Chapter 3
Stochastic Frontier Models with Bounded Inefficiency

Pavlos Almanidis, Junhui Qian, and Robin C. Sickles

JEL Classification: C13, C21, C23, D24, G21

3.1 Introduction

The parametric approach to estimate stochastic production frontiers was introduced by Aigner et al. (1977), Meeusen and van den Broeck (1977), and Battese and Corra (1977). These approaches specified a parametric production function and a two-component error term. One component, reflecting the influence of many unaccountable factors on production as well as measurement error, is considered "noise" and is usually assumed to be normally distributed. The other component describes inefficiency and is assumed to have a one-sided distribution, of which the conventional candidates include the half normal (Aigner et al. 1977), truncated normal (Stevenson 1980), exponential (Meeusen and van den Broeck 1977) and gamma (Greene 1980a,b, 1990; Stevenson 1980). This stochastic frontier production function has become an iconic modeling paradigm in econometric research, rate making

P. Almanidis (✉)
International Tax Services, Ernst & Young LLP, 222 Bay Street, P.O. Box 251, Toronto, ON, M5K 1J7, Canada
e-mail: pavlos.almanidis@ca.ey.com

J. Qian
School of Economics, Shanghai Jiao Tong University,
535 Fa Hua Zhen Road, Shanghai 200052, China
e-mail: jhqian@sjtu.edu.cn

R.C. Sickles
Department of Economics, Rice University, 6100 South Main Street,
Houston, TX 77005-1892, USA
e-mail: rsickles@rice.edu

R.C. Sickles and W.C. Horrace (eds.), *Festschrift in Honor of Peter Schmidt: Econometric Methods and Applications*, DOI 10.1007/978-1-4899-8008-3_3,

decisions in regulated industries across the world, in evaluating outcomes of market reforms in transition economies, and in establishing performance benchmarks for local, state, and federal governmental activities.

In this paper we propose a new class of parametric stochastic frontier models with a more flexible specification of the inefficiency term, which we view as improvement on the basic iconic stochastic frontier production model. Instead of allowing unbounded support for the distribution of productive (cost) inefficiency term in the right (left) tail, we introduce an unobservable upper bound to inefficiencies or a lower bound to the efficiencies, which we call the *inefficiency bound*. The introduction of the inefficiency bound makes the parametric stochastic frontier model more appealing for empirical studies in at least two aspects. First, it is plausible to allow only bounded support in many applications of stochastic frontier models wherein the extremely inefficient firms in a competitive industry of market are eliminated by competition. Bounded inefficiency makes sense in this setting since the extremely inefficient stores will be forced to close and thus individual production units constitute a truncated sample.[1] This is consistent with the arguments of Alchian (1950) and Stigler (1958) wherein firms are at any point in time not in a static long run equilibrium, but rather are tending to that situation as they are buffeted by demand and cost shocks. As a consequence, even if we correctly specify a family of distributions for the inefficiency term, the stochastic frontier model may still be misspecified. This particular setting is one in which the inefficiency bound is informative as an indicator of competitive pressures and/or the extent of supervisory oversight by direct management or by corporate boards. In settings in which firms can successfully differentiate their product, which is the typical market structure and not the exception, or where there are market concentrations that may reflect collusive behavior or conditions for a natural monopoly and regulatory oversight, incentives to fully exploit market power or to instead make satisficing decision are both possible outcomes. Much more likely is that it is not one or the other but some middle ground between the two extremes that would be found empirically.[2]

[1] In addition, the frequent use of balanced panels in empirical studies would in effect eliminate those failing firms from the sample and thus would provide more merit to the bounded inefficiency model.

[2] "The quiet life hypothesis" (QLH) by Hicks (1935) argues that, due to management's subjective cost of reaching the optimal profits, firms use their market power to allow inefficient allocation of resources. Increasing competitive pressure is likely to force management to work harder to reach optimal profits. Another hypothesis that relates market power and efficiency is "the efficient structure hypothesis" (ESH) by Demsetz (1973). ESH argues that firms with superior efficiencies or technologies have lower costs and therefore higher profits. These firms are assumed to gain larger market shares which lead to higher concentration. Recently Kutlu and Sickles (2012) have constructed a model in which the dynamic game is played out and have tested for the alternative outcomes, finding support for the QLH in certain airlines city-pair markets and the ESH in others. Orea and Steinbuks (2012) have also explored the use of such a lower bound in their analysis of market power in the California wholesale electricity market.

A second justification for our introduction of the inefficiency bound into the classical stochastic production frontier model is that our model points to an explanation for the finding of positive skewness in many applied studies using the traditional stochastic frontier, and thus to the potential of our bounded inefficiency model to explain these positive ("wrong") skewness findings.[3] Researchers have often found positive instead of negative skewness in many samples examined in applied work, which may point to the stochastic frontier being incorrectly specified. However, we conjecture that the distribution of the inefficiency term may itself be negatively skewed, which may happen if there is an additional truncation on the right tail of the distribution. One such specification in which this is a natural consequence is when the distribution of the inefficiency term is doubly truncated normal, that is, a normal distribution truncated at a point on the right tail as well as at zero. As normal distributions are symmetric, the doubly truncated normal distribution may exhibit negative skewness if the truncation on the right is closer to the mode than that on the left. We also consider the truncated half normal distribution, which is a special case of the former, and the truncated exponential distribution. Although these two distributions are always positively skewed, the fact that there is a truncation on the right tail makes the skewness very hard to identify empirically. That is to say, when the true distribution of the one-sided inefficiency error is bounded (truncated), the extent to which skewness is present in any finite sample may be substantially reduced, often to the extent that negative sample skewness for the composite error is not statistically significant. Thus the finding of positive skewness may speak to the weak identifiability of skewness properties in a bounded frontier model.

In addition to proposing new parametric forms for the classical stochastic production frontier model, we also show that our models are identifiable, and in which cases the identification is local or global. Initial consistent estimates are based on method of moments estimates, based on explicit analytic expressions which we derive, and which either can be used in a two-step method of scoring or as starting values in solving the normal equations for the relevant sample likelihood, based on the parametric density functions whose expressions we also provide. As the regulatory conditions for maximum likelihood estimation method are satisfied, we employ it in order to obtain consistent and asymptotically efficient estimates of the model parameters, including this of the inefficiency bound. We conduct Monte Carlo experiments to study the finite sample behavior of our estimators. We also extend the model to the panel data setting and allow for a time-varying inefficiency bound. By allowing the inefficiency bound to be time-varying, we contribute another time-varying technical efficiency model to the efficiency literature. Our model differs from those most commonly used in the literature, e.g., Cornwell et al. (1990), Kumbhakar (1990), Battese and Coelli (1992), and Lee and Schmidt (1993) in

[3] The term wrong is set in quotes to point out that the conventional wisdom that positive skewness is inconsistent with the standard stochastic frontier production model errors skewness is not necessarily the correct wisdom.

that, while previous time-varying efficiency models are time-varying in the mean or intercept of individual effects, our model is time-varying in the lower support of the distribution of individual effects.

The outline of this paper is as follows. In Sect. 3.2 we present the new models and derive analytic formula for density functions and expressions that allow us to evaluate inefficiencies. Section 3.3 deals with the positive skewness issue inherent in the traditional stochastic frontier model. Section 3.4 discusses the identification of the new models and the methods of estimation. Section 3.5 presents Monte Carlo results on the finite sample performance of the bounded inefficiency model vis-a-vis classical stochastic frontier estimators. The extension of the new models to panel data settings and specification of the time-varying bound is presented in Sect. 3.6. In Sect. 3.7 we give an illustrative study of the efficiency of US banking industry in 1984–2009. Section 3.8 concludes.

3.2 The Model

We consider the following Cobb-Douglas production model,

$$y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{i,k} + \varepsilon_i \tag{3.1}$$

where

$$\varepsilon_i = v_i - u_i. \tag{3.2}$$

For every production unit i, y_i is the log output, x_{ik} the k-th log input, v_i the noise component, and u_i the (nonnegative) inefficiency component. We maintain the usual assumption that v_i is iid $N(0, \sigma_v^2)$, u_i is iid, and v_i and u_i are independent from each other and from regressors. Clearly we can consider other more flexible functional forms for production (or cost) that are linear or linear in logarithms, such as the generalized Leontief or the transcendental logarithmic, or ones that are nonlinear. The only necessary assumption is that the error process ε_i is additively separable from the functional forms we employ in the stochastic production (cost) frontier.

As described in the introduction, our model differs from the traditional stochastic frontier model in that u_i is of bounded support. Additional to the lower bound, which is zero and which is the frontier, we specify an upper bound to the distribution of u_i (in the case of the cost frontier $\varepsilon_i = v_i + u_i$). In particular, we assume that u_i is distributed as doubly truncated normal, the density of which is given by

$$f(u) = \frac{\frac{1}{\sigma_u}\phi(\frac{u-\mu}{\sigma_u})}{\Phi(\frac{B-\mu}{\sigma_u}) - \Phi(\frac{-\mu}{\sigma_u})}\mathbf{1}_{[0,B]}(u), \ \sigma_u > 0, B > 0 \tag{3.3}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of the standard normal distribution, respectively, and $1_{[0,B]}$ is an indicator function. It is a distribution obtained by truncating $N(\mu, \sigma_u^2)$ at zero and $B > 0$. The parameter B is the upper bound of the distribution of u_i and we may call it the inefficiency bound. The inefficiency bound may be a useful index of competitiveness of a market or an industry.[4] In the banking industry, which we examine in Sect. 3.7, the inefficiency bound may also represent factors that influence the financial health of the industry. It may be natural to extend this specification and treat the bound as a function of individual specific covariates z_i, such as $\exp(\delta' z_i)$, which would allow identification of bank-specific measures of financial health.

Using the usual nomenclature of stochastic frontier models, we may call the model described above the normal-doubly truncated normal model, or simply, the doubly truncated normal model. The doubly truncated normal model is rather flexible. It nests the truncated normal ($B \to \infty$), half normal ($\mu = 0$ and $B \to \infty$), and truncated half normal models ($\mu = 0$). One desirable feature of our model is that the doubly truncated normal distribution may be positively or negatively skewed, depending on the truncation parameter B. This feature provides us with an alternative explanation for the positive skewness problem prevalent in empirical stochastic frontier studies. This will be made more clear later in the paper. Another desirable feature of our model is that, like the truncated normal model, it can describe the scenario that only a few firms in the sector are efficient, a phenomenon that is described in the business press as “few stars, most dogs”, while in the truncated half normal model and the truncated exponential model (in which the distribution of u_i is truncated exponential), most firms are implicitly assumed to be relatively efficient.[5]

In Table 3.1 we provide detailed properties of our model. In particular, we present the density functions for the error term ε_i, which is necessary for maximum likelihood estimation, and the analytic form for $E[u_i|\varepsilon_i]$, which is the best predictor of the inefficiency term u_i under our assumptions, and the conditional distribution of u_i given ε_i, which is useful for making inferences on u_i. The results for the truncated half normal model, a special case of the doubly truncated normal model ($\mu = 0$), are also presented. Finally, we also provide results for the truncated exponential model, in which the inefficiency term u_i is distributed according to the following density function,

$$f(u) = \frac{1}{\sigma_u(1 - e^{-B/\sigma_u})} e^{-\frac{u}{\sigma_u}} \mathbf{1}_{[0,B]}(u), \sigma_u > 0, B > 0 \tag{3.4}$$

The truncated exponential distribution can be further generalized to the truncated gamma distribution, which shares the nice property with the doubly truncated normal distribution that it may be positively or negatively skewed.

[4] The inefficiency bound has a natural role in gauging the tolerance for or ruthlessness against inefficient firms. It is also worth mentioning that, using this bound as the “inefficient frontier,” we may define “inverted” efficiency scores in the same spirit of “Inverted DEA” described in Entani et al. (2002).

[5] We thank C. A. K. Lovell for providing us this link between our econometric methodology and the business press.

