Non-parametric stochastic frontier models^{*}

Subal C. Kumbhakar Department of Economics State University of New York Binghamton, NY 13902, USA. Phone: (607) 777 4762, Fax: (607) 777 2681 E-mail: kkar@binghamton.edu

and

Efthymios G. Tsionas Department of Economics Athens University of Economics and Business 76 Patission Street, 104 34 Athens, Greece Phone: (301) 0820 388, Fax: (301) 0820 3301 E-mail: <u>tsionas@aueb.gr</u>

August 1, 2002

ABSTRACT

In this paper, we use local maximum likelihood (LML) method to estimate stochastic frontier models. This method permits us to remove many of the standard deficiencies of econometric SF models. In particular, we relax the assumption that all firms share the same production technology and provide completely firm-specific parameter estimates and inefficiency measures. We also introduce non-parametric heteroscedasticity in both the noise and inefficiency components, allow for non-parametric inefficiency effects. A cost frontier is estimated for a sample of 3691 U.S. commercial banks for the year 2000 to illustrate the new technique.

JEL Classification No: C14, C50, D23, G21.

Keywords: Anchoring Model; Data Envelopment Analysis; Cost Efficiency; Non-parametric estimation; Local Maximum Likelihood estimation, U.S. Commercial Bank.

^{*} We 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. None, other than us, is responsible for any errors.

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) 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 heavily on the choice of functional forms, distributional assumptions, fixity of parameters of the underlying production technology, and so on. Some of these are strong assumptions and, in practice, one is always subject to the criticism that empirical results depend on these assumptions. For example, 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 theory can be rejected, and estimates of efficiency will be biased.

An alternative to the SF approach is the deterministic non-parametric approach, viz., the Data Envelopment Analysis (DEA) popularized by Charnes et al. (1978). While the SF models assume specific functional forms for the production or cost frontiers, and adopt strong distributional assumptions on the noise and inefficiency components, the DEA models do not make such assumptions. However, it cannot separate 'genuine 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. For DEA models the statistical theory is not well developed (although some progress have been made in terms of bootstrapping (see, for example, Simar and Wilson (2000)), as a result of which most applied researchers are unable to make statements regarding the statistical properties of the estimated functions such as input elasticities, scale economies, efficiency, etc.

Park, Sickles, and Simar (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 non-parametric estimator based on the FDH concept. The new estimator is more robust relative to DEA but it

will not envelope all the data. This is, essentially, a stochastic DEA estimator for which the authors provide an asymptotic theory (Simar and Wilson (2000)).

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 non-parametric stochastic frontier models with arbitrary heteroscedasticity, and arbitrary dependence of efficiency on covariates. We use the local maximum likelihood (LML) method, which is a non-parametric 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 **b** is a non-parametric function $\mathbf{b}(x_i)$, the familiar linear model $y_i = x'_i \mathbf{b} + u_i$ becomes effectively a non-parametric model.

We take advantage of the LML methodology in estimating SF models in such a way that many of the limitations of the SF models originally proposed by Aigner et al. (1977), Meeusen and van den Broeck (1977), and their extensions in the last two and a half decades are relaxed. First, we relax the functional form assumption. By making the parameters of the underlying production technology functions of data, we make the technology completely flexible. Second, we introduce non-parametric heteroscedasticity in the one-sided inefficiency component as well as in the noise component, instead of assuming specific functional forms for heteroscedasticity. Third, we allow for unspecified, non-parametric dependence of inefficiency (both the mean and the variance) on a vector of exogenous variables. By doing so the propose method is able to provide completely non-parametric inefficiency estimates. This is because the observation-specific estimates of inefficiency depend neither on the assumption that all firms share a global technology nor on the assumption that the inefficiency distribution is the same for all producers. Thus, the main contribution of this paper is in the estimation of SF models free from many (if not all) of the restrictive assumptions that are currently used. The removal of all these deficiencies turns SF models into non-parametric models comparable to the DEA. Moreover, we can apply standard econometric tools to perform estimation and draw inferences.

The remainder of the paper is organized as follows. Local estimation is reviewed in section 2. Local ML estimation of SF models is presented in section 3. Some computational and practical issues are discussed in section 4. In section 5 we illustrate the LML technique by

estimating cost frontiers using a sample of U.S. commercial banks. The paper concludes with a summary of the main findings in section 6.

