Empirical Economics (2008) 34:585–602
DOI 10.1007/s00181-007-0137-2

ORIGINAL PAPER

Scale and efficiency measurement using a semiparametric stochastic frontier model: evidence from the U.S. commercial banks

Subal C. Kumbhakar · Efthymios G. Tsionas

Received: 15 May 2006 / Accepted: 15 January 2007 / Published online: 9 March 2007

Abstract In this paper, we use the local maximum likelihood (LML) method proposed by Kumbhakar et al. (J Econom, 2007) to estimate stochastic cost frontier models for a sample of 3,691 U.S. commercial banks. This method relaxes several deficiencies in the econometric estimation of frontier functions. In particular, we relax the assumption that all banks share the same production technology and provide bank-specific measures of returns to scale and cost inefficiency. The LML method is applied to estimate the cost frontiers in which a truncated normal distribution is used to model technical inefficiency. This formulation allows the cost frontier, inefficiency effects and heteroskedasticity in both noise and inefficiency components to be quite flexible.

Keywords Data envelopment analysis · Cost efficiency · Local maximum likelihood estimation · U.S. commercial bank

JEL Classification C14 · C50 · D23 · G211

1 Introduction

Since the publication of the seminal papers by Aigner et al. (1977) and Meeusen and van den Broeck (1977), econometric estimation of stochastic frontier (SF)

S. C. Kumbhakar (✉)
Department of Economics, State University of New York at Binghamton,
Binghamton, NY 13902, USA
e-mail: kkar@binghamton.edu

E. G. Tsionas
Department of Economics, Athens University of Economics and Business,
76 Patission Street, 104 34 Athens, Greece
e-mail: tsionas@aueb.gr

models became a standard practice in efficiency measurement studies. Although SF models can be estimated either by sampling theory or Bayesian techniques, efficiency measurement in these models rely on the choice of functional forms, distributional assumptions, fixity of parameters of the underlying production technology, etc. Some of these assumptions are strong in the sense that empirical results are often sensitive to these assumptions. In a recent survey Yatchew (1998) argues that economic theory rarely, if ever, specifies precise functional forms for production or cost functions. Consequently, its implications are not, strictly speaking, testable when arbitrary parametric functional forms are specified. To the extent that the production or cost functions are misspecified, it is possible that a true model can be rejected, and estimates of efficiency will be biased. Thus, care must be taken in specifying functional forms in empirical research.

An alternative to the SF approach is the deterministic nonparametric approach, viz., the data envelopment analysis (DEA) popularized by Charnes et al. (1978). While the SF models assume specific parametric functional forms for the production or cost frontiers, and use distributional assumptions on the noise and inefficiency components, the DEA models do not make such assumptions. This is clearly an advantage of DEA over SF approach. However, DEA cannot separate ‘inefficiency‘ from ‘noise’. Since the statistical theory is well developed for SF models, one can make statistical inferences about parameters and functions of interest, based on estimated parameters and data, including inefficiency. However, this is not the case in DEA models (although some progress has been made in terms of bootstrapping (see, for example, Simar and Wilson 2000). Thus, applied researchers are unable to make statements regarding the statistical properties of the estimated functions such as input elasticities, scale economies, efficiency, etc., using the DEA.

Park et al. (1998) have considered semi-parametric efficient estimation of SF panel models under alternative assumptions on the joint distribution of random firm effects and the regressors. This approach is certainly useful, provided there is no uncertainty about *linearity* of the model. More recently, Cazals et al. (2002) have proposed a nonparametric estimator based on the free disposable hull (FDH) concept. This estimator is more robust relative to the DEA but it doesn’t envelope all the data. This is, essentially, a stochastic DEA estimator for which the authors provide an asymptotic theory.

Our purpose in this paper is not to improve on estimating techniques for linear stochastic frontier models as in Park et al. (1998) but to propose efficient estimating techniques for semiparametric stochastic frontier models. More specifically we use the local maximum likelihood (LML) method, which is a semiparametric technique in the sense that it makes the parameters of a given parametric model dependent on the covariates via a process of localization. For example, if β is a nonparametric function $\beta(x_i)$, the familiar linear model $y_i = x_i'\beta + u_i$ becomes a semiparametric model. This approach has been introduced in stochastic frontiers by Kumbhakar et al. (2007), henceforth KPST, the focus of which was to prove asymptotic properties of LML estimator. Here our interest is mostly in the application of their model with two modest extensions. First, we examine both Cobb–Douglas and translog functional forms as anchoring models for the semiparametric SF function. Second, we use the truncated normal distribution instead of the half-normal distribution to model the inefficiency component. The advantages of the LML approach are as follows. First, we avoid assuming a global parametric functional form (that holds for all observations) for the cost frontier. Although the anchoring function is linear it is not globally linear because the parameters of the linear function are made observation-specific through localization. Thus, this approach will hopefully avoid parameter inconsistency problem associated with misspecification in functional form. Second, by allowing the location parameter of the truncated normal distribution for the one-sided error to be an arbitrary function of relative prices and outputs we allow for inefficiency effects, which considerably generalizes the model of Kumbhakar et al. (1991) (henceforth KGM) and Battese and Coelli (1995). Third, by allowing the scale parameters of both error terms to be arbitrary functions of relative prices and outputs we allow for heteroskedasticity of quite general form in both inefficiency and noise components. We view these generalizations quite important in drawing robust conclusions about cost efficiency as well as other measures like returns to scale, elasticities, etc., in any empirical application.

The remainder of the paper is organized as follows. Local estimation and its application in SF models are reviewed in Sect. 2. Some computational and practical issues are discussed in Sect. 3. In Sect. 4 we apply the LML technique to estimate cost frontiers using a sample of U.S. commercial banks. The paper concludes with a summary of the main findings in Sect. 5.