Table 3.1 Key results. $f(\varepsilon)$ is the density of $\varepsilon = v - u$, $\mathbb{E}(u|\varepsilon)$ is the conditional mean of u given ε, and $f(u|\varepsilon)$ is the conditional density of u given ε. $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively. And $\mathbf{1}_{[0,B]}(\cdot)$ is an indicator function

Model	$f(\varepsilon)$	$\mathbb{E}(u\|\varepsilon)$	$f(u\|\varepsilon)$
Doubly truncated normal	$\left[\Phi(\frac{B-\mu}{\sigma_u}) - \Phi(\frac{-\mu}{\sigma_u})\right]^{-1} \cdot \left[\frac{1}{\sigma}\phi(\frac{\varepsilon+\mu}{\sigma})\right] \cdot$ $\left[\Phi(\frac{(B+\varepsilon)\lambda+(B-\mu)\lambda^{-1}}{\sigma}) - \Phi(\frac{\varepsilon\lambda-\mu\lambda^{-1}}{\sigma})\right]$	$\mu_* + \sigma_* \frac{\phi(-\frac{\mu*}{\sigma*})-\phi(\frac{B-\mu*}{\sigma*})}{\Phi(\frac{B-\mu*}{\sigma*})-\Phi(-\frac{\mu*}{\sigma*})}$	$\frac{\frac{1}{\sigma*}\phi(\frac{u-\mu*}{\sigma*})}{\Phi(\frac{B-\mu*}{\sigma*})-\Phi(-\frac{\mu*}{\sigma*})}\mathbf{1}_{[0,B]}(u)$
Truncated half normal	$\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$, $\lambda = \sigma_u/\sigma_v$	$\mu_* = \frac{\mu\sigma_v^2-\varepsilon\sigma_u^2}{\sigma^2}$, $\sigma_* = \frac{\sigma_u\sigma_v}{\sigma}$	
	$[\Phi(\frac{B}{\sigma_u}) - 1/2]^{-1} \cdot \frac{1}{\sigma}\phi(\frac{\varepsilon}{\sigma}) \cdot$ $[\Phi(\frac{(B+\varepsilon)\lambda+B\lambda^{-1}}{\sigma}) - \Phi(\frac{\varepsilon\lambda}{\sigma})]$	$\mu_* + \sigma_* \frac{\phi(-\frac{\mu*}{\sigma*})-\phi(\frac{B-\mu*}{\sigma*})}{\Phi(\frac{B-\mu*}{\sigma*})-\Phi(-\frac{\mu*}{\sigma*})}$	$\frac{\frac{1}{\sigma*}\phi(\frac{u-\mu*}{\sigma*})}{\Phi(\frac{B-\mu*}{\sigma*})-\Phi(-\frac{\mu*}{\sigma*})}\mathbf{1}_{[0,B]}(u)$
		$\mu_* = -\frac{\varepsilon\sigma_u^2}{\sigma^2}$, $\sigma_* = \frac{\sigma_u\sigma_v}{\sigma}$	
Truncated exponential	$\frac{e^{\frac{\varepsilon}{\sigma_u}+\frac{\sigma_v^2}{2\sigma_u^2}}[\Phi(\frac{B+\varepsilon}{\sigma_v}+\frac{\sigma_v}{\sigma_u})-\Phi(\frac{\varepsilon}{\sigma_v}+\frac{\sigma_v}{\sigma_u})]}{\sigma_u(1-e^{-B/\sigma_u})}$	$\mu_* + \sigma_v \frac{\phi(-\frac{\mu*}{\sigma_v})-\phi(\frac{B-\mu*}{\sigma_v})}{\Phi(\frac{B-\mu*}{\sigma_v})-\Phi(-\frac{\mu*}{\sigma_v})}$	$\frac{\frac{1}{\sigma_v}\phi(\frac{u-\mu*}{\sigma_v})}{\Phi(\frac{B-\mu*}{\sigma_v})-\Phi(-\frac{\mu*}{\sigma_v})}\mathbf{1}_{[0,B]}(u)$
		$\mu_* = -\varepsilon - \frac{\sigma_v^2}{\sigma_u}$	

For the doubly truncated normal model and the truncated half normal model, the analytic forms of our results use the so-called γ-parametrization, which specifies

$$\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}, \quad \gamma = \sigma_u^2/\sigma^2. \tag{3.5}$$

By definition $\gamma \in [0, 1]$, a compact support, which is desirable for the numerical procedure of maximum likelihood estimation. Another parametrization initially employed by Aigner et al. (1977) is the λ-parametrization

$$\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}, \quad \lambda = \sigma_u/\sigma_v. \tag{3.6}$$

We may check that when $B \to \infty$, the density function for ε_i in the doubly truncated normal model reduces to that of the truncated normal model introduced by Stevenson (1980). Furthermore, if $\mu = 0$, it reduces to the likelihood function for the half normal model introduced by Aigner et al. (1977). Similarly, the truncated exponential model reduces to the exponential model introduced by Meeusen and van den Broeck (1977).

3.3 The Skewness Issue

A common and important methodological problem encountered when dealing with empirical implementation of the stochastic frontier model is that the residuals may be skewed in the wrong direction. In particular, the ordinary least squares (OLS) residuals may show positive skewness even though the composed error term $v - u$ should display negative skewness, in keeping with $u's$ positive skewness. This problem has important consequences for the interpretation of the skewness of the error term as a measure of technological inefficiency. It may imply that a nonrepresentative random sample had been drawn from an inefficiency distribution possessing the correct population skewness (see Carree 2002; Greene 2007; Simar and Wilson 2010; Almanidis and Sickles 2011[6]; Feng et al. 2012). This is considered a finite sample "artifact" and the usual suggestion in the literature and by programs

[6] This paper goes far beyond the topics covered in Almanidis and Sickles (2011). In this paper we are concerned with the set identification of the bounded inefficiency model as well as in its use to better understand the behavior of this lower bound as the banking industry moved towards and through the financial meltdown. Such a pattern of a lower bound for inefficiency during the period prior to the meltdown speaks to the industry becoming lax in its allowance of banks that are not efficient in their provision of intermediation services as they appeared to focus instead on other off-balance sheet activities for which of course we do not have much credible information, as they are off-balance sheet operations. Our paper also shows the advantages of specifying a lower bound and estimating it, along with the other parameters of the model. Our paper is based on substantial efforts in data construction and uses data that has not appeared yet in the literature. Our paper also carries out a much more detailed set of MC experiments.

implementing stochastic frontier models is to treat all firms in the sample as fully efficient and proceed with straightforward OLS based on the results of Olson et al. (1980) and Waldman (1982). As this would suggest setting the variance of the inefficiency term to zero, it would have problematic impacts on estimation and on inference. Simar and Wilson (2010) suggest a bagging method to overcome the inferential problems when a half-normal distribution for inefficiencies is specified. However, a finding of positive skewness in a sample may also indicate that inefficiencies are in fact drawn from a distribution which has positive skewness.[7] Carree (2002) considers one-sided distributions of inefficiencies (u_i) that can have negative or positive skewness. However, Carree (2002) uses the binomial distribution, which is a discrete distribution wherein continuous inefficiencies fall into discrete "inefficiency categories" and which implicitly assumes that only a very small fraction of the firms attain a level of productivity close to the frontier, especially when u_i is negatively skewed.[8]

Our model addresses the positive skewness problem in the spirit of Carree (2002), but with a more appealing distributional specification on the efficiency term. For the doubly truncated normal model, let $\xi_1 = \frac{-\mu}{\sigma_u}$, $\xi_2 = \frac{B-\mu}{\sigma_u}$, and $\eta_k \equiv \frac{\xi_1^k \phi(\xi_1) - \xi_2^k \phi(\xi_2)}{\Phi(\xi_2) - \Phi(\xi_1)}$, $k = 0, 1, \ldots, 4$. Note that η_0 is the inverse Mill's ratio and it is equal to $\sqrt{2/\pi}$ in the half normal model, and that ξ_1 and ξ_2 are the lower and upper truncation points of the standard normal density, respectively. The skewness of the doubly truncated normal distribution is given by

$$S_u = \frac{2\eta_0^3 - \eta_0(3\eta_1 + 1) + \eta_2}{\left(1 - \eta_0^2 + \eta_1\right)^{3/2}}. \tag{3.7}$$

It can be checked that when $B > 2\mu$, S_u is positive and when $B < 2\mu$, S_u is negative. Since $B > 0$ by definition, it is obvious that only when $\mu > 0$ is it possible for u_i to be negatively skewed. The larger μ is, the larger range of values B may take such that u_i is negatively skewed. Consider the limiting case where a normal distribution with $\mu \to \infty$ is truncated at zero and $B > 0$. An infinitely

[7] Simar and Wilson (2010) consider inferences on efficiency conditional on composite error. They propose a bagging method and a bootstrap procedure for interval prediction and show that they are superior over the conventional methods that are based on the estimated conditional distribution. The relation of theirs to our paper is that they show that their methods work even when "wrong skewness" appears, while traditional MLE-based procedures do not. When the latter discovers a "wrong skewness", either (i) obtain a new sample, or (ii) re-specify the model (but not like what we do). What is common between our paper and SW is that both address the skewness problem. But "wrong skewness" in SW is due to finite sample bad luck, while we argue that it may be due to model specification. Larger samples would correct finite sample bad luck, but not if the underlying DGP is doubly truncated as we propose. The skewness problem is not the main issue in SW but their paper does have implications for it. The SW paper focuses on computational matters, while our paper concerns econometric specification and estimation.

[8] A negatively skewed doubly truncated normal inefficiency distribution does not necessarily imply that there are only few units in the population that operate close to the frontier.

large μ means that there is effectively no truncation on the left at all and that any finite truncation on the right gives rise to a negative skewness. Finally, for both the truncated half normal model ($\mu = 0$) and the truncated exponential model, the skewness of u_i is always positive.

Consequently, the doubly truncated normal model has a residual that has an ambiguous sign of the skewness, which depends on an unobservable relationship between the truncation parameter B and μ. We argue that this ambiguity theoretically could explain the prevalence of the positive skewness problem in applied stochastic frontier research. When the underlying data generating process for u_i is based on the doubly truncated normal distribution, increasing sample size does not solve the positive skewness problem. The skewness of the OLS residual ε may be positively skewed even when sample size goes to infinity. Hence the positive skewness problem also may be a large sample problem.[9]

Based on the above discussion, it is clear that the doubly truncated normal model generalizes the stochastic frontier model in a way that allows for positive as well as negative skewness for the residual. In addition, although the truncated half normal and the truncated exponential models have negative (correct) skewness in large samples, the existence of the inefficiency bound reduces the identifiability of negative skewness in finite sample, often to the extent that positive skewness appears. This implies that finding a positive skewness does not necessarily mean that the stochastic frontier model is inapplicable. It may be due to a finite sample "artifact" (Simar and Wilson 2010) or it may be that we are studying a market or an industry in which firms do not fall below some minimal level of efficiency in order to remain in the market or industry. In the latter case, the traditional unbounded support for the inefficiency term would be misspecified and should be substituted with the model of bounded inefficiency.

3.4 Identification and Estimation

3.4.1 Identification

We utilize the set or partial identification concepts that have been revisited (see, for example, Tamer 2010) and that were enunciated early in the production setting by Marschak and Andrews (1944) (see also the critique by Nerlove (1965)). That this has been the relatively recent interest of many econometricians speaks to a cycle of classical econometric study that has defined the production frontier portion of

[9] See Almanidis and Sickles (2011) for more discussion and simulation study on positive skewness issue in parametric stochastic frontier models.

Peter Schmidt's research that our paper develops. We can put it into a historical perspective by looking at the intellectual development of the production function by Paul Samuelson (see his 1979 review of his professor Paul Douglas), his student Lawrence Klein whose classic Textbook of Econometrics (1953) sold at the unheard price of \$6.00 and which provides insights today for those interested in production econometrics, his student Arthur Goldberger (see, for example, "The Interpretation and Estimation of Cobb-Douglas Functions", 1968), his student Jan Kmenta (see, for example, Zellner et al. 1966), and his student Peter Schmidt, whose work on the stochastic frontier production function with Dennis Aigner and C. A. Knox Lovell (1977) is regarded as the seminal research contribution to the field of productive efficiency econometrics. In turn, each of these legacies arguably can be viewed as the most successful student of their respective professor. Our contribution is leveraged by these seminal contributions as well as the selective constraints that economic theory has imposed on their contributions, which we try to address in our stochastic frontier model with bounded inefficiency.