2. Local estimation

Suppose the model is $y_i = f(x_i) + e_i$ where f is an unknown function to be estimated nonparametrically, and x_i is a scalar explanatory variable. The Nadaraya-Watson estimator of the unknown function (Pagan and Ullah (1999), pp. 79-83) minimizes the criterion

$$\sum_{i=1}^{n} (y_i - m)^2 K_h(x_i - x) \text{ with respect to } m \text{ , and provides the solution } \widetilde{m}(x) = \frac{\sum_{i=1}^{n} K_i y_i}{\sum_{i=1}^{n} y_i} \text{ where } \sum_{i=1}^{n} y_i$$

 $K_i \equiv K_h(x_i - x)$. This estimator fits a constant to the data and performs weighted LS to estimate this constant. The weights depend on x, and the model is effectively non-parametric. Alternatively, instead of fitting a constant one can fit a linear model in which case the relevant criterion to minimize would be $\sum_{i=1}^{n} (y_i - a - b_{x_i})^2 K_h(x_i - x)$. The resulting estimates $\tilde{a}(x)$ and $\tilde{b}(x)$ depend on x and are also non-parametric and can be computed using weighted LS across a number of x points.

Fan (1992, 1993), Fan and Gijbels (1992) and Ruppert and Wand (1994) have extensively investigated the local linear estimator.¹ Gozalo and Linton (2000) provided a generalization of the local linear estimator based on an anchoring model f(x; q). Their local nonlinear least locally by minimizing the squares estimator estimates a criterion function $\sum_{i=1}^{n} (y_i - f(x_i; \boldsymbol{q}))^2 \kappa_h(x_i - x)$. They showed that the asymptotic variance and the asymptotic bias of $f(x_i; \mathbf{q})$ do not depend on the particular kernel, and anchoring models that are globally closer to the true non-parametric model (i.e., the distance between f(x; q) and the true model f(x) is small for all x) endow the local estimator with better bias performance. There are, however, many ways to combine parametric and non-parametric information (see, for example, Pagan and Ullah (1999, pp. 106-108)) but local estimation seems particularly well suited for econometric applications. One usually has a good idea what the model should be (for example, a Cobb-Douglas or translog production function) but we cannot claim that this is exactly an appropriate functional form globally. By localizing the parameters of these models it is possible to construct non-parametric estimators of the unknown functional form.

It is not possible to apply directly the local NLS algorithm of Gozalo and Linton (2000) in the case of stochastic frontiers. This is because the distribution of the dependent variable conditional on the parameters and the covariates does not admit a factorization that reduces the model to a specification that can be estimated by local NLS method. As a result of this we consider a LML approach.

To fix ideas, suppose we have a parametric model that specifies the density of an observed dependent variable y_i conditional on a vector of observable covariates $x_i \in X \subseteq \mathbb{R}^k$, a vector of unknown parameters $\mathbf{q} \in \Theta \subseteq \mathbb{R}^m$, and let the density be $l(y_i; x_i, \mathbf{q})$. The parametric ML estimator is given by

$$\widetilde{\boldsymbol{q}} = \underset{\boldsymbol{q} \in \Theta}{\operatorname{argmax}} : \sum_{i=1}^{n} \ln l(y_i; x_i, \boldsymbol{q})$$

The problem with the parametric ML estimator is that it relies heavily on the parametric model that can be incorrect if there is uncertainty regarding the functional form of the model, the density, *etc*. The LML estimation technique is a way to allow for nonparametric effects within the parametric model. A natural way to convert the parametric model to a nonparametric one is to make the parameter q 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

$$\widetilde{\boldsymbol{q}}(x) = \underset{\boldsymbol{q}\in\Theta}{\operatorname{argmax}} : \sum_{i=1}^{n} \ln l(y_i; x_i, \boldsymbol{q}) K_H(x_i - x)$$

where K_H is a kernel that depends on a matrix bandwidth H. The idea behind LML 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

¹ See Hastie and Loader (1993) for a review.

the parametric ML estimator would do. By solving the LML problem for several points $x \in X$, we can construct the function $\tilde{q}(x)$ that is an estimator for q(x), and effectively we have a fully general way to convert the parametric model to a non-parametric approximation to the unknown model.

LML estimation has been proposed by Tibshirani (1984) and has been applied by Gozalo and Linton (2000) in the context of non-parametric estimation of discrete response models, using the probit as an anchoring model (see also, Pagan and Ullah (1999, p. 286)). Their estimator effectively removes the assumption of a particular distributional form. LML estimation is a natural extension of local linear estimation (Pagan and Ullah (1999, pp. 93-106)).

