Regularized Conditional Estimators of Unit Inefficiency in Stochastic Frontier Analysis, with Application to Swedish Electricity Distribution Market

Zangin Zeebari [1, a, c,†], Kristofer Månsson [2,a], Pär Sjölander [3,a], and Magnus Söderberg [4,b]

[a] Department of Economics, Finance and Statistics, Jönköping International Business School (JIBS), Jönköping University, P.O. Box 1026, SE-551 11 Jönköping, Sweden

[b] Ratio Institute, Box 3203, S-103 64 Stockholm, Sweden

[c] Karolinska Institutet, Department of Global Public Health, Stockholm, Sweden

E-mails:

[1] zangin.zeebari@ju.se;

[2] kristofer.mansson@ju.se;

[3] par.sjolander@ju.se;

[4] magnus.soderberg@hh.se

[†] Corresponding author

Abstract

The value of Stochastic Frontier Analysis (SFA) increases when the level of accuracy at which it estimates unit-specific inefficiencies improves. Conventional estimation of SFA unit inefficiency is based on the mean/mode of the inefficiency, conditioned on the composite error. It is known that the conditional mean of inefficiency shrinks towards the mean, rather than the unit inefficiency. In this paper, we analytically prove that the conditional mode cannot accurately estimate unit inefficiency, either. We propose regularized estimators of unit inefficiency that restrict the unit inefficiency estimators to satisfy some a priori assumptions, and derive the closed form regularized conditional mode estimators for the three most commonly used inefficiency densities. Extensive simulations show that, under common empirical situations, e.g., regarding sample size and signal-to-noise ratio, the regularized estimators outperform the conventional (unregularized) estimators when the inefficiency is greater than its mean/mode. Based on real data from the electricity distribution sector in Sweden, we demonstrate that the conventional conditional estimators and our regularized conditional estimators provide substantially different results for highly inefficient companies.

Keywords:

Uncertainty modelling; Productivity; Regularized Estimators; Constrained Estimators; Conditional Estimators

1. Introduction

Since the publication of the papers by Aigner et al. (1977) and Meeusen and Van Den Broeck (1977), stochastic frontier analysis (SFA) has been a common approach to gain deeper insights into the potential for productivity improvement (Kumbhakar et al., 2020) and cost reduction in monopolized markets (Bogetoft and Otto, 2011).

For unit inefficiency, the standard estimation approach was developed by Jondrow, Lovell, Materov and Schmidt (1982), acronymed “JLMS” in the SFA literature. The JLMS estimator is based on the mean (and the mode) of the inefficiency conditioned on the composite error when the inefficiency is drawn from a half-normal distribution. Later studies have extended the JLMS estimator to situations when inefficiencies are drawn from an exponential distribution (Kumbhakar and Lovell, 2000) and a truncated normal distribution (Battese and Coelli, 1988).

Despite its widespread use, the JLMS estimator has been criticized. Specifically, Wang and Schmidt (2009) explain that it shrinks the inefficiency towards its mean, leading to a distribution that is different from that of the unconditional inefficiency. Naturally, the mean and mode are not fully representative characteristics of the conditional distribution of the inefficiency, especially if each unit is observed only once. Thus, in the cross-sectional case, each conditional estimator produces an inconsistent estimator of the inefficiency. Moreover, this estimator is conditioned on an estimated composite error rather than on the composite error itself, as explained by Horrace (2005). More details are provided by Kumbhakar et al. (2015) and Kumbhakar et al. (2018). Therefore, the sampling distribution of the conditional estimator is different from the theoretically assumed conditional distribution of the inefficiency. Consequently, the inefficiencies are inaccurately estimated, something regulatory agencies have stated as an impediment for the practical use of SFA (e.g., Badunenko et al., 2012; Stone, 2002 and Tsionas, 2017). This is also illustrated in a simulation study by Andor et al. (2019), where they show that both the SFA and Data envelopment analysis (DEA) methods used by regulators underestimate the true efficiency values. One way to reduce this problem is to combine the SFA and the DEA (Andor et al. (2019) and Tsionas (2021)), but such combinatory approaches are not able to eliminate the underestimation problem.

The approach presented in this paper is similar to the combinatory approach in that it can be viewed as a weighted average of unit inefficiency estimators, but in contrast, it is a weighted average of the sample (industry), solely based on the SFA approach. The proposed regularized estimators can be used as stand-alone estimators along with any other estimators in a combinatory approach. In addition, the regularized estimators described here can be used in a variety of situations but in this paper, we limit ourselves to studying unit inefficiency estimation in a cross-sectional context, using the classical stochastic frontier model suggested by Aigner et al. (1977).

We propose a regularized (constrained) estimator based on Bayesian risk (expected loss) that restricts the inefficiencies to satisfy some underlying theoretical (and/or intuitive) conditions. Conditions on the moments are common options for the imposed constraints upon the likelihood functions (e.g., Hall and Presnell (1999)). Our regularized estimators are easily calculated, e.g., they can be the JLMS estimators, with imposed constraints on the first and the second moments of the conditional distribution of the inefficiency.

The proposed methodology is different from other recent contributions in the field. For example, Kumbhakar et al. (1991) suggest a single step procedure for the estimation of unit inefficiency when they deploy firm-specific determinants of the inefficiency in the maximum likelihood estimation of the SFA model. They show that ignoring the determinants would lead to biased and inconsistent estimators. However, firm-specific determinants are often unobserved, and even unknown. Another recent contribution is the use of non-parametric and semi-parametric estimation methods. However, these methods are different from what we do their incorporation into SFA, for example by Kumbhakar et al. (2007), use the JLMS estimator for estimating firm-specific inefficiency. Another avenue of research is the use of quantile regression into the estimation of the production function (Bernini et al. (2004), Wang et al. (2014), and Behr (2010)). However, this approach introduces a new challenge, specifically that one needs to pay more attention to the selection of appropriate quantiles which can be different for distinct densities of the composite error (Jradi et all. (2019)). In addition, no post-estimation of firm-specific inefficiencies exists when using the quantile regression approach (Kumbhakar et al., 2020).

Under mild assumptions, e.g., the log-concavity of distributions that covers most of the distributions used in the SFA literature, we analytically investigate some properties of the conditional mode (maximum *a posteriori* probability estimator) and give a general formula for the conditional mode and its functions that can be used with any inefficiency density. Next, we derive a regularized conditional mode estimator with the three most commonly used inefficiency densities, i.e., the half-normal, truncated normal and exponential distributions. The proposed unit inefficiency estimation is considered a restricted or penalized estimation method that improves the estimation of unit inefficiency based on the conditional mean/mode.

An extensive simulation study is conducted, with varying factors, such as the sample size, inefficiency density and signal-to-noise ratio (relative variation of the inefficiency to the variation of random shocks). The simulation results show that the regularized estimators outperform the conventional (unregularized) estimators when the inefficiencies are greater than their mean/mode, especially with a larger signal-to-noise ratio. As the unregularized conditional mean/mode shrinks towards the mean/mode, the simulation results show that the regularized conditional mean/mode shrinks less towards the mean/mode, especially for larger inefficiency scores.

We apply both unregularized and regularized estimators to data from the Swedish electricity distribution sector. The results show that the estimated inefficiencies from the two regularized and unregularized estimators are substantially different, particularly for firms that are in the right tail of the inefficiency distribution. Considering the results from the simulation study, supported analytically (Theorem 3), we recommend that regulators use the results from the regularized estimators for highly inefficient firms.

The remainder of this paper is structured as follows. In section 2, we derive a general formula for the conditional mode of the inefficiency and analytically investigate some of its properties under mild distributional assumptions of unconditional inefficiency. Next, the regularized estimator is discussed and formally derived for both production and cost functions. Derivations are presented for inefficiency under three different distributional assumptions. In section 4, both regularized and unregularized estimators are evaluated using extensive Monte Carlo simulations. In section 5, we present an application based on real data. The data represent electricity distribution firms in Sweden, and we estimate the cost inefficiency, which is used by the Energy Markets Inspectorate as an input in their revenue cap regulation. Section 5 concludes the paper and discusses avenues for future research.

2. Theory

A stochastic frontier, cross-sectional, production model can be formulated as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - u_i,$$

where i indicates the unit, y_i is the observed output, $\boldsymbol{x}_i$ is the given $k \times 1$ vector of inputs, u_i is the unobserved inefficiency, v_i is the unobserved noise and $\boldsymbol{\beta}$ is an unknown $k \times 1$ vector of functional parameters.

The conventions of a simple parametric cross-section SFA assume i.i.d. random noise terms with a density function $g_v(v)$ that is symmetric around zero and i.i.d. nonnegative inefficiencies with a density function $f_u(u)$. For example, the most common (semistandard) $g_v(v)$ is assumed to be the density of a zero-mean normal distribution $N(0, \sigma_v^2)$[1], and the equivalent candidates for $f_u(u)$ are assumed to be the densities of a half-normal distribution $N^+(0, \sigma_u^2)$, an exponential distribution $Exp(\sigma_u)$ with scale parameter σ_u, and a truncated normal distribution $N^+(\mu, \sigma_u^2)$ with a general μ that can take any real number.

The maximum likelihood estimation of an SFA model is based on maximizing the likelihood of the i.i.d. composite errors $\varepsilon_i = v_i - u_i$ with the density function

[1] Other zero-mean symmetric distributions have been suggested, such as Laplace (Horrace and Parmeter, 2018; Nguyen, 2010), but they are less common in applications.

$$h_\varepsilon(\varepsilon) = \int_0^{+\infty} f_u(u)\, g_v(u+\varepsilon)\, du$$

where the composite error ε_i is $\varepsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$.

It has been argued, for example, by Greene (1990) and Ruggiero (1999), that the selection of different inefficiency density functions should not result in noticeable differences between the fit of the SFA models, or the ranks of the estimated conditional unit inefficiency scores. However, they may differ in the magnitude of the inefficiency scores, especially for highly inefficient units.

As mentioned in the Introduction section, the most common way of scoring unit inefficiency is through the method proposed by Jondrow et al. (1982). For the i^{th} unit, the inefficiency is scored as $\hat{u}_i = E(u|\varepsilon_i)$ or $\hat{u}_i = Mode(u|\varepsilon_i)$ using the following conditional density function of inefficiency u given a composite error ε.

$$f_{u|\varepsilon}(u) = \frac{f_u(u)\, g_v(u+\varepsilon)}{h_\varepsilon(\varepsilon)}$$

However, as stated by Kumbhakar et al. (2020), the conditional score of the inefficiency is an estimator of a characteristic (mean or mode) of the conditional inefficiency rather than of the inefficiency itself. Such a distinction between the two remains unchanged regardless of the sample size. In fact, it depends on the size of the noise variance rather than on the sample size. This fact is proven by Wang and Schmidt (2009) for the conditional mean when the inefficiency follows a half-normal distribution, and they argue that it also holds when the inefficiencies are drawn from exponential and general truncated normal distributions. However, such argument has not been proven for the conditional mode, although there is a general belief in the SFA literature that the JLMS estimators, whether mean or mode, are shrinkage estimators.

