ECONOMIC DISCUSSION PAPERS

Efficiency Series Paper 06/2003

# Recent Developments in Stochastic Frontier Modeling

Subal Kumbhakar y Efthymios Tsionas

Departamento de Economía

Universidad de Oviedo

Available online at: www.uniovi.es/economia/edp.htm

# UNIVERSIDAD DE OVIEDO

# DEPARTAMENTO DE ECONOMÍA

## PERMANENT SEMINAR ON EFFICIENCY AND PRODUCTIVITY

## RECENT DEVELOPMENTS IN STOCHASTIC FRONTIER MODELING*

**Subal C. Kumbhakar♣ and Efthymios G. Tsionas♠**

**Efficiency Series Paper 06/2003**

**Abstract:** Some of the recent developments in the efficiency measurement area using stochastic frontier models are: A. Estimation of the IO model, B. Latent class models to examine technological heterogeneity as well as heterogeneity in economic behavior, C. Estimation of stochastic frontier models using LML, D. Non-constant parameters: Random coefficient models with and without inefficiency, Markov switching stochastic frontier models, E. Estimation of cost/profit system with technical and allocative inefficiency using alternative techniques. We consider these as "open problems". In the past, estimation of some of these models was considered to be too difficult. Advances have been made in recent years to estimate some of these so-called difficult models. In this paper we will focus on the first three of the above topics. There are some papers that deal with issues listed under D and E. Both Bayesian and classical approaches are used to address these issues.

---

* Keynote address to the 8th European Workshop on Efficiency and Productivity Analysis held at the University of Oviedo, Spain, Sept. 24-27th, 2003. This address is drawn from three papers that are available from the authors upon request. The first paper is “Estimation of Stochastic Frontier Production Functions with Input-Oriented Technical Efficiency”, the second paper is “Do Firms Maximize Profit? A Latent Class Approach” and the third paper is “Non-parametric stochastic frontier models” which is undergoing a major overhaul with two additional co-authors (L. Simar and B. Park). Some other papers that address the issues listed under D and E are available upon request.

♣ Department of Economics, State University of New York, Binghamton, USA.

♠ Athens University of Economics and Business, Athens, Greece.

E-mails: kkar@binghamton.edu; tsionas@aueb.gr.

## 1. Introduction

In this paper we focus on three issues. First, we discuss issues (mostly econometric) related to input-oriented (IO) and output-oriented (OO) measures of technical inefficiency and talk about the estimation of production functions with IO technical inefficiency. We discuss implications of the IO and OO measures from both the primal and dual perspectives. Second, the latent class (finite mixing) modeling approach is extended to accommodate behavioral heterogeneity. Specifically we consider profit (revenue) maximizing and cost minimizing behaviors with technical inefficiency. In our mixing/latent class model first we consider a system approach in which some producers maximize profit while others simply minimize cost, and then we use a distance function approach and mix the input and output distance functions (in which it is assumed, at least implicitly, that some producers maximize revenue while others minimize cost). In the distance function approach the behavioral assumptions are not explicitly taken into account. The prior probability in favor of profit (revenue) maximizing behavior is assumed to depend on some exogenous variables. Third, we consider stochastic frontier (SF) models that are estimated using local maximum likelihood (LML) method to address the flexibility issue (functional form, heteroskedasticity and determinants of technical inefficiency).

## 2. The IO and OO controversy

The technology (with or without inefficiency) can be looked at from either a primal or a dual perspective. In a primal setup two measures of technical efficiency are mostly used in the efficiency literature. These are: (i) input-oriented (IO) technical inefficiency, (ii) output oriented (OO) technical inefficiency.[1] There are some basic differences between the IO and OO models so far as features of the technology are concerned. Although some of these differences and their implications are well known no one has estimated a stochastic production frontier model econometrically with IO technical

[1] Another measure is hyperbolic technical inefficiency that combines both the IO and OO measures in a special way (see, e.g., Fare et al. (1985), Cuesta and Zofio (1999), Orea, Roibás and Wall (2003)). This measure is not as popular as the other two.

inefficiency using cross-sectional data.[2] Here we consider estimation of a translog production model with IO technical inefficiency.

### 2.1 The IO and OO Models

Consider a single output production technology where Y is a scalar output and *X* is a vector of inputs. Then the production technology with the IO measure of technical inefficiency can be expressed as

$$Y_i = f(X_i \cdot \Theta_i),\ i = 1, \ldots, n, \tag{1}$$

where $Y_i$ is a scalar output, $\Theta_i \le 1$ is IO efficiency (a scalar), $X_i$ is the $J \times 1$ vector of inputs, and *i* indexes firms. The IO technical inefficiency for firm *i* is defined as $\ln \Theta_i \le 0$ and is interpreted as the rate at which all the inputs can be reduced without reducing output. On the other hand, the technology with the OO measure of technical inefficiency is specified as

$$Y_i = f(X_i) \cdot \Lambda_i \tag{2}$$

where $\Lambda_i \le 1$ represents OO efficiency ( a scalar), and $\ln \Lambda_i \le 0$ is defined as OO technical inefficiency. It shows the percent by which actual output could be increased without increasing inputs.

It is clear from (1) and (2) that if $f(\cdot)$ is homogeneous of degree *r* then $\Theta_i^{\ r} = \Lambda_i$ that is independent of *X* and *Y*. If homogeneity is not present their relationship will depend on the input quantities and the parametric form of $f(\cdot)$.

We now show the IO and OO measures of technical efficiency graphically. The observed production plan (Y, X) is indicated by the point A. The vertical length AB measures OO technical inefficiency, while the horizontal distance AC measures IO technical inefficiency. Since the former measures percentage loss of output while the latter measures percentage increase in input usage in moving to the production frontier starting from the inefficient production plan indicated by point A, these two measures are, in general, not directly comparable. If the production function is homogeneous then

[2] On the contrary the OO model has been estimated by many authors using DEA (see, eg, Ray (2003) and references cited in there).

one measure is a constant multiple of the other, and they are the same if the degree of homogeneity is one.

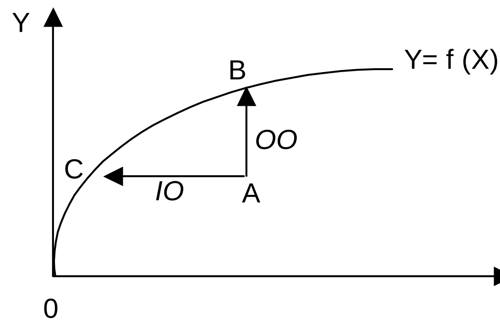



**Figure 1: IO and OO technical inefficiency**

In the more general case, they are related in the following manner: $f(X)\cdot\Lambda = f(X\,\Theta)$. Although we consider technologies with a single output, the IO and OO inefficiency can be discussed in the context of multiple output technologies as well.

