and profitability in selected U.S. manufacturing industries, *Review of Industrial Organization* **16**(4), 327–342.
- Banker, R. D., Gadh, V. M. and Gorr, W. L.: 1993, A monte carlo comparison of two production frontier estimation methods: Corrected ordinary least squares and data envelopment analysis, *European Journal of Operational Research* **167**(3), 332–343.
- Berndt, E. R. and Morrison, C. J.: 1995, High-tech capital formation and economic performance in U.S. manufacturing industries: An exploratory analysis, *Journal of Econometrics* **65**(1), 9–43.
- Bojanic, A. N., Caudill, S. B. and Ford, J. M.: 1998, Small-sample properties of ML, COLS, and DEA estimators of frontier models in the presence of heteroscedasticity, *European Journal of Operational Research* **180**(1), 140–148.
- Charnes, A., Cooper, W. W. and Rhodes, E.: 1978, Measuring the efficiency of decision-making units, *European Journal of Operational Research* **84**(4), 655–676.
- Cubbin, J.: 2005, Efficiency in the water industry, *Utilities Policy* **13**(4), 289–293.
- Cubbin, J. and Tzanidakis, G.: 1998, Regression versus Data Envelopment Analysis for efficiency measurement: an application to the England and Wales regulated water industry, *Utilities Policy* **7**(2), 63–119.
- Dassler, T., Parker, D. and Saal, D. S.: 2006, Methods and trends of performance benchmarking in UK utility regulation, *Utilities Policy* **14**(3), 166–174.
- Fan, Y., Li, Q. and Weersink, A.: 1996, Semiparametric estimation of stochastic production frontier models, *Journal of Business & Economic Statistics* **14**(4), 460–468.
- Farrell, M. J.: 1957, The measurement of productive efficiency, *Journal of the Royal Statistical Society. Series A (General)* **120**(3), 253–290.
- Gong, B.-H. and Sickles, R.: 1992, Finite sample evidence on the performance of stochastic frontiers and Data Envelopment Analysis using panel data, *Journal of Econometrics* **51**(1-2), 259–284.
- Guilkey, D. K., Lovell, C. A. K. and Sickles, R. C.: 1983, A comparison of the performance of three flexible functional forms, *International Economic Review* **24**(3), 591–616.

- Hastie, T. J. and Tibshirani, R. J.: 1993, Varying-coefficient models, *Journal of the Royal Statistical Society Series B-Methodological* **55**(4), 757–796.
- Hayfield, T. and Racine, J. S.: 2008, Nonparametric econometrics: The np package, *Journal of Statistical Software* **27**(5), 1–32.  
**URL:** <http://www.jstatsoft.org/v27/i05/>
- Henderson, D. J. and Russell, R. R.: 2005, Human capital and convergence: A production frontier approach, *International Economic Review* **46**(4), 1167–1205.
- Horrace, W. C. and Schmidt, P.: 1996, Confidence statements for efficiency estimates from stochastic frontier models, *Journal of Productivity Analysis* **7**(2-3), 257–282.
- Hurvich, C. M., Simonoff, J. S. and Tsai, C. L.: 1998, Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B* **60**(2), 271–293.
- Jondrow, J., Knox Lovell, C. A., Materov, I. S. and Schmidt, P.: 1982, On the estimation of technical inefficiency in the stochastic frontier production function model, *Journal of Econometrics* **19**(2-3), 233–238.
- Kneip, A. and Simar, L.: 1996, A general framework for frontier estimation with panel data, *Journal of Productivity Analysis* **7**(2-3), 187–212.
- Kneip, A., Simar, L. and Wilson, P. W.: 2008, Asymptotics for dea estimators in non-parametric frontier models, *Econometric Theory* **24**, 1663–1697.
- Kumbhakar, S. C. and Lovell, C. A. K.: 2003, *Stochastic Frontier Analysis*, number 9780521666633 in *Cambridge Books*, Cambridge University Press.
- Kumbhakar, S. C. and Tsionas, E. G.: 2011, Stochastic error specification in primal and dual production systems, *Journal of Applied Econometrics* **26**(2), 270–297.
- Lien, G., Størdal, S. and Baardsen, S.: 2007, Technical efficiency in timber production and effects of other income sources, *Small-Scale Forestry* **6**(1), 65–78.
- Meeusen, W. and van den Broeck, J.: 1977, Efficiency estimation from Cobb-Douglas production functions with composed error, *International Economic Review* **18**(2), 435–444.
- Morrison Paul, C. J., Johnston, W. E. and Frengley, G. A. G.: 2000, Efficiency in New Zealand sheep and beef farming: The impacts of regulatory reform, *Review of Economics and Statistics* **82**(2), 325–337.
- Park, S.-U. and Lesourd, J.-B.: 2000, The efficiency of conventional fuel power plants in South Korea: A comparison of parametric and non-parametric approaches, *International Journal Production Economics* **63**(1), 59–67.