2 Local estimation in stochastic frontier models

KPST introduced the notion of local likelihood estimation for the SF model and derived the asymptotic properties of the estimators. To describe the model very briefly, suppose we have a parametric model that specifies the density of an observed dependent variable y_i $(i = 1, \ldots, n)$ conditional on a vector of observable covariates $x_i \in X \subseteq R^k$, a vector of unknown parameters $\theta \in \Theta \subseteq R^m$, and let the probability density function of y_i given x_i be $l(y_i; x_i, \theta)$. The parametric ML estimator is then given by

$$\tilde{\theta} = \underset{\theta \in \Theta}{\arg\max} : \sum_{i=1}^{n} \ln l(y_i; x_i, \theta)$$

The LML estimation technique is a way to allow for nonparametric effects within the parametric model (Tibshirani 1984). A natural way to convert the parametric model to a nonparametric one is to make the parameter θ function of the covariates x_i. Within LML this is accomplished as follows. For an arbitrary $x \in X$, the LML estimator solves the problem

$$\tilde{\theta}(x) = \underset{\theta \in \Theta}{\arg\max} : \sum_{i=1}^{n} \ln l(y_i; x_i, \theta) K_H(x_i - x)$$

where K_H is a kernel function that depends on a matrix bandwidth H. The idea is to choose an anchoring parametric model and maximize a weighted log-likelihood function that places more weight to observations near x rather than weight each observation equally, as the parametric ML estimator would do. By solving the LML problem for each point $x \in X$, we can construct the function $\tilde{\theta}(x)$ that is an estimator for $\theta(x)$. This is a general way of converting the parametric model to a nonparametric approximation to the unknown model.[1] To proceed further, suppose we have the following stochastic frontier cost model

$$y_i = x_i'\beta + v_i + u_i; \quad v_i \sim i.i.d.\,(0, \sigma^2), \quad u_i \sim i.i.d.\,(\mu, \omega^2), \quad u_i \geq 0$$
$$\text{for } i = 1, \ldots, n, \beta \in R^k$$

where y is log cost and x_i is a vector of input prices and outputs (in logs);[2] v_i and u_i are the noise and inefficiency components, respectively. Furthermore, v_i and u_i are assumed to be mutually independent as well as independent of x_i. The main shortcoming of this model, part from being linear in x_i and making distributional assumptions on the noise term (v) and inefficiency term (u), is that the parameter vector β that describes the underlying production technology is constant. That is, neither β nor μ and ω depend on x_i. Although μ and ω are often made functions of covariates (thereby adding some flexibility into the model), specific parametric functions are to be used to estimate the model using the standard ML procedure. However, the β coefficients are assumed to be either constant (for all observations) or random (with a constant mean and constant variance). Given that estimated efficiency depends to a great extent on the chosen functional form, it is desirable to use a model that is flexible and robust.

To make the frontier model more flexible, we adopt the following strategy. Consider the usual parametric ML estimator for the normal (v) and truncated normal (u) stochastic cost frontier model that solves the following problem (Stevenson 1980):

$$\tilde{\theta} = \underset{\theta \in \Theta}{\arg\max} : \sum_{i=1}^{n} \ln l(y_i; x_i, \theta)$$

where

$$l(y_i; x_i, \theta) = [\Phi(\psi)]^{-1} \Phi\left[\frac{\sigma^2\psi + \omega(y_i - x_i'\beta)}{\sigma(\omega^2 + \sigma^2)^{1/2}}\right] \left[2\pi(\omega^2 + \sigma^2)\right]^{-1/2}$$
$$\times \exp\left[-\frac{(y_i - x_i'\beta - \mu)^2}{2\left(\omega^2 + \sigma^2\right)}\right],$$

[1] Another way to deal with heterogeneity in frontier models is the random coefficient approach (Tsionas 2002) which is, however, heavily parametric.

[2] The cost function specification is discussed in details in Sect. 5.2.

$\psi = \mu/\omega$, and Φ denotes the standard normal cumulative distribution function. The parameter vector is $\theta = [\beta, \sigma, \omega, \psi]$ and the parameter space is $\Theta = R^k \times R_+ \times R_+ \times R$. Local ML estimation of the corresponding above model involves the following steps. First, we choose a kernel function. A reasonable choice is

$$K_H(d) = (2\pi)^{-m/2}|H|^{-1/2} \exp\left(-\frac{1}{2}d'H^{-1}d\right), \quad d \in R^m,$$

where m is the dimensionality of $\theta, H = h \cdot S, h > 0$ is a scalar bandwidth, and S is the sample covariance matrix of x_i. Second, for a particular point $x \in X$, we solve the following problem:

$$\tilde{\theta}(x) = \underset{\theta\in\Theta}{\arg\max} : \sum_{i=1}^{n} \left\{ -\ln\Phi(\psi) + \ln\Phi\left[\frac{\sigma^2\psi + \omega(y_i - x_i'\beta)}{\sigma(\omega^2+\sigma^2)^{1/2}}\right] \right.$$
$$\left. -\frac{1}{2}\ln\left(\omega^2+\sigma^2\right) - \frac{1}{2}\frac{(y_i - x_i'\beta - \mu)^2}{\left(\omega^2+\sigma^2\right)} \right\} K_H(x_i - x)$$

A solution to this problem provides the LML parameter estimates of $\tilde{\beta}(x)$, $\tilde{\sigma}(x)$, $\tilde{\omega}(x)$ and $\tilde{\psi}(x)$. It allows for nonparametric heteroskedasticity in both error components unlike Caudill et al. (1995), Hadri (1999) and Wang (2002). It is well known that inefficiency effects and heteroskedasticity in the one-sided error component, if ignored, yield inconsistent parameter estimates. In these situations, the LML technique provides a better way to obtain heteroskedasticity-corrected parameter estimates (from the point of view of large sample theory) without making any functional form assumptions on the form of heteroskedasticity.