**2.2 Economic implications of the IO and OO models**

Here we ask two questions. First, does it matter whether one uses the IO or the OO representation so far as estimation of the technology is concerned? That means, whether features of the estimated technology such as elasticities, returns to scale, etc., are invariant to the choice of efficiency orientation. Second, are efficiency rankings of firms invariant to the choice of efficiency orientation? That is, does one get the same efficiency measures (converted in terms of either output loss or increase in costs) in both cases? It is not possible to provide general theoretical answers to these questions. These are clearly empirical issues so it is necessary to engage in applied research to get a feel for the similarities and differences of the two approaches.Answers to these questions depend on the form of the production technology. If it is homogeneous then there is no difference between these two models econometrically. This is because for a homogeneous function $r\ln\Theta_i \;= \ln\Lambda_i$ where *r* is the degree of homogeneity. Thus, rankings of firms with respect to $\ln\Lambda_i$ and $\ln\Theta_i$ will be exactly the same (one being a constant multiple of the other). Moreover, since $f(X) \cdot \Lambda = f(X)\Theta^r$, the input elasticities as well as returns to scale measures based on these two specifications of the technology will be the same.[3]

This is, however, not the case if the technology is non-homogenous. In the OO model the elasticities and returns to scale will be independent of the technical inefficiency because technical efficiency (that is assumed to be independent of inputs) enters multiplicatively into the production function. This is not true for the IO model, where technical inefficiency enters multiplicatively with the inputs. This will be shown explicitly later for a non-homogeneous translog production function.

### 2.3 Econometric modeling and efficiency measurement

Using the lower case letters to indicate the log of a variable, and assuming that $f(.)$ has a translog form the IO model can be expressed as:

$$\begin{aligned} y_i = \beta_0 + (x_i - \theta_i 1_J)'\beta + \tfrac{1}{2}(x_i - \theta_i 1_J)'\Gamma(x_i - \theta_i 1_J) + \beta_T T_i + \tfrac{1}{2}\beta_{TT} T_i^2 \\ + T_i(x_i - \theta_i 1_J)'\varphi + v_i, \quad \mathrm{i} = 1,\ldots,\mathrm{n}, \end{aligned} \tag{3}$$

where $y_i$ is the log of output, $1_J$ denotes the $J \times 1$ vector of ones, $x_i$ is the $J \times 1$ vector of inputs in log terms, $T_i$ is the trend/shift variable, $\beta_0$, $\beta_T$ and $\beta_{TT}$ are scalar parameters, $\beta$, $\varphi$ are $J \times 1$ parameter vectors, $\Gamma$ is a $J \times J$ symmetric matrix containing parameters, and $v_i$ is the noise term. To make $\theta$ non-negative we defined it as $-\ln\Theta = \theta$.

We rewrite the IO model above as :

$$y_i = \left(\beta_0 + x_i'\beta + \frac{1}{2}x_i'\Gamma x_i + \beta_T T_i + \frac{1}{2}\beta_{TT} T_i^2 + x_i'\varphi T_i\right) - g(\theta_i, x_i) + v_i, \; i = 1, . n, \tag{4}$$

where $g(\theta_i, x_i) = -[\frac{1}{2}\theta_i^2\Psi - \theta\,\Xi_i]$, $\Psi = 1_J'\Gamma 1_J$, and $\Xi_i = 1_J'(\beta + \Gamma x_i + \varphi T_i)$, $i = 1, . n$. Note that if the production function is homogeneous of degree *r*, then $\Gamma 1_J = 0, 1_J'\beta = r,$ and $1_J'\varphi = 0$. In such a case the $g(\theta_i, x_i)$ function becomes a constant multiple of $\theta$, (viz., $[\frac{1}{2}\theta_i^2\Psi - \theta\,\Xi_i] = -r\theta_i$), and consequently, the IO model

[3] Alvarez, Arias and Kumbhakar (2003) addressed these issues in a panel data framework with time invariant technical inefficiency (using fixed effects models).

cannot be distinguished from the OO model. The $g(\theta_i, x_i)$ function shows the percent by which output is lost due to technical inefficiency. For a well-behaved production function $g(\theta_i, x_i) \geq 0$ for each *i*.

The OO model, on the other hand, takes a much simpler form, viz.,

$$y_i = \left(\beta_0 + x_i'\beta + \tfrac{1}{2} x_i'\Gamma x_i + \beta_T T_i + \tfrac{1}{2}\beta_{TT} T_i^2 + x_i'\varphi T_i\right) - \lambda_i + v_i, \; i = 1, . n, \tag{5}$$

where we defined $-\ln \Lambda = \lambda$ to make it non-negative.[4] The OO model in this form is the one introduced by Aigner et al. (1977), Meeusen and van den Broeck (1977) and since then it has been used extensively in the efficiency literature. Since the IO model is never estimated in a primal framework, especially when θ is random, we now focus our attention on it.[5]

We write (5) more compactly as

$$y_i = z_i'\alpha + \tfrac{1}{2}\theta_i^2 \Psi - \theta\, \Xi_i + v_i, \; i = 1, . n. \tag{6}$$

Both $\Psi$ and $\Xi_i$ are functions of the original parameters, and $\Xi_i$ also depends on the data ( $x_i$ and $T_i$).

Under the assumption that $v_i \sim IN(0, \sigma^2)$, and $\theta_i$ is distributed independently of $v_i$ with the density function $f(\theta_i; \omega)$ where $\omega$ is a parameter, the probability density function of $y_i$ can be expressed as

$$f(y_i; \mu) = (2\pi\sigma^2)^{-1/2} \int_0^{\infty} \exp\left[-\frac{(y_i - z_i'\alpha - \frac{1}{2}\theta_i^2\Psi + \theta_i \Xi_i)^2}{2\sigma^2}\right] f(\theta_i; \omega) d\theta_i, \; i = 1, . n, \tag{7}$$

where $\mu$ denotes the entire parameter vector.

[4] The above equation gives the IO model (when the production function is homogeneous) by labeling $\lambda_i = r\theta_i$ .

[5] Alvarez, Arias and Kumbhakar (2003) estimated an IO primal model in a panel data model where technical inefficiency is assumed to be fixed and parametric.

We consider a half-normal and an exponential specification for the density $f(\theta_i;\omega)$, namely,

$$f(\theta_i;\omega) = (\pi\omega^2/2)^{-1/2}\exp\left(-\frac{\theta_i^2}{2\omega^2}\right),\ \theta_i \geq 0, \tag{8}$$

and,

$$f(\theta_i;\omega) = \omega\exp(-\omega\theta_i),\ \theta_i \geq 0. \tag{9}$$

The likelihood function of the model is then

$$L(\mu; y, X) = \prod_{i=1}^{n} f(y_i;\mu),$$

where $f(y_i;\mu)$ has been defined above. Since the integral defining $f(y_i;\mu)$ is not available in closed form we cannot find an analytical expression for the likelihood function. However, we can approximate the integrals using a simulation as follows. Suppose $\theta_{i,(s)}, s=1,..S.$ is a random sample from $f(\theta_i;\omega)$. Then it is clear that

$$f(y_i;\mu) \approx \tilde{f}(y_i;\mu) \equiv S^{-1}\sum_{s=1}^{S}\exp\left[-\frac{(y_i - z_i'\alpha - \frac{1}{2}\theta_{i,(s)}^2\Psi + \theta_{i,(s)}\Xi_i)^2}{2\sigma^2}\right], \tag{10}$$

and an approximation of the log-likelihood function is given by

$$\log L \approx \sum_{i=1}^{n}\log\tilde{f}(y_i;\mu), \tag{11}$$

which can be maximized by numerical optimization procedures to obtain the ML estimator. For the distributions we adopted, random number generation is trivial so implementing the SML estimator is straightforward.[6]