Properties of the LML estimator are analogous to the properties of local nonlinear least squares (Gozalo and Linton, 2000) or the local likelihood estimator of a density (Chapter 2 in Pagan and Ullah, 1999). Furthermore, standard normal asymptotics apply to the functional fits. More specifically, the asymptotic variance of the estimated function $\tilde{f}(x; \tilde{q}(x))$ is independent of the anchoring parametric model, so it should be the same as the variance of the Nadaraya-Watson and local linear estimators. Naturally, the asymptotic variance depends on the bandwidth parameter h. However, it does not depend on the joint distribution of regressors so it is design-adaptive. The behavior of the bias depends on the distance of the anchoring model f(x; q) from the nonparametric model, f(x). For example, if the true function is close to a functional form g(x), local estimation anchoring on g(x) will have better bias performance relative to the linear form for example. An important property is that if the anchoring model is approximately true (for some parameter value and for every x) then there is no upper bound on bandwidth parameter and, therefore, one could choose higher bandwidth values to get faster converge to the asymptotic distribution. Gozalo and Linton (2000) illustrate these properties nicely in the context of local likelihood analysis with an anchoring probit model.²

 $^{^{2}}$ Hall and Simar (2002), show that there can be no unique solution to the non-parametric frontier problem in the presence of measurement error. However, they argue that a useful non-parametric approach can be developed when measurement error variance is small. This result holds when error distributions are completely unknown. Our approach differs from Hall and Simar since we maintain normality assumptions on error terms (although we allow for arbitrary heteroscedasticity and inefficiency effects), and use a parametric anchoring model that is globally "close" to the frontier.

3. Local Maximum Likelihood estimation of stochastic frontier models

Suppose we have the following stochastic frontier cost model

$$y_i = x'_i \boldsymbol{b} + v_i + u_i; v_i \sim \mathbb{N}(0, \boldsymbol{s}^2), u_i \sim \mathbb{N}(\boldsymbol{m} \boldsymbol{w}^2), u_i \geq 0 \text{ for } i = 1, ..., n, \boldsymbol{b} \in \mathbb{R}^k$$

where y is log cost and x_i is a vector of input prices and outputs³; 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 . This model is heavily parametric. First of all, it is linear in x_i , although one can make it non-linear without any major problem. Second, it makes strong distributional assumptions on the two-sided (v) and one-sided (u) error terms. Third, it assumes that the parameter vector **b** that describes the underlying production technology, and more importantly **m** and **w** do not depend on x_i . Although some SF models assume that **m** and **w** are linear or log-linear functions of some covariates, these specifications are *ad hoc*. It is well known that the end results (parameter estimates as well as estimated efficiency) depend to a great extent on functional form assumptions, as well as assumptions about the covariates entering in these functions. For these reasons, many empirical researchers are relactant to use the SF models in efficiency studies and adopt DEA formulations instead.

To make the frontier model non-parametric, 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):

$$\widetilde{\boldsymbol{q}} = \underset{\boldsymbol{q}\in\Theta}{\operatorname{argmax}} : \sum_{i=1}^{n} \ln l(y_i; x_i, \boldsymbol{q})$$

where

$$l(y_i; x_i, \boldsymbol{q}) = [\Phi(\boldsymbol{y})]^{-1} \Phi \left[\frac{\boldsymbol{s}^2 \boldsymbol{y} + \boldsymbol{w}(y_i - x_i' \boldsymbol{b})}{\boldsymbol{s}(\boldsymbol{w}^2 + \boldsymbol{s}^2)^{1/2}} \right] \left[2\boldsymbol{p}(\boldsymbol{w}^2 + \boldsymbol{s}^2) \right]^{-1/2} \exp \left[-\frac{(y_i - x_i' \boldsymbol{b} - \boldsymbol{m})^2}{2(\boldsymbol{w}^2 + \boldsymbol{s}^2)} \right],$$

 $\mathbf{y} = \mathbf{m}/\mathbf{w}$, and Φ denotes the standard normal cumulative distribution function. The parameter vector is $\boldsymbol{q} = [\boldsymbol{b}, \boldsymbol{s}, \boldsymbol{w}, \boldsymbol{y}]$ and the parameter space is $\Theta = R^k \times R_+ \times R_+ \times R_-$. Local ML estimation of the corresponding non-parametric model involves the following steps. First, we choose a kernel function. A reasonable choice is

$$K_{H}(d) = (2\mathbf{p})^{-m/2} |H|^{1/2} \exp\left(-\frac{1}{2}d'H^{-1}d\right), d \in \mathbb{R}^{m},$$

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

$$\widetilde{\boldsymbol{q}}(x) = \underset{\boldsymbol{q} \in \Theta}{\operatorname{arg max}} :$$

$$\sum_{i=1}^{n} \left\{ -\ln \Phi(\boldsymbol{y}) + \ln \Phi\left[\frac{\boldsymbol{s}^{2}\boldsymbol{y} + \boldsymbol{w}(y_{i} - x_{i}'\boldsymbol{b})}{\boldsymbol{s}(\boldsymbol{w}^{2} + \boldsymbol{s}^{2})^{1/2}}\right] - \frac{1}{2}\ln \left(\boldsymbol{w}^{2} + \boldsymbol{s}^{2}\right) - \frac{1}{2}\frac{(y_{i} - x_{i}'\boldsymbol{b} - \boldsymbol{m})^{2}}{(\boldsymbol{w}^{2} + \boldsymbol{s}^{2})}\right\} K_{H}(x_{i} - x)$$