- R Development Core Team: 2008, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.  
ISBN 3-900051-07-0.  
**URL:** <http://www.R-project.org>
- Ruggiero, J.: 2007, A comparison of dea and stochastic frontier model using panel data, *International Transactions in Operational Research* **14**(3), 259–266.
- Schmidt, P. and Lin, T.-F.: 1984, Simple tests of alternative specifications in stochastic frontier models, *Journal of Econometrics* **24**(3), 349–361.
- Shephard, R. W.: 1970, *Theory of Cost and Production Functions*, Princeton University Press, Princeton, N. J.
- Simar, L. and Wilson, P.: 1998, Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* **44**(1), 49–61.
- Simar, L. and Wilson, P.: 2000, A general methodology for bootstrapping in non-parametric frontier models, *Journal of Applied Statistics* **27**(6), 779–802.
- Smith, P. C. and Street, A.: 2005, Measuring the efficiency of public services, *Journal of the Royal Statistical Society, Series A* **168**(2), 401–417.
- Stone, M.: 2002, How not to measure the efficiency of public services (and how one might), *Journal of the Royal Statistical Society, Series A* **165**(3), 405–434.
- Thanassoulis, E.: 1993, A comparison of regression analysis and data envelopment analysis as alternative methods for performance assessments, *Journal of the Operational Research Society* **44**(11), 1129–1144.
- Thanassoulis, E.: 2000a, DEA and its use in the regulation of water companies, *European Journal of Operational Research* **127**(1), 1–13.
- Thanassoulis, E.: 2000b, The use of Data Envelopment Analysis in the regulation of UK water utilities: Water distribution, *European Journal of Operational Research* **126**(2), 436–453.
- Wilson, P. W.: 2008, FEAR: A software package for frontier efficiency analysis with R, *Socio-Economic Planning Sciences* **42**, 247–254.

Table 1: Combinations of error and inefficiency used in the Monte Carlo simulations

	$\sigma_u = 0.01$	$\sigma_u = 0.05$
$\sigma_v = 0.01$	s1: $\lambda = 1.0$	s2: $\lambda = 5.0$
$\sigma_v = 0.05$	s3: $\lambda = 0.2$	s4: $\lambda = 1.0$

Table 2: Finite sample performance of the efficiency estimates under a Cobb-Douglas production technology.<sup>a</sup>

	ECA <sup>b</sup>		Technical Efficiency							
	95% CI		Bias		RMSE		Upward Bias <sup>c</sup>		Correlation <sup>d</sup>	
	FLW <sup>e</sup>	KSW	FLW	KSW	FLW	KSW	FLW	KSW	FLW	KSW
<b>s1:</b> $\sigma_v = 0.01, \sigma_u = 0.01, \lambda = 1$										
n=50	0.92	0.46	-0.0034	-0.0098	0.0044	0.0098	0.3200	0.2000	0.4335	0.3072
n=100	0.95	0.34	-0.0030	-0.0115	0.0039	0.0115	0.3200	0.1600	0.4485	0.3603
n=200	0.96	0.24	-0.0031	-0.0136	0.0591	0.1166	0.3150	0.1100	0.4623	0.4001
<b>s2:</b> $\sigma_v = 0.01, \sigma_u = 0.05, \lambda = 5$										
n=50	0.92	0.72	-0.0107	0.0028	0.0133	0.0049	0.2800	0.4400	0.7616	0.5978
n=100	0.94	0.62	-0.0127	-0.0023	0.0132	0.0035	0.2100	0.3500	0.7846	0.6884
n=200	0.94	0.49	-0.0148	-0.0057	0.0148	0.0057	0.1600	0.2700	0.8026	0.7494
<b>s3:</b> $\sigma_v = 0.05, \sigma_u = 0.01, \lambda = 0.2$										
n=50	0.78	0.22	-0.0379	-0.0584	0.0379	0.0584	0.0200	0.1000	0.1216	0.0882
n=100	0.85	0.14	-0.0331	-0.0715	0.0331	0.0715	0.0200	0.0600	0.1200	0.0987
n=200	0.88	0.09	-0.0278	-0.0831	0.0278	0.0831	0.0250	0.0400	0.1206	0.1042
<b>s4:</b> $\sigma_v = 0.05, \sigma_u = 0.05, \lambda = 1$										
n=50	0.92	0.34	-0.0134	-0.0318	0.0196	0.0318	0.3400	0.2600	0.4563	0.3528
n=100	0.95	0.24	-0.0092	-0.0476	0.0160	0.0476	0.3600	0.1700	0.4675	0.3885
n=200	0.96	0.16	-0.0059	-0.0621	0.0124	0.0621	0.3950	0.1100	0.4694	0.4110

<sup>a</sup> **A Cobb-Douglas production function:**  $y_i = x_{1i}^\alpha \cdot x_{2i}^{\gamma-\alpha} \exp(v_i - u_i)$ ,  $\gamma = 1.0$ ,  $\alpha = 1/3$ ,  $u \sim N^+(0, \sigma_u)$ ;

<sup>b</sup> Empirical Coverage Accuracy is the share of true technical efficiencies that are within bounds of predicted 95 percent confidence interval for estimated technical efficiency. Reported in this table is the median of such shares across all Monte Carlo simulations;

<sup>c</sup> Upward Bias is the share of predicted technical efficiencies strictly larger than the true ones. The desired value of upward bias is 0.5. The values less (greater) than 0.5 indicates systematic underestimation (overestimation) of technical efficiencies. Reported in this table is the median of such shares across all Monte Carlo simulations;

<sup>d</sup> Kendall correlation coefficient between predicted and true technical efficiencies. Reported in this table is the median of such coefficients across all Monte Carlo simulations;

<sup>e</sup> FLW represents the SFA estimator. KSW represents the DEA estimator..