Inefficiency estimation is accomplished by considering the distribution of $\theta_i$ conditional on the data and estimated parameters

$$f(\theta_i \mid \tilde{\mu}, D_i) \propto \exp\left[-\frac{(y_i - z_i'\tilde{\alpha} - \frac{1}{2}\theta_i^2\tilde{\Psi} + \tilde{\Xi}_i\theta_i)^2}{2\tilde{\sigma}^2}\right] f(\theta_i;\tilde{\omega}),\ i=1,..n, \tag{12}$$

[6] Greene (2003) used SML for the OO normal-gamma model.

where a tilde denotes the ML estimate, and $D_i = [x_i, T_i]$ denotes the data. For example, when $f(\theta_i; \tilde{\omega})$ is half-normal we get

$$f(\theta_i \mid \tilde{\mu}, y, X) \propto \exp\left[ -\frac{(y_i - z_i'\tilde{\alpha} - \frac{1}{2}\theta_i^2\tilde{\Psi} + \theta_i \Xi_i)^2}{2\tilde{\sigma}^2} - \frac{\theta_i^2}{2\tilde{\omega}^2} \right], \ \theta_i \geq 0, \ i = 1, . n, \quad (13)$$

This is not a known density, and even the normalizing constant cannot be obtained in closed form. However, the first two moments and the normalizing constant can be obtained by numerical integration, for example, using Simpson's rule.

To make inferences on *efficiency*, define efficiency as $r_i = \exp(-\theta_i)$ and obtain the distribution of $r_i$ and its moments by changing the variable from $\theta_i$ to $r_i$. This yields

$$f_r(r_i \mid \tilde{\mu}, D_i) = r_i^{-1} f(-\ln r_i \mid \tilde{\mu}, y, X), \ 0 < r_i \leq 1, \ i = 1, . n. \quad (14)$$

We used *Win GAUSS* 3.2.38 to perform the computations.[7] Numerical optimization is performed with a conjugate gradients algorithm using 1000 pseudo-random draws per evaluation of each one of the $n$ components of the log-likelihood function. Numerical integration for efficiency measurement is implemented, again, using Simpson's rule. We define a fixed sequence of $r_i$'s ranging from 0.1 to 1, and use 1,000 points in between to approximate the normalizing constant and the first two moments of the conditional distribution $f_r(r_i \mid \tilde{\mu}, D_i)$.

The likelihood function for the OO model is given in Aigner, Lovell and Schmidt (1977) (hereafter ALS).[8] The Maximum likelihood method for estimating the parameters of the production function in the OO models are straightforward and have been used extensively in the literature starting from ALS.[9] Once the parameters are estimated, technical inefficiency (λ) is estimated from $E(\lambda \mid (v - \lambda))$ -- the Jondrow *et al*. (1982)

---

7 The GAUSS programs are available from the authors upon request.

8 See also Kumbhakar and Lovell (2000, pp. 74-82) for the log-likelihood functions under both half-normal and exponential distributions for the OO technical inefficiency term.

9 It is not necessary to use the simulated ML method to estimate the parameters of the frontier models if the technical inefficiency component is distributed as half-normal, truncated normal or exponential along with the normality assumption on the noise component. For other distributions, for example, gamma for technical inefficiency and normal for the noise component the standard ML method may not be ideal (see Greene (2003) who used the simulated ML method to estimate OO technical efficiency in the gamma-normal model).

formula. Alternatively, one can estimate technical efficiency from $E(e^{-\lambda} \mid (v-\lambda))$ using the Battese-Coelli (1988) formula.

**2.4 Application to Spanish dairy farms**

The empirical analysis is based on a balanced panel data set of 80 Spanish dairy farms for the years 1996 to 1998. These are all small family farms. We consider one output (liters of milk) and four variable inputs (viz., number of cows, kilograms of concentrates, hectares of land and labor (measured in man-equivalent units)). Since there are only 80 farms in the data in our cross-sectional model we use all 240 observations without using any panel feature of the data. In this approach each observation is treated as a separate farm. However, we used time as an additional regressor to capture technical change. Thus, our IO model is the one specified in equation (4) in which the *x* variables are the four inputs, and *T* is the time trend variable. The OO model is given in (5).

Here we mainly focus on estimates of technical efficiency (TE). Note that TE in the OO model is $\Lambda \equiv e^{-\lambda} \leq 1$, where $\lambda$ is OO technical inefficiency (that measures output shortfall/loss, given inputs). On the other hand TE in the OO model is $\Theta \equiv e^{-\theta} \leq 1$ where $\theta$, the IO technical inefficiency, measures excess input-usage (given the output level) due to technical inefficiency. We report (in Figure 1a) TE measures from the IO model with half-normal and exponential distributions on IO technical inefficiency, $\theta$. It can be seen that the TE distributions are quite similar, although the half-normal distribution generates a slightly tighter TE distribution. Average TE is found to be around 92%.

We noted before that TE in the IO and OO models (i.e., $\Lambda$ and $\Theta$) are not the same unless the returns to scale is unity (in which case the percentage decrease in inputs and output are the same). To make these measures comparable, we convert the IO measure in terms of the OO measure (i.e., a computed OO technical efficiency labeled as TE_OO(CO)) using the formula TE_OO(CO) = exp(– $g(\theta, x, T)$), where $g(\theta, x, T)$ is defined beneath equation (4). By doing so we can compare it directly with TE_OO because both the measures are ratios of actual to frontier output. The distributions of these two efficiency measures (each for half normal and exponential distributions) are plotted in Figure 1b. It can be seen from these figures (kernel density plots) that efficiency distributions generated from the IO and OO models (especially with half normal distributions) are not the same. Their means, variances, skewness and kurtosis are all different.[10]

### 2.5 Open questions

- How to decide which specification is right/appropriate for the data at hand?

In a sense this is not a question that the data can help to resolve because the choice of orientation is made by the econometrician. Since the I-O model is heteroskedastic while the O-O model is homoskedastic, one may think of econometrically testing for heteroscedasticity as a way out. But the heteroskedasticity is of a very special form and under the null of homoskedasticity the IO model not only reduces to an IO model but the production function also becomes homogeneous.

- Can one generalize the specification to obtain the IO, OO and the in-between cases?