In in Theorem 3, we provide a proof that, under mild distributional assumptions, the conditional mode of the inefficiency analogously shrinks the inefficiency score towards the mode of the inefficiency. This means that the conditional mode estimator underestimates large inefficiencies. It also overestimates the inefficiencies of almost fully efficient firms when the inefficiency mode is a positive number (as it is the case for a truncated normal distribution with location parameter $\mu > 0$).

The conditional mode score is the maximum *a posteriori* probability estimator, which is the mode of the *a posteriori* distribution, i.e., $Mode(u|\varepsilon_i) = \underset{u\in\mathbb{R}^+}{\text{Argmax}}\, f_{u|\varepsilon}(u) = \underset{u\in\mathbb{R}^+}{\text{Argmax}}\big(f_u(u)g_v(u+\varepsilon)\big)$, where $\mathbb{R}^+$ denotes the nonnegative real numbers. According to the laws of total mean and variance, we have $E(E(u|\varepsilon_i)\,) = E(u)$, but $Var(E(u|\varepsilon_i)\,) = Var(u) - E\big(Var(u|\varepsilon_i)\big) < Var(u)$. However, arguments analogous to the total mean and variance do not generally hold for conditional mode scores. In other words, the mode (or mean) of the conditional mode score is not generally equal to the mode (or mean) of the inefficiency itself. The variance of the conditional mode score can also be larger than the variance of the inefficiency itself.[2] In Theorem 1, we give a general formula to calculate the conditional mode of the inefficiency for any inefficiency density function that fulfills the mild assumptions stated below.

Theorem 1: Suppose the noise of the production function in (1) is $v\sim N(0,\sigma_v^2)$. If the log-concave inefficiency density $f_u(u)$ is nonzero and continuously infinitely differentiable (analytic function) for all $u \geq 0$, such that $\left|\frac{\mathrm{d}^2 \ln[f_u(u)]}{\mathrm{d}u^2}\right| < \sigma_v^{-2}$, then the inefficiency mode score conditioned on the composite error ε is $\hat{u} = Mode(u|\varepsilon) = \max\{0, \tilde{u}\}$, where

$$\tilde{u} = -\varepsilon - \sum_{k=1}^{+\infty} \frac{(\sigma_v^2)^k}{k!} \left[\frac{\partial}{\partial\varepsilon}\right]^{k-1} \left\{-\frac{f_u'(-\varepsilon)}{f_u(-\varepsilon)}\right\}^k$$

Proof:

Let $\tilde{u} = \underset{u\in\mathbb{R}^+}{\text{Argmax}}\big(f_u(u)g_v(u+\varepsilon)\big)$. Then, for $u \geq 0$, we can write

$$f_u'(\tilde{u})\, g_v(\tilde{u}+\varepsilon) + g_v'(\tilde{u}+\varepsilon)\, f_u(\tilde{u}) = 0$$

[2] For example, with exponentially distributed inefficiencies the conditional mode has a variance equal to the variance of the composite error, hence it is larger than the variance of the unconditional inefficiency. A similar argument holds with half normally distributed inefficiencies when $\sigma_u^2/\sigma_v^2 = \pi/(4-\pi)$.

$$f_u'(\tilde{u})\, g_v(\tilde{u}+\varepsilon) - \frac{(\tilde{u}+\varepsilon)}{\sigma_v^2}\, g_v(\tilde{u}+\varepsilon)\, f_u(\tilde{u}) = 0$$

$$\tilde{u} = -\varepsilon + \sigma_v^2 \frac{f_u'(\tilde{u})}{f_u(\tilde{u})} \qquad (2)$$

If $\left|\frac{\mathrm{d}^2 \ln[f_u(u)]}{\mathrm{d}u^2}\right| < \sigma_v^{-2}$, there is a unique solution of $\tilde{u}$ in terms of bounded ε in (2) since $\sigma_v^2 \frac{f_u'(\tilde{u})}{f_u(\tilde{u})} - \tilde{u} = \varepsilon$ becomes a monotonically decreasing function of $\tilde{u} \geq 0$. Then, $\left|\sigma_v^2 \frac{f_u'(\tilde{u})}{f_u(\tilde{u})}\right| < M\tilde{u}$ for a constant $M > 0$, and the bilateral Laplace transform $\mathcal{L}\left\{\sigma_v^2 \frac{f_u'(\tilde{u})}{f_u(\tilde{u})}\right\}(s)$ exists, for some $s > 0$. Using the Lagrange reversion theorem (see Whittaker and Watson (1927), pp. 132-133, and Grossman (2005)[3]), any differentiable function of $\tilde{u}$, including $\tilde{u}$ itself, is uniquely expressed in terms of the same function with $-\varepsilon$ as its argument and a power series of σ_v^2. This completes the proof. ■

Note that the commonly used inefficiency densities of half normal, exponential, general truncated normal and gamma (with shape parameter ≥1) are log-concave distributions. When the noise v and the inefficiency u are distributed as assumed in Theorem 1, for each of the density and distribution functions $q \in \{\tilde{f}_{\hat{u}|u},\ \tilde{F}_{\hat{u}|u},\ \tilde{f}_{\hat{u}|\varepsilon},\ \tilde{F}_{\hat{u}|\varepsilon},\ \tilde{f}_{u|\hat{u}},\ \tilde{F}_{u|\hat{u}},\ f_u,\ F_u,\ \tilde{f}_{\hat{u}},\ \tilde{F}_{\hat{u}}\}$, and any other differentiable function of $\hat{u} \geq 0$, we have,

$$q(\hat{u}) = q(-\varepsilon) - \sum_{k=1}^{+\infty} \frac{(\sigma_v^2)^k}{k!} \left[\frac{\partial}{\partial\varepsilon}\right]^{k-1} \left\{\left(-\frac{f_u'(-\varepsilon)}{f_u(-\varepsilon)}\right)^k q'(-\varepsilon)\right\}.$$

In general, the conditional mode, theoretically and empirically, is less covered in the SAF literature when JLMS estimators are used, in favor of the conditional mean. To the best of authors' knowledge, the article by Papadopoulos (2021) is an exception, in that the author elaborates on the conditional model and proves its monotonicity in terms of the composite error when the inefficiency follows a generalized exponential distribution. Monotonicity of the conditional mode in terms of the composite error is important in that both (mean/mode) JLMS estimators must rank the unit inefficiencies identically. If so, using the conditional

[3] Contrast to Grossman (2005) in which the proof is based on unilateral Laplace and the necessary condition $f_u'(0) = 0$, in our proof such condition $f_u'(0) = 0$ is not necessary as we use bilateral Laplace transform.

modes, the inefficiencies can be ranked based on their corresponding composite errors, i.e., there must be a -1 coefficient of ranked correlation between the two. A similar argument holds for conditional mean scores, as shown by Bera and Sharma (1999) and Ondrich and Ruggiero (2001). In Theorem 2, we show that under mild distributional assumptions, the monotonicity of the conditional mode in terms of the composite error is generalizable to any other inefficiency distribution.

Theorem 2: Suppose $v \sim N(0, \sigma_v^2)$. The inefficiency density $f_u(u)$ is nonzero, twice differentiable and log-concave at $u \geq 0$. The inefficiency mode score conditioned on the composite error ε is a monotonically decreasing function of the composite error.

Proof:

Since $f_u(u)$ is log-concave, we have $\frac{\mathrm{d}^2 \ln[f_u(u)]}{(\mathrm{d}u)^2} \leq 0$ for all $u \geq 0$. As shown in Theorem 1, we can write equation (2) as $\hat{u} = -\varepsilon + \sigma_v^2 \frac{\mathrm{d} \ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}}$. Then, by the chain rule of derivatives, we have

$$\frac{\partial \hat{u}}{\partial \varepsilon} = -1 + \sigma_v^2 \frac{\mathrm{d}^2 \ln[f_u(\hat{u})]}{(\mathrm{d}\hat{u})^2} \frac{\partial \hat{u}}{\partial \varepsilon}$$

$$\frac{\partial \hat{u}}{\partial \varepsilon} = \frac{-1}{1 - \sigma_v^2 \frac{\mathrm{d}^2 \ln[f_u(\hat{u})]}{(\mathrm{d}\hat{u})^2}} < 0$$

The above negative derivative would imply strict monotonicity if negative scores were acceptable. Since they are restricted to be zero scores, monotonicity is not strict, in general. Thus, the proof is complete. ■

As mentioned in the Introduction section, Wang and Schmidt (2009) show that the conditional mean is a shrinkage estimator of the unit inefficiency in that it shrinks towards the mean of inefficiency rather than towards the unit inefficiency itself. This property is disadvantageous to the unit inefficiencies that depart from the mean inefficiency since it underestimates highly inefficient firms and overestimates the inefficiencies lower than the mean. It is also a disadvantage of the conditional mean for the regulators to accurately estimate the inefficiency in the lower and, especially, in the upper tail of the inefficiency distribution. Although being able to rank the units based on their inefficiencies is of regulators' interest, in some cases the magnitude of inefficiency is of crucial importance, for instance, the EU countries' (in)efficiencies in their climate plans to cut the emissions of the greenhouse gases.

In Theorem 3, we prove that the conditional mode has a similar property, in that it is a shrinkage estimator towards the inefficiency mode rather than towards the inefficiency itself. With such property, although the conditional mode would outperform the conditional mean in estimating the lower tail of an inefficiency distribution with its mode in a narrow positive neighborhood of zero, it is still a poor estimator for highly inefficient firms, i.e., the right tail of the distribution.

Theorem 3: Suppose $v \sim N(0, \sigma_v^2)$ and, the inefficiency density $f_u(u)$ is nonzero, twice differentiable, log-concave for $u \geq 0$ and with $m = Mode(u) = \underset{u \in \mathbb{R}^+}{\operatorname{Argmax}} f_u(u)$. Let the inefficiency score be $\hat{u} = Mode(u|\varepsilon)$.