Table 3: Finite sample performance of the efficiency estimates under a constant elasticity of substitution production technology.<sup>a</sup>

	ECA <sup>b</sup>		Technical Efficiency							
	95% CI		Bias		RMSE		Upward Bias <sup>c</sup>		Correlation <sup>d</sup>	
	FLW <sup>e</sup>	KSW	FLW	KSW	FLW	KSW	FLW	KSW	FLW	KSW
<b>s1:</b> $\sigma_v = 0.01, \sigma_u = 0.01, \lambda = 1$										
n=50	0.92	0.44	-0.0045	-0.0109	0.0052	0.0109	0.2800	0.2000	0.4498	0.3085
n=100	0.94	0.31	-0.0040	-0.0125	0.0045	0.0125	0.2800	0.1400	0.4582	0.3671
n=200	0.96	0.21	-0.0040	-0.0146	0.0650	0.1210	0.2700	0.1000	0.4672	0.4015
<b>s2:</b> $\sigma_v = 0.01, \sigma_u = 0.05, \lambda = 5$										
n=50	0.92	0.70	-0.0120	0.0015	0.0142	0.0046	0.2200	0.4000	0.7796	0.6016
n=100	0.92	0.59	-0.0148	-0.0033	0.0151	0.0039	0.1700	0.3200	0.8012	0.6912
n=200	0.92	0.47	-0.0168	-0.0069	0.0168	0.0069	0.1300	0.2400	0.8141	0.7516
<b>s3:</b> $\sigma_v = 0.05, \sigma_u = 0.01, \lambda = 0.2$										
n=50	0.78	0.22	-0.0397	-0.0605	0.0397	0.0605	0.0200	0.1000	0.1135	0.0890
n=100	0.85	0.14	-0.0334	-0.0729	0.0334	0.0729	0.0200	0.0600	0.1139	0.0933
n=200	0.88	0.09	-0.0280	-0.0851	0.0280	0.0851	0.0250	0.0400	0.1201	0.1057
<b>s4:</b> $\sigma_v = 0.05, \sigma_u = 0.05, \lambda = 1$										
n=50	0.92	0.34	-0.0157	-0.0323	0.0209	0.0323	0.3200	0.2600	0.4661	0.3528
n=100	0.95	0.24	-0.0105	-0.0486	0.0167	0.0486	0.3500	0.1700	0.4683	0.3883
n=200	0.96	0.15	-0.0068	-0.0633	0.0133	0.0633	0.3800	0.1050	0.4724	0.4137

<sup>a</sup> **A constant elasticity of substitution production function:**  $y_i = [\beta x_{1i}^\rho + (1 - \beta)x_{2i}^\rho]^{\gamma/\rho} \exp(v_i - u_i)$ ,  $\gamma = 1.0$ ,  $\beta = 2/3$ ,  $\rho = 0.5$ ,  $u \sim N^+(0, \sigma_u)$ ;

<sup>b</sup> Empirical Coverage Accuracy is the share of true technical efficiencies that are within bounds of predicted 95 percent confidence interval for estimated technical efficiency. Reported in this table is the median of such shares across all Monte Carlo simulations;

<sup>c</sup> Upward Bias is the share of predicted technical efficiencies strictly larger than the true ones. The desired value of upward bias is 0.5. The values less (greater) than 0.5 indicates systematic underestimation (overestimation) of technical efficiencies. Reported in this table is the median of such shares across all Monte Carlo simulations;

<sup>d</sup> Kendall correlation coefficient between predicted and true technical efficiencies. Reported in this table is the median of such coefficients across all Monte Carlo simulations;

<sup>e</sup> FLW represents the SFA estimator. KSW represents the DEA estimator..



Table 4: Finite sample properties of  $\hat{\lambda}$  under a Cobb-Douglas production technology.<sup>a</sup>

	$\hat{\lambda}_i - \lambda_i$		$(\hat{\lambda}_i - \lambda_i)^2$	Size of the test at $\alpha$ level <sup>b</sup>		
	median	IQR	median	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
<b>s1:</b> $\sigma_v = 0.01, \sigma_u = 0.01, \lambda = 1$						
n=50	0.5629	1.0838	0.4006	0.0145	0.0545	0.1040
n=100	0.4280	0.7819	0.2282	0.0105	0.0415	0.0884
n=200	0.4066	0.5981	0.1924	0.0161	0.0537	0.1069
<b>s2:</b> $\sigma_v = 0.01, \sigma_u = 0.05, \lambda = 5$						
n=50	-1.3368	2.8774	4.0912	0.0125	0.0529	0.0979
n=100	-1.1435	2.0947	2.3719	0.0095	0.0484	0.0868
n=200	-0.8765	1.6037	1.2408	0.0145	0.0464	0.0863
<b>s3:</b> $\sigma_v = 0.05, \sigma_u = 0.01, \lambda = 0.2$						
n=50	0.9956	0.9488	0.9913	0.0126	0.0537	0.1054
n=100	0.7728	0.5979	0.5973	0.0151	0.0563	0.0996
n=200	0.5994	0.4601	0.3592	0.0087	0.0471	0.0952
<b>s4:</b> $\sigma_v = 0.05, \sigma_u = 0.05, \lambda = 1$						
n=50	0.4958	1.0889	0.3368	0.0055	0.0415	0.0855
n=100	0.2627	0.7710	0.1617	0.0130	0.0586	0.1088
n=200	0.1581	0.5498	0.0846	0.0136	0.0522	0.0989

<sup>a</sup> **A Cobb-Douglas production function:**  $y_i = x_{1i}^\alpha \cdot x_{2i}^{\gamma-\alpha} \exp(v_i - u_i)$ ,  $\gamma = 1.0$ ,  $\alpha = 1/3$ ,  $u \sim N^+(0, \sigma_u)$ ;

<sup>b</sup> Size of the test for the null that estimated value of  $\lambda$  is equal to the true value of  $\lambda$ . We use 500 bootstrap replications in each of our 2000 Monte Carlo simulations.