A solution to this problem provides the LML parameter estimates $\tilde{\boldsymbol{b}}(x), \tilde{\boldsymbol{s}}(x), \tilde{\boldsymbol{w}}(x)$ and $\tilde{\mathbf{y}}(x)$. Also notice that the weights $K_{\mu}(x_i - x)$ 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.⁴

Following are some of the reasons why the LML estimate of the SF models is an improvement over the existing alternatives. First, the parameter estimates $\tilde{b}(x)$ depend on x so we completely solve the functional form misspecification problem in stochastic frontier

$$\widetilde{\boldsymbol{q}}(z) = \underset{\boldsymbol{q} \in \Theta}{\operatorname{arg max}}:$$

$$\sum_{i=1}^{n} \left\{ -\ln \Phi(\boldsymbol{y}) + \ln \Phi\left[\frac{\boldsymbol{s}^{2}\boldsymbol{y} + \boldsymbol{w}(y_{i} - x_{i}'\boldsymbol{b})}{\boldsymbol{s}(\boldsymbol{w}^{2} + \boldsymbol{s}^{2})^{1/2}}\right] - \frac{1}{2}\ln\left(\boldsymbol{w}^{2} + \boldsymbol{s}^{2}\right) - \frac{1}{2}\frac{(y_{i} - x_{i}'\boldsymbol{b} - \boldsymbol{m})^{2}}{(\boldsymbol{w}^{2} + \boldsymbol{s}^{2})}\right\} K_{H}(z_{i} - z)$$

where z are the given values for the vector of exogenous variables. The main feature of this formulation is that the **b** parameters as well as S, W, and Y will now be functions of z instead of x.

 $^{^{3}}$ The cost function specification is discussed in details in section 5.2. 4 An alternative, that could be relevant in some applications, is to localize based on a vector of exogenous variables z_i instead of the x_i 's. In that case, the LML problem becomes

models in the following sense. If we have a regression model $y_i = x_i' \boldsymbol{b}(x_i) + e_i$ with $e_i \sim IN(0, \boldsymbol{S}^2(x_i))$ where $\boldsymbol{b}(x_i)$ and $\boldsymbol{S}(x_i)$ are non-parametric functions of x, then the model is effectively non-parametric.⁵

Second, variances of both u and v (i.e., s^2 and w^2) are made functions of x and are estimated non-parametrically. This means that effectively we have heteroscedasticity of unknown form in both the noise and inefficiency components. Thus the present formulation generalizes Caudill, Ford and Gropper (1995), Hadri (1999), Kumbhakar and Lovell (2000) in the non-parametric direction without imposing any functional form assumptions on the structure of heteroscedasticity so far as the variance of the inefficiency component is concerned. The variance of the noise term is often viewed as risk. That is, a producer with higher variance of the noise component v is considered to be riskier (compared to an otherwise identical producer) from production/cost point of view. Such risks can often be explained by some specific inputs (Kumbhakar and Tveterås, 2002). Furthermore, it is likely that such risks vary among producers. Since s^2 is a non-parametric function of x, we can claim that our model captures producer-specific production/cost risk so long as the covariates are producer-specific. One can also examine effects of covariates on risk without assuming any functional form⁶ on the risk function s^2 . Such marginal effects are producer-specific and also vary with covariates.

Third, since y is made a function of x, we have inefficiency effects of non-parametric form. Thus the present model generalizes Kumbhakar, Ghosh and McGuckin (1991) and Battese and Coelli (1995) formulation of determinants of inefficiency in the non-parametric direction.

Fourth, the model generalizes the "thick frontier" concept (Berger and Humphrey (1991)). The thick frontier model fits a parametric model (for example the translog cost function) to quartiles of average cost and, therefore, it provides parameter estimates (of the usual translog cost function) that are specific to quartiles. In the context of the present specification, we are able to make all parameters (not just regression parameters) observation-specific. A

⁵ The model also generalizes the random coefficient stochastic frontier model of Tsionas (2002) without making any strong distributional assumptions on the coefficients or assuming that the coefficients do not depend on covariates.

⁶ Following Just and Pope (1978), Kumbhakar and Tveterås (2002) assumed specific functional for the risk function in estimating production functions without taking inefficiency into account.

disadvantage of thick frontiers is the assumption that all firms within a given quartile share the same technology, and face the same set of parameters of inefficiency estimates. Furthermore, it is not possible to test any hypothesis using results from different quartiles.