Then,

a) as $\sigma_v^2 \to 0$, $\hat{u} \to_p u$,

b) as $\sigma_v^2 \to 0$, $\hat{u} \to_d u$,

c) as $\sigma_v^2 \to 0$, $\frac{\hat{u}-u}{\sigma_v} \to_d N(0,1)$

d) as $\sigma_v^2 \to \infty$, $\hat{u} \to_p m = Mode(u)$.

e) as $\sigma_v^2 \to \infty$, $\sigma_v^2 [\ln(f_u(m))]' + \left(\sigma_v^2 [\ln(f_u(m))]'' - 1\right)(\hat{u} - m) \to_d (\varepsilon + m)$.

Proof:

By assumption, $f_u(u)$ is differentiable and nonzero for $u \geq 0$. Then, $\frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}}$ is bounded, as shown in Theorem 1. Since $\hat{u} = \underset{u \in \mathbb{R}^+}{\operatorname{Argmax}}\left(f_u(u).\, g_v(u+\varepsilon)\right)$, then for $\hat{u} > 0$,

$$f_u'(\hat{u})\, g_v(\hat{u}+\varepsilon) + f_u(\hat{u})\, g_v'(\hat{u}+\varepsilon) = 0$$

$$\frac{1}{\sigma_v^2} = \frac{1}{\hat{u}+\varepsilon} \frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}}$$

In addition, as $\sigma_v^2 \to 0$, the normal density tends to Dirac's delta function with its mass concentrated at the mean, i.e., $g_v(v) \to_p \delta(E(v))$. Then, $v \to_p E(v) = 0$.

a) As $\sigma_v^2 \to 0, v \to_p E(v) = 0$. Then $\varepsilon \to_p (-u)$. Additionally, as $\sigma_v^2 \to 0, \frac{1}{\hat{u}+\varepsilon}\frac{\mathrm{d}ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}} \to \infty$. Since $\frac{\mathrm{d}ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}}$ is bounded, then $\hat{u} + \varepsilon_i \to 0$, or $\hat{u} \to -\varepsilon_i$. It means $\hat{u} \to_p u$.

b) Although the convergence in probability, as shown in point (a), automatically implies the convergence in distribution, another direct proof, independent from the result of point (a) above, can be as follows.

$$\varepsilon = \sigma_v^2 \frac{f_u'(\hat{u})}{f_u(\hat{u})} - \hat{u}$$

Then,

$$\left|\frac{\mathrm{d}\varepsilon}{\mathrm{d}\hat{u}}\right| = \left|\sigma_v^2 \frac{\mathrm{d}^2 \ln[f_u(\hat{u})]}{(\mathrm{d}\hat{u})^2} - 1\right|$$

$$\tilde{f}_{\hat{u}}(\hat{u}) = h_\varepsilon\left(\varepsilon = \sigma_v^2 \frac{f_u'(\hat{u})}{f_u(\hat{u})} - \hat{u}\right)\left|\frac{\mathrm{d}\varepsilon}{\mathrm{d}\hat{u}}\right|$$

$$= \left|\sigma_v^2 \frac{\mathrm{d}^2 \ln[f_u(\hat{u})]}{(\mathrm{d}\hat{u})^2} - 1\right| \int_0^{+\infty} f_u(u)\, g_v\left(u + \sigma_v^2 \frac{f_u'(\hat{u})}{f_u(\hat{u})} - \hat{u}\right) du$$

As $\sigma_v^2 \to 0$,

$$\tilde{f}_{\hat{u}}(\hat{u}) \to_d \lim_{\sigma_v^2} h_\varepsilon\left(\varepsilon = \sigma_v^2 \frac{f_u'(\hat{u})}{f_u(\hat{u})} - \hat{u}\right)\left|\frac{\mathrm{d}\varepsilon}{\mathrm{d}\hat{u}}\right|$$

$$= \lim_{\sigma_v^2} \left|\sigma_v^2 \frac{\mathrm{d}^2 \ln[f_u(\hat{u})]}{(\mathrm{d}\hat{u})^2} - 1\right| \int_0^{+\infty} f_u(u)\, g_v\left(u + \sigma_v^2 \frac{f_u'(\hat{u})}{f_u(\hat{u})} - \hat{u}\right) du$$

$$= \int_0^{+\infty} f_u(u)\, \delta(u - \hat{u}) du$$

$$= f_u(\hat{u})$$

c) Since $v \sim N(0, \sigma_v^2)$, then $-\frac{v}{\sigma_v} \sim N(0,1)$. It means

$$-\frac{v}{\sigma_v} = -\frac{u + \varepsilon}{\sigma_v} = \frac{\hat{u} - u - \sigma_v^2 \frac{f_u'(\hat{u})}{f_u(\hat{u})}}{\sigma_v} \sim N(0,1)$$

As $\sigma_v^2 \to 0, \frac{\hat{u}-u-\sigma_v^2\frac{f_u'(\hat{u})}{f_u(\hat{u})}}{\sigma_v} \to_p \frac{\hat{u}-u}{\sigma_v}$, then $\frac{\hat{u}-u}{\sigma_v} \to_d N(0,1)$.

d) Since we have $\frac{1}{\sigma_v^2} = \frac{1}{\hat{u}+\varepsilon}\frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}}$, for a finite value of ε, as $\sigma_v^2 \to \infty$, it implies two possibilities. First, if $\frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}} = 0$, then $\hat{u} = Mode(u)$ since $f_u(u)$ is unimodal (log-concave). Second, if $\frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}} \neq 0$, then

$\frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}} < 0$ and $\hat{u} \to -\infty$, which is restricted to $\hat{u} = 0$ (note: at $\hat{u} \to +\infty$, $\frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}} \ngtr 0$). In such case, $\hat{u}$ again is the mode of u since $f_u(u)$ must be strictly monotonically decreasing.

e) From equation (2), we have $\tilde{u} = -\varepsilon + \sigma_v^2 \frac{\mathrm{d}\ln[f_u(\hat{u})]}{\mathrm{d}\hat{u}}$. As $\sigma_v^2 \to \infty$, we ca use the fact in point d and the mean value theorem around the mode m to write $\lim_{\sigma_v^2 \to \infty} \hat{u} \to \lim_{\sigma_v^2 \to \infty} \frac{-\varepsilon + \sigma_v^2 \left[\ln\left(f_u(m)\right)\right]' - \sigma_v^2\, m \left[\ln\left(f_u(m)\right)\right]''}{1 - \sigma_v^2 \left[\ln\left(f_u(m)\right)\right]''}$. This means, as $\sigma_v^2 \to \infty$, $\sigma_v^2\left[\ln\left(f_u(m)\right)\right]' + \left(\sigma_v^2\left[\ln\left(f_u(m)\right)\right]'' - 1\right)(\hat{u} - m) \to_d (\varepsilon + m)$. ■

Note that in point e of Theorem 3, for half normal and general truncated normal densities, the first derivate evaluated at the mode $\left[\ln\left(f_u(m)\right)\right]' = 0$, while for the exponential density, only the second derivate at the mode $\left[\ln\left(f_u(m)\right)\right]'' = 0$. Theorem 3 indicates that the conditional model score of inefficiency is a measure of unconditional inefficiency only in the absence of a random shock (noise). In addition, when the random shock variance increases the conditional mode score decreases towards the mode of the inefficiency itself.

A question might arise is, does a larger sample prevent the shrinkage of the JLMS estimators? In cross-sectional context, the simple answer to this question in the literature is 'no', for several reasons. First, since the inefficiencies are unobservable, the conditional estimators cannot be improved by learning from more data in contrast to regression models. Second, the productivity of each unit is observed only once, therefore, due to the assumption of independence between the units, conditional estimator of each unit inefficiency is conditioned on a single composite error corresponding to the unit itself. Third, due to lack of replication, the JLMS estimator is based on a guess (a typical value, like the mean or the mode) from the conditional distribution of the inefficiency, conditioned on a single composite error. Therefore, inconsistency and high uncertainty of the JLMS estimators are expected in the cross-sectional context. In econometrics literature, it is known that regularization increases the accuracy of an estimator by reducing its variance. The accuracy of a regularized estimator is due to a trade-off between decreased variance and increased bias.

3. Regularization

It has been shown in the literature that a maximum likelihood estimator is improved by maximizing an *a posteriori* or a penalized (regularized) likelihood function; see for example, Cox and O'Sullivan (1990) and Flynn et al. (2013). One can consider the conditional mode and the conditional mean from Bayes expected loss and the Bayes risk minimization perspective. For example, for the conditional mean, the loss function is $(u-\hat{u})^2$, whose risk minimization yields $\underset{\hat{u}\in\mathbb{R}^+}{\operatorname{Argmin}} E\{(u-\hat{u})^2|\varepsilon\} = E(u|\varepsilon)$. For the conditional mode, a the loss function is a zero-one indicator function $(I(u \neq \hat{u}) - 1)$, whose risk minimization yields $\underset{\hat{u}\in\mathbb{R}^+}{\operatorname{Argmin}} E\{I(u\neq\hat{u})-1|\varepsilon\} = \underset{\hat{u}\in\mathbb{R}^+}{\operatorname{Argmin}}[-f_{u|\varepsilon}(\hat{u})] = \underset{\hat{u}\in\mathbb{R}^+}{\operatorname{Argmax}} f_{u|\varepsilon}(\hat{u}) = Mode(u|\varepsilon)$.

A regularization of the risk minimization is achieved by adding extra information to, or imposing more constraints on, the risk function (expected loss). Suppose the constraints are a set of m zero-equality equations of twice differentiable functions $\boldsymbol{R}(u)$, i.e., $\boldsymbol{R}(u) = \boldsymbol{0}_{m\times 1}$. The regularized conditional mean of the inefficiency is the solution to the following constrained objective function.

$$\min_{\hat{u}\in\mathbb{R}^+} E\{(u-\hat{u})^2|\varepsilon\}$$
$$\text{Subject to: } \boldsymbol{R}(\hat{u}) = \boldsymbol{0}$$

The solution is $\underset{\hat{u}\in\mathbb{R}^+}{\operatorname{Argmin}} E\{(u-\hat{u})^2|\varepsilon\} + \boldsymbol{\lambda}'\boldsymbol{R}(\hat{u})$, where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. The regularized conditional mean is the solution to the following system of equations.

$$\begin{cases} \hat{u} - E(u|\varepsilon) + 0.5\boldsymbol{\lambda}'\nabla\boldsymbol{R}(\hat{u}) = 0 \\ \boldsymbol{R}(\hat{u}) = \boldsymbol{0} \end{cases}$$

For the conditional mode of the inefficiency, the objective function and the constraints are as follows.

$$\max_{\hat{u}\in\mathbb{R}^+} f_{u|\varepsilon}(\hat{u})$$
$$\text{Subject to: } \boldsymbol{R}(\hat{u}) = \boldsymbol{0}$$

The regularized conditional mode is the solution to the following system of equations.

$$\begin{cases} f'_{u|\varepsilon}(\hat{u}) + \boldsymbol{\lambda}'\nabla\boldsymbol{R}(\hat{u}) = 0 \\ \boldsymbol{R}(\hat{u}) = \boldsymbol{0} \end{cases}$$

The regularized JLMS estimators can be developed for both the mean and the mode. However, in the next section, we develop only the regularized conditional mode estimators for the three most commonly used inefficiency densities, which are the half normal, exponential and general truncated normal. The conditional mode is explicitly the maximum likelihood estimator of the inefficiency for its joint density with the composite error, or for its density conditioned on the composite error.