To elaborate this, let's start from the model that combines both the IO and OO technical inefficiencies, viz.,

$$\begin{aligned} y_i &= \beta_0 + (x_i - \theta_i 1_J)'\beta + \tfrac{1}{2}(x_i - \theta_i 1_J)'\Gamma(x_i - \theta_i 1_J) + \beta_T T_i + \tfrac{1}{2}\beta_{TT} T_i^2 \\ &\quad + T_i (x_i - \theta_i 1_J)'\varphi - \lambda_i + v_i, \quad \mathrm{i} = 1,\ldots,\mathrm{n}, \end{aligned} \tag{15}$$

and sets values of $\theta_i \in (0,\infty)$ then the resulting models will be all OO models. One can then estimate all these OO models to obtain the corresponding $\lambda_i$. This technique will, however, not help to determine the relationship between $\theta_i$ and $\lambda_i$ empirically because the estimated technology changes with different values of $\theta_i$.

[10] Orea, Roibás and Wall (2003) used a model selection procedure (Vuong test, J test, and JA test) to select the best model in a panel data framework when technical inefficiency is assumed to be time-invariant (fixed parameters).

### 2.6 The IO and OO controversy and the Distance function approach

Instead of using a production function approach one can use distance functions to estimate technical inefficiency. Two types of distance functions are used for this.

The output distance function is defined by

$$\mathrm{D_O}(X, Y) = \inf\{\varphi > 0 : (X, Y/\varphi) \in \mathrm{T}(X, Y)\}$$

where the technology is defined by $\mathrm{T}(X,Y) = \{(X,Y) : X \text{ can produce } Y\}$.

The output distance function ranges from $0 < D_O(X, Y) \leq 1$. $D_O(X,Y)$ is homogeneous of degree one in outputs, non–decreasing in outputs and increasing in inputs.

Using one of the outputs as the numeraire to impose linear homogeneity, the stochastic output distance function (in logarithms) can be written as

$$-\ln Y_1 = \ln D_O(X, Y/Y_1) + u + v \qquad (16)$$

where $u = -\ln D_O \geq 0$ and v is the stochastic noise component. The argument $Y/Y_1$ = $(Y_2/Y_1, \ldots, Y_M/Y_1)$ when there are $M$ outputs. Thus the above formulation is not different (qualitatively) from a stochastic production function model (after multiplying both sides by – 1). The only advantage is that one can accommodate multiple outputs.

The input distance function is defined as

$$\mathrm{D_I}(X,Y) = \sup\{\phi > 0 : (X/\phi, Y) \in \mathrm{T}(X, Y)\}.$$

It ranges from $1 \leq D_I(X, Y) \leq +\infty$, and is linearly homogeneous of degree one and non-decreasing in inputs, and increasing in outputs.

Using the linear homogeneity property the stochastic input distance function can be expressed as

$$-\ln X_1 = \ln D_I(X/X_1, Y) - u + v \qquad (17)$$

where $u = \ln D_I \geq 0$ and $v$ is the stochastic noise component. The argument $X/X_1$ = $(X_2/X_1, \ldots, X_J/X_1)$ when there are $J$ outputs. By labeling inputs as outputs and vice-versa (and multiplying both sides by –1) the above model can be viewed as a SF model (with a change in the sign on the one-sided term *u*).

Since both the input and output distance functions are similar to the OO SF functions, from the estimation point of view there is a clear advantage. However, the main concern is: Can the technology be invariant (theoretically and empirically) to the choice of input and output distance functions? Alternatively, how can one decide whether the input or the output distance function is appropriate for the data at hand? Since the left hand side variable is not the same in the input and output distance functions, the estimated technology is likely to differ even if there is constant (or even unitary) returns to scale. (Note: regressing Y on X versus X on Y in the standard regression framework).

**Other important issues:**

- Is there any behavioral assumption behind the input/output distance function?
- What is the relationship between inefficiency in the $D_I$ and $D_O$ models? How to recover one from the other both theoretically and empirically?

The input and output distance functions can be obtained from a multiple-input, multiple-output production function, $F(X, Y) = \alpha$, using the same homogeneity restrictions that are used with the distance function to obtain the input and output distance functions. However, the homogeneity restrictions in the production function do not follow from the theory.

## 2.7 Looking through the dual cost functions

### *2.7.1 The IO approach*

We now examine the IO and OO models when behavioral assumptions are explicitly introduced. First, we examine the models when producers minimize cost to produce the given level of output(s). The objective of a producer is to

$$\text{Min. } w'X \text{ subject to } Y = f(X \cdot \Theta)$$

from which conditional input demand functions can be derived. The corresponding cost function can then be expressed as

$$w'X = C^a = C(w, Y)/\Theta\,,$$

where $C(w,Y)$ is the minimum cost function (cost frontier) and $C^a$ is the actual cost. Finally, one can use Shephard's lemma to obtain $X_j^a = X_j^*(w,Y)/\Theta \geq X_j^*(w,Y) \ \forall \ j,$ where the superscripts *a* and * indicate actual and cost minimizing levels of input $X_j$.

Thus, the IO model implies (i) a neutral shift in the cost function which in turn implies that RTS and input elasticities are unchanged due to technical inefficiency, (ii) a equi-proportional increase (at the rate given by $\theta$) in the use of all inputs due to technical inefficiency, irrespective of the output level and input prices.

**Comments:** Result (i) is just the opposite of what we obtained in the primal case (see Arias and Alvarez, 1998). Result (ii) states that when inefficiency is reduced firms will move horizontally to the frontier (as expected by the IO model).

***2.7.2 The OO Model:***

Here the objective function is written as

$$\text{Min. } w'X \text{ subject to } \ Y = f(X) \cdot \Lambda$$

from which conditional input demand functions can be derived. The corresponding cost function can then be expressed as

$$w'X = C^a = C(w, Y/\Lambda) \equiv C(w,Y) \cdot q(w,Y,\Lambda)\,, \tag{18}$$

where as before $C(w,Y)$ is the minimum cost function (cost frontier) and $C^a$ is the actual cost. Finally, $q(.) = C(w,Y/\Lambda)/C(w,Y) \geq 1$. One can then use Shephard's lemma to obtain

$$X_j^a = X_j^*(w,Y)[q(.) + \{C(w,Y)/X_j^*\}\{\partial q(.)/\partial w_j\}] \geq X_j^*(w,Y) \ \forall \ j, \tag{19}$$

where the last inequality will hold if the cost function is well behaved. Note that $X_j^a \neq X_j^*(w,Y) \ \forall \ j$ unless $q(.)$ is a constant.

Thus, the results from the OO model are just the opposite from those of the IO model. Here (i) inefficiency shifts the cost function non-neutrally (meaning that $q(.)$ depends on output and input prices as well as $\Lambda$; (ii) increases in input use are not equi-proportional (depends on output and input prices); (iii) the cost shares are **not** independent of technical inefficiency, (iv) the model is harder to estimate (similar to the IO model in the primal case).[11]

More importantly, the result in (i) is just the opposite of what we reported in the primal case. Result (ii) is not what the OO model predicts (increase in output) when inefficiency is eliminated. Since output is exogenously given in a cost-minimizing framework, input use has to be reduced when inefficiency is eliminated.

Conclusion: The results from the dual cost function models are just the opposite of what the primal models predict. Since the estimated technologies using cost functions are different in the IO and OO models, as in the primal case, we do not repeat the results based on the production/distance functions results here.