4. Some computational/practical issues

The LML method proposed here is somewhat computationally intensive ($O(n^2)$ -intensive), especially localization is performed at $x = x_i$ for all i = 1, ..., n. Since for each x we have good starting values from the parametric ML estimation convergence of nonlinear estimation algorithms7 will typically be fast. In practice when the sample contains a large number of observations one may make a choice of "interesting" points $x \in X$ where the LML estimator is computed. For example, first, we may classify the dependent variable y_i into deciles/percentiles, and find the corresponding x_i 's for the given decile/percentile. Then we choose x_i as the median of x_i 's for the given decile/percentile, and solve the LML optimization problem for each one of these x's. Effectively, we have parameter estimates that decile/percentile-specific provided that medians explanatory are of variables are representative for the given decile/percentile. In this way, we can reduce computational costs significantly since it is required is to solve only ten/hundred LML optimization problems. Since good starting values are available from the parametric ML estimator, this is unlikely to place enormous computational burden upon empirical research.

Another practical issue 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 with all data except for observation *j*, and define for some $\overline{x} \in X$,

$$\widetilde{\boldsymbol{q}}^{(j)}(x,h) = \operatorname*{argmax}_{\boldsymbol{q}\in\Theta} : \sum_{i\neq j} \ln l(y_i; x_i, \boldsymbol{q}) K_H(x_i - \overline{x})$$

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

⁷ Widely used algorithms are BHHH and BFGS.

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

where $\tilde{y}_j(h)$ denotes the fitted value of y_j based on *h*. For stochastic frontier models, this problem is particularly easy because cross-validation can be implemented without actually solving the LML optimization problem.⁸

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 Cobb-Douglas or translog models depending on whichever model specification provides a better fit of the data. The choice will also influence computational burden since translog models involve many parameters. Another important consideration is that anchoring models must be able to incorporate parametric curvature and monotonicity restrictions. This is straightforward for the Cobb-Douglas but more complicated for the translog, where such restrictions have to be imposed at each observed data point.

So far as the choice of an anchoring model for the one-sided error is concerned, one can choose from the half-normal, truncated normal, exponential, and gamma distributions. The half-normal distribution is a special case of the truncated normal distribution when y = m = 0. Gamma distributions (Greene (1990), Ritter and Simar (1997), Tsionas (2000)) are difficult to work with and, therefore, may not be well suited as anchoring models in non-parametric stochastic frontier models since iterative non-linear estimation algorithms may fail during the course of fitting the model to a particular point. An exponential distribution (special case of the gamma distribution) would be a reasonable competitor of a half-normal specification. Therefore, in terms of 'well-behaved' models, the truncated normal specification is the most general and has the added advantage that it allows to parameterize the mean in terms of the explanatory variables in a non-parametric fashion. In practice, the likelihood functions resulting from a truncated normal distribution for the one-sided error tend to be flat in the direction of y (Greene (1994), Ritter and Simar (1997)) that might cause convergence problem (it might converge to unreasonable values). One way to solve this

⁸ It is known that cross-validation is not a panacea in bandwidth selection. For larger values of the bandwidth parameter h, we are effectively placing more weight on distant points from x, and in the limit as $h \rightarrow \infty$ we recover the parametric ML estimator. Therefore, it is a good idea to keep the bandwidth parameter relatively "small" in order to recover the local properties of the true non-parametric function. Gozalo and Linton (2000) also recommend bandwidth selection based on the asymptotic distribution of functional fits.

problem is to adopt a pseudo-prior distribution for y as in van den Broeck, Koop, Osiewalski and Steel (1994), which is to assume that $y \sim N(0, a^2)$ where a > 0 is the "prior" standard deviation of the y parameter. This results in a quasi Bayes estimator. Local quasi Bayes estimators result when the anchoring quasi Bayes estimator is localized to each observation or to a group of observations by some rule. We find that this choice makes the optimization problem more regular and convergence is much faster. Since we have more than 3,600 observations in our application, the pseudo-prior should have a minimal effect on final estimates. The introduction of pseudo-prior should not make the empirical researchers, especially the non-practitioners of Bayesian methods, feel uneasy given its advantage in regularizing the LML optimization problems. Hamilton (1994, p. 689) employed similar methods in the context of estimation of finite normal mixture models using sampling theory.

5. An application to U.S. commercial banks

The above methodology is applied to analyze 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. 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? Answer to these questions depends on the estimation technique (parametric vs. non-parametric) used, functional form chosen, etc.⁹ Since the banking industry consists of large number of small banks and assets are highly concentrated in a few very large banks, heteroscedasticity is likely to be present in both the noise and inefficiency components.¹⁰ Moreover, the production technology among banks is likely to differ.¹¹ These problems are avoided in the

⁹ There are numerous studies that address scale economies and efficiency. See, e.g., McAllister and McManus (1993), Berger and Mester (1997), Berger and Humphrey (1992), Boyd and Graham (1991), Mukherjee et al. (2001), Wheelock and Wilson (2001), among others.
¹⁰ It is well known that if inefficiency component is heteroscedastic and one ignores it, both parameter estimates

¹⁰ It is well known that if inefficiency component is heteroscedastic and one ignores it, both parameter estimates and estimated inefficiencies will be inconsistent (see Kumbhakar and Lovell (2000, Chapter 3.4)). Consequently, estimated of economies of scale are likely to be wrong.