Table 5: Finite sample properties of  $\hat{\lambda}$  under a constant elasticity of substitution production technology.<sup>a</sup>

	$\hat{\lambda}_i - \lambda_i$		$(\hat{\lambda}_i - \lambda_i)^2$	Size of the test at $\alpha$ level <sup>b</sup>		
	median	IQR	median	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
<b>s1:</b> $\sigma_v = 0.01, \sigma_u = 0.01, \lambda = 1$						
n=50	0.6296	1.1817	0.4571	0.0111	0.0453	0.0850
n=100	0.4849	0.8343	0.2745	0.0145	0.0511	0.0952
n=200	0.4349	0.6412	0.2128	0.0161	0.0614	0.1172
<b>s2:</b> $\sigma_v = 0.01, \sigma_u = 0.05, \lambda = 5$						
n=50	-1.0636	3.3007	4.1587	0.0125	0.0507	0.1009
n=100	-0.7875	2.5632	2.4105	0.0135	0.0556	0.1093
n=200	-0.6807	1.7407	1.1920	0.0120	0.0531	0.1138
<b>s3:</b> $\sigma_v = 0.05, \sigma_u = 0.01, \lambda = 0.2$						
n=50	1.0356	0.9705	1.0724	0.0086	0.0499	0.0943
n=100	0.7696	0.6270	0.5922	0.0096	0.0494	0.0984
n=200	0.5977	0.4284	0.3573	0.0103	0.0405	0.0893
<b>s4:</b> $\sigma_v = 0.05, \sigma_u = 0.05, \lambda = 1$						
n=50	0.5690	1.1489	0.4275	0.0121	0.0564	0.1058
n=100	0.2881	0.7917	0.1716	0.0141	0.0552	0.1035
n=200	0.1465	0.5706	0.0900	0.0141	0.0510	0.1081

<sup>a</sup> **A constant elasticity of substitution production function:**

$$y_i = [\beta x_{1i}^\rho + (1 - \beta)x_{2i}^\rho]^{\gamma/\rho} \exp(v_i - u_i), \quad \gamma = 1.0, \beta = 2/3, \rho = 0.5,$$

$$u \sim N^+(0, \sigma_u);$$

<sup>b</sup> Size of the test for the null that estimated value of  $\lambda$  is equal to the true value of  $\lambda$ . We use 500 bootstrap replications in each of our 2000 Monte Carlo simulations.

Table 6: Relative finite sample performance of the efficiency estimates at the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles of the efficiency distribution<sup>a</sup>

	$\sigma_u = 0.01$	$\sigma_u = 0.05$
$\sigma_v = 0.01$	FLW $\approx$ KSW < Truth	FLW < KSW $\approx$ Truth
	5: FLW	5: KSW
	50: FLW	50: KSW
	95: KSW	95: KSW
$\sigma_v = 0.05$	FLW $\approx$ KSW < Truth	FLW $\approx$ KSW < Truth
	5: FLW	5: FLW
	50: FLW	50: FLW
	95: KSW	95: KSW

<sup>a</sup> FLW represents the SFA estimator. KSW represents the DEA estimator.

Table 7: Reliability of efficiency rankings<sup>a</sup> (same conclusion for both FLW and KSW estimators)

	$\sigma_u = 0.01$	$\sigma_u = 0.05$
$\sigma_v = 0.01$	5: fair	5: good
	50: poor	50: good
	95: fair	95: good
$\sigma_v = 0.05$	5: poor	5: fair
	50: poor	50: poor
	95: poor	95: fair

<sup>a</sup> FLW represents the SFA estimator. KSW represents the DEA estimator.

Table 8: Estimates from the first stage regression of the FLW estimator<sup>a</sup>

	$R^{2b}$	$\lambda$	$\sigma_\lambda^c$	$\sigma_u$	$\sigma_v$	$\mu$
Data 1	0.8235	4.0961	1.3206	0.4336	0.1059	0.3460
Data 2	0.9139	0.0016	0.7271	0.0003	0.1693	0.0002
Data 3	0.9339	1.2222	0.5157	0.1697	0.1388	0.1354

<sup>a</sup> FLW represents the SFA estimator.

<sup>b</sup>  $R^2$  of the first stage regression in FLW is defined as the Kendall correlation coefficient between the predicted and true output.

<sup>c</sup>  $\sigma_\lambda$  is calculated via 1000 bootstrap replications.

Table 9: Technical efficiency descriptives

	Min.	Q1	Median	Mean	Q3	Max.
Data 1						
FLW	0.3131	0.6180	0.7627	0.7288	0.8597	0.9620
KSW	0.2796	0.6438	0.7426	0.7325	0.8829	0.9633
Data 2						
FLW	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
KSW	0.0774	0.1481	0.1725	0.1792	0.1996	0.7621
Data 3						
FLW	0.5518	0.8641	0.8852	0.8774	0.9075	0.9778
KSW	0.1524	0.3454	0.3834	0.3871	0.3932	0.9638

<sup>a</sup> FLW represents the SFA estimator. KSW represents the DEA estimator.

Table 10: Correlation coefficients between FLW and KSW efficiency estimates

	Pearson	Kendall	Spearman
Data 1	0.8653	0.6722	0.8121
Data 2	0.6494	0.5726	0.7597
Data 3	-0.8735	-0.6131	-0.7926

<sup>a</sup> FLW represents the SFA estimator. KSW represents the DEA estimator.