Restricted moments are common imposed constraints upon the likelihood functions (e.g., Hall and Presnell (1999)). The constraint can, for instance, be on the sum of the inefficiencies or the sum of squared inefficiencies. These are considered as constraints on the first and second moment of inefficiencies, respectively.

3.1 First- and Second-Moment Constraints

Inconsistency and high uncertainty of the JLMS estimators are expected in the cross-sectional context since any JLMS estimator of a unit inefficiency is conditioned on a single composite error corresponding to the unit itself. For each unit inefficiency estimation, we can also exploit extra information from other composite errors. For example, one can impose a restriction on all estimated inefficiencies such that their sample mean equals to the sample mean of the composite errors. Such a restriction is equivalent to a sample zero-mean constraint on the random shocks.

In terms of economic theory, the zero-mean random shock constraint is interpreted as a condition where the unit's productivity is invariant to the random shocks in the market. Let us take a production frontier model such as the Cobb-Douglas or a translog model with inefficiency as the single source of shortfall.

$$y_i = f(\boldsymbol{x}_i;\boldsymbol{\beta})\; e^{-u_i}$$

If firm i experiences a random shock (v_i), its production can expand or shrink, depending on the sign of v_i.

$$y_i = f(\boldsymbol{x}_i;\boldsymbol{\beta})\; e^{-u_i}\; e^{v_i}$$

Random shocks for some units can cover part of their inefficiencies, while for others, they might worsen their productivities, depending on whether the random shocks and the firm specific inefficiencies are in the same or opposite directions. An assumption can be that if a firm were consecutively exposed to the shocks from the whole market, its productivity would eventually return to the same level.

$$y_i = f(\boldsymbol{x}_i;\boldsymbol{\beta}).\, e^{-u_i} = f(\boldsymbol{x}_i;\boldsymbol{\beta}).\, e^{-u_i}.\, e^{v_1} \ldots e^{v_i} \ldots e^{v_n} = f(\boldsymbol{x}_i;\boldsymbol{\beta}).\, e^{-u_i}$$

Imposing the above-mentioned market (industry) shock-invariance assumption on the conditional mode of the inefficiency and the constraint that $\varepsilon_i = v_i - u_i$, for $i = 1, \dots, n$, with a sample of n units, is translated into the inefficiency sum (mean) restriction. The regularized conditional mode is the solution of the following constrained objective function.

$$\max_{u_1,\dots,u_n} \left\{ \sum_{i=1}^{n} \ln[g_v(u_i + \varepsilon_i)] + \ln[f_u(u_i)] \right\}$$

subject to:

$$\sum_{i=1}^{n} (u_i + \varepsilon_i) = 0$$

Using the Lagrange multiplier method, the above constrained objective function is written as

$$m_0(u_1, \dots, u_n | \varepsilon_1, \dots, \varepsilon_n) = \underset{u_1,\dots,u_n}{\operatorname{Argmax}} \left\{ \sum_{i=1}^{n} \ln[g_v(u_i + \varepsilon_i)] + \ln[f_u(u_i)] + \lambda \sum_{i=1}^{n} (u_i + \varepsilon_i) \right\}$$

Then, the estimated inefficiencies are forced to fulfill the constraint $\sum_{i=1}^{n} v_i = 0$. We can extend the number of restrictions, for example, by adding a restriction on the variance or the sum of squares of the estimated conditional modes, as follows.

$$\max_{u_1,\dots,u_n} \left\{ \sum_{i=1}^{n} \ln[g_v(u_i + \varepsilon_i)] + \ln[f_u(u_i)] \right\}$$

Subject to:

$$\sum_{i=1}^{n} (u_i + \varepsilon_i) = 0$$

$$\sum_{i=1}^{n} u_i^2 = c$$

The c on the right side of the second constraint can be, for example, $c = nE(u^2)$. With the Lagrange multiplier method, the above constrained objective function is written as

$$m_0(u_1, \dots, u_n | \varepsilon_1, \dots, \varepsilon_n) = \underset{u_1,\dots,u_n}{\operatorname{Argmax}} \left\{ \sum_{i=1}^{n} \ln[g_v(u_i + \varepsilon_i)] + \ln[f_u(u_i)] + \lambda \sum_{i=1}^{n} (u_i + \varepsilon_i) + \theta \left[\sum_{i=1}^{n} u_i^2 - c \right] \right\}.$$

Table 1: The unit inefficiency estimator $\hat{u} = \max\{0, -\tilde{u}\}$, with $\tilde{u}$ given in the cells of the following table.

Measure	**Half Normal,** $\boldsymbol{u_i \sim N^+(0, \sigma_u^2)}$	**Truncated Normal,** $\boldsymbol{u_i \sim N^+(\mu, \sigma_u^2)}$	**Exponential,** $\boldsymbol{u_i \sim Exp(\sqrt{\sigma_u^2})}$
Unregularized Mode$(u \mid \varepsilon_i)$	$\frac{\sigma_u^2 \varepsilon}{\sigma_v^2 + \sigma_u^2}$	$\frac{\sigma_u^2 \varepsilon - \mu \sigma_v^2}{\sigma_v^2 + \sigma_u^2}$	$\varepsilon + \frac{\sigma_v^2}{\sqrt{\sigma_u^2}}$
Mode$(u \mid \varepsilon_i)$ with 1st Moment Constraint	$\frac{\sigma_u^2 \varepsilon + \sigma_v^2 \bar{\varepsilon}}{\sigma_v^2 + \sigma_u^2}$	$\frac{\sigma_u^2 \varepsilon + \sigma_v^2 \bar{\varepsilon}}{\sigma_v^2 + \sigma_u^2}$	ε
Mode$(u \mid \varepsilon_i)$ with 1st & 2nd Moment Constraints	$\frac{\varepsilon + \bar{\varepsilon}\left(\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sigma_u^2 - \bar{\varepsilon}^2}} - 1\right)}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sigma_u^2 - \bar{\varepsilon}^2}}}$	$\frac{\varepsilon_i + \bar{\varepsilon}\left(\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{E(u^2) - \bar{\varepsilon}^2}} - 1\right)}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{E(u^2) - \bar{\varepsilon}^2}}}$	$\frac{\varepsilon + \bar{\varepsilon}\left(\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{2\sigma_u^2 - \bar{\varepsilon}^2}} - 1\right)}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{2\sigma_u^2 - \bar{\varepsilon}^2}}}$

Table 1 shows the regularized conditional mode estimators of the unit inefficiency in a production model when inefficiencies follow the three most commonly used inefficiency densities: half normal, truncated normal and exponential densities. Note that; $\bar{\varepsilon} = \frac{\sum_{i=1}^{n} \varepsilon_i}{n}$ and $\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=0}^{n} (\varepsilon_i - \bar{\varepsilon})^2}{n}$, and for $u_i \sim N^+(\mu, \sigma_u^2)$,

$$E(u^2) = \sigma_u^2 \left(1 + \frac{\mu}{\sqrt{\sigma_u^2}} \frac{\phi\left(\frac{\mu}{\sqrt{\sigma_u^2}}\right)}{\Phi\left(\frac{\mu}{\sqrt{\sigma_u^2}}\right)} + \frac{\mu^2}{\sigma_u^2} \right).$$

For half-normal and exponential distributions, $E(u^2)$ is σ_u^2 and $2\sigma_u^2$, respectively. Thus, with the first- and second-moment constraints, $\tilde{u}$ has the same closed-form solution in terms of $E(u^2)$. The conditional mean $E(u \mid \varepsilon_i)$ with each of the densities shown in Table 1, has the following general form:

$$E(u \mid \varepsilon_i) = \tilde{\sigma} \frac{\phi\left(\frac{\tilde{\mu}}{\tilde{\sigma}}\right)}{\Phi\left(\frac{\tilde{\mu}}{\tilde{\sigma}}\right)} - \tilde{\mu}$$

where $\tilde{\mu}$ is the negative of the cells of the 1st row corresponding to the unregularized mode ($-\tilde{u}$) in Table 1, and $\tilde{\sigma} = \sigma_v$ for the exponential density and $\tilde{\sigma} = \frac{\sigma_v \sigma_u}{\sigma}$ for each of the half-normal and truncated normal densities, where $\sigma^2 = \sigma_v^2 + \sigma_u^2$.

As stated in Theorem 3, the conditional mode of inefficiency shrinks towards the mode of inefficiency in response to any noise variance inflation. From Table 1, we realize that the regularized estimators (explicitly with the first-moment restriction) serve to hold the unit inefficiency estimators away from the inefficiency mode by adding fractions of the noise variance to the conditional mode estimators, i.e., they serve to reduce the shrinkage of the conditional mode estimator towards the mode of inefficiency (0 or $\mu > 0$).

For each of the above inefficiency densities and the set of the constraints, the same estimators are developed for a cost function. To save space, they are not presented here, but they are obtained straightforwardly by altering the signs of ε and $\bar{\varepsilon}$ inside the above closed-form formulae in Table 1. The purpose of presenting regularized conditional mode estimators is to introduce the methodology with closed-form mathematical expressions. Analogous to the conditional mode, the methodology can also be applied to regularized conditional mean estimators, with properly selected constraints[4].

4. Simulations

An extensive simulation study is conducted to assess the performance of the proposed methodology relative to the classical methods of the conditional mean/mode of inefficiency. The varying factors of the simulation study are (i) the sample size, (ii) the inefficiency distribution, (iii) the noise variance, (iv) the inefficiency variance and (iv) the location parameter, only when the inefficiency follows a truncated normal distribution. Samples of sizes 20, 30, 50, 100 and 250 were simulated. A Cobb-Douglas production model was assumed, and each sample consisted of three simulated variables: production output and labor and capital inputs. The values for the intercept, elasticities, and means and variances of labor and capital were selected to imitate a production model originally used in the Cobb-Douglas (1928) article. Specifically, the regression coefficients were selected as $\boldsymbol{\beta} = \{-0.25, 0.25, 075\}$ and labor and capital were drawn from the bivariate normal distribution $\boldsymbol{x} \sim N_2\left([5.5, 5], \begin{bmatrix} 0.25 & 0 \\ 0 & 0.04 \end{bmatrix}\right)$.