¹¹ 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.

non-parametric LML model that makes parameters bank-specific without using any ad hoc specification.

5.1 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 and selected a sample of 3691 commercial banks. Median value of assets of these banks is 76 million dollars. The distributions of bank assets and banks are shown in Figure 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 very likely that the parameters of the underlying technology (cost function in our case) will differ among banks.



Figure 1: Distribution of assets/banks

In banking literature there is 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 individuals for (to

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-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 number of full time 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.

5.2. Results from the localized Cobb-Douglas model

We choose a Cobb-Douglas functional form primarily because a simple OLS fit of a Cobb-Douglas cost function resulted in a reasonably good fit (R^2 of about 0.93). We have also fitted a translog, but the Schwarz criterion strongly favored the Cobb-Douglas specification. Therefore, for the data at hand, the Cobb-Douglas cost function provides an acceptable local fit. Moreover, 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. Since we localize the parameters at each point, flexibility is not a problem. In other words, the use of the CD function gives a clear meaning to each and every coefficient and each of these coefficients are made bank-specific through localization. We choose the *h* parameter by using cross-validation in the relevant range of that parameter. To minimize computational costs, we perform cross-validation using median values of variables by deciles of the dependent variable as our target variables. Therefore, for each value of h we performed only ten local ML estimations.

We experimented with both half-normal and truncated normal distributions on the one-sided error 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 cost function (note the change the notation of the dependant varia ble), viz.,

 $C_i = x_i' \mathbf{b} + v_i + u_i$, where as before $v_i \sim IN(0, \mathbf{S}^2)$ and $u_i \sim IN(\mathbf{m} \mathbf{w}^2)$, $u_i \ge 0$ i = 1,..., n, $\mathbf{b} \in \mathbb{R}^{k+m}$. Here *C* 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 total cost and the input prices by one input price (w₃) before taking logs. Thus, the estimated cost function is

$$C = \boldsymbol{b}_{0} + \sum_{i} \boldsymbol{b}_{yi} \ln y_{i} + \sum_{j \neq 3} \boldsymbol{b}_{wj} \ln(w_{j} / w_{3}) + v_{i} + u_{i}$$

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

We report the frequency distribution of estimated parameters in Figure 2. The histograms for the parameters show different patterns (some are unimodal while others are bimodal but none is symmetric). For example, the cost elasticities with respect to outputs (\mathbf{b}_{yi} , i = 1,...,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 an 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 \mathbf{s}_v and \mathbf{y} show large variations while the opposite is true for \mathbf{s}_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 $SCE = \sum_{i=1}^{5} \partial \ln C / \partial \ln y_i = \sum_{i=1}^{5} \mathbf{b}_{yi}(y, x)$. Since all the parameters are observation-specific, the SCE measure is bank-specific as well. Thus, although we start from a CD cost function, the SCE measure is fully flexible. The SCE

measures are reported in Figure 3 in a histogram. It can be easily seen from the histogram that economies of scale is not exhausted (SCE being less than unity thereby meaning that returns to scale is 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 Figure 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 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.

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 $\mathbf{b}_{(j)}$, $\mathbf{s}_{(j)}$, $\mathbf{m}_{(j)}$, $\mathbf{w}_{(j)}$. Since $u_i \sim N(\mathbf{m}, \mathbf{w}^2)$, $u_i \ge 0$ the conditional distribution of u_i given the data has mean given by

$$m_{i,(j)} = \frac{s_{(j)} l_{(j)}}{1 + l_{(j)}^2} \left[\frac{f(z_{i,(j)})}{\Phi(z_{i,(j)})} - z_{i,(j)} \right],$$

where $z_{i,(j)} = \frac{e_{i,(j)} \mathbf{l}_{(j)}}{\mathbf{s}_{(j)}} + \frac{\mathbf{m}_{(j)}}{\mathbf{s}_{(j)} \mathbf{l}_{(j)}}$, $\mathbf{l}_{(j)} = \mathbf{w}_{(j)} / \mathbf{s}_{(j)}$, $e_{i,(j)} = y_i - x'_i \mathbf{b}_{(j)}$, for each i = 1,..., n, and \mathbf{f}, Φ denote the standard normal probability density and distribution function respectively. Therefore, $m_{i,(j)}$ is the inefficiency measure¹² for observation i when we localize with respect to observation j. A reasonable inefficiency measure for observation iis provided by $m_i^* = \sum_{i=1}^n m_{i,(j)} W_i$ which is a weighted average of all $m_{i,(j)}$ based on the LML weights. Naturally, the dominating element in this average will be $m_{i,(j)}$, the inefficiency measure of a particular observation when we localize with respect to this observation. This inefficiency estimate is derived completely from firm-specific parameter estimates of $\mathbf{b}, \mathbf{m}, \mathbf{s}$ and **w** and can be viewed as a non-parametric estimate of inefficiency for the particular observation. The firm-specific cost efficiency measures can be obtained from $\exp(-m_i^*)$.