Identification using first and second order moments is a well-accepted methodology. Our models are not identified by such moments alone and require higher order moments. The use of higher order moments to identify and estimate econometric models is well-known and has proven quite important in parametric econometric modeling (see, for example, Cragg 1997; Dagenais and Dagenais 1997). Identification strategies that utilize the properties of the underlying joint distribution function for the exponential class, requiring the identification of distributions defined by third and forth order moments, have been the mainstay of recent work in nonparametric identification (Newey and Powell 2003; Matzkin 2012). Alternative approaches have also been introduced to utilize other types of information, such as heteroskedastic covariance restrictions to obtain point and set identification for parametric and semiparametric models (Lewbel 2012). We explore the sensitivity of the use of such higher order moments restrictions in our Monte Carlo experiments.

Identification of our model may be done in two parts. The first part is concerned with the parameters describing the technology, and the second part identifies the distributional parameters using the information contained in the distribution of the residual. For models without an intercept term the identification conditions for the first part are well known and are satisfied in most of the cases. The structural parameters can be consistently obtained by applying straightforward OLS. However, for models containing an intercept term there is a need to bias correction it using the distributional parameters since $E[\varepsilon] = -E[u] \neq 0$ (see Afriat 1972; Richmond 1974). Therefore, the identification of the second part, which is based on method-of-moments requires a closer examination. Table 3.2 lists the population (central) moments of (ε_i) for the doubly truncated normal model and the truncated exponential model. The moments of the truncated half normal model can be obtained by setting $\mu = 0$ in the doubly truncated normal model. These results are essential for the discussion of identification and the method of moments estimation.

Table 3.2 Central moments of ε

Moment	Doubly-truncated-normal
ψ_1	$-\mu - \sigma_u \eta_0$
ψ_2	$\sigma_u^2 \left(1 - \eta_0^2 + \eta_1\right) + \sigma_v^2$
ψ_3	$-\sigma_u^3 \left(2\eta_0^3 - 3\eta_1 \eta_0 - \eta_0 + \eta_2\right)$
ψ_4	$\sigma_u^4 \left(3 + 3\eta_1 + \eta_3 - 2\eta_0^2 - 4\eta_0\eta_2 + 6\eta_0^2\eta_1 - 3\eta_0^4\right) + 6\sigma_u^2\sigma_v^2 \left(1 - \eta_0^2 + \eta_1\right) + 3\sigma_v^4$
ψ_5	$-10\sigma_v^2\sigma_u^3 \left(2\eta_0^3 - 3\eta_1\eta_0 - \eta_0 + \eta_2\right)$
	$-\sigma_u^5 \left(\eta_4 + 4\eta_2 - 5\eta_0\eta_3 + 10\eta_0^2\eta_2 - 10\eta_0^3\eta_1 + 10\eta_0^3 - 15\eta_0\eta_1 + 4\eta_0^5 - 7\eta_0\right)$
	See the text for the definitions of $\eta_k, k = 0, \ldots, 4$
	Truncated-exp.
ψ_1	$-\sigma_u \left(1 - \frac{\kappa}{e^\kappa - 1}\right)$
ψ_2	$\sigma_v^2 + \sigma_u^2 \frac{e^{2\kappa} - \left(\kappa^2 + 2\right)e^\kappa + 1}{e^{2\kappa} - 2e^\kappa + 1}$
ψ_3	$-\sigma_u^3 \frac{2e^{3\kappa} - \left(\kappa^3 + 6\right)e^{2\kappa} + \left(6 - \kappa^3\right)e^\kappa - 2}{e^{3\kappa} - 3e^{2\kappa} + 3e^\kappa - 1}$
ψ_4	$\sigma_u^4 \frac{-9e^{4\kappa} + 36e^{3\kappa} - 54e^{2\kappa} + 36e^\kappa - 9 + 6\kappa^2 e^\kappa (e^{2\kappa} - 2e^\kappa + 1) + \kappa^4 e^\kappa (e^{2\kappa} + e^\kappa + 1)}{-e^{4\kappa} + 4e^{3\kappa} - 6e^{2\kappa} + 4e^\kappa - 1}$
	$+6\sigma_v^2\sigma_u^2 \frac{e^{2\kappa} - \left(\kappa^2 + 2\right)e^\kappa + 1}{e^{2\kappa} - 2e^\kappa + 1} + 3\sigma_v^4, \;\; \kappa = B/\sigma_u$

To examine the identification of the second part we note that under the assumption of independence of the noise and inefficiency term the following equality holds

$$\begin{aligned} & E\left[(\varepsilon - E(\varepsilon))^4\right] - 3\left(E\left[(\varepsilon - E(\varepsilon))^2\right]\right)^2 \\ & \quad = \psi_4 - 3\psi_2^2 = E\left[(u - E(u))^4\right] - 3\left(E\left[(u - E(u))^2\right]\right)^2 \end{aligned} \tag{3.8}$$

This is a measure of excess kurtosis and for the truncated half-normal model is derived as

$$\psi_4 - 3\psi_2^2 = \sigma_u^4(-\xi^3\tilde{\eta}_0 + 3\xi\tilde{\eta}_0 - 4\xi^2\tilde{\eta}_0^2 - 4\tilde{\eta}_0^2 - 3\xi^2\tilde{\eta}_0^2 - 12\xi\tilde{\eta}_0^3) \tag{3.9}$$

where $\tilde{\eta}_0 = \frac{(2\pi)^{-1/2} - \xi\phi(\xi)}{\Phi(\xi) - \frac{1}{2}}$. Notice that for normal distribution $\tilde{\eta}_0 = 0$ and thus the excess kurtosis is also zero.

After multiplying (3.9) by $\psi_3^{-4/3}$ we eliminate σ_u and the resulting function, which we denote by g has only one argument ξ

$$g(\xi) = \frac{-\xi^3\tilde{\eta}_0 + 3\xi\tilde{\eta}_0 - 4\xi^2\tilde{\eta}_0^2 - 4\tilde{\eta}_0^2 - 3\xi^2\tilde{\eta}_0^2 - 12\xi\tilde{\eta}_0^3}{\left(2\tilde{\eta}_0^3 - 3\xi\tilde{\eta}_0^2 - \tilde{\eta}_0 + \xi^2\tilde{\eta}_0\right)^{-4/3}} \tag{3.10}$$

The weak law of large numbers implies that

$$\text{plim}\frac{1}{n}\sum\nolimits_i \hat{\varepsilon}_i^k = m_k = \psi_k \tag{3.11}$$

The first order moment is zero by definition and thus is not useful for identification purposes. By employing the Slutsky theorem we can specify the following function G

$$\begin{aligned} g(\xi) &= \frac{m_4 - 3m_2^2}{m^{4/3}} \\ &\Longrightarrow \\ G(\xi) &= g(\xi) - \frac{m_4 - 3m_2^2}{m^{4/3}} \end{aligned}$$

Similarly, we can derive the function G for the normal-truncated exponential model with function g expressed by

$$g(\xi) = \frac{36e^{2\xi} - 24e^{\xi} - 24e^{3\xi} + 6e^{4\xi} - \xi^4 e^{\xi} - 4\xi^4 e^{2\xi} - \xi^4 e^{3\xi} + 6}{(6e^{2\xi} - 4e^{\xi} - 4e^{3\xi} + e^{4\xi} + 1)(-\frac{2e^{3\xi} - (\xi^3+6)e^{2\xi} + (6-\xi^3)e^{\xi} - 2}{e^{3\xi} - 3e^{2\xi} + 3e^{\xi} - 1})^{4/3}} \tag{3.12}$$

Both the truncated half normal model and the truncated exponential model are globally identified. To see this, we can examine the monotonicity of the function G with respect to the parameter ξ which will allow us to express this parameter (implicitly) as a function of sample moments and data. This condition provides the necessary and sufficient condition for global identification ala Rothenberg (1971). For the truncated half normal model, G is monotonically decreasing and for the truncated exponential model, G is monotonically increasing. Hence, in both cases, G is invertible and ξ can be identified. The identification of other parameters then follows from the third order moment of least squares residuals. Note, however, that for large values of ξ (e.g., $\xi > 5$ for the normal-truncated half-normal model and $\xi > 20$ for the normal-truncated exponential model), the curve $g(\xi)$ is nearly flat and gives poor identification. ξ can be large for two reasons: either σ_u goes to zero or the bound parameter is large. In the first case the distribution of the inefficiency process approaches the Dirac-delta distribution which makes it very hard for the distributional parameters to be identified. This limiting case is discussed in Wang and Schmidt (2008). In the second case the distribution of the inefficiency term becomes unbounded as in the standard stochastic frontier models for which it is straightforward to show that the model is globally identified (see Aigner et al. 1977; Olson et al. 1980).

It is not clear, however, that the doubly truncated normal model is globally identifiable. However, local identification can be verified. We may examine $\psi_3^{-4/3}(\psi_4 - 3\psi_2^2)$ and $\psi_3^{-5/3}(\psi_5 - 10\psi_2\psi_3)$, both of which are functions of ξ_1 and ξ_2 only and we denote them as $g_1(\xi_1, \xi_2)$ and $g_2(\xi_1, \xi_2)$, respectively. Let $\hat{g}_1$ and $\hat{g}_2$ be the sample versions of g_1 and g_2, respectively, we have the following system of identification equations,

$$\begin{aligned} G_1(\xi_1, \xi_2) &\equiv g_1(\xi_1, \xi_2) - \hat{g}_1 = 0 \\ G_2(\xi_1, \xi_2) &\equiv g_2(\xi_1, \xi_2) - \hat{g}_2 = 0. \end{aligned}$$

By the implicit function theorem (or Rothenberg 1971), the identification of ξ_1 and ξ_2 depends on the rank of the matrix

$$H = \begin{pmatrix} \frac{\partial g_1}{\partial \xi_1} & \frac{\partial g_1}{\partial \xi_2} \\ \frac{\partial g_2}{\partial \xi_1} & \frac{\partial g_2}{\partial \xi_2} \end{pmatrix}.$$

If H is of full rank, then ξ_1 and ξ_2 can be written as functions of $\hat{g}_1$ and $\hat{g}_2$; the identification of the model then follows. The analytic form of H is very complicated, but we may examine the invertibility of H by numerically evaluating g_1 and g_2 and inferring the sign of each element in H. It can be verified that the determinant of H is nonzero in neighborhoods within I_1, I_2, and I_4, the definitions of which are given as follows,

(i) $I_1 \equiv \{(\mu, B) | \mu \leq 0, B > 0\}$
(ii) $I_2 \equiv \{(\mu, B) | \mu > 0, B \in (0, 2\mu)\}$
(iii) $I_3 \equiv \{(\mu, B) | B = 2\mu > 0\}$
(iv) $I_4 \equiv \{(\mu, B) | \mu > 0, B > 2\mu\}$.

The line $I_3 \equiv \{(\mu, B) | B = 2\mu > 0\}$ corresponds to the case where $B = 2\mu$ and $\psi_3 = 0$, hence the functions g_1 and g_2 are not continuous and the implicit function theorem is not applicable. Nonetheless, simulation results in the next section show that when the true values of B and μ satisfy $B = 2\mu$, both B and μ are consistently estimated. This may indicate that the restricted ($B = 2\mu$) model may be nested in the unrestricted model and the model is locally identifiable on $I_2 \bigcup I_3 \bigcup I_4$.

We may treat the doubly truncated normal model as a collection of different sub-models corresponding to the different domains of parameters. Treated separately, each of the sub-models is globally identified. In maximum likelihood estimation, the separate treatment is easily achieved by constrained optimization on each parameter subset. For example, on the line of $\{(\mu, B) | \mu = 0, B > 0\} \subset I_1$, the doubly truncated normal model reduces to the truncated half normal model. As another useful example, the line I_2 corresponds to a sub-model that has positive skewness even asymptotically.