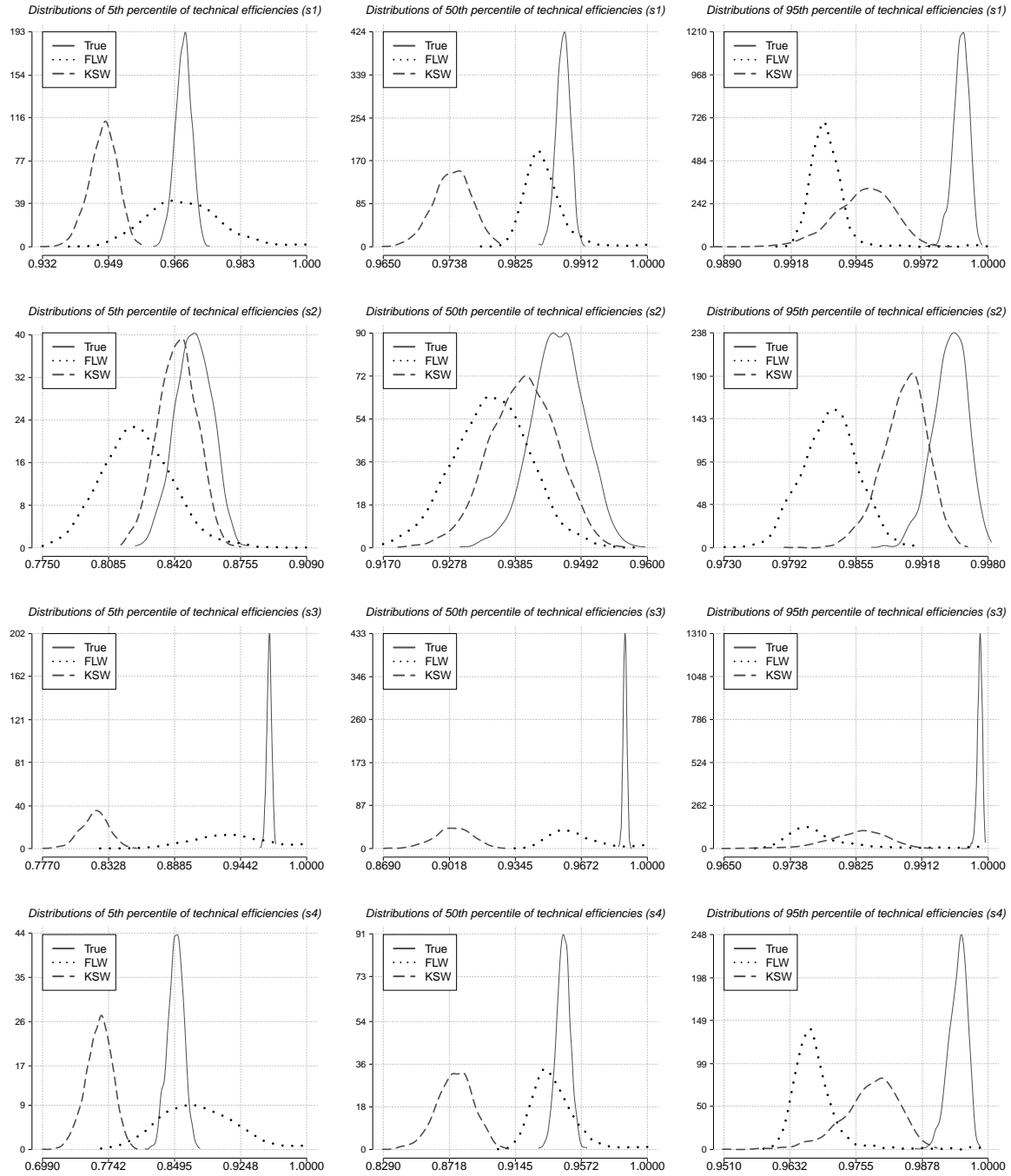


Figure 1: Distributions (estimated kernel densities) of 5th, 50th, and 95th percentiles of technical efficiency estimates under a Cobb-Douglas ( $y_i = x_{1i}^\alpha \cdot x_{2i}^{\gamma-\alpha} \exp(v_i - u_i)$ ,  $\gamma = 1.0$ ,  $\alpha = 1/3$ ,  $u \sim N^+(0, \sigma_u)$ ) production technology.

Four scenarios:

**s1:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.01$ ,  $\lambda = 1$ ;

**s2:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.05$ ,  $\lambda = 5$ ;

**s3:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.01$ ,  $\lambda = 0.2$ ;

**s4:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.05$ ,  $\lambda = 1$ .

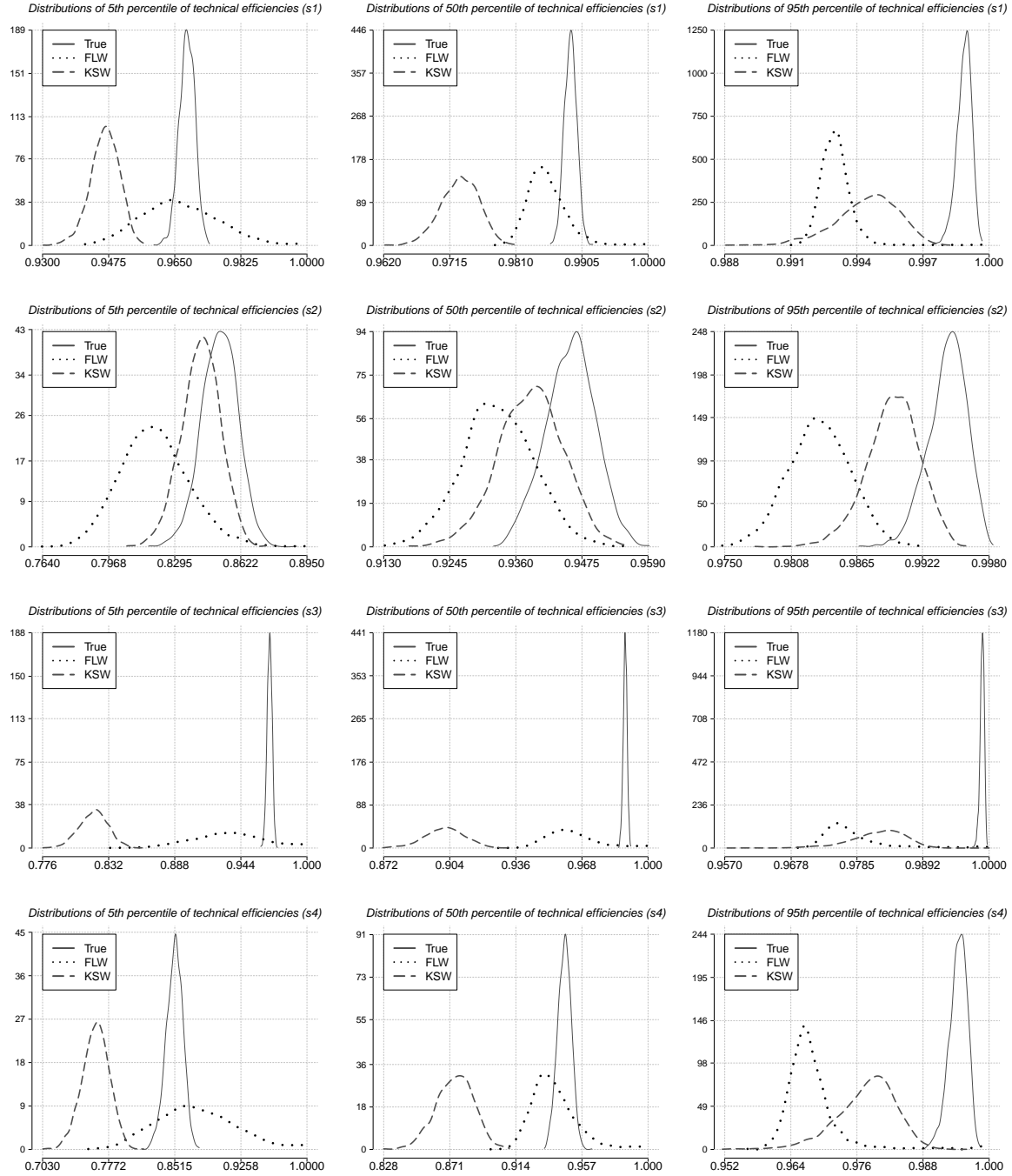


Figure 2: Distributions (estimated kernel densities) of 5th, 50th, and 95th percentiles of technical efficiency estimates under a constant elasticity of substitutionm ( $y_i = [\beta x_{1i}^\rho + (1 - \beta)x_{2i}^\rho]^{\gamma/\rho} \exp(v_i - u_i)$ ,  $\gamma = 1.0$ ,  $\beta = 2/3$ ,  $\rho = 0.5$ ,  $u \sim N^+(0, \sigma_u)$ ) production technology.

Four scenarios:

**s1:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.01$ ,  $\lambda = 1$ ;

**s2:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.05$ ,  $\lambda = 5$ ;

**s3:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.01$ ,  $\lambda = 0.2$ ;

**s4:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.05$ ,  $\lambda = 1$ .

FLW represents the SFA estimator. KSW represents the DEA estimator.