[4] For instance, a quadratic interpolated polynomial through the points 0, maximum deterministic inefficiency (the negative of the minimum composite error in a production function) and the estimated mean of inefficiency density.

Random shocks from the distribution $N(0, \sigma_v^2)$ and inefficiencies from distributions $N^+(0, \sigma_u^2)$, $N^+(\mu, \sigma_u^2)$ and $Exp(\sqrt{\sigma_u^2})$ were simulated and used to simulate model (1). The noise variance was given values of $\sigma_v^2 = 0.1,\ 0.5$, and 0.9, and the inefficiency variance σ_u^2 was selected such that $Var(u) = 1 - \sigma_v^2$, i.e., the variance of the composite error was kept at unity with each simulated sample ($Var(\varepsilon) = 1$). For the truncated normal inefficiency, the variance is also affected by the location parameter μ. In the simulations, $\mu = 0$ (for half normal) and $\mu = 0.1$, and 0.2 (for truncated normal).

The simulations were implemented as follows. For each sample size, the simulated design matrix was fixed across all simulations. To assess the performance of each estimator across different ranks of inefficiency, we considered two different scenarios. The first scenario is to rank the firms constantly based on their inefficiencies such that the first simulated firm always receives the lowest simulated inefficiency, and the last simulated firm always receives the largest inefficiency. The second scenario is to randomly rank the firms based on their inefficiencies. The results of the two scenarios were consistent; hence, the second scenario was followed to avoid any potential effect due to differences in production input across the firms. This process was repeated 100 times, i.e., 100 samples of ranked inefficiencies were simulated from the above-mentioned inefficiency probability distributions. For each of the 100 samples of inefficiencies, 100 samples of noise terms were randomly generated from the above-mentioned normal distributions. This resulted in 10000 replications for each of the 60 combinations of the above factors (sample size, probability distribution, σ_v^2, σ_u^2 and μ).

With each replication of the simulation process, the four measures of unit-level inefficiencies were calculated, which were the conditional mean, conditional mode, conditional mode with first-moment constraint and conditional mode with first- and second-moment constraint. The Mean Squared Error (MSE) for the i^{th} firm's inefficiency was calculated as follows.

$$MSE(\hat{u}_i) = \frac{\sum_{k=1}^{100} \sum_{j=1}^{100} \left(\hat{u}_{ji} - u_{ki}\right)^2}{10000}$$

The squared bias for the i^{th} firm's inefficiency measure was calculated as follows.

$$Bias^2(\hat{u}_i) = \frac{\sum_{k=1}^{100}\left(\frac{\sum_{j=1}^{100}\hat{u}_{ji}}{100} - u_{ki}\right)^2}{100}$$

Each measure's relative efficiency to the conditional mean efficiency was calculated as

$$Relative\ MSE(\hat{u}_i) = \frac{MSE\big(E(u|\varepsilon_i)\big)}{MSE(\hat{u}_i)}.$$

In the above formulae, i represents the unit, j is the noise replication and k represents inefficiency replications. Some results of the relative MSE are shown in Figure 1, Figure 2 and Figure 3. The rest is presented in the Supplementary Materials appendix. In the graphs, the x-axis represents the rank of the inefficiency[5] and the y-axis represents the relative MSE and the bias squared. All the simulations and calculations were run in STATA/SE 16 for Windows 64 bit using the *sfcross* command by Belotti et al. (2013).

The results of the simulations in Figures 1-3 show that when estimating large inefficiencies, the regularized conditional mode estimator, especially with the first-moment constraint, outperforms the unregularized conditional mean and mode estimators as the signal-to-noise ratio (σ_u/σ_v) increases. This is an important finding since government agencies responsible for regulating local monopoly markets want to estimate the inefficiency score of the most inefficient firms with as high accuracy as possible. While the signal-to-noise ratio seems to be more decisive for the relative performance of the regularized estimator than the sample size and distributional assumption, its performance improves further when inefficiencies are exponentially distributed and when the sample size is not very large. Some points can be listed as follows:

- The unregularized conditional mode is almost always the most accurate estimator for units with no or small inefficiencies— a result that is expected due to its shrinkage-towards-mode property. [6]

[5] An alternative is to use the actual scores of inefficiencies or the technical efficiencies. However, the conclusion from the graphs would be the same and only on different scales.
[6] The modes are zero (for half-normal and exponential distributions) and 0.1 for truncated normal distributions.

- The unregularized conditional mean is the most accurate estimator of the unit inefficiency for middle ranks since it is a shrinkage estimator towards the mean.
- Regularized conditional mode estimators, especially the one with the first-moment constraint, are the most accurate estimators of unit inefficiencies that are more to the right tail of the distribution (highly ranked), unless it is a case with low signal-to-noise ratio, in which the unconditional mode (for lower ranks) and unconditional mean (for higher ranks) outperform the regularized estimators.
- A summary of the above 3 points is that the analysts should make an effort to learn the characterizing conditions of the application at hand since the preferred estimation approach changes as the signal-to-noise ratio changes and depends on how and where the inefficiencies are distributed. Therefore, depending on the signal-to-noise ratio and the rank of the inefficiency (simply based on the estimated composite errors), the optimal estimator can be a mixture (or weighted sum) of the conditional mode (for lower inefficiency ranks), conditional mean (for middle inefficiency ranks) and regularized conditional mode, especially the one subject to the first-moment constraint (for high inefficiency ranks).

Figure 1: Normal—Half-Normal Model: Relative Inefficiency (MSE Ratio) compared to $E(u|\varepsilon)$

- - - - Conditional mean $E(u|\varepsilon)$, reference
- - - - Conditional $Mode(u|\varepsilon)$
—— $Mode(u|\varepsilon)$, restricted with $\sum u$
—— $Mode(u|\varepsilon)$, regularized with $\sum u$ & $\sum u^2$

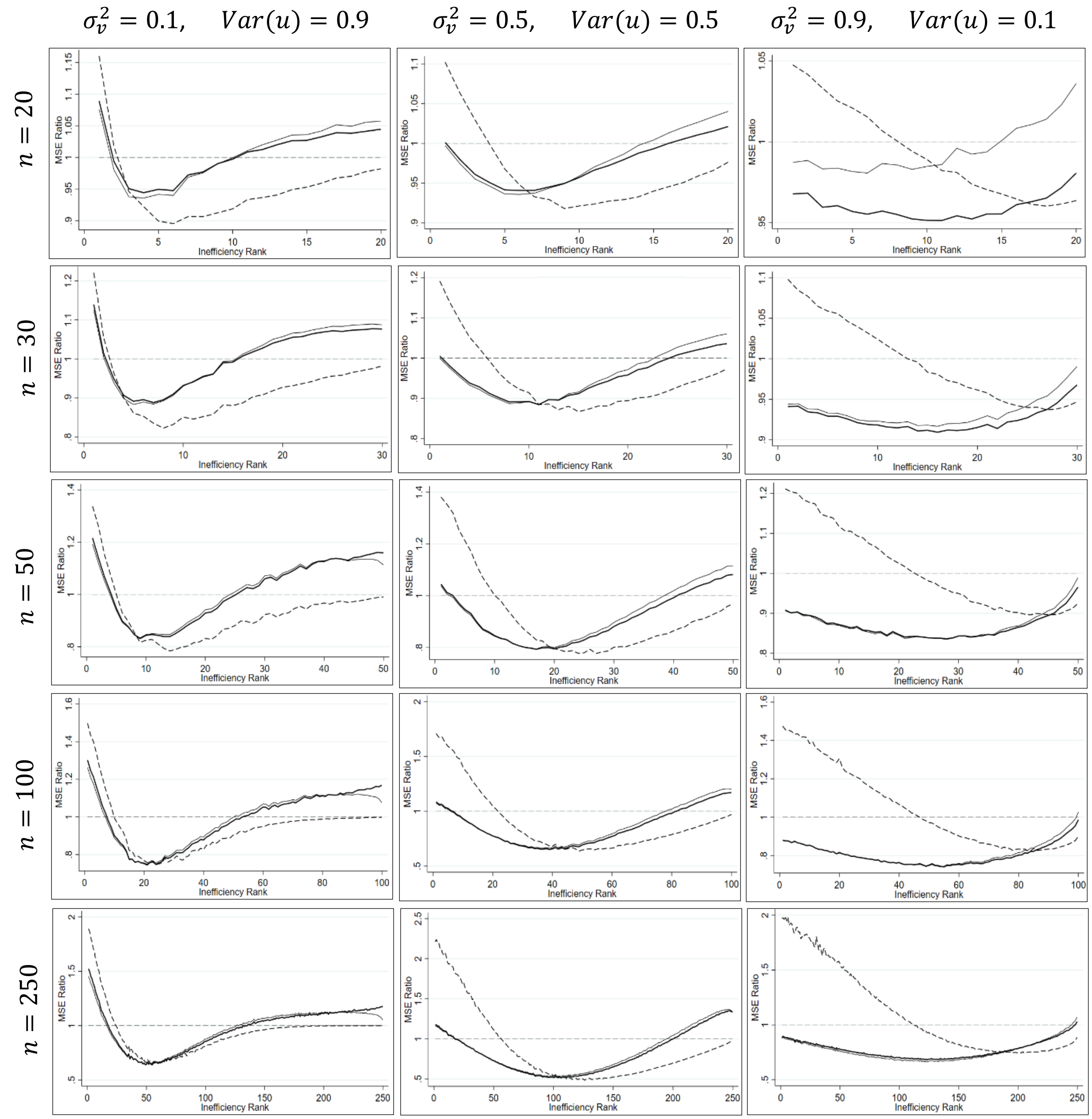
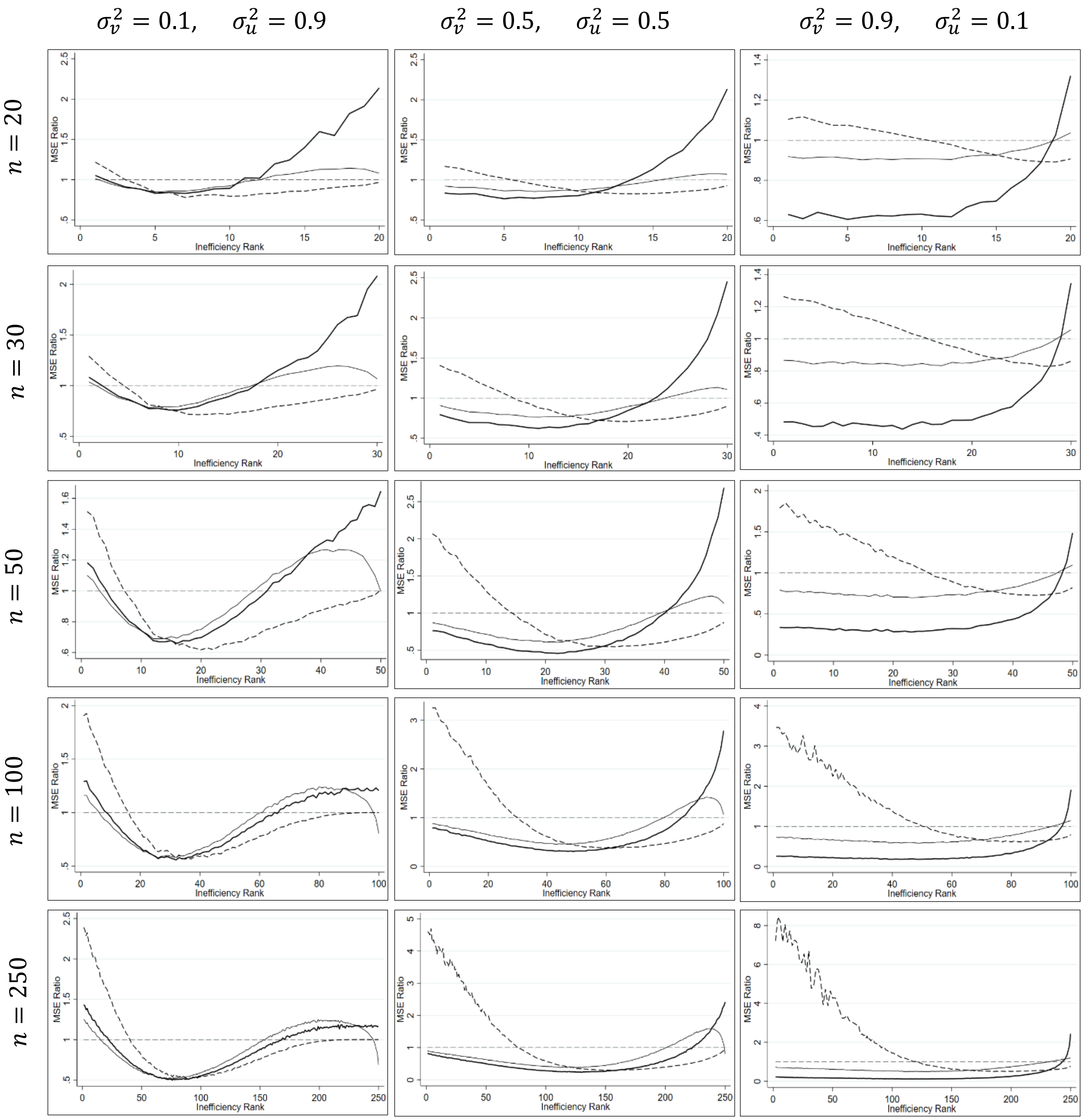