3.4.2 *Method of Moment Estimation*

The method-of-moments (Olson et al. 1980) may be employed to estimate our model or to obtain initial values for maximum likelihood estimation. In the first step of this approach, OLS is used to obtain consistent estimates of the parameters describing the technology, apart from the intercept. In the second step, using the distributional assumptions on the residual, equations of moment conditions are solved to obtain estimates of the parameters describing the distribution of the residual.

More specifically, we may rewrite the production frontier model in (3.1) and (3.2) as

$$y_i = (\alpha_0 - \mathbb{E}u_i) + \sum_{k=1}^{K} \alpha_k x_{i,k} + \varepsilon_i^*,$$

where $\varepsilon_i^* = \varepsilon_i + (\mathbb{E}u_i)$ has zero mean and constant variance σ_ε^2. Hence OLS yields consistent estimates for ε_i^* and α_k, $k = 1, \ldots, K$. Equating the sample moments of estimated residuals $(\hat{\varepsilon}_i^*)$ to the population moments, one can solve for the parameters associated with the distribution of (ε_i^*).

3.4.3 Maximum Likelihood Estimation

For more efficient estimation, we may use maximum likelihood estimation (MLE). Note that with the presence of a noise term v_i, the range of residual is unbounded and does not depend on the parameter. No other standard regularity conditions might be questioned. In the remainder of this section we provide the log-likelihood functions for the bounded inefficiency model for the three parametric distributions we have considered. Note that in practice we may also need the gradients of the log likelihood function. The gradients are complicated in form but straightforward to derive. These are provided in the appendix.

In addition to the γ-parametrization discussed earlier, we re-parametrize the bound parameter with another parameter $\tilde{B} = \exp(-B)$. Unlike the bound, $\tilde{B}$ takes values in compact unit interval which facilitates the numerical procedure of maximum likelihood estimation as well as establishing the asymptotic normality of this parameter. When $\tilde{B}$ lies in the interior of parameter space, the MLE estimator is asymptotically normal (see Rao 1973; Davidson and MacKinnon 1993 among others).

The log-likelihood function for the doubly truncated normal model with γ-parameterization is given by

$$\ln L = -n \ln \left[\Phi(\frac{-\ln \tilde{B} - \mu}{\sigma_u(\sigma, \gamma)}) - \Phi(\frac{-\mu}{\sigma_u(\sigma, \gamma)}) \right]$$

$$-n \ln \sigma - \frac{n}{2} \ln(2\pi) - \sum_{i=1}^{n} \frac{(\varepsilon_i + \mu)^2}{2\sigma^2}$$

$$+\sum_{i=1}^{n}\ln\left\{\Phi\left(\frac{(-\ln\tilde{B}+\varepsilon_i)\sqrt{\gamma/(1-\gamma)}-(\ln\tilde{B}+\mu)\sqrt{(1-\gamma)/\gamma}}{\sigma}\right)\right.$$
$$\left.-\Phi\left(\frac{\varepsilon_i\sqrt{\gamma/(1-\gamma)}-\mu\sqrt{(1-\gamma)/\gamma}}{\sigma}\right)\right\}, \tag{3.13}$$

where $\varepsilon_i = y_i - x_i\alpha$, $x_i = (1, x_{ik})$, and $\alpha = (\alpha_0, \alpha_k)'$.

$$\sigma_u(\sigma,\gamma) = \sigma\sqrt{\gamma}. \tag{3.14}$$

This can be expressed in terms of the λ-parametrization as in Aigner et al. (1977) by substituting γ in (3.13) with

$$\gamma(\lambda) = \frac{\lambda^2}{1+\lambda^2}. \tag{3.15}$$

The log-likelihood function for the truncated half normal model is

$$\ln L = -n\ln\left(\Phi\left(\frac{-\ln\tilde{B}}{\sigma_u(\sigma,\gamma)}\right)-\frac{1}{2}\right)-n\ln\sigma-\frac{n}{2}\ln(2\pi)$$
$$-\sum_{i=1}^{n}\frac{\varepsilon_i^2}{2\sigma^2}+\sum_{i=1}^{n}\ln\left\{\Phi\left(\frac{(-\ln\tilde{B}+\varepsilon_i)\sqrt{\gamma/(1-\gamma)}-\ln\tilde{B}\sqrt{(1-\gamma)/\gamma}}{\sigma}\right)\right.$$
$$\left.-\Phi\left(\frac{\varepsilon_i\sqrt{\gamma/(1-\gamma)}}{\sigma}\right)\right\}, \tag{3.16}$$

Again, substituting γ into (3.16) with $\gamma(\lambda)$ in (3.15), we get the $\log L$ with λ-parametrization.

Finally, the log-likelihood function for the truncated exponential model with γ-parametrization is given by

$$\ln L = -\frac{n}{2}\ln\gamma - n\ln\sigma - n\ln\left(1-e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}\right)+\frac{n}{2}\frac{1-\gamma}{\gamma}+\frac{\gamma^{-1/2}}{\sigma}\sum_{i=1}^{n}\varepsilon_i$$
$$+\sum_{i=1}^{n}\ln\left[\Phi\left(\frac{(-\ln\tilde{B}+\varepsilon_i)(1-\gamma)^{-1/2}}{\sigma}+\sqrt{\frac{1-\gamma}{\gamma}}\right)\right.$$
$$\left.-\Phi\left(\frac{\varepsilon_i(1-\gamma)^{-1/2}}{\sigma}+\sqrt{\frac{1-\gamma}{\gamma}}\right)\right], \tag{3.17}$$

where $\varepsilon_i = y - x_i\alpha$.

After estimating the model, we can estimate the composed error term ε_i:

$$\hat{\varepsilon}_i = y_i - \hat{\alpha}_0 - \sum x_{i,k}\hat{\alpha}_k, i = 1, \cdots, n. \tag{3.18}$$

From this we can estimate the inefficiency term u_i using the formula for $E(u_i|\varepsilon_i)$ in Table 3.1.

One reasonable question is whether or not one can test for the absence or the presence of the bound ($H_0 : \tilde{B} = 0$ vs. $H_1 : \tilde{B} > 0$), which one may wish to test since this would suggest that the proper specification would be the standard SF model which assumes no bound as a special case of our more general bounded SF model. The test procedure is slightly complicated but still feasible. The first complication arises from the fact that $\tilde{B}$ lies on the boundary of the parameter space under the null. Second, it is obvious from the log-likelihood functions provided above that the bound is not identified in this case and it can be shown that any finite order derivative of the log-likelihood function with respect to $\tilde{B}$ is zero. Thus the conventional Wald and Lagrange Multiplier (LM) statistics are not defined and the Likelihood Ratio (LR) statistic has a nonstandard asymptotic distribution that strictly would dominate the $\chi^2_{(1)}$ distribution. Lee (1993) derives the asymptotic distribution of such an estimate as a mixture of χ^2 distributions under the null that its value is zero, focusing in particular on the SF model under the assumption of half-normally distributed inefficiencies. Here λ is globally identified, which can also be seen using the method-of-moments estimator provided in Aigner et al. (1977). Lee (1993) provides useful one-to-one reparametrization which transform the singular information matrix into a nonsingular one. However, since the bound in our model case is not identified in this situation, there is no such re-parametrization and hence this procedure cannot be used. An alternative is to apply the bootstrap procedure proposed by Hansen (1996, 1999) to construct asymptotically equivalent p-values to make an inference. To implement the test we treat the $\hat{\varepsilon}_i$ $(i = 1, \ldots, n)$ as a sample from which the bootstrap samples $\hat{\varepsilon}_i^{(m)}$ $(i = 1, \ldots, n; m = 1, \ldots, M)$ are drawn with replacement. Using the bootstrap sample we estimate the model under the null and the alternative of bounded inefficiency and construct the corresponding LR statistic. We repeat this procedure M times and calculate the percentage of times the bootstrap LR exceeds the actual one. This provides us with the bootstrap estimate of the asymptotic p-value of LR under the null.

3.5 Panel Data

In the same spirit as Schmidt and Sickles (1984) and Cornwell et al. (1990), we may specify a panel data model of bounded inefficiencies:

$$y_{it} = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{it,k} + \varepsilon_{it} \tag{3.19}$$

where

$$\varepsilon_{it} = v_{it} - u_{it}. \tag{3.20}$$

We assume that the inefficiency components (u_{it}) are positive, independent from the regressors, and are independently drawn from a time-varying distribution with upper bound B_t. We may set B_t to be time-invariant. However, it is certainly more plausible to assume otherwise, as the market or industry may well become more or less forgiving as time goes by, especially in settings in which market reforms are being introduced or firms are adjusting to a phased transition from regulation to deregulation.

Note that since u_{it} is time-varying, the above panel data model is in effect a time-varying technical efficiency model. Our model differs from the existing literature in that, while previous time-varying efficiency models, notably Cornwell et al. (1990), Kumbhakar (1990), Battese and Coelli (1992), and Lee and Schmidt (1993), are time-varying in the mean or intercept of individual effects, our model is time-varying in the upper support of the distribution of inefficiency term u_i.

The assumption that u_{it} is independent over time simplifies estimation and analysis considerably. In particular, the covariance matrix of $\varepsilon_i \equiv (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$ is diagonal. This enables us to treat the panel model as a collection of cross-section models in the chronological order. We may certainly impose more structure on the sample path of the upper bound of u_{it}, B_t, without incurring heavy costs in terms of analytic difficulty. For example, we may impose smoothness conditions on B_t. This is empirically plausible, indeed, since changes in the market competitive conditions may come gradually. And it is also technically desirable, since imposing smoothness conditions gives us more degree of freedom in estimation, hence better estimators of model parameters. A natural way of doing this is to let B_t be a sum of weighted polynomials,

$$B_t = \sum_{i=0}^{K} b_i (t/T)^i, \quad t = 1, \ldots, T, \tag{3.21}$$

where (b_i) are constants. We may also use trigonometric series, splines, among others, in the modeling of B_t. For an extensive survey of efforts to generalize such heterogeneities in efficiencies see Sickles et al. (2013).

3.6 Simulations

To examine the finite sample performance of the MLE estimator of the doubly truncated normal model,[10] we run a series of Monte Carlo experiments in the standard cross-sectional setting. The data generating process is (3.1) and (3.2) with

[10] The results for the truncated half-normal and truncated exponential models are available upon the request.

one regressor x and no constant term and is based on the data generating process utilized in study 2 of Aigner, Lovell, and Schmidt. We maintain the assumption that v_i is iid $N(0, \sigma_v^2)$, u_i is iid, and v_i and u_i are independent from each other and from regressors. The number of repetitions is 1,000. Throughout we keep the coefficient α on the single regressor technology parameter set at 0.6 and examine performances in terms of bias and mean absolute error as we change in each of the distributional parameters (σ, γ, μ, and B). As the SF benchmark we use the singly truncated normal model (Stevenson 1980) on the simulated data. We report average estimates and mean absolute errors (MAE) in Tables 3.3–3.6. Each of these sets of experiments selectively change the distributional parameters. We draw the following conclusions from these experiments.[11]

First, all parameters in the doubly truncated normal model appear to be well-estimated, with biases and MAE's that fall as sample sizes rise. The biases are generally small, and the MAE's of almost all estimates decrease at $\sqrt{N}$ rate as N increases, except that of $\hat{\mu}$ in a couple of particular cases. More specifically, when σ is small (i.e., the variation in the composite error is small), $\hat{\mu}$ does not converge at the optimal rate as N increases (see Table 3.3). The same happens when B is large (see Table 3.5). This observation is connected with the well-known difficulty of identifying μ in the singly truncated model ($B \to \infty$) from finite sample. As is well known, the technological parameter α in the singly truncated normal model is consistently estimated. However, estimates of distributional parameters in the singly truncated model are not well-defined and thus we do not calculate the corresponding MAE's.

Second, Table 3.3 shows that as σ becomes smaller, the MAE of $\hat{\alpha}$ is monotone decreasing, while the MAE's of $\hat{\sigma}$, $\hat{\gamma}$, and $\hat{\mu}$ is monotone increasing. To reconcile the apparent divergence, note that the composite error ε is noise for the technological parameters, but signal for distributional parameters. The effect of σ on the MAE of $\hat{B}$ is ambiguous, which decreases at first and then increases as σ becomes smaller.