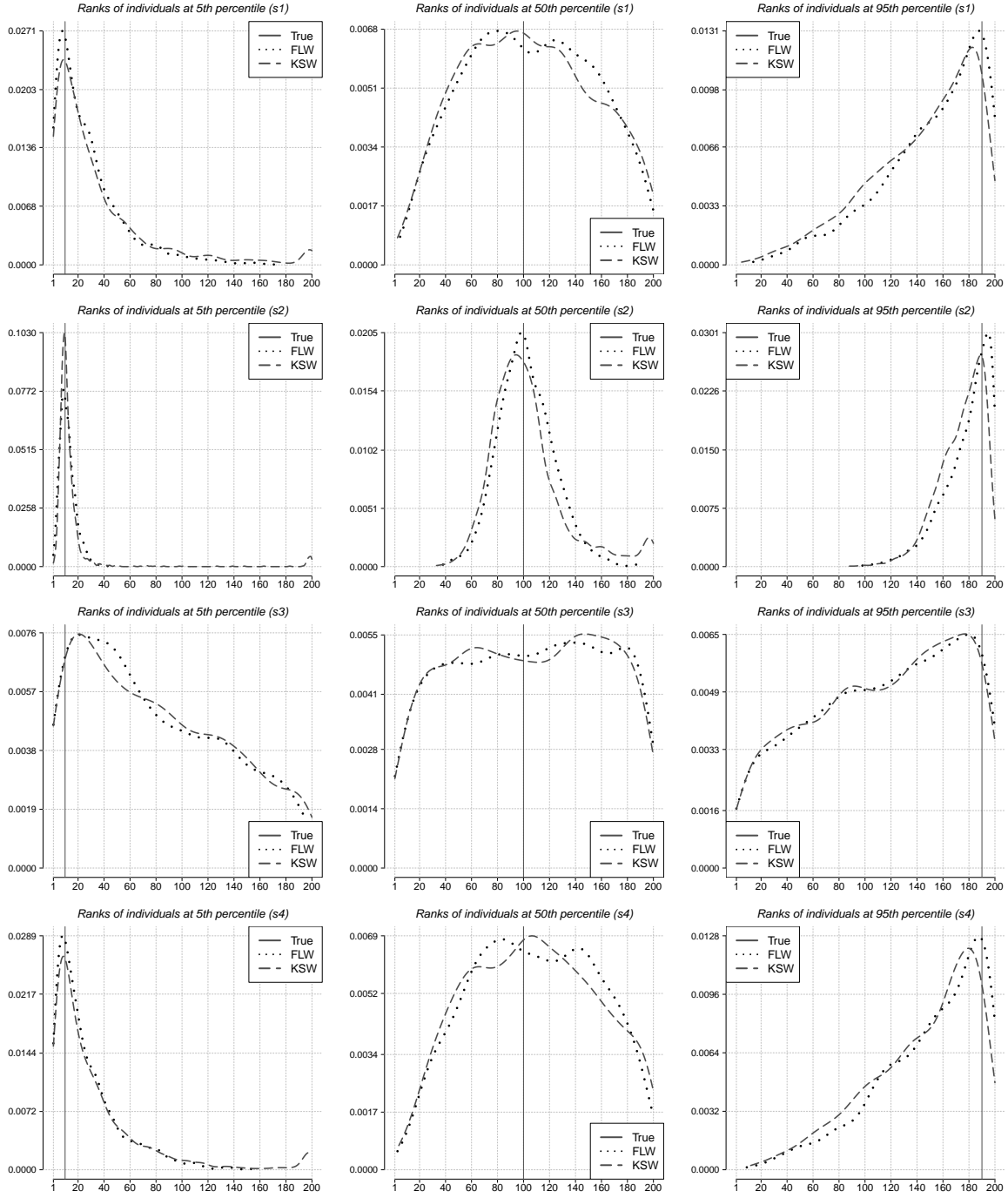


Figure 3: Distributions (estimated kernel densities) of 5th, 50th, and 95th percentiles of ranks of technical efficiency estimates under a Cobb-Douglas ( $y_i = x_{1i}^\alpha \cdot x_{2i}^{\gamma-\alpha} \exp(v_i - u_i)$ ,  $\gamma = 1.0$ ,  $\alpha = 1/3$ ,  $u \sim N^+(0, \sigma_u)$ ) production technology.

Four scenarios:

**s1:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.01$ ,  $\lambda = 1$ ;

**s2:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.05$ ,  $\lambda = 5$ ;

**s3:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.01$ ,  $\lambda = 0.2$ ;

**s4:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.05$ ,  $\lambda = 1$ .

FLW represents the SFA estimator. KSW represents the DEA estimator.