Figure 2: Normal—Exponential Model: Relative Inefficiency (MSE Ratio) compared to $E(u|\varepsilon)$

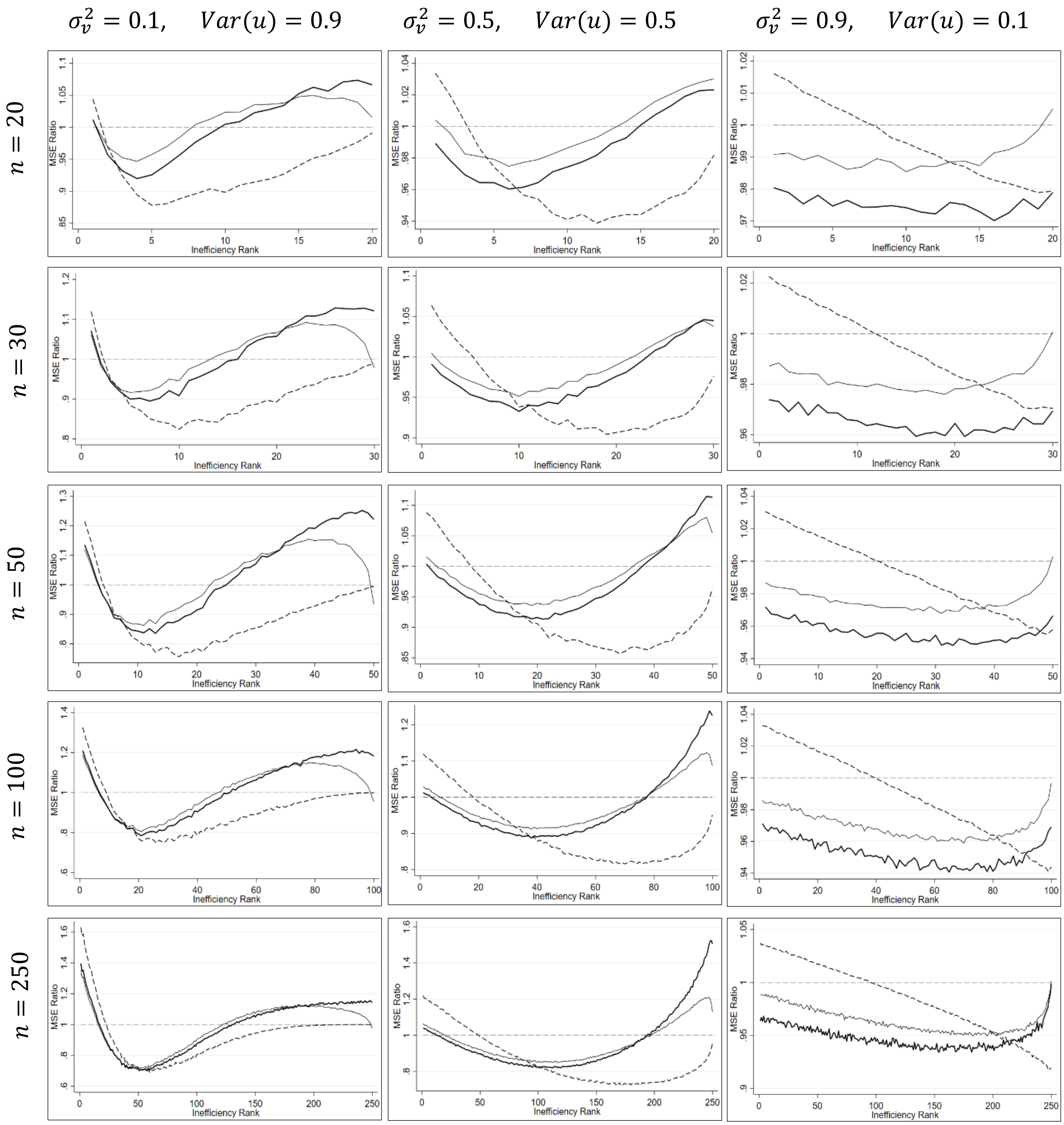
- - - - Conditional mean $E(u|\varepsilon)$, reference
- - - - Conditional $Mode(u|\varepsilon)$
—— $Mode(u|\varepsilon)$, restricted with $\sum u$
—— $Mode(u|\varepsilon)$, restricted with $\sum u$ & $\sum u^2$



Note: The larger the noise variance, the choppier the curve of the unregularized conditional mode is.

Figure 3: Normal—Truncated Normal Model (with $\mu = 0.1$): Relative Inefficiency (MSE Ratio) compared to $E(u|\varepsilon)$

- - - - Conditional mean $E(u|\varepsilon)$, reference
- - - - Conditional $Mode(u|\varepsilon)$
—— $Mode(u|\varepsilon)$, constrained with $\sum u$
—— $Mode(u|\varepsilon)$, constrained with $\sum u$ & $\sum u^2$



Note: The larger the noise variance, the choppier the curves of the regularized conditional mode estimators are.

5. Application

We consider the Swedish electricity distribution market that consisted of 154 local monopolies with complete data in 2013. The regulator wants to know the extent to which each firm can improve relative to the efficient frontier. For that purpose, we specify and estimate a variable cost (c) function where the number of customers/connection points (s) is the relevant output variable and the price of labor (l) and electricity (e) are the corresponding input prices. This production process is similar to what has been used in the past in this field; see e.g., Söderberg (2008), p. 65-66, for an extensive literature review. The price of electricity is included because firms purchase electricity to cover network losses and pay for transit on the high voltage network. The electricity price is calculated as the total costs of transit and the losses divided by the sum of the losses and high voltage deliveries.

Since the estimation of the unit inefficiency is a post-estimation procedure in SFA, entering the discussion of the selection between different productivity models, for instance between Cobb-Douglas and translog, might divert our attention away from the purpose of our proposed regularized estimators. Therefore, to save space, we only assume a Cobb-Douglas production model, and specify the variable cost function as $c_i = \alpha s_i^{\beta} e_i^{\gamma} l_i^{\delta}$, where i denotes the firm. The homogeneity restriction can be imposed by normalizing c_i and l_i by e_i, which after natural logarithm transform allows us to write the model as:

$$ln\left(\frac{c_i}{e_i}\right) = \beta_0 + \beta_1 \ln(s_i) + \beta_2\, ln\left(\frac{l_i}{e_i}\right).$$

This expression has normal Cobb-Douglas properties, e.g., β_1 reveals the nature of the scale of production. Specifically, if $\beta_1 < 1$, then there are economies of scale; if $\beta_1 = 1$, then there is constant returns to scale; and if $\beta_1 > 1$, there are diseconomies of scale. It is straightforward to extend this Cobb-Douglas model to a stochastic frontier setting with inefficiency (u) and idiosyncratic error (ν) terms (Coelli et al., 2005):

$$ln\left(\frac{c_i}{e_i}\right) = \beta_0 + \beta_1 \ln(s_i) + \beta_2\, ln\left(\frac{l_i}{e_i}\right) + \nu_i + u_i.$$

Data on variable costs (Opex), the number of customers, and the price of electricity are collected from the Swedish energy regulator (the Energy Markets Inspectorate). The price of labor, which measures the average regional salary for employees in the public sector, is collected from Statistics Sweden. Data are cross-sectional from the year 2013. Because the objective function, or type of customers, can be different for different ownership forms, as shown by Meade and Söderberg (2020), we argue that the regulator has to restrict the benchmark to the firms that have the same type of owners. For Swedish electricity distribution, therefore, we need three different benchmark samples: (i) municipality owned firms (*n*=99), (ii) cooperatively owned firms (*n*=32), and (iii) firms owned by private investors (*n*=23). Some descriptive statistics of the data are presented in Table A1 in the Appendix.