Third, if we mistakenly estimate a singly truncated model on a DGP with double truncation, we tend to underestimate the average technical efficiency (ATE). This is understandable since the singly truncated model may treat some extreme (negative) measurement errors as inefficiencies. Within the doubly truncated model, it is also clear that as B becomes larger, the ATE decreases (See Table 3.5). However, our simulation results show that the efficiency ranking would not be affected if we estimate a misspecified model.

Finally, as is expected, MLE correctly estimates the doubly truncated normal model when the composite error has positive population skewness. This is evident in Table 3.6, where the third case ($\mu = 0.3$, $B = 0.5$) corresponds to negative (positive) skewness in u (ε). In all cases, the double truncation in the DGP of u makes finite-sample positive skewness more probable, resulting in many zero $\hat{\gamma}$'s (super-efficiency) from the misspecified (singly truncated) model. Hence the average $\hat{\gamma}$'s in the misspecified model are generally much lower than the true value.

[11] We have similar limited Monte Carlo results based on two regressors with varying correlations and our results are qualitatively similar. Results are available on request.

Table 3.3 Monte Carlo simulation of the doubly truncated normal model: on the dimension of σ. $\gamma = 0.95$, $\mu = 0$, $B = 0.5$. The number of repetitions is 1,000. Mean absolute errors (MAE) are given in parentheses

		Doubly truncated						Singly truncated				
σ	N	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	$\hat{B}$	ATE	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	ATE
2	500	0.599(0.0170)	1.993(0.0099)	0.950(0.0025)	0.00371(0.1720)	0.500(0.0351)	0.887	0.599(0.0170)	0.515	0.208	0.0446	0.826
	1,000	0.600(0.0124)	1.995(0.0069)	0.950(0.0017)	0.000001(0.1250)	0.501(0.0239)	0.882	0.600(0.0125)	0.504	0.188	0.0980	0.825
	5,000	0.600(0.0054)	1.998(0.0029)	0.950(0.00078)	0.006(0.0585)	0.500(0.0111)	0.873	0.600(0.0054)	0.488	0.147	0.175	0.824
$\sqrt{2}$	500	0.600(0.0124)	1.405(0.0111)	0.950(0.0029)	−0.0253(0.3380)	0.502(0.0273)	0.860	0.600(0.0124)	0.369	0.205	0.177	0.842
	1,000	0.600(0.0092)	1.409(0.0071)	0.950(0.0019)	−0.00600(0.2850)	0.500(0.0189)	0.857	0.600(0.0092)	0.362	0.176	0.203	0.849
	5,000	0.600(0.0037)	1.412(0.0030)	0.950(0.0009)	0.0129(0.1280)	0.500(0.0094)	0.848	0.600(0.0037)	0.354	0.138	0.229	0.856
1	500	0.600(0.0093)	0.884(0.1180)	0.929(0.0220)	0.0120(0.4170)	0.532(0.0545)	0.820	0.600(0.0093)	0.276	0.227	0.219	0.853
	1,000	0.600(0.0071)	0.917(0.0844)	0.938(0.0127)	−0.013(0.4070)	0.519(0.0386)	0.822	0.600(0.0071)	0.274	0.221	0.225	0.845
	5,000	0.600(0.0029)	0.984(0.0171)	0.948(0.0022)	−0.00413(0.2990)	0.504(0.0172)	0.826	0.600(0.0029)	0.271	0.198	0.234	0.839

Table 3.4 Monte Carlo simulation of the doubly truncated normal model: on the dimension of γ. $\sigma = \sqrt{2}$, $\mu = 0$, $B = 0.5$. The number of repetitions is 1,000. Mean absolute errors (MAE) are given in parentheses

		Doubly truncated						Singly truncated				
γ	N	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	$\hat{B}$	ATE	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	ATE
0.8	500	0.600(0.0170)	1.409(0.0099)	0.799(0.0025)	0.00540(0.1720)	0.500(0.0351)	0.912	0.600(0.0232)	0.729	0.234	−0.238	0.806
	1,000	0.599(0.0124)	1.411(0.0069)	0.799(0.0017)	0.0154(0.1250)	0.500(0.0239)	0.908	0.599(0.0162)	0.723	0.223	−0.194	0.798
	5,000	0.600(0.0054)	1.413(0.0029)	0.800(0.0008)	−0.00171(0.0585)	0.500(0.0111)	0.898	0.600(0.0073)	0.726	0.245	−0.203	0.737
0.5	500	0.602(0.0379)	1.411(0.0190)	0.500(0.0127)	−0.00521(0.0887)	0.495(0.0731)	0.941	0.602(0.0378)	1.124	0.225	−0.797	0.822
	1,000	0.600(0.0268)	1.412(0.0139)	0.500(0.0093)	0.00735(0.0709)	0.503(0.0536)	0.936	0.600(0.0268)	1.133	0.242	−0.846	0.793
	5,000	0.600(0.0112)	1.414(0.0060)	0.500(0.0042)	−0.00347(0.0243)	0.500(0.0234)	0.930	0.600(0.0112)	1.138	0.247	−0.878	0.748
0.3	500	0.601(0.0382)	1.410(0.0225)	0.303(0.0152)	0.0231(0.0935)	0.500(0.0906)	0.947	0.601(0.0382)	1.293	0.182	−0.877	0.847
	1,000	0.600(0.0301)	1.413(0.0149)	0.301(0.0101)	0.0100(0.0693)	0.498(0.0658)	0.946	0.600(0.0301)	1.303	0.191	−0.928	0.832
	5,000	0.600(0.0130)	1.414(0.0067)	0.301(0.0046)	−0.00432(0.0268)	0.501(0.0288)	0.939	0.600(0.0131)	1.309	0.200	−0.984	0.795

Table 3.5 Monte Carlo simulation of the doubly truncated normal model: on the dimension of B. $\sigma = \sqrt{2}$, $\gamma = 0.95$, $\mu = 0$. The number of repetitions is 1,000. Mean absolute errors (MAE) are given in parentheses

		Doubly truncated						Singly truncated				
B	N	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	$\hat{B}$	ATE	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	ATE
0.3	500	0.599(0.0111)	1.409(0.0077)	0.950(0.0022)	0.000963(0.0910)	0.298(0.0251)	0.939	0.599(0.0111)	0.370	0.222	−0.0808	0.882
	1,000	0.600(0.0078)	1.411(0.0051)	0.950(0.0015)	0.0110(0.0744)	0.300(0.0169)	0.935	0.600(0.0078)	0.359	0.191	−0.0113	0.884
	5,000	0.600(0.0037)	1.413(0.0022)	0.950(0.0007)	−0.00142(0.0311)	0.300(0.0076)	0.929	0.600(0.0037)	0.345	0.149	0.0688	0.883
0.5	500	0.600(0.0126)	1.405(0.0120)	0.949(0.0030)	−0.0248(0.3440)	0.503(0.0280)	0.860	0.600(0.0126)	0.370	0.206	0.176	0.842
	1,000	0.600(0.0090)	1.408(0.0074)	0.950(0.0019)	−0.0177(0.2870)	0.502(0.0202)	0.856	0.600(0.0090)	0.362	0.181	0.201	0.846
	5,000	0.600(0.0037)	1.412(0.0031)	0.950(0.0009)	0.00974(0.1370)	0.500(0.0093)	0.848	0.600(0.0037)	0.354	0.137	0.229	0.857
0.7	500	0.600(0.0129)	1.316(0.1010)	0.940(0.01230)	−0.0184(0.5450)	0.732(0.0633)	0.766	0.600(0.0129)	0.392	0.244	0.299	0.791
	1,000	0.600(0.0097)	1.364(0.0524)	0.946(0.0060)	−0.00766(0.5040)	0.715(0.0401)	0.770	0.600(0.0097)	0.387	0.218	0.312	0.797
	5,000	0.600(0.0043)	1.407(0.0091)	0.950(0.0011)	0.00365(0.3010)	0.702(0.0178)	0.769	0.600(0.0044)	0.382	0.202	0.326	0.779

Table 3.6 Monte Carlo simulation of the doubly truncated normal model: On the dimension of μ. $\sigma = \sqrt{2}$, $\gamma = 0.95$, $B = 0.5$. The number of repetitions is 1,000. Mean absolute errors (MAE) are given in parentheses

		Doubly truncated						Singly truncated				
μ	N	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	$\hat{B}$	ATE	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	ATE
−0.2	500	0.601(0.0128)	1.405(0.0117)	0.949(0.0029)	−0.0903(0.3320)	0.498(0.0277)	0.862	0.601(0.0128)	0.370	0.211	0.168	0.839
	1,000	0.600(0.0089)	1.409(0.0068)	0.950(0.0018)	−0.0941(0.2710)	0.498(0.0190)	0.857	0.600(0.0089)	0.362	0.184	0.198	0.842
	5,000	0.600(0.0040)	1.412(0.0031)	0.950(0.0009)	−0.122(0.1490)	0.499(0.0094)	0.848	0.600(0.0040)	0.354	0.138	0.226	0.856
0.2	500	0.599(0.0128)	1.405(0.0117)	0.950(0.0029)	0.0386(0.3480)	0.506(0.0271)	0.859	0.599(0.0127)	0.368	0.206	0.183	0.840
	1,000	0.600(0.0088)	1.409(0.0070)	0.950(0.0019)	0.0626(0.2820)	0.503(0.0193)	0.856	0.600(0.0088)	0.361	0.176	0.208	0.849
	5,000	0.600(0.0038)	1.412(0.0028)	0.950(0.0008)	0.156(0.1070)	0.501(0.0088)	0.849	0.600(0.0038)	0.353	0.133	0.233	0.860
0.3	500	0.600(0.01270)	1.407(0.0095)	0.950(0.0025)	0.0658(0.3550)	0.505(0.0263)	0.859	0.600(0.0127)	0.368	0.206	0.184	0.840
	1,000	0.600(0.0081)	1.409(0.0071)	0.950(0.0019)	0.121(0.2790)	0.504(0.0201)	0.856	0.600(0.0081)	0.361	0.182	0.207	0.844
	5,000	0.600(0.0041)	1.412(0.0033)	0.950(0.00084)	0.215(0.1310)	0.502(0.0095)	0.848	0.600(0.0041)	0.353	0.134	0.233	0.860

3.7 An Empirical Illustration to Analyze US Banking Industry Dynamics

3.7.1 Empirical Model and Data

We now apply the bounded inefficiency (BIE) model to an analysis of the US banking industry, which underwent a series of deregulatory reforms in the early 1980s and 1990s, and experienced an adverse economic environment in the last few turbulent years of 2000s.[12] Our analysis covers a lengthy period between 1984 and 2009 and our illustration aims to use the panel variant of our BIE model to capture efficiency trends of the US banking sector during these years as well as how the lower bound of inefficiency also changed as the market became more or less competitive vis-a-vis inefficient firms.

Following Adams et al. (1999) and Kneip et al. (2012), we specify a multi-output/multi-input stochastic output distance frontier model as[13]

$$Y_{it} = Y_{it}^{*\prime}\gamma + X_{it}^{\prime}\beta + v_{it} - u_{it}, \tag{3.22}$$

where Y_{it} is the log of real estate loans; X_{it} is the negative of log of inputs, which include demand deposit (dd), time and savings deposit (dep), labor (lab), capital (cap), and purchased funds (purf).[14] Y_{it}^{*} includes the log of commercial and industrial loans/real estate loans (ciln) and installment loans/real estate loans (inln). In order to account for the riskiness and heterogeneity of the banks we include the log of the ratio of equity to total assets $(eqrt)$ which usually measures the risk of insolvency of the banks in banking literature.[15] The lower the ratio the more riskier a bank is considered. We assume the v_{it} are iid across i and t, and for each t, u_{it} has a upper bound B_t. Then we can treat this model as a generic panel data bounded inefficiency model as discussed in Sect. 3.5. Once the individual effects u_{it} are estimated, technical efficiency for a particular firm at time t is calculated as $TE = \exp(u_{it} - \max_{1\leq j\leq N} u_{jt})$.