We report estimates of cost efficiency in Figure 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 Figure 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 merger of large banks may not be very high (Berger and Humphrey (1992)), especially from efficiency point of view.

5.3 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 (year 2000) in which we assume truncated normal distribution for the inefficiency component and normal distribution for the noise component. Heteroscedasticity is not included in any of the error components.¹³ We find evidence of scale economies for majority of banks (see Figure A.1 that shows the histogram of SCE, and Figure A.2 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 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

¹² This is the well-known Jondrow et al. (1982) estimator.

¹³ 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.

economies for banks of all sizes.¹⁴ We also estimated the localized translog cost function and obtained similar results.¹⁵

To compare the estimated efficiencies derived from the LML and global translog models, first, we compare the frequency distributions (reported in Figures 5 and A.3 as well as Figures 6 and A.4). 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.¹⁶ Since the LML model is more flexible and it accommodates heteroscedasticity associated with both error components, the LML results are robust to functional form misspecification, heteroscedasticity, etc. This is, however, not the case with the global translog cost functions that suffers from all the problems associated with the SF models. Thus, we credit the LML for its flexibility, which in turn gives more precise results on both scale economies and efficiency compared to the global translog cost frontier.¹⁷

We conclude this section with the following remarks. The parametric models used to estimate scale economies and cost efficiency of banks often led to results that are contrary to conventional wisdom. 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 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 bank-specific then using a single cost function almost surely introduces bias in parameter estimates. These biases are likely to give inaccurate estimates of scale economies and cost efficiency (McAllister and McManus (1993)).

¹⁴ 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.
¹⁵ Space constrain doesn't permit us to report all these results, which can be obtained from the authors upon

¹⁵ Space constrain doesn't permit us to report all these results, which can be obtained from the authors upon request.

¹⁶ The global translog model show large spread in efficiency among the very large and very small banks.

6. Conclusions

In this paper, we relaxed many rigidities/assumptions associated with estimation of stochastic frontier models. First, we made the parametric stochastic frontier (SF) models completely non-parametric by using the principle of local maximum likelihood (LML) estimation. This technique permitted us to remove the assumption of a rigid functional form for the technology, and provide completely firm-specific parameter estimates and inefficiency measures that are not dependent on the assumption that all firms share the same technology. Second, we introduced non-parametric heteroscedasticity in both the noise and inefficiency components in the composed error SF models. Third, we allowed for non-parametric inefficiency effects thereby relaxing the assumption that inefficiency effects are log-linear.

We used both the Cobb-Douglas and translog localized models to estimate the stochastic cost frontier using a sample of 3691 U.S. commercial banks for the year 2000. 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 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 don't always support conventional wisdom believed by many bankers.

¹⁷ 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.

References

Aigner D, Lovell CAK, Schmidt P., 1977, Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* **6**: 21-37.

Battese, G. E., and T. J. Coelli, 1995, A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data, *Empirical Economics* 20, 325-332.

Berger A. and D. Humphrey, 1991, The dominance of inefficiencies over scale and product mix economies in banking, *Journal of Monetary Economics*, 28, 117-148.

Berger A. and D. Humphrey, 1992, Megamergers in banking and the use of cost efficiency as an antitrust defense, *The Antitrust Bulletin* 37, 541-600.

Berger, A. N, and L. J. Mester, 1997, Inside the Black Box: What Explain Differences in the Efficiency of Financial Institutions?, *Journal of Banking and Finance*, 21, 895-947.

Boyd, J.H. and S.L. Graham, 1991, Investigating the banking consolidation trend, *Quarterly Review*, Federal Bank of Minneapolis, 3-15.

Broeck van den J, Koop G, Osiewalski J., M.F.J. Steel, 1994, Stochastic frontier models: A Bayesian perspective, *Journal of Econometrics* **61**: 273-303.

Caudill S.B., J.M. Ford, and D.M. Gropper, 1995, Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity, *Journal of Business and Economic Statistics* 13, 105-111.

Cazals, C., J.-P. Florens, and L. Simar, 2002, Nonparametric frontier estimation: a robust approach, *Journal of Econometrics* 106, 1-25.

Charnes, A., W. W. Cooper and E. Rhodes, 1978, Measuring the Efficiency of Decision-Making Units, *European Journal of Operational Research* 2:6, 429-44.

Fan, J., 1992, Design-adaptive nonparametric regression, *Journal of the American Statistical Association* 87, 998-1004.