Table 2: Model Estimates

Estimate (S.E.)		Half Normal	Exponential	Truncated Normal
Municipality Owned, $n = 99$	β_0	-8.510 (0.3309)	-8.4294 (0.329)	-8.5354 (0.4585)
	β_1	.8087 (0.0255)	0.8075 (0.0259)	.8090 (0.0257)
	β_2	1.1079 (0.0325)	1.1078 (0.0328)	1.1079 (0.0325)
	σ_u^2	.0529 (0.0301)	.013 (0.0104)	.0459 (0.0669)
	σ_v^2	.0222 (0.0097)	.0287 (0.0090)	.0204 (0.0243)
	μ	-	-	.0827 (0.8431)
	$H_0: No\ ineficiency$	LR test, $\bar{\chi}^2$ (p-value): 1.58 (0.104)	LR test, $\bar{\chi}^2$ (p-value): 1.24 (0.133)	Z test, z (p-value): 1.24 (0.108)
Cooperatively Owned, $n = 32$	β_0	-9.122 (0.3741)	-9.0215 (0.3026)	-9.0329 (0.3117)
	β_1	.8854 (0.0279)	.8912 (0.0275)	.8903 (0.0278)
	β_2	1.1198 (0.0336)	1.1101 (0.0276)	1.1113 (0.0286)
	σ_u^2	.0925 (0.0393)	.0358 (0.018)	.5632 (2.3897)
	σ_v^2	.0064 (0.0090)	.0087 (0.005)	.0083 (0.0051)
	μ	-	-	-2.5088 (12.5981)
	$H_0: No\ inefficiency$	LR test, $\bar{\chi}^2$ (p-value): 5.67 (0.009)	LR test, $\bar{\chi}^2$ (p-value): 6.47 (0.005)	Z test, z (p-value): 2.321 (0.010)
Private Investors, $n = 23$	β_0	-8.0991 (1.2217)	-7.9061 (0.6246)	-7.919 (0.6574)
	β_1	.9424 (0.0435)	.9441 (0.028)	.9442 (0.0283)
	β_2	.9589 (0.1033)	.9513 (0.0679)	.9514 (0.0685)
	σ_u^2	.0201 (0.1272)	.0367 (0.0357)	.8233 (11.2301)
	σ_v^2	.1549 (0.4719)	.0388 (0.023)	.0375 (0.029)
	μ	-	-	-3.6906 (58.5673)
	$H_0: No\ inefficiency$	LR test, $\bar{\chi}^2$ (p-value): 1.22 (0.134)	LR test, $\bar{\chi}^2$ (p-value): 1.24 (0.132)	Z test, z (p-value): 1.163 (0.122)

In Figure 4, we see that the regularized estimators suggest that the highly inefficient firms have less technical efficiency (or equivalently larger inefficiency scores) compared to what the unregularized estimators estimate. Any inference regarding unit inefficiency can be poor when only a single sample is available, as it is in a cross-sectional context. However, we know that the conditional mean and the conditional mode (Theorem 3) are shrinking estimators, i.e., they underestimate larger inefficiencies.

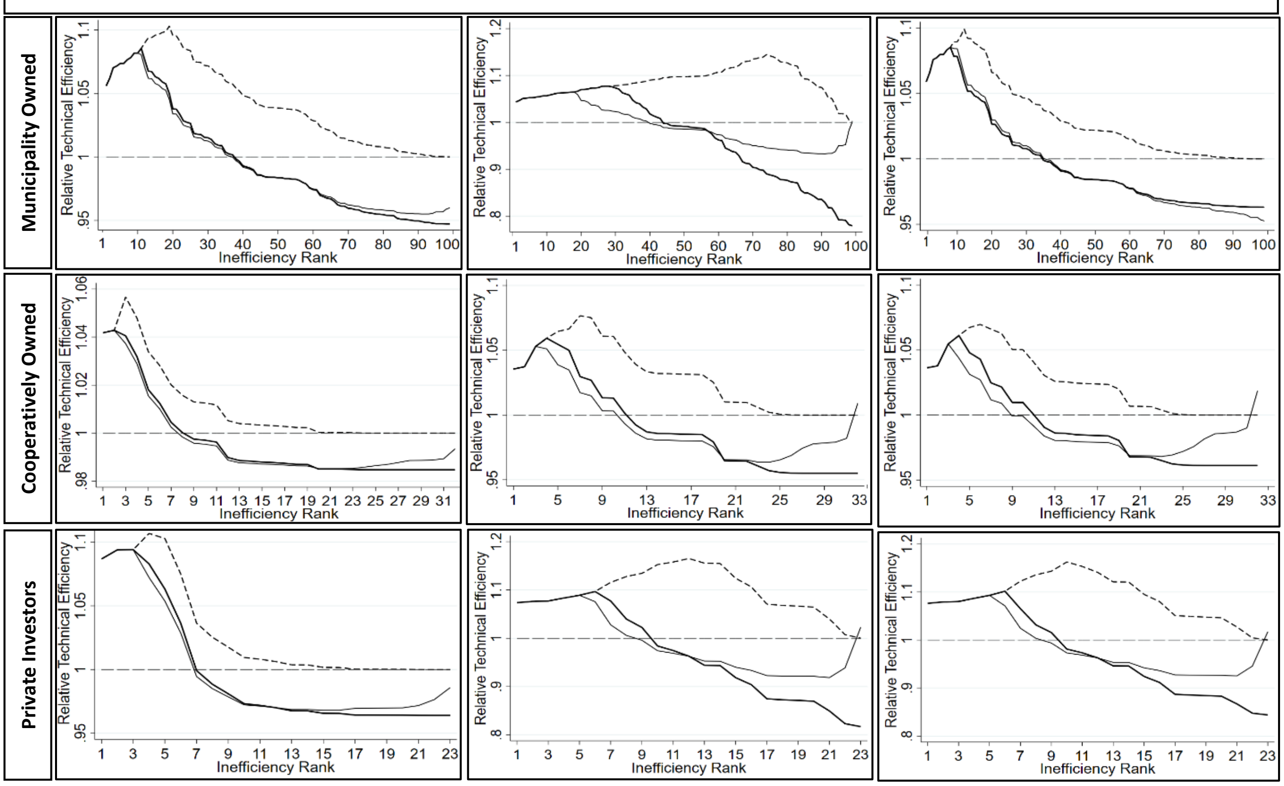
Therefore, our regularized estimators behave better in that sense, i.e., they estimated larger inefficiencies further from the mean/mode compared to the unregularized estimators. In addition, they have desired properties in the sense that they follow the theoretical first and second moments of inefficiency, i.e., their sample mean and variance are close to the estimated industry mean and variance.

Figure 4: Relative technical efficiency based on restricted and unrestricted conditional mode estimates, compared to technical efficiency based on unrestricted $E(u|\varepsilon)$ estimtes.

- - - - - Conditional mean $E(u|\varepsilon)$, reference — - - - - - Conditional $Mode(u|\varepsilon)$

——— $Mode(u|\varepsilon)$, restricted with $\sum u$ — ——— $Mode(u|\varepsilon)$, restricted with $\sum u$ & $\sum u^2$

Half-Normal **Exponential** **Truncated Normal**



The inference with a single sample is challenging. We checked the relative performance of each estimator by running a simulation with the same sampled data (number of customers and prices) but with the costs generated from the estimated parameters ($\hat{\sigma}_v^2$, $\hat{\sigma}_u^2$, $\hat{\mu}$ and $\widehat{\boldsymbol{\beta}}$) in Table 2. The simulation procedure was the same as that explained in the simulation section (Figure A1 in the Appendix).

6. Conclusions

The conditional mean/mode estimator of unit inefficiency is a shrinkage estimator towards the inefficiency mode (mean), depending on the noise variance (or signal-to-noise ratio). It is mostly different from the firm's inefficiency itself unless there is no noise in the productivity model. The proposed regularized conditional mode estimators outperform the classical conditional mode/mean estimators, especially for highly inefficient units.

The constraints used in this paper were imposed on the first and the second moments of the inefficiencies when estimating the conditional mode of inefficiency. The idea can be further generalized to other sorts of constraints, distributions other than those used in this paper, or constraints on the conditional mean. In this article, the methodology is discussed in a cross-sectional context. However, it can be directly applied to a panel data context wherever the conditional mode/mean of the unit inefficiency is estimated. According to Tsionas (2017), one issue that continues to plague SFA is the endogeneity of the inputs. Our methodology is also directly applicable to the SFA methods dealing with the endogeneity. It can be applied to the endogeneity situation discussed by Amsler et al. (2016) and the semiparametric estimation method in Fan et al. (1996). And most importantly, the proposed regularized estimators are beneficial to regulators for accurately estimating high unit inefficiencies since the benchmark methods systematically underestimate the inefficiency of less efficient units.

Acknowledgments

The authors appreciate the valuable comments on the manuscript in the 2021 virtual North American Productivity Workshop in Miami, which helped us to rewrite the manuscript in its current version.

Conflict of Interest

The Authors do not have any conflict of interest to declare.