The output distance function is known as a Young index (ratio of the geometric mean of the outputs to the geometric mean of the inputs) described in Balk (2008), which leads to the Cobb-Douglas specification of the distance function

[12] These deregulations gradually allowed banks in different states to merge with other banks across the state borders. The Reigle-Neal Act that was passed by the Congress in 1994 also allowed the branching by banks across the state lines.

[13] For more discussion on stochastic distance frontiers see Lovell et al. (1994).

[14] Purchased funds include federal funds purchased and securities sold under agreements to repurchase, time deposits in $100,000 denominations, mortgage debt, bank's liability on acceptances, and other liabilities that are not demand deposits and retail time and savings deposits.

[15] We exclude from the sample banks with $eqrt$ less that 0.02. Typically, these banks are close to failure and estimation of their efficiency scores require special treatments (see Wheelock and Wilson 2000; Almanidis 2013 for more discussion).

introduced by Klein (1953). Although this functional form has been criticized for its separability and curvature properties it remains a reasonable and parsimonious first-order local approximation to the true function (Coelli 2000) and we use it in our limited empirical illustration of the bounded stochastic frontier model. We use the parsimonious Cobb-Douglas model as well to allow comparisons with the results from our Monte Carlo simulations, which due to the need to estimate highly nonlinear models, have been somewhat limited by computational and time constraints to a relatively simple linear in logs specification.[16] Translog distance function estimates, which one may view as more general, have their own attendant problems due to multicollinearity in the second order terms of the four-output/five-input technology. This typically is addressed by utilizing additional restrictions, such as those imposed by cost minimization or profit maximization, in order to be empirically identify the translog parameters.[17] We do not use these side conditions to empirically identify the parameters due to our use of a stochastic frontier model that admits to technical inefficiency but does not attempt to trace this inefficiency to its logical implication in the first order conditions of cost minimization or profit maximization (the so-called "Greene problem", Kutlu 2013). Utilization of side conditions to address errors in the optimization of allocations is beyond the scope of this paper. That said, our translog estimates have provided qualitatively similar results, which are available on request.

We use US commercial banking data from 1984 first quarter through 2009 third quarter. There are several ways in which data can be merged or deleted depending on whether or not banks continued as independent entities during the sample period we consider in our illustration of the insights gained by the bounded inefficiency model.

[16] The empirical illustration is used in part to link the use of the Cobb–Douglas functional form in expressing the provision of banking intermediation services to Peter Schmidt's intellectual predecessors, whom we have discussed above, and who used the Cobb-Douglas functional form substantially. It also has been the predominate functional form used by the NBER's Productivity Program in their seminal work on productivity and growth. We understand the limitations of the Cobb–Douglas functional form. Indeed, one of the authors has been writing on the topic for 30 years (Guilkey et al. 1983). Recent work on banking efficiency and returns to scale by Wheelock and Wilson (2012) have fitted local linear and local quadratic estimator with on the order of one million parameters to a cost relationship and use duality theory to link the cost estimates to the returns to scale in the banking industry and utilize multi-step bootstrapping methods to assess significance. It is unclear what has been estimated in such an exercise as standard regularity conditions for the function to indeed be a cost function have not been checked, nor it is clear how such a test would be conducted. Obviously, with such an overparameterized model, they overwhelmingly reject generalizations of the Cobb–Douglas, such as second-order Taylor series expansions in logs, such as the translog functional form. Without the regularity conditions met by at least some of the observations their results are meaningless. Moreover, it is not even clear that their use of the bootstrap in the multi-step algorithms they use is even valid. We find that regularity conditions are met by a substantial portion of the data we use and do find little qualitative difference in terms of the efficiency patterns, which is of course what the paper focuses on, between those generated by the Cobb–Douglas and the translog.

[17] For an example of the use of such side conditions and with just such justifications in the multi-output cost function setting see Hughes and Mester (1993).

One approach is to express the data for a bank on a pro-forma basis that goes back in time to account for mergers. For example, if a bank in 2008 is the result of a merger in 2008 then the pre-2008 data is merged on a pro-forma basis wherein the non-surviving bank's data is viewed as part of the surviving bank in earlier time periods. The Federal Reserve uses this approach in estimating risk measurement models, such as the Charge-off at Risk Model (Frye and Pelz 2008), which is the basis of risk dashboards used for centralized bank supervision. This sample design reflects methodologies used by banks in calibrating credit risk models, such as those used for Basel III and for Comprehensive Capital Analysis and Review (CCAR).[18] An alternative to the retroactive merging in of legacy banks is to utilize an unbalanced design wherein banks simply attrit from the sample when their ownership changes. Although at first blush this would seem to address the problem of selection in cases when weaker banks get taken over, there are also many cases of mergers-of-equals as well (e.g., JP Morgan and Bank One merger). Roughly 84 % of banks in our sample ceased their operation due to reasons other than failure, such as merger or voluntary liquidation, or remained inactive, or were no longer regulated by the Federal Reserve. Almanidis and Sickles (2012) have proposed a general model that combines the mixture hazard model with the canonical stochastic frontier model to investigate the main determinants of the probability and time to failure of a panel of US commercial banks during the financial distress that began in August of 2007. In their analysis they focused on banks failures, not on ownership changes or changes in regulatory oversight that were not due to liquidation due to financial distress. Unlike the standard hazard model, which would assume that all banks in the sample eventually experience the event (failure), the mixture hazard model distinguishes between healthy (long-term survivors) and at-risk banks. Almanidis and Sickles did not find that selection on banks per se impacted their estimates in any significant way. Moreover, their formal mixture hazard framework is far removed from the basic modeling issues addressed in this paper, namely the introduction of a different stochastic frontier paradigm that acknowledges a lower bound to inefficient firm operating practices. In order to maintain comparability between our results and those from many other studies using stochastic frontier analysis and to find some middle ground between the pro-forma merging algorithm practiced by the Federal Reserve and the deletion of firms from the sample that attrit and the potential misspecification due to the many potential ways (unobserved in our sample) in which such attrition may have occurred, we utilize a balanced panel and study only firms that have remained in business during our sample period.

The data is a balanced panel of 4,193 commercial banks and was compiled from the Consolidated Reports of Condition and Income (Call Report) and the FDIC Summary of Deposits. The data set includes 431,879 observations for 103 quarterly periods. This is a fairly long panel and thus the assumption of time-invariant inefficiencies does not seem tenable. For this reason we compare the estimates from our BIE model to the estimates from other time-varying effects

[18] For more discussion of this issue and the use of similar data in models of risk aggregation see Inanoglu and Jacobs (2009).

Table 3.7 Descriptive statistics for bank-specific variables

Variable name	Mean	Median	Standard deviation	Min	Max
Real estate loans	212,968	17,549	4,341,501	145	4.61E+08
Commercial and industrial loans	103,272	4,908	2,143,974	46	1.82E+08
Installment loans	58,869	4,360	1,417,908	86	1.51E+08
Demand deposits	54,913	7,282	912,761	186	1.03E+08
Time and savings deposits	449,003	46,954	1.00E+07	1,446	9.93E+08
Labor	186	29	2,960	4	215,670
Capital	8,196	913	129,778	9	1.16E+07
Purchased funds	163,785	13,698	3,322,838	286	3.37E+08
Ratio of equity to total assets	0.1007	0.0936	0.0312	0.0210	0.7459

models such as CSSW (the within variant of Cornwell et al. (1990)) and BC (Battese and Coelli 1992) models, along with the baseline fixed effect estimator (FIX) of Schmidt and Sickles (1984). Descriptive statistics for the bank-level variables are given in Table 3.7, where all nominal values are converted to reflect 2000 year values.

3.7.2 Results

Table 3.8 compares the parameter estimates of the bounded inefficiency (BIE) model with that of FIX, CSSW, and BC.[19] The structural parameters are statistically significant at the 1 % level and have the expected sign for all four models. The adjusted Bera and Premaratne (2001) skewness test statistic is calculated to be 990.26, leading to rejection of the null hypothesis of symmetry at any conventional significance level. The asymmetry of the least squares residuals is also verified by quantile-quantile plot representation in Fig. 3.1. The technology parameters from BIE model are somewhat different from those obtained from other models. The negative value of the coefficient of the $eqrt$ implies that riskier firms tend to produce more loans, and especially real estate loans that are considered of high risk. The positive sign of the estimate of the time trend shows technological progress on average. There is a slight difference between the distributional parameters of BIE and BC model which are also statistically significant at any conventional significance level. We also tested (not reported here) other distributional specifications for BIE discussed above. The distributional parameters obtained from normal-truncated half-normal model did not differ very much from that reported in the table, but those obtained from normal-truncated exponential model did. However, this is not a specific to bounded inefficiency models. Similar differences have been documented in unbounded SF models as well.

[19] We estimate the normal-doubly truncated normal model in order to be able to compare it with the BC model which specifies the inefficiencies to follow the truncated normal distribution.

Table 3.8 Comparisons of various estimators. Estimates and standard errors (in parentheses) for each model parameters from competing models (FIX, CSSW, BC, BIE)

	FIX	CSSW	BC	BIE
ciln	0.2407(0.0015)	0.2971(0.0014)	0.2284(0.0013)	0.2838(0.0012)
inln	0.2206(0.0013)	0.1715(0.0012)	0.2043(0.0013)	0.2609(0.0013)
dd	−0.0940(0.0024)	−0.0935(0.0020)	−0.1197(0.0024)	−0.0996(0.0020)
dep	−0.3999(0.0048)	−0.4037(0.0051)	−0.4368(0.0048)	−0.4053(0.0034)
lab	−0.3104(0.0046)	−0.2219(0.0042)	−0.1610(0.0044)	−0.1892(0.0020)
cap	−0.0460(0.0016)	−0.0464(0.0014)	−0.0510(0.0015)	−0.0965(0.0015)
purf	−0.1507(0.0034)	−0.1658(0.0029)	−0.1627(0.0034)	−0.1665(0.0031)
time	0.0057(0.0001)	–	0.0020(0.0001)	0.0021(0.0001)
eqrt	−0.1369(0.0045)	−0.1189(0.0041)	−0.0975(0.0044)	−0.1088(0.0039)
γ	0	0	0.7980(0.0115)	0.7690(0.0058)
σ	0.2210(0.0034)	0.2070(0.0020)	0.2733(0.0045)	0.2712(0.0022)
μ	–	–	0.3240(0.0139)	0.3518(0.0630)
B	–	–	–	1.5186
ATE	0.5853	0.6470	0.6410	0.6998

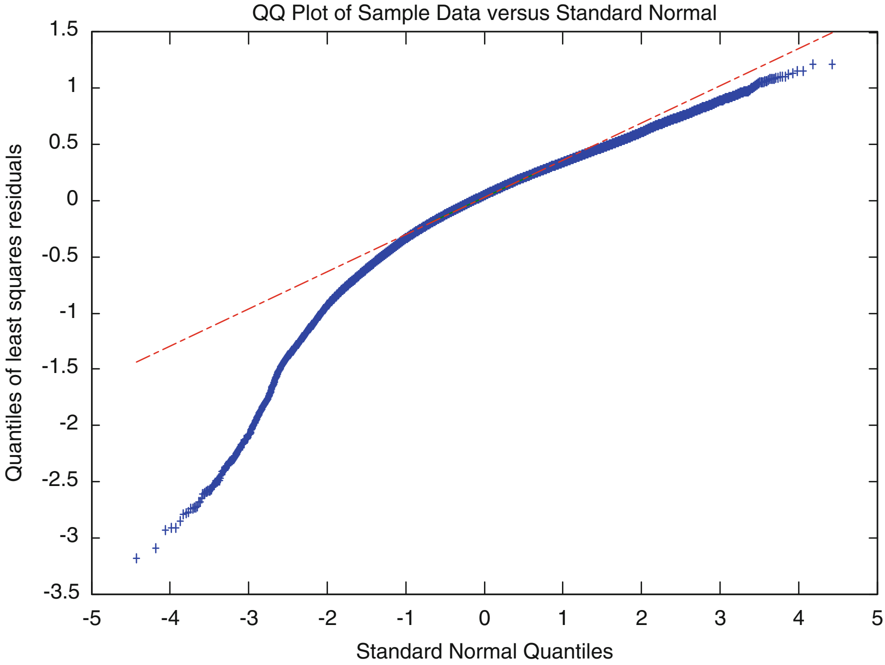

Fig. 3.1 Quantile-quantile plot

We also estimate the time-varying inefficiency bound, B, using two approaches. First we estimate the bound for the panel data model without imposing any restriction on its sample path. In the second approach we specify the bound as