Finally, it should be noticed that the kernel weights $K_H(x_i - x)$ do not involve unknown parameters (if h is known) so they can be computed in advance and, therefore, the estimator can be programmed in any standard econometric software. However, one downside of this approach has to chose a kernel function and bandwidth parameter, h. Optimal bandwidth choice through cross-validation technique removes arbitrariness on the choice of h.

3 Computational/practical issues

An important practical issue in estimation is the choice of the bandwidth parameter h. This parameter can be chosen by cross-validation. To do this, first we solve the LML problem at all data points except for observation j, and define for some $\bar{x} \in X$,

$$\tilde{\theta}^{(j)}(x,h) = \underset{\theta\in\Theta}{\arg\max} : \sum_{i\neq j} \ln l(y_i; x_i, \theta) K_H(x_i - \bar{x})$$

for all $j = 1, \ldots, n$. The point $\bar{x}$ can be the overall median of the data. Then we choose h to minimize

$$\sum_{j=1}^{n} \left(y_j - \tilde{y}_j(h)\right)^2$$

where $\tilde{y}_j(h)$ denotes the fitted value of y_j based on h.

Other practical issues are related to the specification of an anchoring model for the regression part as well as anchoring models for the one-sided error term. One can either fit either Cobb–Douglas or translog models depending on size of the data and goodness of fit desired. The choice will also influence computational burden since translog models usually gives a better fit but involve many parameters to estimate. Another consideration is that anchoring functions should satisfy curvature and monotonicity restrictions. This is straightforward for the Cobb–Douglas functions but more complicated for the translog functions, where such restrictions have to be imposed at each observed data point.

4 An application to U.S. commercial banks

We use the above technique to examine cost efficiency of the U.S. commercial banks. The commercial banking industry is one of the largest and most important sectors of the U.S. economy. The structure of the banking industry has undergone rapid changes in the last two decades, mostly due to extensive consolidation. The number of commercial banks has declined over time and concentration at the national level has increased. The number and size of large banks has also increased through acquisitions and mergers. Justification of mergers and acquisitions is often provided in terms of economies of scale and efficiency. Thus, it is important to ask: (i) Are large banks necessarily more efficient? (ii) Do large banks operate beyond their efficient scale? Answers to these questions depend on the estimation technique (parametric vs. nonparametric) used, functional form chosen, etc.[3] Since the banking industry consists of a large number of small banks and assets are highly concentrated in a few very large banks, heteroskedasticity is likely to be present in both the noise and inefficiency components.[4] Moreover, the production technology among banks is likely to differ.[5] These problems can be avoided using the LML approach that makes parameters bank-specific without using any ad hoc specification.

[3] There are numerous studies that address scale economies and efficiency. See, e.g., McAllister and McManus (1993), Berger and Mester (1997), Berger and Humphrey (1991), Boyd and Graham (1991), Mukherjee et al. (2001), Wheelock and Wilson (2001), among others.

[4] It is well known that if the inefficiency component is heteroscedastic and one ignores it, both parameter estimates and estimated inefficiencies will be inconsistent (see Kumbhakar and Lovell 2000, Chap. 3.4). Consequently, estimates of economies of scale are likely to be wrong.

[5] Although, in a parametric setting one can test this using the Chow test for structural change (parameter stability) in which banks are grouped under small, medium, large, etc., there is no universally accepted criterion for grouping banks and deciding how many groups are to be chosen. McAllister and McManus (1993) argued that returns to scale estimates are biased when one fits a single cost function for all the banks.

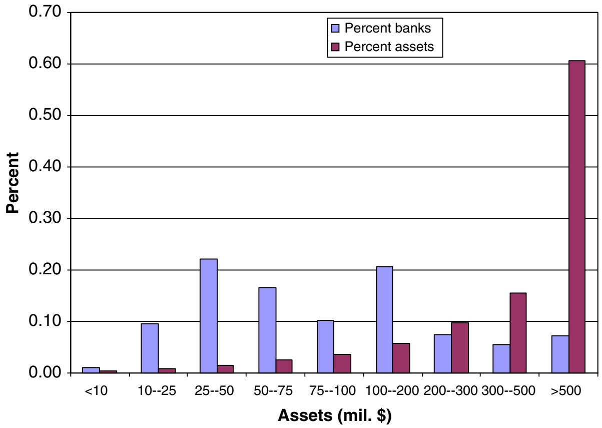

Fig. 1 Distribution of assets

4.1 Data

The data for this study is taken from the commercial bank and bank holding company database managed by the Federal Reserve Bank of Chicago. It is based on the Report of Condition and Income (Call Report) for all U.S. commercial banks that report to the Federal Reserve banks and the FDIC. In this paper we used the data for the year 2000[6] and selected a sample of 3,691 commercial banks. Median value of assets of these banks is 76 million dollars. The distributions of bank assets and banks are shown in Fig. 1. The top 7% of the banks control more than 60% of the total assets while the bottom 10% of the banks control about 1% of total bank assets. About 20% of the top banks control more than 85% of the assets. Thus, the distribution of assets across banks is highly skewed. As a result of this, it is very likely that the parameters of the underlying technology (cost function in our case) will differ among banks.