**2.8 Looking through the dual profit functions**

***2.8.1 The IO Model***

Here we assume that the objective of a producer is to

$$\text{Max.}\,\pi = p \cdot Y - w'X \equiv p \cdot Y - (w/\Theta)'X \cdot \Theta \text{ subject to } Y = f(X \cdot \Theta)$$

from which unconditional input demand and supply functions can be derived. Since the above problem reduces to a standard neo-classical profit maximizing problem when *X* is replaced by $X \cdot \Theta$ and *w* is replaced by $w/\Theta$, the corresponding profit function can be expressed as

$$\pi^a = p \cdot Y - (w/\Theta)'X \cdot \Theta = \pi(w/\Theta, p) \equiv \pi(w, p) \cdot h(w, p, \Theta) \leq \pi(w, p), \qquad (20)$$

where $\pi^a$ is actual profit, $\pi(w, p)$ is the profit frontier (homogeneous of degree one in *w* and *p*) and $h(w, p, \Theta) = \pi(w/\Theta, p)/\pi(w, p) \leq 1$ is profit inefficiency. Note that the $h(w, p, \Theta)$ function depends on *w*, *p* and $\Theta$ in general. Application of Hotelling's lemma yields the following expressions for the output supply and input demand functions:

[11] Atkinson and Cornwell (1994) estimated translog cost functions with both input and output oriented technical inefficiency using panel data. They assumed technical inefficiency to be fixed and time-invariant. See also Orea, Roibás and Wall (2003).

$$Y^a = Y^*(w,p)[h(.) + (\pi(w,p)/Y^*)(\partial h(.)/\partial p)] \le Y^*(w,p) \quad ,$$
$$X_j^a = X_j^*(w,p)[h(.) - (\pi(w,p)/X_j^*)(\partial h(.)/\partial w_j)] \le X_j^*(w,p) \;\; \forall\, j, \qquad (21)$$

where the superscripts *a* and * indicate actual and optimum levels of output *Y* and inputs $X_j$. The last inequality in the above equations will hold if the underlying production technology is well behaved.

***2.8.2 The OO Model***

Here the objective function can be written as

$$\text{Max.}\, \pi = p \cdot Y - w'X \equiv p \cdot \Lambda \cdot Y / \Lambda - w' X \cdot \Theta \text{ subject to } Y = f(X) \cdot \Lambda$$

which can be viewed as a standard neo-classical profit maximizing problem when *Y* is replaced by $Y/\Lambda$ and *p* is replaced by $p \cdot \Lambda$, the corresponding profit function can be expressed as

$$\pi^a = p \cdot \Lambda \cdot Y / \Lambda - w' X = \pi(w, p \cdot \Lambda) \equiv \pi(w,p) \cdot g(w,p,\Lambda) \le \pi(w,p), \qquad (22)$$

where $g(w,p,\Lambda) = \pi(w, p \cdot \Lambda)/\pi(w,p) \le 1$. Similar to the IO model using Hotelling's lemma we get

$$Y^a = Y^*(w,p)[g(.) + (\pi(w,p)/Y^*)(\partial g(.)/\partial p)] \le Y^*(w,p) \quad ,$$
$$X_j^a = X_j^*(w,p)[g(.) - (\pi(w,p)/X_j^*)(\partial g(.)/\partial w_j)] \le X_j^*(w,p) \;\; \forall\, j. \qquad (23)$$

The last inequality in the above equations will hold if the underlying production technology is well behaved.

Summary of the basic results: (i) A shift in the profit functions for both the IO and OO models is non-neutral. Therefore, estimated elasticities, RTS, etc., are affected by the presence of technical inefficiency, no matter what form is used. (ii) Technical inefficiency leads to a decrease in the production of output and decreases in input use in both models, however, prediction of the reduction in input use and production of output are not the same under both the models.

Conclusions: Even under profit maximization that recognizes endogeneity of both inputs and output, it matters which model is used to represent the technology!! (Any intuition??) These results are different from those obtained under the primal models and from the cost minimization framework.

**Overall conclusion:** It matters (both theoretically and empirically) whether one uses an input or output oriented measure of technical inefficiency.

## 3. Latent Class Models

### 3.1 Modeling technological heterogeneity

In modeling production technology we almost always assume that all the producers use the same technology. In other words, we don't allow the possibility that there might be more than one technology being used by the producers in the sample. Furthermore, the analyst may not know who is using what technology. Recently, a few studies have combined the stochastic frontier approach with the latent class structure in order to estimate a mixture of several technologies (frontier functions). Greene (2001, 2002) proposes a maximum likelihood for a latent class stochastic frontier with more than two classes. Caudill (2003) introduces an expectation-maximization (EM) algorithm to estimate a mixture of two stochastic cost frontiers with two classes.[12] Orea and Kumbhakar (2003) estimated a four-class stochastic frontier cost function (translog) with time-varying technical inefficiency.

Following the notations of Greene (2001, 2002) we specify the technology for class *j* as

$$\ln y_i = \ln f(x_i, z_i, \beta_j)|_j + v_i|_j - u_i|_j \quad (24)$$

where $u_i|_j$ is a non-negative random term added to the production function to accommodate technical inefficiency.

We assume that the noise term for class *j* follows a normal distribution with mean zero and constant variance, $\sigma_{vj}^2$. The inefficiency term $u_t|_j$ is modeled as a half-normal random variable following standard practice in the frontier literature, viz.,

[12] See, in addition, Beard, Caudill and Gropper (1991, 1997) for applications using a non-frontier approach. For applications in social sciences, see, Hagenaars, J.A. and McCutcheon (2002). Statistical aspects of the mixing models are dicussed in details in Mclachlan and Peel (2000)

$$u_i\,|_j \sim N_+(0,\omega_j^2) \quad u_i\,|_j \geq 0 \tag{25}$$

i.e., a half-normal distribution with scale parameter $\omega_j$ for each class.

With these distributional assumptions, the likelihood for firm *i*, if it belongs to class *j*, can be written as (Kumbhakar and Lovell (2000))

$$l(i\,|\,j) = \frac{2}{\sigma_j}\phi\left\{\frac{\varepsilon(i\,|\,j)}{\sigma_j}\right\}\Phi\left\{-\frac{\lambda_j\,\varepsilon(i\,|\,j)}{\sigma_j}\right\} \tag{26}$$

where $\sigma_j^2 = \omega_j^2 + \sigma_j^2$, $\lambda_j = \omega_j / \sigma_{vj}$ and $\varepsilon(i\,|\,j) = \ln y_i - \ln f(x_i, z_i, \beta_j)\,|_j$. Finally, $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a standard normal variable.