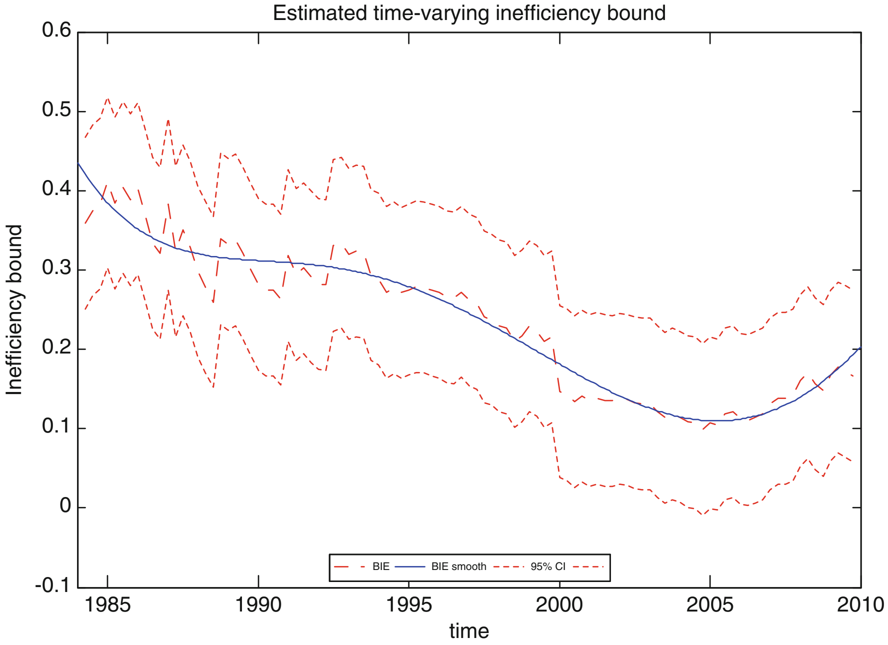

Fig. 3.2 Estimated and smoothed inefficiency bound

a sum of weighted time polynomials. We choose to fit a fifth degree polynomial the coefficients of which are estimated by MLE along with the rest parameters of the model.[20] Both approaches are illustrated in Fig. 3.2 with their respective 95 % confidence intervals. It can be seen that the inefficiency bound has had a decreasing trend up to year 2005, when the financial crisis (informally) began, and then it is increasing for the remaining periods through the third quarter of 2009. One interpretation of this trend can be that the deregulations in 1980s and 1990s increased competitive pressures and forced many inefficient banks to exit the industry, reducing the upper limit of inefficiency that banks could sustain and still remain in their particular niche market in the larger banking industry. The new upward trend can be attributed to the adverse economic environment and an increase in the proportion of banks that are characterized as “too big to fail.”

Of course, for time-varying efficiency models such as CSSW, BC, and BIE, average efficiencies change over time.[21] These are illustrated in Fig. 3.3 along with their

[20] The choice of degrees of the time polynomial was based on a simple likelihood-ratio (LR) test and degrees of the polynomial ranging from 1 to 10. The maximum likelihood estimates of coefficients for this polynomial are given by

$b_0 = -3.9477e - 007$, $b_1 = 0.0039509^{**}$, $b_2 = -15.816^{***}$, $b_3 = 31656^{**}$, $b_4 = -3.168e + 007^{*}$, $b_5 = 1.2682e + 010$.

[21] We trimmed the top and bottom 5 % of inefficiencies to remove the effects of outliers.

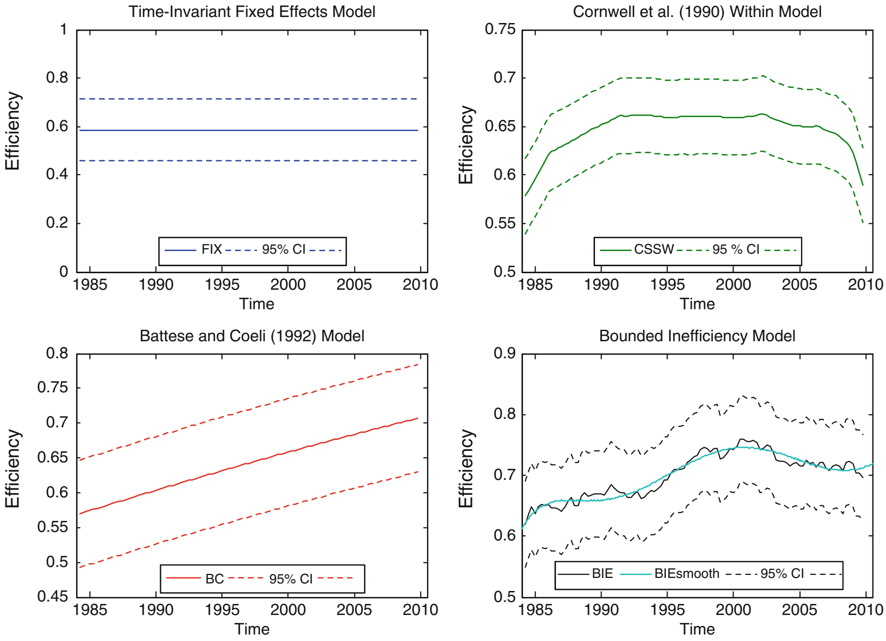

Fig. 3.3 Averaged efficiencies from each estimator

95 % confidence bounds. The BIE averaged efficiencies (panel 4) are significantly higher than those obtained from the fixed effect time-invariant model. However, the differences are small compared to BC and CSSW models. These small differences are not unexpected, however, since the existence of the inefficiency bound implies that the mean conditional distribution of inefficiencies is also bounded from above, resulting in higher average efficiencies. Failing to take the bound into account could possibly yield underestimated mean and individual efficiency scores (see Table 3.1). We smooth the BIE averaged efficiencies by fitting ninth degree polynomial of time in order to capture their trend and also to be able to compare them with other two time-varying averaged efficiency estimates. These are represented by a curve labeled BIEsmooth. It can be seen that the efficiency trend for the BIE model is in close agreement with the CSSW model and better reflects the deregulatory reforms and consolidation of the US commercial banking industry. It is increasing initially and then falls soon after the saving and loans (S&L) crisis of early 1990s began. It has the decreasing pace and reaches its minimum in 1993 a year before Congress passed the Reigle-Neal Act which allowed commercial banks to merge with and acquire banks across the state lines. This spurred a new era of interstate banking and branching, which along with the Gramm-Leach-Billey Act that granted broad-based securities and insurance power to commercial banks, substantially decreased the number of banks operated in the US from 10,453 in 1994 to 8,315 by the end of the millennium. After 1994 the banking industry witnessed a rapid increase in

averaged efficiencies of its institutions due in part to the disappearance of inefficient banks previously sheltered from competitive pressure and due to the expansion of large banks that both financially and geographically diversified their products. The increasing trend continues until the new recessionary period of 2001 and then steadily falls thereafter until the rapid decline illustrating the effects of the 2007–2009 crisis. The CSSW model is able to show the weakness of the banking industry as early as 2005. This weakness is illustrated by the estimated inefficiency bound from the BIE model. On the other hand, the BC model shows a slight, statistically non-significant, upward efficiency trend for all these periods ($\eta = 0.0066$).

In sum, Figs. 3.2 and 3.3 display an interesting findings: on one hand, an upward trend is observed for the average efficiency of the industry, presumably benefiting from the deregulations in the 1980s and 1990s; on the other hand, the industry appears to be more "tolerant" of less efficient banks in the last decade. Possibly, these banks have a characteristic that we have not properly controlled for and we are currently examining this issue. Given the recent experiences in the credit markets due in part to the poor oversight lending authorities gave in their mortgage and other lending activities, our results also may be indicative of a backsliding in the toleration of inefficiency that could have contributed to the problems the financial services industry faces today.

3.8 Conclusions

In this paper we have introduced a series of parametric stochastic frontier models that have upper (lower) bounds on the inefficiency (efficiency). The model parameters can be estimated by maximum likelihood, including the inefficiency bound. The models are easily applicable for both cross-section and panel data settings. In the panel data setting, we set the inefficiency bound to be varying over time, hence contributing another time-varying efficiency model to the literature. We have examined the finite sample performance of the maximum likelihood estimator in the cross-sectional setting. We also have showed how the positive skewness problem inherent in traditional stochastic frontier model can be avoided when the bound is taken into account. An empirical illustration focusing on the US banking industry using the new model revealed intuitive and revealing trends in efficiency scores.

Acknowledgements The idea of addressing the skewness problem in stochastic frontier models via the use of our new Bounded Stochastic Frontier was conjectured by C. A. Knox Lovell in discussions at the presentation of a very preliminary draft of this paper at the Tenth European Workshop on Efficiency and Productivity, Lille, France, June, 2007. Subsequent versions have been presented at the Texas Econometrics Camp XV, Montgomery, Texas, February, 2010; the Efficiency and Productivity Workshop, University of Auckland, New Zealand, February, 2010; the North American Productivity Workshop, Houston, Texas, June 2010; the International Econometrics Workshop, Guanghua Campus (SWUFE), Chengdu, August 12, 2010; and the 10th World Congress of the Econometric Society, Shanghai, August 21, 2010. We would like to thank participants of those conferences and workshops for their helpful comments and insights. We thank Carlos Martins-Filho for his helpful suggestions and criticism. We would especially like to thank

Robert Adams at the Board of Governors of the Federal Reserve System for his guidance and Rob Kuvinka for his excellent research assistance that was essential in our development of the banking data set that we analyze in our empirical illustration. The views expressed by the first author are independent of those of Ernst&Young LLP. The usual caveat applies.

Appendix

First-Order Derivatives of Log-Likelihood Function

The scores for the normal-doubly-truncated normal model that can either be used in a generalized method of moments estimation or in standard mle (3.13) under the γ-parametrization and the $\tilde{B}$-parametrization are:

$$\frac{\partial \ln L}{\partial a} = \sum_{i=1}^{n} \frac{(\varepsilon_i + \mu)x_i}{\sigma^2} + \frac{\sqrt{\gamma}/(1-\gamma)}{\sigma} \sum_{i=1}^{n} x_i \frac{\phi(z_{4i}) - \phi(z_{3i})}{\Phi(z_{3i}) - \Phi(z_{4i})}$$

$$\frac{\partial \ln L}{\partial \gnapprox^2} = \frac{n}{2\sigma^2} \frac{[(z_1\phi(z_1) - z_2\phi(z_2)]}{\Phi(z_1) - \Phi(z_2)} - \frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(\varepsilon_i + \mu)^2}{2\sigma^4} + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{[z_{4i}\phi(z_{4i}) - z_{3i}\varphi(z_{3i})]}{\Phi(z_{3i}) - \Phi(z_{4i})}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \lambda} = {} & \frac{n}{\gneq} \frac{[(z_1\phi(z_1) - z_2\phi(z_2)]}{\Phi(z_1) - \Phi(z_2)} \\ & + \frac{1}{\sigma} \sum_{i=1}^{n} \frac{1}{\Phi(z_{3i}) - \Phi(z_{4i})} \Bigg\{ ((-\ln(\tilde{B}) + \varepsilon_i) \frac{1}{(1-\gamma)^2} \sqrt{(1-\gamma)/\gamma} \\ & + (\ln(\tilde{B}) + \mu) \frac{1}{\gamma^2} \sqrt{\gamma/(1-\gamma)}) \phi(z_{3i}) \\ & - (\varepsilon_i \frac{1}{(1-\gamma)^2} \sqrt{(1-\gamma)/\gamma} - \mu\lambda \frac{1}{\gamma^2} \sqrt{\gamma/(1-\gamma)}) \phi(z_{4i}) \Bigg\}\end{aligned}$$

$$\frac{\partial \ln(L)}{\partial \mu} = \frac{n}{\sigma\sqrt{\gamma}} \frac{\phi(z_1) - \phi(z_2)}{\Phi(z_1) - \Phi(z_2)} - \sum_{i=1}^{n} \frac{(\varepsilon_i + \mu)}{\sigma^2} + \frac{\sqrt{(1-\gamma)/\gamma}}{\sigma} \sum_{i=1}^{n} \frac{\phi(z_{4i}) - \varphi(z_{3i})}{\Phi(z_{3i}) - \Phi(z_{4i})}$$

$$\frac{\partial \ln(L)}{\partial \tilde{B}} = \frac{n}{\tilde{B}\sigma\sqrt{\gamma}} \frac{\phi(z_1)}{\Phi(z_1) - \Phi(z_2)} - \frac{1}{\tilde{B}\sigma\sqrt{(1-\gamma)\gamma}} \sum_{i=1}^{n} \frac{\phi(z_{3i})}{\Phi(z_{3i}) - \Phi(z_{3i})}$$

where $z_1 = -\frac{(\ln(\tilde{B})+\mu)}{\sigma\sqrt{\gamma}}$, $z_2 = \frac{-\mu}{\sigma\sqrt{\gamma}}$, $z_{3i} = -\frac{(\ln(\tilde{B})-\varepsilon_i)\sqrt{\gamma/(1-\gamma)}+(\ln(\tilde{B})+\mu)\sqrt{(1-\gamma)/\gamma}}{\sigma}$, and $z_{4i} = \frac{\varepsilon_i\sqrt{\gamma/(1-\gamma)}-\mu\sqrt{(1-\gamma)/\gamma}}{\sigma}$. The scores for the normal-truncated half-normal model are obtained after substituting $\mu = 0$ in the above expressions.