Fan, J., 1993, Local linear regression smoothers and their minimax efficiencies, *Annals of Statistics* 21, 196-216.

Fan, J., and I. Gijbels, 1992, Variable bandwidth and local linear regression smoothers, *Annals of Statistics* 20, 2008-2036.

Gozalo, P.L., and O. Linton, 2000, Local nonlinear least squares estimation: using parametric information in nonparametric regression, *Journal of Econometrics* 99, 63-106.

Greene, W.H., 1990, A gamma-distributed stochastic frontier model, *Journal of Econometrics* 46, 141-163.

Greene, W.H., 1993, The econometric approach to efficiency analysis, in H.O. Fried, C.A.K. Lovell and S.S. Schmidt (eds), *The measurement of productive efficiency: Techniques and applications*, Oxford: Oxford University Press.

Hadri, K., 1999, Estimation of a doubly heteroscedastic stochastic frontier cost function, *Journal of Business and Economic Statistics*, 17, 359-363.

Hall, P., and L. Simar, 2002, Estimating a change point, boundary, or frontier in the presence of observation error, *Journal of the American Statistical Association*, forthcoming.

Hamilton, J.D., 1994, Time Series Analysis, Princeton, Princeton University Press.

Hastie, T., and C. Loader, 1993, Local regression: automatic kernel carpentry, *Statistical Science* 8, 120-143.

Jondrow, J., C. A. K. Lovell, I. S. Materov and P. Schmidt, 1982, On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model, *Journal of Econometrics* 19:2/3 (August), 233-38.

Just, R. E., and Pope, R. D., 1978, Stochastic Specification of Production Functions and Economic Implications, *Journal of Econometrics*, **7**, 67-86.

Kaparakis, E.I., Miller, S.M., and A. Noulas, 1994, *Journal of Money, Credit and Banking* 26, 875-893.

Kumbhakar, S., S. Ghosh, and T. McGuckin, 1991, A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms, *Journal of Business and Economic Statistics*, 279-286.

Kumbhakar, S. and R. Tveterås, 2002, Production Risk, Risk Preference and Firm-Heterogeneity, mimeo., State University of New York, Binghamton, New York.

McManus, D.A., 1994a, Making the Cobb-Douglas functional form an efficient nonparametric estimator through localization, manuscript, Board of Governors of the Federal Reserve Bank.

McManus, D.A., 1994b, The nonparametric translog with application to banking scale and scope economies, Proceedings of the Business and Economic Statistics Section, American Statistical Association.

McAllister, P.H. and D.A. McManus, 1993, , Resolving the scale efficiency puzzle in banking, *Journal of Banking and Finance* 17, 389-405.

Meeusen W. and van den Broeck J., 1977, Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* **8**: 435-444.

Mukherjee, K., Ray, S.C. and S.M. Miller, 2001, Productivity growth in large US commercial banks: The initial post-deregulation experience, *Journal of Banking and Finance* 25, 913-939.

Pagan, A., and A. Ullah, 1999, *Nonparametric econometrics*, Cambridge, Cambridge University Press, NY.

Park, B.U., R.C. Sickles, and L. Simar, 1998, Stochastic panel frontiers: A semiparametric approach, *Journal of Econometrics* 84, 273-301.

Ritter, C. and L. Simar, 1997, Pitfalls of normal-gamma stochastic frontier models, *Journal of Productivity Analysis* 8, 167-182.

Ruppert, D., and M.P. Wand, 1994, Multivariate weighted least squares regression, *Annals of Statistics* 22, 1346-1370.

Simar, L., and P.W. Wilson, 2000, Statistical inference in nonparametric frontier models: the state of the art, *Journal of Productivity Analysis* 13, 49-78.

Stevenson R.E, 1990. Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics* **13**: 57-66.

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

Tsionas, E.G., 2000, Full likelihood inference in normal gamma stochastic frontier modek, *Journal of Productivity Analysis*, 13, 179-201.

Tsionas, E.G., 2002, Stochastic frontier models with random coefficients, *Journal of Applied Econometrics* 17, 127-147.

Wheelock, D.C. and Wilson, P.W. (2001), New evidence on returns to scale and product mix among U.S. commercial banks, *Journal of Monetary Economics* 47, 653-674.

Yatchew, A., 1998, Nonparametric regression techniques in economics, *Journal of Economic Literature* 36, 669-721.





 \mathfrak{c}_1



Figure 3: Histogram of SCE (local ML)





3-

Figure 5: Histogram of cost-efficiency (local ML)



 \mathcal{V}



Figure 6: Plot of cost-inefficiency against log(assets) (local ML)

9



Figure A.1: Histogram of SCE (global TL)

~

scale economies









Figure A.3: Histogram of cost-efficiency (global TL)

A 3