In the banking literature there is a controversy regarding the choice of inputs and outputs. Here we follow the intermediation approach [Kaparakis et al. (1994) in which banks are viewed as financial firms transforming various financial and physical resources into loans and investments. The output variables are: installment loans (to individuals for personal/household expenses) (y_1), real estate loans (y_2), business loans (y_3), federal funds sold and securities purchased under agreements to resell (y_4), other assets (assets that cannot be properly included in any other asset items in the balance sheet) (y_5). The input variables are: labor (x_1), capital (x_2), purchased funds (x_3), interest-

[6] It is possible to extend the data beyond 2000. However, the panel extension of the KPST (2007) model is not trivial. One obvious complication in a panel set-up is to model the temporal behavior of inefficiency. This will affect consistency and asymptotic normality of the LML parameter estimates. Because of this, we decided to stick to the cross-sectional data. However, instead of using the 2000 data, we could have used the 2005 data. Preliminary investigation of the 2005 data showed that the results are qualitatively similar to those from the 2000 data.

bearing deposits in total transaction accounts (x_4) and interest-bearing deposits in total nontransaction accounts (x_5). The input prices are calculated in the usual way. The price of labor (w_1) is the average wage/salary per employee and is obtained from expenses on salaries and benefits divided by the total number of full time equivalent employees. Similarly, the price of physical capital, $w_2 =$ (expenses on premises and fixed assets)/the dollar value of premises and fixed assets; the price of purchased funds, $w_3 =$ (interest expense on money market deposit accounts + expense of federal funds purchased and securities sold under agreements to repurchase + interest expense on demand notes issued to U.S. Treasury and other borrowed money)/dollar value of purchased funds], price of interest-bearing deposits, $w_4 =$ (interest expense on interest-bearing categories of total transaction accounts/dollar value of interest-bearing categories in total transaction accounts, the price of interest-bearing deposits in total nontransaction accounts, $w_5 =$ (interest expense on total deposits − interest expense on interest-bearing categories in total transaction accounts − interest expense on money market deposit accounts)/dollar value of interest-bearing deposits in total nontransaction account. Total cost is then defined as the sum of cost of these five inputs.

4.2 Results from the localized Cobb–Douglas model

Here we present results from the Cobb–Douglas (CD) functional form because the coefficients of the CD function are easy to interpret. Furthermore, the use of the CD function avoids the muticollinearity problem that arises with a flexible functional form such as the translog and the Fourier functional forms.[7] Since we localize the parameters at each point, flexibility is not an issue. In other words, the use of the CD function gives a clear economic meaning to each and every coefficient that is made bank-specific through localization. We choose the h parameter by using cross-validation method.

We experimented with both half-normal and truncated normal distributions on the one-sided inefficiency term. Results from the truncated normal specification are found to be better than those from the half-normal specification. Because of this result we report results based on the truncated normal distribution on the inefficiency component. The results are based on a CD anchoring function, i.e., the cost function is specified as

$$y_i = x_i'\beta + v_i + u_i$$

where as before $v_i \sim i.i.d.(0, \sigma^2)$ and $u_i \sim i.i.d.(\mu, \omega^2), u_i \geq 0 i = 1, \ldots, n$, $\beta \in R^{k+m}$. Here y is total cost (in natural log) and the x variables contain m (5) outputs and k (5) input prices (all in natural log). Furthermore, to impose linear homogeneity (in input prices) restrictions on the cost function, we normalize

[7] For example, Wheelock and Wilson (2001) found that a global translog cost function violates regularity conditions for most of the banks. This might be the result of either a wrong functional form or fitting a parametric function globally.

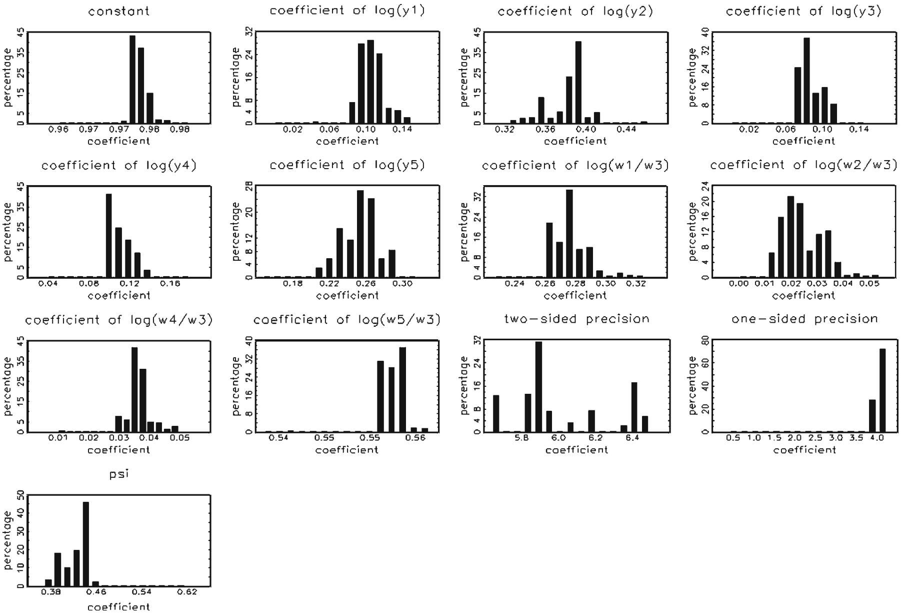

Fig. 2 Histogram of parameter estimates

total cost and the input prices by one input price (w$_3$) before taking logs. Thus, the estimated cost function is

$$\ln C = \beta_0 + \sum_i \beta_{yi} \ln y_i + \sum_{j \neq 3} \beta_{wj} \ln(w_j/w_3) + v_i + u_i$$

when $C = (\text{total cost}/w_3)$. Total number of parameters in β (i.e., $k + m$) is 10.

We report the frequency distribution of estimated parameters in Fig. 2. The histograms for the parameters show different patterns (some are unimodal while others are bimodal but none are symmetric). For example, the cost elasticities with respect to outputs ($\beta_{yi},\ i = 1, \ldots, 5$) are skewed to the right for y_1, y_3, y_4 and y_5. The distribution is bimodal for y_2, y_3 and y_5. The estimated elasticities vary substantially among banks, sometimes as much as 100% from the smallest to the highest. A similar picture comes out of the cost elasticities with respect to input prices (with the exception of w_5 that shows minimum variation among banks). Two of the three parameters associated with the distributions of the noise and inefficiency components show large variations among banks. The estimates of σ_v and ψ show large variations while the opposite is true for σ_u. These large variations in estimated coefficients show why estimating a single set of parameters for all banks might not be a good idea.