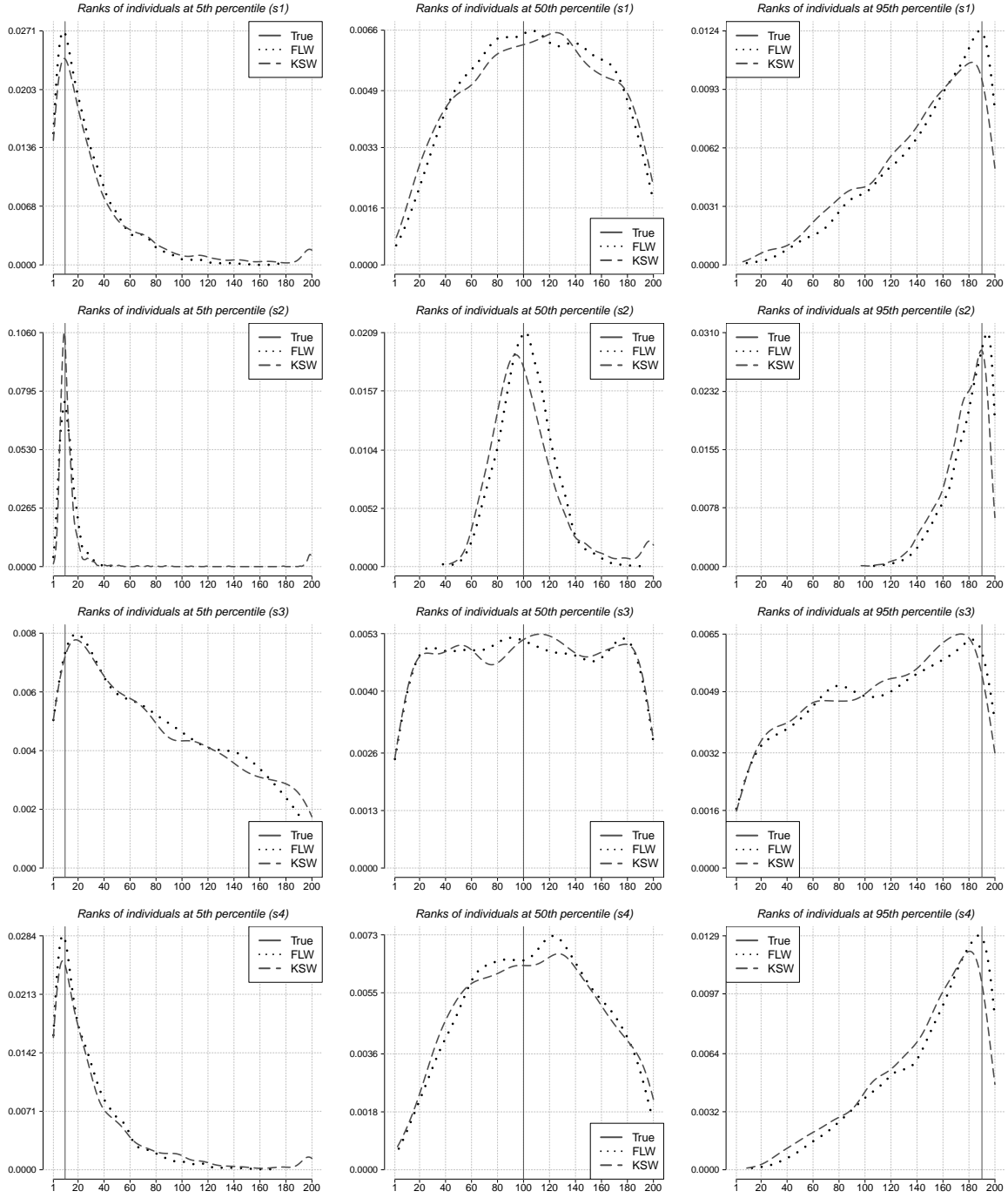


Figure 4: Distributions (estimated kernel densities) of 5th, 50th, and 95th percentiles of ranks of technical efficiency estimates under a constant elasticity of substitution ( $y_i = [\beta x_{1i}^\rho + (1 - \beta)x_{2i}^\rho]^{\gamma/\rho} \exp(v_i - u_i)$ ,  $\gamma = 1.0$ ,  $\beta = 2/3$ ,  $\rho = 0.5$ ,  $u \sim N^+(0, \sigma_u)$ ) production technology.

Four scenarios:

**s1:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.01$ ,  $\lambda = 1$ ;

**s2:**  $\sigma_v = 0.01$ ,  $\sigma_u = 0.05$ ,  $\lambda = 5$ ;

**s3:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.01$ ,  $\lambda = 0.2$ ;

**s4:**  $\sigma_v = 0.05$ ,  $\sigma_u = 0.05$ ,  $\lambda = 1$ .

FLW represents the SFA estimator. KSW represents the DEA estimator.

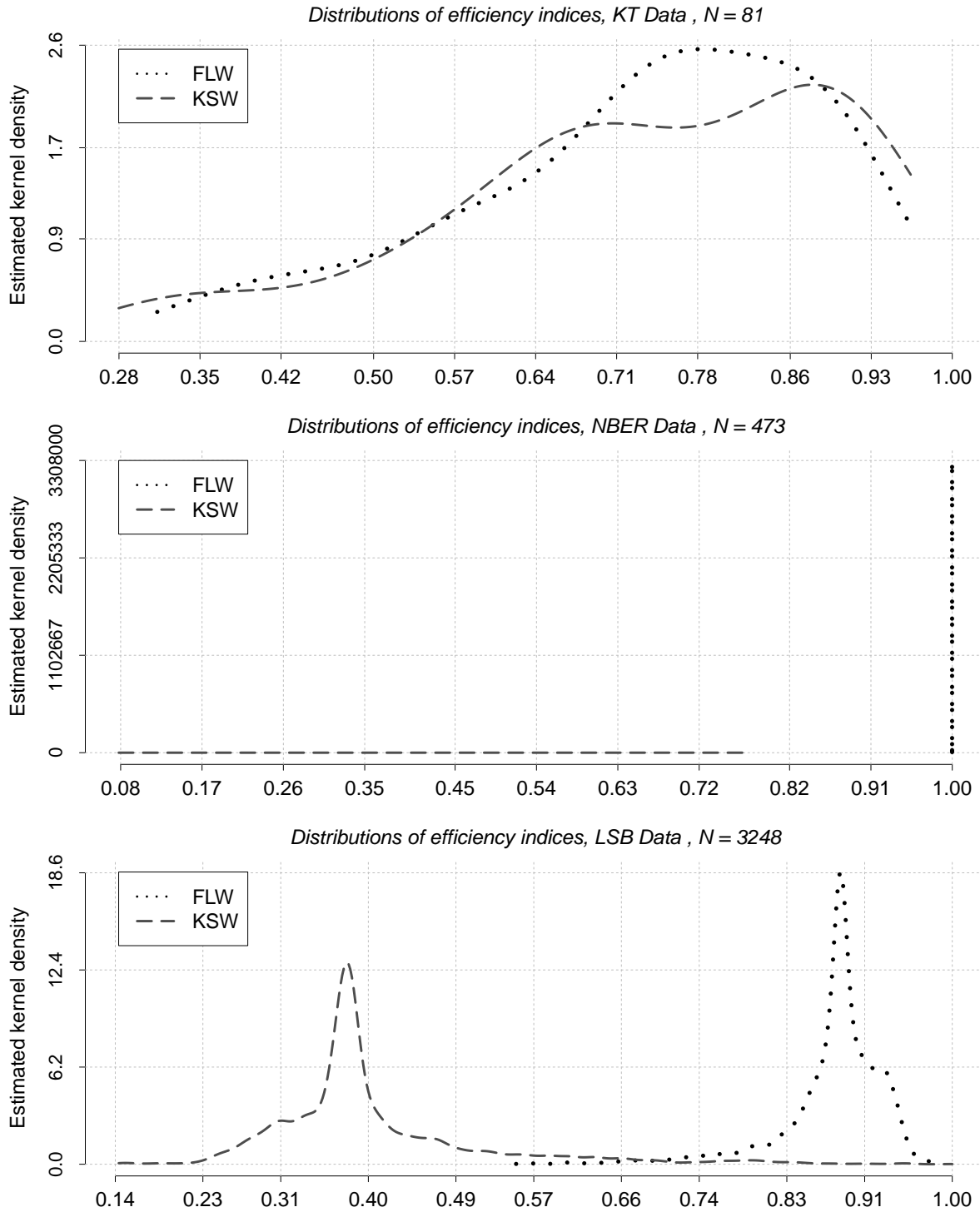


Figure 5: Distributions of technical efficiency estimates. FLW represents the SFA estimator. KSW represents the DEA estimator.