The unconditional likelihood for firm *i* is obtained as the weighted sum of their *j*-class likelihood functions, where the weights are the prior probabilities of class membership. That is,

$$l(i) = \sum_{j=1}^{J} l(i\,|\,j)\cdot\pi_{ij} \quad , \quad 0 \leq \pi_{ij} \leq 1 \quad , \quad \sum\nolimits_j \pi_{ijt} = 1 \tag{27}$$

where  the class probabilities can be parameterized by, for example, a logistic function. Finally, the log likelihood function is:

$$\ln L = \sum_{i=1}^{n} \ln l(i) = \sum_{i=1}^{n} \ln\left\{\sum_{j=1}^{J} l(i\,|\,j)\cdot\pi_{ij}\right\} \tag{28}$$

The estimated parameters can be used to compute the conditional posterior class probabilites. Using Bayes' theorem (see Greene (2001, 2002) and Orea and Kumbhakar (2003)) the posterior class probabilities can be obtained from

$$P(j\,|\,i) = \frac{l(i\,|\,j)\cdot\pi_{ij}}{\sum_{j=1}^{J} l(i\,|\,j)\cdot\pi_{ij}} \tag{29}$$

This expression shows that the posterior class probabilities depend not only on the estimated parameters in $\pi_{ij}$, but also on parameters of the production frontier and the data. This means that a latent class model classifies the sample into several groups even when the $\pi_{ij}$ are fixed parameters (independent of *i*).

In the standard stochastic frontier approach where the frontier function is the same for every firm, we estimate inefficiency relative to the frontier for all observations, viz, inefficiency from $E(u_i \mid \varepsilon_i)$ and efficiency from $E[exp(-u_i)|\varepsilon_i]$. In the present case, we estimate as many frontiers as the number of classes. So the question is how to measure the efficiency level of an individual firm when there is no unique technology against which inefficiency is to be computed. This is solved by using the following method,

$$\ln EF_i = \sum_{j=1}^{J} P(j \mid i) \cdot \ln EF_i(j) \qquad (30)$$

where $P(j|i)$ is the posterior probability to be in the $j^{th}$ class for a given firm $i$ (defined in (29)), and $EF_i(j)$ is its efficiency using the technology of class $j$ as the reference technology. Note that here we don't have a single reference technology. It takes into account technologies from every class. The efficiency results obtained by using (30) would be different from those based on the most likely frontier and using it as the reference technology. The magnitude of the difference depends on the relative importance of the posterior probability of the most likely cost frontier, the higher the posterior probability the smaller the differences.

### 3.2 Modeling behavioral heterogeneity: Profit Maximization and Cost Minimization

Cost minimization and profit maximization are the two most widely used behavioral assumptions in the theory of firm. Most often researchers favor using a cost minimizing behavior without much justification for it from either theoretical or empirical viewpoints. One may formally test whether the producers in the given sample are profit maximizers or not, for example, following the methodology developed by Schankerman and Nadiri (1986), Based on the test results, one will be using either a cost or a profit function formulation. This implictly assumes that all producers in the sample behave in the same way. In reality, firms in a particular industry, although using the same technology, may differ in terms of their behavior. For example, some producers might find it costly to adjust output to the profit maximizing level due to high adjustment costs, while for others it might be optimal to maximize profit. In such a case, the estimated technology by imposing profit maximization behavior on producers who are not maximizing profit maxizers and vice versa, is likely to be biased. Consequently, estimates of returns to scale, elasticities, technical change, *etc*., will be wrong.

If one knows which producers are maximizing profit and which are not, then one can split the sample into two classes. A profit function approach is estimated using the sample observations in the first class, and a cost function is used for the producers in the second class. This procedure is not *efficient* because the above approach doesn't take into account the fact that the underlying technology is exactly the same for all producers. The other practical problem is that no one knows in advance, which producers are profit maximizers and which are not. Consequently, this approach cannot be used in practice.

The advantage of the LCM is that it is not necessary to impose *a priori* criterion to identify which producers are in what class. Moreover, we can formally examine whether some exogenous factors are responsible for the presence or absence of profit maximizing (cost minimizing) behavior by making the probabilities functions of exogenous variables. When panel data is available, we do not need to assume that producers behave like profit maximizers all the time, so we can accommodate switching behavior, and determine when they behaved like profit maximizers and when not.

***3.2.1 The cost system***

First, we consider the cost-share system with technical inefficiency that is represented by the following system of equations.

$$\begin{aligned} \ln C_i &= \ln C(\ln p_i, \ln y_i) + v_{1i} + u_i \\ S_{1i} &= S_1(\ln p_i, \ln y_i) + v_{2i} \\ &\vdots \\ S_{M-1,i} &= S_{M-1}(\ln p_i, \ln y_i) + v_{Mi}, \end{aligned} \tag{31}$$

where $\ln C_i$ is the log of expenditure, $S_{1i}, \ldots S_{M-1,i}$ denote the $M-1$ cost shares[13], $p_i$ is the $M \times 1$ vector of input prices, $y_i$ is the $Q \times 1$ vector of outputs, and $v_i = [v_{1i}, \ldots v_{Mi}]'$

[13] One cost share is dropped to avoid the singularity problem.

represents the noise terms. The subscript $i$ ($i=1,\ldots N.$) indicates producers/firms and $u_i \geq 0$ represents technical inefficiency. For the ML method we make the following distributional assumptions: $v_i \sim IN_M(0_M, \Omega)$ where $\Omega$ is the $M \times M$ covariance matrix, and $u_i \sim N_+(0, \omega^2)$ distributed independently of $v_i$. Let $\Omega^{-1} = \begin{bmatrix} \omega^{11} & \phi' \\ \phi & \Omega^{22} \end{bmatrix}$, where $\omega^{11}$ is a scalar, $\phi$ is $(M-1)\times 1$, and $\Omega^{22}$ is an $(M-1)\times(M-1)$ matrix. The joint density of the endogenous variables ($Z_i = [\ln \hat{C}_i, S_{1i}, \ldots S_{M-1,i}]'$) has been presented before only for the special case $\phi = 0_{M-1}$ (Kumbhakar and Lovell, 2000, pages 150 and 156). Here, we take up the more general case.

The joint density of the cost system in the present case can be written as

$$f_{Z_i}(Z_i) = 2^{1/2}(2\pi)^{-M/2}(1+\omega^{11}\omega^2)^{-1/2}\left(\det\Omega^{-1}\right)^{1/2}\exp\left(-\frac{Q_i-\mu_{*i}^2}{2\sigma_*^2}\right)\Phi(\mu_{*i}/\sigma_*) \quad , \quad (32)$$

where

$$\sigma_*^2 = \frac{\omega^2}{1+\omega^{11}\omega^2}\,,\ \mu_{*i} = \sigma_*^2(\omega^{11}R_{1i} + R'_{*i}\phi)\,,\ Q_i = \sigma_*^2(\omega^{11}R_{1i}^2 + 2R_{1i}R'_{*i}\phi + R'_{*i}\Omega^{22}R_{*i})\,,$$

$$R_i = \begin{bmatrix} \ln C_i - \ln C(\ln p_i, \ln y_i) \\ S_{1i} - S_1(\ln p_i, \ln y_i) \\ \vdots \\ S_{M-1,i} - S_{M-1}(\ln p_i, \ln y_i) \end{bmatrix} = \begin{bmatrix} R_{1i} \\ R_{*i} \end{bmatrix},\ i=1,\ldots N.,$$

and $R_{1i}$ is a scalar (cost function "residual"), $R_{*i}$ is a $(M-1)\times 1$ vector of "cost share residuals", and $\Phi$ denotes the cumulative distribution function of a standard normal variable. The log-likelihood function can be maximized to obtain the ML parameter estimates.