We compute scale economies (SCE) as $SCE = \sum_{i=1}^{5} \partial \ln C / \partial \ln y_i = \sum_{i=1}^{5} \beta_{yi}(y, x)$. Since all the parameters are observation-specific, the SCE measure is bank-specific as well. Thus, although we start from a CD cost function,

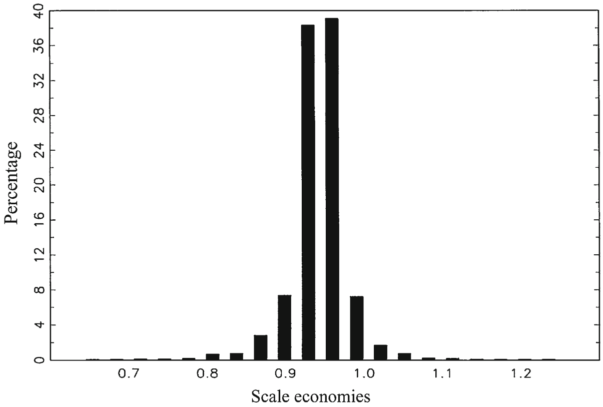

Fig. 3 Histogram of SCE (local ML)

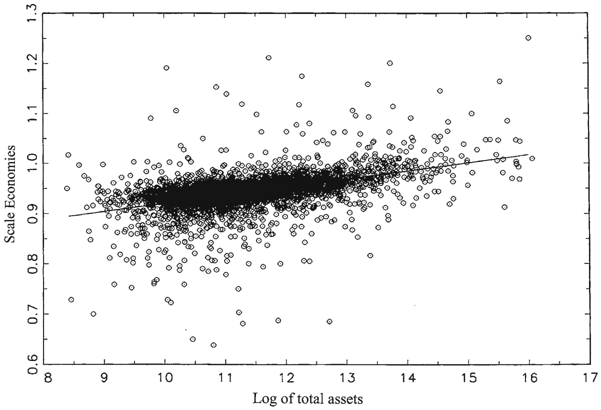

Fig. 4 Plot of SCE against log(asset) (local ML)

the SCE measure is quite flexible. The SCE measures are reported in Fig. 3 in a histogram. It can be easily seen from the histogram that economies of scale are not exhausted (SCE being less than unity thereby meaning that returns to scale are greater than unity) for most of the banks. Returns to scale (RTS = 1/SCE) is less than unity for less than 5% of the banks. This result contradicts some earlier studies that show little or no scale economies left for medium and larger banks. From Fig. 4 that plots SCE against assets (in logarithm) we find that the benefits of scale economies tend to be lower (in general) for large banks. This can be seen from the scatter plot that shows a positive relationship between SCE and log assets. However, we find that RTS is above unity (SCE < 1) for most of the banks. Examining the scatter plot above the line with SCE = 1 (not drawn) (i.e., banks for which RTS < 1), we find no clear pattern between SCE and log assets. That means no strong evidence is found to support the finding (mostly from parametric studies that use a single cost function for all banks) that large/very large banks are operating beyond their optimum size. In other words, our results support the conventional wisdom that justifies bank mergers to exploit benefits of scale economies. Using a global translog cost system Wang and Kumbhakar (2006) found that most of the banks have exhausted their scale economies. They also found that when banks are clustered in terms of their business strategies, scale economoies are found for banks in different groups. While the LML approach goes much beyond grouping banks (since the coefficients are bank-specific), their results confirm that counter-intuitive results might be obtained if wrong functional form (one technology for all the banks) is used.

Now we consider measurement of inefficiency. Suppose we localize with respect to observation $j$and denote the resulting LML estimates of the frontier parameter parameters by $\beta_{(j)}, \sigma_{(j)}, \mu_{(j)}, \omega_{(j)}$. Since $u_i \sim N(\mu, \omega^2), u_i \geq 0$ the conditional distribution of u_i given the data has mean given by

$$m_{i,(j)} = \frac{\sigma_{(j)}\lambda_{(j)}}{1+\lambda^2_{(j)}}\left[\frac{\phi(z_{i,(j)})}{\Phi(z_{i,(j)})} - z_{i,(j)}\right],$$

where $z_{i,(j)} = \frac{e_{i,(j)}\lambda_{(j)}}{\sigma_{(j)}} + \frac{\mu_{(j)}}{\sigma_{(j)}\lambda_{(j)}}, \lambda_{(j)} = \omega_{(j)}/\sigma_{(j)}, e_{i,(j)} = y_i - x_i'\beta_{(j)}$, for each $i = 1, \ldots, n$, and ϕ, Φ denote the standard normal probability density and distribution functions, respectively. Therefore, $m_{i,(j)}$ is the inefficiency measure[8] for observation i when we localize with respect to observation j. A reasonable inefficiency measure for observation i is provided by $m_i^* = \sum_{j=1}^{n} m_{i,(j)} W_j$, which is a weighted average of all $m_{i,(j)}$ based on the LML weights $W_j = K_H(x_j - x)$. Naturally, the dominating element in this average is $m_{i,(i)}$, the inefficiency measure of a particular observation (bank) when we localize with respect to this observation. Since inefficiency estimate is based on bank-specific parameter estimates of β, μ, σ and ω our estimate of inefficiency for the particular bank is quite flexible.[9] The firm-specific cost efficiency measures can then be obtained from $\exp(-m_i^*)$.

We report estimates of cost efficiency in Fig. 5. Modal efficiency is found to be quite high and about half of the banks are found to be operating at the efficiency level of 90% or more. To explore this issue further we plot estimates of cost inefficiency against log assets in Fig. 6. From the scatter plot of banks we find some (weak) evidence to support the hypothesis that large banks are more efficient (a weak inverse relationship between inefficiency and log assets is observed from the scatter plot). Thus, one could argue that the cost advantage from mergers of large banks may not be very high (Berger and Humphrey 1991), especially from an efficiency point of view.