The scores for normal-truncated exponential model are derived from (3.17) as

$$\frac{\partial \ln L}{\partial a} = -\frac{\gamma^{-1/2}}{\sigma}\sum_{i=1}^{n} x_i + \frac{(1-\gamma)^{-1/2}}{\sigma}\sum_{i=1}^{n}\frac{\phi(\tilde{z}_{2i})-\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i})-\Phi(\tilde{z}_{2i})}x_i$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} = &-\frac{n}{\sigma} - \frac{n\ln\tilde{B}\gamma^{-1/2}}{\sigma^2}\frac{e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}}{1-e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}}\\ &+\frac{(1-\gamma)^{-1/2}}{\sigma^2}\sum_{i=1}^{n}\left\{\frac{\phi(\tilde{z}_{2i})-\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i})-\Phi(\tilde{z}_{2i})}\varepsilon_i + \frac{\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i})-\Phi(\tilde{z}_{2i})}\ln\tilde{B}\right\}\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} = &-\frac{n}{2\gamma} - \frac{n\ln\tilde{B}}{2\gamma^{3/2}}\frac{e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}}{1-e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}} - \frac{n}{2\gamma^2} - \frac{1}{2\gamma^{3/2}}\sum_{i=1}^{n}\varepsilon_i\\ &-\frac{1}{2}\sum_{i=1}^{n}\left\{\frac{\phi(\tilde{z}_{2i})-\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i})-\Phi(\tilde{z}_{2i})}\left(\frac{\varepsilon_i}{\sigma(1-\gamma)^{3/2}} - \frac{1}{\gamma^2}\sqrt{\frac{\gamma}{1-\gamma}}\right)\right.\\ &\left.-\frac{\ln\tilde{B}}{\sigma(1-\gamma)^{3/2}}\frac{\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i})-\Phi(\tilde{z}_{2i})}\right\}\end{aligned}$$

$$\frac{\partial \ln L}{\partial \tilde{B}} = \frac{n\gamma^{-1/2}}{\sigma\tilde{B}}\frac{e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}}{1-e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}} - \frac{(1-\gamma)^{-1/2}}{\tilde{B}\sigma}\sum_{i=1}^{n}\frac{\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i})-\Phi(\tilde{z}_{2i})}$$

where $\tilde{z}_{1i} = \frac{(-\ln\tilde{B}+\varepsilon_i)(1-\gamma)^{-1/2}}{\sigma} + \sqrt{\frac{1-\gamma}{\gamma}}$ and $\tilde{z}_{1i} = \frac{\varepsilon_i(1-\gamma)^{-1/2}}{\sigma} + \sqrt{\frac{1-\gamma}{\gamma}}$.

References

Adams RM, Berger AN, Sickles RC (1999) Semiparametric approaches to stochastic panel frontiers with applications in the banking industry. J Bus Econ Stat 17:349–358

Afriat SN (1972) Efficiency estimation of production functions. Int Econ Rev 13:568–598

Aigner D, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econom 6:21–37

Alchian AA (1950) Uncertainty, evolution, and economic theory. J Political Econom 58:211–221

Almanidis P (2013) Accounting for heterogeneous technologies in the banking industry: a time-varying stochastic frontier model with threshold effects. J Product Anal 39(2):191–205

Almanidis P, Sickles RC (2011) Skewness problem in stochastic frontier models: fact or fiction? In: Van Keilegom I, Wilson P (eds) Exploring research frontiers in contemporary statistics and econometrics: a festschrift in Honor of Leopold Simar. Springer, New York

Almanidis P, Sickles RC (2012) Banking crises, early warning models, and efficiency, Mimeo, Rice University

Balk BM (2008) Price and quantity index numbers: models for measuring aggregate change and difference. Cambridge University, New York

Battese GE, Coelli TJ (1992) Frontier production functions, technical efficiency and panel data, with application to paddy farmers in India. J Product Anal 3:153–169

Battese GE, Corra G (1977) Estimation of a production frontier model: with application to the pastoral zone of eastern Australia. Aust J Agric Econ 21:167–179

Bera AK, Premaratne G (2001) Adjusting the tests for skewness and kurtosis for distributional misspecifications. UIUC-CBA Research Working Paper No. 01-0116

Carree MA (2002) Technological inefficiency and the skewness of the error component in stochastic frontier analysis. Econ Lett 77:101–107

Coelli T (2000) On the econometric estimation of the distance function representation of a production technology. Center for Operations Research & Econometrics, Universite Catholique de Louvain

Cornwell C, Schmidt P, Sickles RC (1990) Production frontiers with cross-sectional and time series variation in efficiency levels. J Econom 46:185–200

Cragg JG (1997) Using higher moments to estimate the simple errors-in-variables model. RAND J Econ 28(Special Issue in Honor of Richard E. Quandt):S71–S91

Dagenais M, Dagenais D (1997) Higher moment estimators for linear regression models with errors in the variables. J Econom 76:193–221

Davidson R, MacKinnon JG (1993) Estimation and inference in econometrics. Oxford University Press, New York

Demsetz H (1973) Industry structure, market rivalry, and public policy. J Law Econ 16:1–9

Entani T, Maeda Y, Tanaka H (2002) Dual models of interval DEA and its extension to interval data. Eur J Oper Res 136:32–45

Feng Q, Horrace W, Wu GL (2012) Wrong skewness and finite sample correction in parametric stochastic frontier models. Mimeo

Frye J, Pelz E (2008) BankCaR (bank capital-at-risk): US commercial bank chargeoffs. Working paper # 3, Federal Reserve Bank of Chicago

Goldberger A (1968) The interpretation and estimation of Cobb-Douglas functions. Econometrica 35:464–472

Greene WH (1980a) Maximum likelihood estimation of econometric frontier functions. J Econom 13:27–56

Greene WH (1980b) On the estimation of a flexible frontier production model. J Econom 13: 101–115

Greene WH (1990) A gamma distributed stochastic frontier model. J Econom 46:141–164

Greene WH (2007) The econometric approach to efficiency analysis In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency: techniques and applications. Oxford University Press, New York

Guilkey DK, Lovell CAK, Sickles RC (1983) A comparison of the performance of three flexible functional forms. Int Econ Rev 24:59l–6l6

Hansen BE (1996) Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64:413–430

Hansen BE (1999) Threshold effects in non-dynamic panels: estimation, testing, and inference. J Econom 93:345–368

Hicks JR (1935) Annual survey of economic theory: the theory of monopoly. Econometrica 3:1–20

Hughes JP, Mester LJ (1993) A quality and risk-adjusted cost function for banks: evidence on the "too-big-to-fail" doctrine. J Product Anal 4:293–315

Inanoglu H, Jacobs M Jr (2009) Models for risk aggregation and sensitivity analysis: an application to bank economic capital. J Risk Financ Manag 2:118–189

Klein L (1953) Textbook of econometrics. Row, Peterson and Company, New York

Kneip A, Sickles RC, Song W (2012) A new panel data treatment for heterogeneity in time trends. Econom Theory 28:590–628

Kumbhakar SC (1990) Production frontiers, panel data, and time-varying technical efficiency. J Econom 46:201–212

Kutlu L (2013) Misspecification in allocative efficiency: a simulation study. Econ Lett 118: 151–154

Kutlu L, Sickles RC (2012) Estimation of market power in the presence of firm level inefficiencies. J Econom 168:141–155

Lee L (1993) Asymptotic distribution of the maximum likelihood estimator for a stochastic frontier function model with a singular information matrix. Econom Theory 9:413–430

Lee YH, Schmidt P (1993) A production frontier model with flexible temporal variation in technical efficiency. In: Fried HO, Lovell CAK, Schmidt P (eds) The measurement of productive efficiency: techniques and applications. Oxford University Press, New York

Lewbel A (2012) Using heteroskedasticity to identify and estimate mismeasured and endogenous regressor models. J Bus Econ Stat 30:67–80

Lovell CAK, Richardson S, Travers P, Wood L (1994) Resources and functionings: a new view of inequality in Australia. In: Eichhorn W (ed) Models and measurements of welfare and inequality. Springer, Berlin

Marschak J, Andrews WH (1944) Random simultaneous equations and the theory of production. Econometrica 12:143–203

Matzkin R (2012) Nonparametric identification, Mimeo, UCLA Department of Economics

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production function with composed error. Int Econ Rev 18:435–444

Nerlove M (1965) Estimation and identification of Cobb-Douglas production functions. Rand McNally, Chicago

Newey WK, Powell JL (2003) Instrumental variable estimation of nonparametric models. Econometrica 71:1565–1578

Olson JA, Schmidt P, Waldman DM (1980) A Monte Carlo study of estimators of the stochastic frontier production function. J Econom 13:67–82

Orea L, Steinbuks J (2012) Estimating market power in homogenous product markets using a composed error model: application to the California electricity market. EPRG Working Paper 1210, CWPE Working Paper 1220

Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York

Richmond J (1974) Estimating the efficiency of production. Int Econ Rev 15:515–521

Rothenberg TJ (1971) Identification in parametric models. Econometrica 39(3):577–591

Samuelson PA (1979) Paul Douglas's measurement of production functions and marginal productivities. J Political Econ 87(Part 1):923–939

Schmidt P, Sickles RC (1984) Production frontiers and panel data. J Bus Econ Stat 2:367–374

Sickles RC, Hao J, Shang C (2013) Productivity and panel data. In: Baltagi B (ed) Chapter 17 of Oxford handbook of panel data. Oxford University Press, New York (forthcoming)

Simar L, Wilson PW (2010) Estimation and inference in cross-sectional, stochastic frontier models. Econom Rev 29:62–98

Stevenson RE (1980) Likelihood functions for generalized stochastic frontier estimation. J Econom 13:57–66

Stigler GS (1958) The economies of scale. J Law Econ 1:54–71

Tamer E (2010) Partial identification in econometrics. Annu Rev Econ 2:167–195

Waldman DM (1982) A stationary point for the stochastic frontier likelihood. J Econom 18: 275–279

Wang WS, Schmidt P (2008) On the distribution of estimated technical efficiency in stochastic frontier models. J Econom 148:36–45

Wheelock D, Wilson P (2000) Why do banks disappear? The determinants of US bank failures and acquisitions. Rev Econ Stat 82(1):127–138

Wheelock DC, Wilson PW (2012) Do large banks have lower costs? New estimates of returns to scale for US banks. J Money Credit Bank 44:171–199

Zellner A, Kmenta J, Dreze J (1966) Specification and estimation of Cobb-Douglas production function models. Econometrica 34:784–795