It can be shown that the conditional distribution of $u_i$ given $R_i$ is truncated normal, i.e., $u_i \mid R_i \sim N_+(\mu_{*i}, \sigma_*^2)$. This result can then be used to obtain a firm-specific measure of technical inefficiency, for example, the mean of the conditional distribution of $u_i$, $\tilde{u}_i = \mu_{*i} + \sigma_*\varphi(\mu_{*i}/\sigma_*)/\Phi(\mu_{*i}/\sigma_*)$, where $\varphi$ denotes the standard normal *pdf*.

### *3.2.2 The profit system*

Now we consider the profit system represented by the following system of equations

$$\begin{aligned}
\ln C_i &= \ln C(\ln p_i, \ln y_i) + v_{1i} + u_i \\
S_{1i} &= S_1(\ln p_i, \ln y_i) + v_{2i} \\
&\vdots \\
S_{M-1,i} &= S_{M-1}(\ln p_i, \ln y_i) + v_{Mi},
\end{aligned}$$

$$\ln y_{ji} = \ln C_i - \ln q_{ji} + \ln\left(ecy_{ji}(\ln p_i, \ln y_i)\right) + v_{M+j,i}, \quad j = 1,\ldots,Q, \quad i = 1,\ldots,N, \qquad (33)$$

where the last $Q$ equations follow from the Q additional conditions, viz., $q_j = \partial C / \partial y_j$ ($j = 1,\ldots,Q$) where $q_j$ is the price of output $y_j$ and $ecy_{ji}(\ln p_i, \ln y_i) = \frac{\partial \ln C(\ln p_i, \ln y_i)}{\partial \ln y_{ji}}$.

These conditions (first-order conditions for profit maximization) state that output allocation is optimal when output price equals marginal cost.

For the ML method we assume that $v_i = [v_{1i},\ldots,v_{M+Q,i}]' \sim IN_{M+Q}(0,\Sigma)$ and $u_i \sim N_+(0,\omega^2)$, distributed independently of $v_i$. We partition the inverse covariance matrix as $\Sigma^{-1} = \begin{bmatrix} \sigma^{11} & \lambda' \\ \lambda & \Sigma^{22} \end{bmatrix}$, where $\sigma^{11}$ is a scalar, $\lambda$ is $(M+Q-1)\times 1$, and $\Sigma^{22}$ is $(M+Q-1)\times(M+Q-1)$. We utilize the results for the cost-share system to show that the density of endogenous variables ($\Xi_i = [\ln C_i, S_i', y_i']'$) can be written as

$$f_{\Xi_i}(\Xi_i) = 2^{1/2}(2\pi)^{-(M+Q)/2}(1+\sigma^{11}\omega^2)^{-1/2}\left(\det \Sigma^{-1}\right)^{1/2} \exp\left(-\frac{Q_i - \mu_{*i}^2}{2\sigma_*^2}\right)\Phi(\mu_{*i}/\sigma_*)\cdot|J_i|, \quad (34)$$

where $\sigma_*^2 = \frac{\omega^2}{1+\sigma^{11}\omega^2}$, $\mu_{*i} = \sigma_*^2(\sigma^{11}R_{1i} + R_{*i}'\lambda)$, $Q_i = \sigma_*^2(\sigma^{11}R_{1i}^2 + 2R_{1i}R_{*i}'\lambda + R_{*i}'\Sigma^{22}R_{*i})$,

$$R_i = \begin{bmatrix} \ln C_i - \ln C(\ln p_i, \ln y_i) \\ S_{1i} - S_1(\ln p_i, \ln y_i) \\ \vdots \\ S_{M-1,i} - S_{M-1}(\ln p_i, \ln y_i) \\ \ln y_{1i} - \ln C_i + \ln q_{1i} + \ln\left(ecy_{1i}(\ln p_i, \ln y_i)\right) \\ \vdots \\ \ln y_{Qi} - \ln C_i + \ln q_{Qi} + \ln\left(ecy_{Qi}(\ln p_i, \ln y_i)\right) \end{bmatrix} = \begin{bmatrix} R_{1i} \\ R_{*i} \end{bmatrix}, \; i=1, . N..$$

Finally, the last term comes from the Jacobian of the transformation from ($u$+$v_1$, $v_2$, …, $v_{M+Q}$) to $[\ln C_i, S_i', y_i']$, viz.,

$$J_i(.) = \frac{\partial(u_i + v_{1i}, v_{2i}, ..., v_{M+Q,i})}{\partial(\ln C_i, S_i', y_i')} = \begin{bmatrix} 1 & 0'_{M-1} & -ecy_i' \\ 0_{M-1} & I_{M-1} & -\dfrac{\partial S_i}{\partial \Xi_i} \\ -1_Q & 0'_{Q\times(M-1)} & I_Q - \dfrac{\partial ecy_i}{\partial \Xi_i} ecy_i^{-1} \end{bmatrix}$$

where $ecy_i = diag[ecy_{1i},...,ecy_{Qi}]'$, $ecy_i^{-1} = diag[1/ecy_{1i},...,1/ecy_{Qi}]'$, and $1_Q$ is the $Q \times 1$ unit vector.

The log likelihood function based on the above joint density function can be maximized to obtain the ML estimates of all the parameters. The conditional distribution of $u_i$ can then be used to obtain firm-specific estimates of technical inefficiency.

### 3.3 The mixing model

Here we assume that every producer is potentially a profit maximizer (with some probability). The probability of being a profit maximizer is specified by a logistic function that depends on some exogenous variables. This gives us a finite mixture model where the density of endogenous variables is given by

$$f(\Xi_i;\theta) = \pi_i f_{\Xi_i}(\Xi_i;\theta) + (1-\pi_i) f_{Z_i}(Z_i;\theta), \; i=1, . N. \qquad (35)$$

where $\pi_i$ is the probability that the $i$th firm behaves as if it were maximizing profit. Given a set of predetermined variables, $W_i$, we specify $\pi_i$ as

$$\pi_i = \frac{\exp(w_i'\delta)}{1+\exp(w_i'\delta)},\ i=1,\ldots N. \qquad (36)$$

This parameterization guarantees that $\pi_i$ is between zero and one. It also provides a direct interpretation of $\delta$.

Based on (7) we can formulate the log-likelihood function

$$\ln L(\theta,\Sigma^{-1},\Omega^{-1},\delta;\Xi,W) = \sum_{i=1}^{N} \ln\left[\pi_i f_{\Xi_i}(\Xi_i;\theta) + (1-\pi_i) f_{Z_i}(Z_i;\theta)\right] \qquad (37)$$

We can maximize this function to obtain FIML estimates of all parameters.