[8] This is the well-known Jondrow et al. (1982) estimator.

[9] It is not fully nonparametric because of the CD assumption on the anchoring function, distributional assumptions on efficiency and noise components.

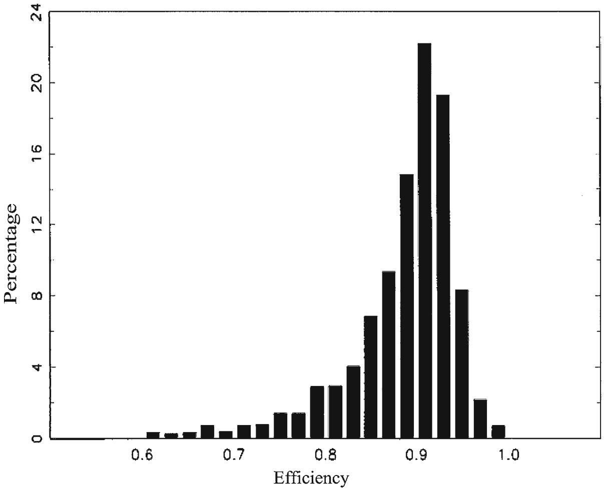

Fig. 5 Histogram of cost-efficiency (local ML)

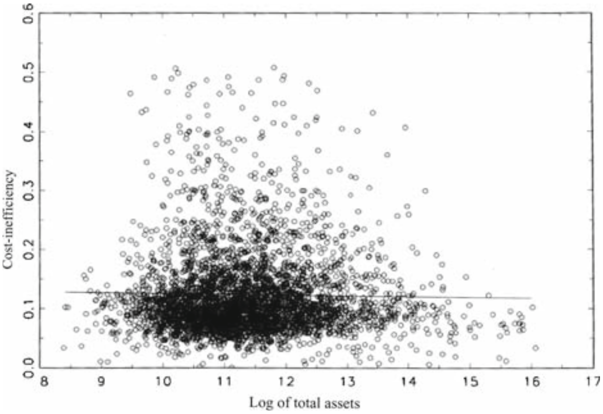

Fig. 6 Plot of cost-inefficiency against log(assets) (local ML)

4.3 The Cobb–Dougals LML and the global translog results: a comparison

McAllister and McManus (1993) fitted a parametric translog cost function to the entire data set for the year 1989 and found that (i) scale economies were absent for most of the medium and large banks, and (ii) extreme scale economies (diseconomies) were found for very small (very large) banks. In comparison, their localized translog model showed much smaller variations in scale economies. For the sake of comparison, we fit a single translog cost frontier for the entire data set in which we assume a truncated normal distribution for the inefficiency component and a normal distribution for the noise component. Heteroskedasticity is not included in any of the error components.[10] We find evidence of scale economies for majority of banks (see Fig. 8 that shows the histogram of SCE, and Fig. 9 that graphs scale economies against log assets). Scale diseconomies are found for the banks with assets more than 1.2 billions of dollars. Thus, the presence of scale economies for most of the banks is observed when a global translog cost frontier is fitted to the entire data set. In contrast, the localized CD cost function results show the presence of scale economies for banks of all sizes.[11] We also estimated the localized translog cost function and obtained similar results.

To compare the estimated efficiencies derived from the LML and global translog models, first, we compare the frequency distributions (reported in Figs. 5 and 10 as well as Figs. 6 and 11). It can be easily seen that these frequency distributions are quite similar. There are, however, differences in levels and spread. For example, the mean efficiency is higher in the LML model and the spread is smaller compared to the global translog model. In the LML model we find evidence to support that very large banks are as efficient as most of the small banks (and in general these banks are more efficient than some of the medium banks.[12] Since the LML model is more flexible and it accommodates heteroskedasticity associated with both error components, the LML results are robust to functional form misspecification, heteroskedasticity, etc. These advantages in turn give more precise results on both scale economies and efficiency compared to the global translog cost frontier.[13]

We conclude this section with the following remarks. The global parametric models used to estimate scale economies and cost efficiency of banks often led to results that are contrary to conventional wisdom (Wheelock and Wilson 2001; Wang and Kumbhakar 2006, among others). For example, the common sense argument used in favor of merger is that large banks take advantage of economies of scale. On the contrary, empirical findings (based on global parametric models) show that the large banks have exhausted economies of scale and they are generally less efficient than their smaller counterparts. Some of these findings might have resulted from assuming a single parametric cost function applicable to all the banks (small, medium, large, etc.) in the sample. If the cost function parameters are either group-specific (Wang and Kumbhakar 2006) or bank-specific (as in LML) then using a single cost function is likely to introduce

[10] Note that we model inefficiency following the stochastic frontier approach whereas McAllister and McManus (1993) did not, and our LML uses all the observations at every point of evaluation whereas they did it for only 25% of the observations.

[11] There are only a few banks for which we observe diseconomies of scale, and these banks are from all assets categories. That is, the banks operating beyond their efficient scale show no strong correlation with assets.

[12] The global translog model show large spread in efficiency among the very large and very small banks.

[13] Again the efficiency results based on the translog LML are similar to the Cobb–Douglas LML results. Since we also find similar result for scale economies, one can perhaps argue that the functional form for the anchoring model is not that important, at least for the present data.

bias in parameter estimates. These biases are likely to give inaccurate estimates of scale economies and cost efficiency (McAllister and McManus 1993).