Appendix

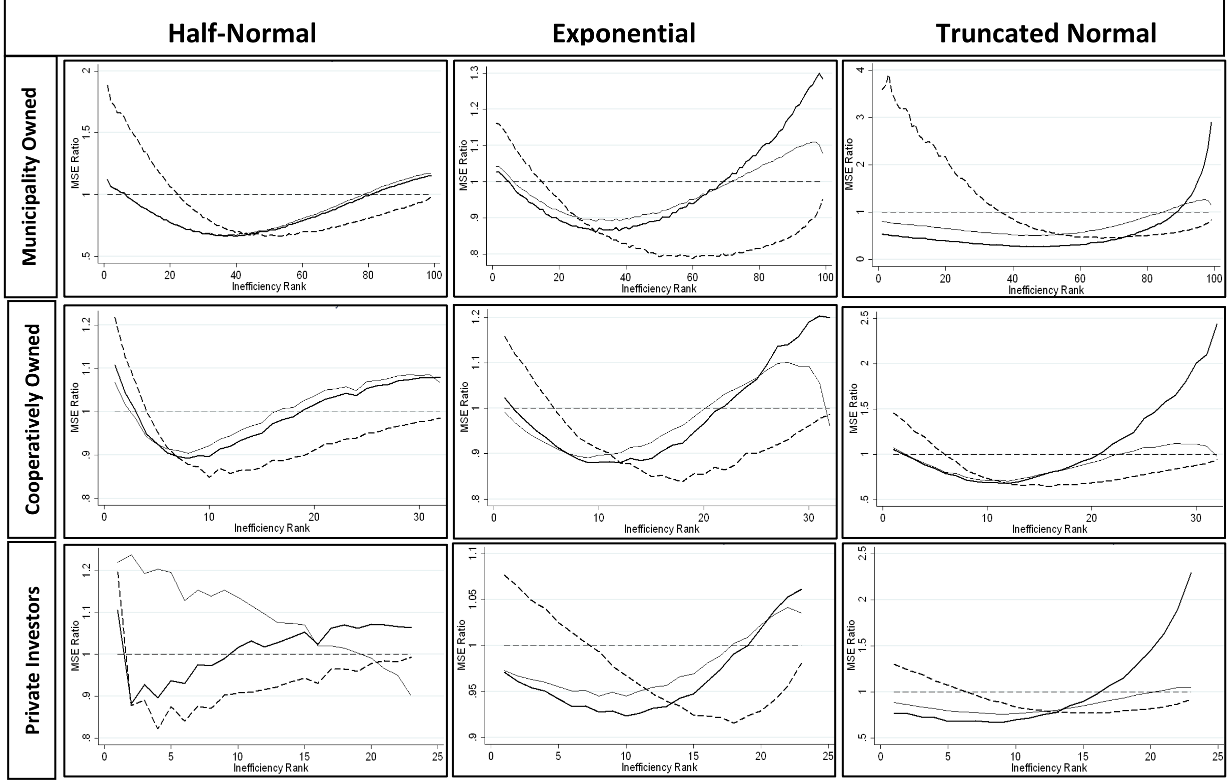
Table A1. Descriptive statistics of the data used in the Application section

Variable	Mean	S.D.	Min	Max
Panel A: municipality owned (n=99)				
Variable cost (c)	57 593	75 404	7 207	612 999
Number of customers (s)	22 160	31 480	2 303	256 549
Price of electricity (e)	0.3673	0.3379	0.0390	1.9990
Price of labor (l)	21 746	356	21 250	23 160
Panel B: cooperatively owned (n=32)				
Variable cost (c)	18 602	14 536	2 851	57 421
Number of customers (s)	4 916	4 599	808	19 120
Price of electricity (e)	1.3177	1.0874	0.0660	4.4560
Price of labor (l)	21 860	390	21 250	23 160
Panel C: owned by private investors (n=23)				
Variable cost (c)	242 954	506 236	509	2 304 885
Number of customers (s)	89 709	193 788	158	802 484
Price of electricity (e)	0.5683	0.5803	0.0380	2.5280
Price of labor (l)	21 777	586	21 250	23 160

Figure A1: The MSE ratio of unrestricted and restricted conditional mode estimators, compared to the unrestricted $E(u|\varepsilon)$ with the simulations based on the data in the Application section

- - - - - Conditional mean $E(u|\varepsilon)$, reference — - - - - Conditional $Mode(u|\varepsilon)$

—— $Mode(u|\varepsilon)$, restricted with $\sum u$ — —— $Mode(u|\varepsilon)$, restricted with $\sum u$ & $\sum u^2$



References

[1] Aigner, D., Lovell, C. A. K., and Schmidt, P. (1977), "Formulation and Estimation of Stochastic Frontier Production Function Models", *Journal of Econometrics*, 6, 21–37.

[2] Amsler, C., Prokhorov, A. and Schmidt, P., (2016), "Endogenous Stochastic Frontier Models", *Journal of Econometrics*, 190, 280-288.

[3] Andor, M.A., Parmeter, C. and Sommer, S. (2019). Combining uncertainty with uncertainty to get certainty? Efficiency analysis for regulation purposes. *European Journal of Operational Research*, 274(1): 240-252.

[4] Badunenko, O., Henderson, D. J. and Kumbhakar, C. S. (2012), "When, where and how to perform efficiency estimation", *Journal of the Royal Statistical Society, Series A* 175: 863–892.

[5] Battese, G. E. and Coelli, T. J. (1988), "Prediction of Firm-Level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data", *Journal of Econometrics*, 38, 387–99.

[6] Behr, A. (2010). Quantile regression for robust bank efficiency score estimation. *European Journal of Operational Research*. 200: 568–58.

[7] Belotti, F., Daidone, S., Ilardiand, G. and Atelaa, V. (2013), "Stochastic frontier analysis using Stata", *The Stata Journal*, 13(4): 719-758.

[8] Bera, A. and Sharma, S. (1999), "Estimating Production Uncertainty in Stochastic Frontier Production Function Models", *Journal of Productivity Analysis*, 12: 187–210.

[9] Bernini, C., Freo, M. & Gardini, A. (2004). Quantile estimation of frontier production function. *Empirical Economics*, 29: 373-381.

[10] Bogetoft, P. and Otto, L., (2011), "Benchmarking with DEA, SFA, and R", *International Series in Operations Research and Management Science*, Vol. 157. Springer http://dx.doi.org/10.1007/978-1-4419–7961-2_10.

[11] Cobb, C. W. and Douglas, P. H. (1928), "A Theory of Production, American Economic Review", *Supplementary, Papers and Proceedings of the Fortieth Annual Meeting of the American Economic Association*, 18(1): 139–165.

[12] Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., and Battese, G. E. (2005), *An introduction to efficiency and productivity analysis*. Springer Science and Business Media.

[13] Cox, D. and O'Sullivan, F. (1990), "Asymptotic analysis of penalized likelihood and related estimators", *Annals of Statistics*, 18:1676–1695.

[14] Fan, Y., Li, Q. and Weersink, A. (1996), "Semiparametric Estimation of Stochastic Production Frontier Models", *Journal of Business and Economic Statistics*, 14: 460–468.

[15] Flynn, C.J., Hurvich, C.M. and Simonoff, J.S. (2013), "Efficiency for Regularization Parameter Selection in Penalized Likelihood Estimation of Misspecified Models", *Journal of American Statistical Association*, 108:503, 1031-1043, DOI:10.1080/01621459.2013.801775.

[16] Greene, W. H. (1980), "On the estimation of a flexible frontier production model", *Journal of Econometrics*, 13(1), 101–15.

[17] Greene, W. H. (1990), "A gamma-distributed stochastic frontier model", *Journal of Econometrics*, 46(1–2): 141–164.

[18] Grossman, N. (2005). A C^{∞} Lagrange Inversion Theorem, *The American Mathematical Monthly*, 112 (6): 512-514

[19] Hall, P. & Presnell, B. (1999) Density Estimation under Constraints, *Journal of Computational and Graphical Statistics*, 8:2, 259-277, DOI: 10.1080/10618600.1999.10474813

[20] Horrace, W. C. (2005), "On ranking and selection from independent truncated normal distributions", *Journal of Econometrics*,126, 335–354.

[21] Horrace, W. C. and Parmeter, C. F. (2018), "A Laplace stochastic frontier model", *Econometric Reviews*, 37:3, 260-280.

[22] Jondrow, J., Lovell, C. A. K., Materov, I. S., and Schmidt, P. (1982). On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. *Journal of Econometrics*, 19, 233–8.
[23] Jradia, S., Parmeter, C. F. & Ruggiero, J. (2019). Quantile Estimation of the Stochastic Frontier Model. *Economics Letters*. 182: 15-18.
[24] Kumhakar, S. C., Ghosh, S. & McGuckin, J. T. (1991). A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms, *Journal of Business & Economic Statistics*, 9,3: 279-286.
[25] Kumbhakar, S. C., Park, B. U., Simar, L., Tsionas, E. G. (2007). Nonparametric stochastic frontiers: A local maximum likelihood approach, *Journal of Econometrics*, 137, 1: 1-27.
[26] Kumbhakar, S., Wang, H. J., and Horncastle, A. (2015), *A Practitioner Guide to Stochastic Frontier Analysis using Stata*. New York: Cambridge University Press.
[27] Kumbhakar, S. C. and Lovell, C. A. K. (2000), *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press.
[28] Kumbhakar S.C., Parmeter C.F., Zelenyuk V. (2020), "Stochastic Frontier Analysis: Foundations and Advances I". *In: Ray S., Chambers R., Kumbhakar S. (eds) Handbook of Production Economics*. pp. 1-40, Springer, Singapore.
[29] Meade, R., and Söderberg, M. (2020), "Is welfare higher when utilities are owned by customers instead of investors? Evidence from electricity distribution in New Zealand", *Energy Economics*, Vol. 86, p. 104700.
[30] Meeusen, W., and van den Broeck, J. (1977), "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error", *International Economic Review*, 18, 435–44.
[31] Nguyen, N. B. (2010), *Estimation of Technical Efficiency in Stochastic Frontier Analysis*, Doctoral Dissertation, Bowling Green State University.
[32] Ondrich, J., and Ruggiero, J. (2001), "Efficiency measurement in the stochastic frontier model", *European Journal of Operational Research*, 129(2): 434–442.
[33] Papadopoulos, A. (2021). "Stochastic Frontier Models Using the Generalized Exponential Distribution". *Journal of Productivity Analysis*, 55: 15-29, https://doi.org/10.1007/s11123-020-00591-9
[34] Ruggiero, J. (1999), "Efficiency estimation and error decomposition in the stochastic frontier model: A Monte Carlo analysis", *European Journal of Operational Research*, 115(6):555–563.
[35] Stevenson, R. E. (1980), "Likelihood Functions for Generalized Stochastic Frontier Estimation", *Journal of Econometrics*, 13, 57–66
[36] 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, Part 3, 405-34.
[37] Söderberg, M. (2008), *Four essays on efficiency in Swedish electricity distribution*. PhD thesis University of Gothenburg.
[38] Tsionas, M. G. (2017), "When, Where, and How of Efficiency Estimation: Improved Procedures for Stochastic Frontier Modeling", *Journal of the American Statistical Association*, 112:519, 948-965.
[39] Tsionas, M. G. (2021). Optimal combinations of stochastic frontier and data envelopment analysis models. *European Journal of Operational Research*, 294 (**2**): 790-800.
[40] Wang, W. S. and Schmidt, P. (2009), "On the distribution of estimated technical efficiency in stochastic frontier models", *Journal of Econometrics*, 148: 35-45.
[41] Wang, Y., Wang, S., Dang, C., Ge, W. (2014). Nonparametric quantile frontier estimation under shape restriction. *European Journal of Operational Research*. 232: 671–6
[42] Whittaker, E. T. and Watson, G. N. (1927), *A Course of Modern Analysis: An Introduction to the General Theory of Infinite Processes and of Analytic Functions with an Account of the Principal Transcendental Functions*, Cambridge University Press, 4th Edition, New York.