4.4 A comparison between the Cobb–Dougals LML and the DEA results

Instead of using the CD anchoring function one can use other functions that are more flexible (such as the translog). However, such additional flexibility might not be necessary because the parameters are observation specific and therefore the cost function is quite flexible. Use of flexible functional form such as the translog often violates regularity conditions.[14] To check robustness of our CD LML results we compare them with the DEA efficiency scores. Since DEA is a nonparametric method of estimating efficiency, it might be argued that the DEA results should be comparable to results obtained from LML. That is, one might expect the DEA results closer to the LML results than those obtained from the global parametric models. Although the DEA is nonparametric it has two main differences with the LML approach. First, while estimating efficiency the DEA cannot take noise into account. Thus, for example good luck (negative noise) will be considered as efficiency in DEA cost efficiency measure. Consequently, the DEA cost frontier in the cost-output space will be below the cost frontier estimated using the SF approach and the LML. In other words, DEA efficiency scores will be lower than those of SFA and LML. Second, the LML can accommodate heterogeneity in the data by making the variance of the noise as well as the inefficiency components observation-specific. Ignoring heteroskedasticity might bias efficiency results even though the model specification (functional form, for example) is correct. Because of these two major differences the DEA and LML efficiency scores are likely to differ.

The distributions of efficiency scores (density plots) based on LML, DEA and the global translog model are reported in Fig. 7. It can be seen that the LML gives the highest efficiency score followed by the global translog and DEA. Since some observations (by construction) in DEA are 100% efficient, the distribution of DEA scores shows a jump towards the end. As expected the DEA efficiency scores are in most cases lower than those in global translog and LML (mainly because DEA cannot discriminate noise from inefficiency). The LML scores are, in general, higher than those from global translog mainly because of inflexibility of the translog to accommodate heterogeneous technology and heteroskedasticity in the noise and inefficiency. The magnitude of the difference will, however, depend on the data—perhaps directly linked to the degree of heterogeneity in the production units.

[14] In fact, we used translog as the anchoring function and found that results are not much different from the CD. Because of this we are not reporting results from the translog model. Moreover, the translog function has much more parameters and often violates regularity conditions on many data points. This is another reason why a simple functional for such as the translog might be preferred for most of the applications.

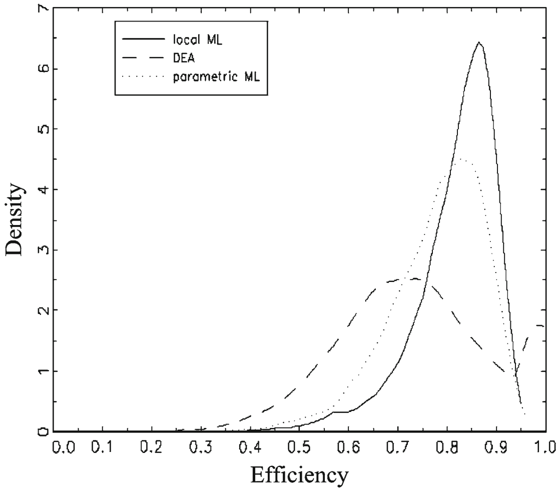

Fig. 7 Distribution of efficiency in DEA, LML and global translog

5 Conclusions

In this paper, we relaxed some rigidities/assumptions associated with estimation of stochastic frontier (SF) models and applied the KPST (2007) local ML estimator for SF models. We introduce a truncated normal distribution instead of a half-normal distribution for the one-sided inefficiency component in the estimation of a stochastic cost frontier using a sample of 3,691 U.S. commercial banks for the year 2000. This approach allows us to model inefficiency effects in the spirit of KGM (1991) and Battese and Coelli (1995). These estimates are, however, more flexible because the underlying model is semiparametric.

Empirically, we find that (i) cost elasticities with respect to outputs and inputs vary substantially among banks; (ii) scale economies are present for most of the banks. Furthermore, we don't find any evidence to support that large banks are less efficient compared to the small banks. Thus, in general we find evidence to support conventional wisdom (i.e., large banks are more efficient and can exploit economies of scale). Although a flexible parametric cost function generates observation-specific elasticities, scale economies, cost efficiency, etc., these so called globally flexible functions are found to violate properties of cost functions at many points, and often give unreliable estimates of scale economies. Results from these models do not always support conventional wisdom believed by many bankers. The semiparametric cost model and the use of LML makes the technology quite flexible even with Cobb–Douglas anchoring function.

Acknowledgements We thank two anonymous referees and Baldev Raj for their constructive comments. We would also like to thank W. Greene, R. Sickles, P. Schmidt, L. Simar, L. Orea and the participants of NAPWII at Schenectady, NY and ACEP at Taipei, Taiwan for their comments on an earlier draft of the paper. None, other than us, is responsible for any errors.

Appendix

Appendix Figs. 8, 9, 10, 11.

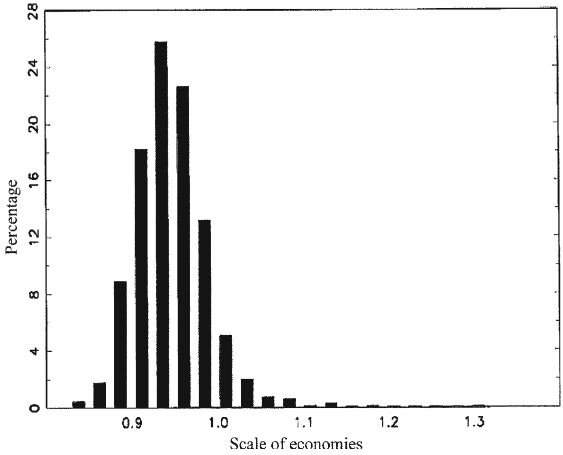

Fig. 8 Histogram of SCE (global TL)

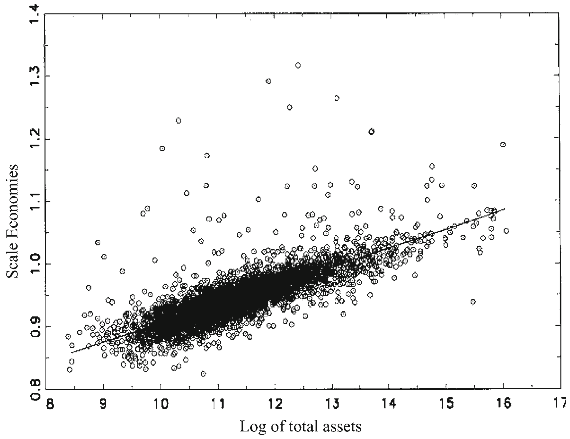

Fig. 9 Plot of SCE against log(assets) (global TL)

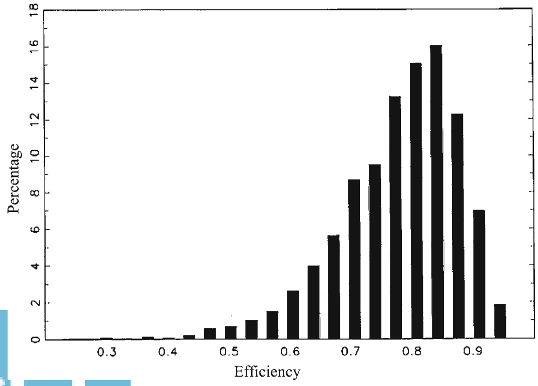

Fig. 10 Histogram of cost-efficiency (global TL)

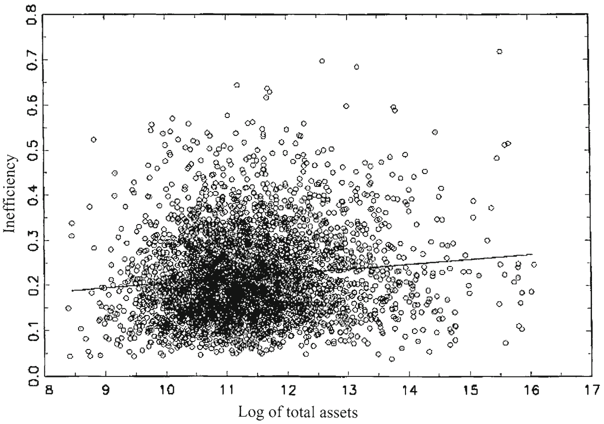

Fig. 11 Plot of cost-inefficiency against log(assets) (global TL)

References

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

Battese GE, Coelli TJ (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. Empir Econ 20:325–332

Berger A, Humphrey D (1991) The dominance of inefficiencies over scale and product mix economies in banking. J Monetary Econ 28:117–148

Berger AN, Mester LJ (1997) Inside the black box: what explain differences in the efficiency of financial institutions? J Bank Financ 21:895–947

Boyd JH, Graham SL (1991) Investigating the banking consolidation trend. Q Rev Federal Bank Minneap 3–15

Broeck van den J, Koop G, Osiewalski J, Steel MFJ (1994) Stochastic frontier models: a Bayesian perspective. J Econom 61:273–303

Caudill SB, Ford JM, Gropper DM (1995) Frontier estimation and firm-specific inefficiency measures in the presence of heteroskedasticity. J Bus Econ Stat 13:105–111

Cazals C, Florens J-P, Simar L (2002) Nonparametric frontier estimation: a robust approach. J Econom 106:1–25

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision-making units. Eur J Oper Res 2:429–444

Hadri K (1999) Estimation of a doubly heteroscedastic stochastic frontier cost function. J Bus Econ Stat 17:359–363

Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. J Econom 19:233–238

Kaparakis EI, Miller SM, Noulas A (1994) Short-run cost-inefficiency of commercial banks: a flexible stochastic frontier approach. J Money Credit Bank 26:875–893

Kumbhakar SC, Ghosh S, McGuckin T (1991) A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. J Bus Econ Stat 9:279–286

Kumbhakar SC, Lovell CAK (2000) Stochastic frontier analysis. Cambridge University Press, New York

Kumbhakar SC, Park BU, Simar L, Tsionas EG (2007) Nonparametric stochastic frontiers: a local maximum likelihood approach. J Econ 137: 1–27

McManus DA (1994a) Making the Cobb–Douglas functional form an efficient nonparametric estimator through localization. Manuscript, Board of Governors of the Federal Reserve Bank

McManus DA (1994b) The nonparametric translog with application to banking scale and scope economies. In: Proceedings of the Business and Economic Statistics Section, Am Stat Assoc

McAllister PH, McManus DA (1993) Resolving the scale efficiency puzzle in banking. J Bank Financ 17:389–405

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb–Douglas production functions with composed error. Int Econ Rev 8:435–444

Mukherjee K, Ray SC, Miller SM (2001) Productivity growth in large US commercial banks: the initial post-deregulation experience. J Bank Financ 25:913–939

Pagan A, Ullah A (1999) Nonparametric econometrics. Cambridge University Press, Cambridge

Park BU, Sickles RC, Simar L (1998) Stochastic panel frontiers: a semiparametric approach. J Econom 84:273–301

Simar L, Wilson PW (2000) Statistical inference in nonparametric frontier models: the state of the art. J Product Anal 13:49–78

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

Tibshirani R (1984) Local likelihood estimation. Ph.D. thesis, Stanford University

Tsionas EG (2002) Stochastic frontier models with random coefficients. J Appl Econom 17:127–147

Wang H-J (2002) Heteroskedasticity and non-monotonic efficiency effects of a stochastic frontier model. J Product Anal 18:241–253

Wang D, Kumbhakar SC (2006) Strategic groups and heterogeneous technologies: An application to the US banking industry. Manuscript, SUNY Binghamton, New York

Wheelock DC, Wilson PW (2001) New evidence on returns to scale and product mix among U.S. commercial banks. J Monet Econ 47:653–674

Yatchew A (1998) Nonparametric regression techniques in economics. J Econ Lit 36:669–721