Straightforward application of Bayes' theorem yields an estimate of the posterior probability that the $i$th firm maximizes profit:

$$\tilde{Q}_i = \frac{\pi_i(\tilde{\delta}) f_{\Xi_i}(\Xi_i;\tilde{\theta})}{\pi_i(\tilde{\delta}) f_{\Xi_i}(\Xi_i;\tilde{\theta}) + (1-\pi_i(\tilde{\delta})) f_{Z_i}(Z_i;\tilde{\theta})},\ i=1,\ldots N., \qquad (38)$$

where the FIML estimates were substituted for $\theta$ and $\delta$, and $\pi_i(\tilde{\delta}) \equiv \frac{\exp(w_i'\tilde{\delta})}{1+\exp(w_i'\tilde{\delta})}$, $i=1,\ldots N.$. These posterior probabilities are firm-specific even when $\pi_i$ is a parameter. Clearly, the estimated posterior probabilities summarize all the evidence for or against profit maximization. Ideally, we would like to have $Q_i$ equal to either zero or one (or nearly so) so that the choice in favor or against profit maximization is more or less clear. Empirically, we cannot always expect that, and $Q_i$ could be anywhere between these limits. In such cases, one could say that a firm is likely to be profit maximizing provided $\tilde{Q}_i > \frac{1}{2}$.

### 3.4 Application to U.S. airlines

To illustrate the technique proposed in the preceding sections, we use an unbalanced panel data set[14] consisting of annual observations on the domestic operations of 23 US airlines over the period 1971-1986. A total of 268 observations are used here. Variable

[14] For details regarding the data, see Appendix A of Baltagi, Griffin and Vadali (1998).

inputs are labor ($L$), materials ($M$) and fuel ($F$). Capital ($K$) it treated as a quasi-fixed factor. To control for firm-heterogeneity, we also include 22 airline dummies in the cost function.

The empirical results for the CM, PM, and LCM models that allow for technical inefficiency derived from the different models and across the two regimes are presented in Figures 3a-3c. These distributions are quite different pointing to the direction that behavioral assumptions are critical for inefficiency measurement. We find some significant differences in inefficiency in the pre- and post-deregulation period in terms of the tail behavior: Before deregulation the probability of relatively high inefficiency is comparatively larger so most airlines seemed to perform better in the post-deregulation period. The most notable difference in inefficiency distributions is between the CM model (Figure 3a) and the PM model (Figure 3b). The former is asymmetric to the right implying mean inefficiency close to 5% while the latter is symmetric and its mean is close to 1.2%. Another difference is in the tail behavior. According to the CM model technical inefficiency is less than 10% almost surely, while values in excess of about 1.5% are quite improbable according to the PM model. Notably, the distribution derived from the LCM (Figure 3c) is much closer to the CM result (Figure 3a) than to the PM result (Figure 3b) despite the fact that most airlines are profit maximizers according to the posterior probability criterion. The reason is that most airlines still have a non-zero probability to be in the non-profit-maximizing regime and this fact must be accounted for in deriving the inefficiency distribution from the LCM.

Prior and posterior probabilities in favor of profit maximization are presented in Figure 4. It is evident that in the pre-deregulation period each airline was equally likely to be a profit maximizer versus a non-profit-maximizer while in the post-deregulation period the probability of failing to maximize profit drops sharply, and about 80% of the observations belong to the PM regime.

### 3.5 Mixing distance functions

While using distance functions we take it for granted that the analyst knows whether to use the input or the output distance function. In doing so, one implicitly assumes that all the produces behave exactly the same way (either minimizes cost or maximizes revenue). Here we follow the argument that the use of an input (output) distance function means that the producers are cost minimizers (revenue maximizers), although these behavioral assumptions are not built into the model explicitly (the way we built in the cost minimizing and profit maximizing behaviors in the models discussed in Sections 3.1 and 3.2).

Here we depart from the traditional approach that assumes that the analyst knows beforehand whether to use the input and output distance functions. We argue that some producers might be minimizing cost (that justifies the use of input distance function) while others maximize revenue (that justifies the use of output distance function). Unfortunately, no information to identify these two groups of producers is available. So we propose the use of a mixture of input and output distance functions[15], meaning that some producers might be minimizing cost while others might be maximizing revenue.

The output distance function is specified as

$$-\ln y_M = \ln f(y/y_M, x, \beta) + v + u^+ \tag{39}$$

while the input distance function is

$$-\ln x_N = \ln f(y, x/x_N, \beta_I) + v - u^+ \tag{40}$$

The determination of the efficiency orientation for each firm is addressed by adopting a latent class structure. In this formulation, the likelihood function for a particular firm is obtained as the weighted sum of both output-oriented and input-oriented likelihood functions, where the weights are the prior probabilities of class membership. That is,

$$g(y,x,\theta,\delta) = g_O(y,x,\theta_O)\cdot\pi_O(\delta_O) + g_I(y,x,\theta_I)\cdot\pi_I(\delta_I) \tag{41}$$

where $0 \le \pi_j \le 1$ (j=O,I), and $\pi_O + \pi_I = 1$, $\theta = (\theta_O, \theta_I)$, $\delta = (\delta_O, \delta_I)$ and the probabilities of being in classes O and I. These class probabilities are parameterized as

[15] This section is based on by Kumbhakar, Orea, Rodriguez and Tsionas (2003).

$$\pi_j(\delta_j) = \frac{\exp(\delta_j' q)}{\sum_j \exp(\delta_j' q)} \quad , \quad j = O, I \tag{42}$$

where $q$ is a vector of firm-specific variables. Under the mantained assumptions, maximum likelihood techniques will give asymptotically efficient estimates of all the parameters.

The estimated parameters can be used to compute posterior probabilites as:

$$P(j \mid y, x) = \frac{g_j(y, x, \theta_j) \cdot \pi_j(\delta_j)}{\sum_j g_j(y, x, \theta_j) \cdot \pi_j(\delta_j)} \quad , \quad j = O, I \tag{43}$$

These posterior probabilities can be used to classify firms. That is, if $\pi_O > \pi_I$ for a firm then we can include it in the O class and vice versa. Once the classification is done we can estimate technical efficiency of the firms in the O class from

$$O\hat{T}E = E\left[\exp(-u^+ \mid e)\right] \tag{44}$$

Similarly, technical efficiency of the firms in the I class can be obtained from

$$I\hat{T}E = 1/E\left[\exp(u^+ \mid e)\right]. \tag{45}$$

For an application of this model to European railroads see Kumbhakar, Orea, Rodriguez and Tsionas (2003).

## 4. Relaxing functional form assumptions (SF model with LML)

In this section we introduce 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. Removal of all these deficiencies generalizes the SF models and makes them comparable to the DEA models. Moreover, we can apply standard econometric tools to perform estimation and draw inferences.

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

$x_i \in X \subseteq R^k$, a vector of unknown parameters $\theta \in \Theta \subseteq R^m$, and let the density be $l(y_i; x_i, \theta)$. The parametric ML estimator is given by

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

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*. A natural way to convert the parametric model to a nonparametric one is to make the parameter $\theta$ a 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

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

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 the parametric ML estimator would do.[16] By solving the LML problem for several points $x \in X$, we can construct the function $\tilde{\theta}(x)$ that is an estimator for $\theta(x)$, and effectively we have a fully general way to convert the parametric model to a non-parametric approximation to the unknown model.

Suppose we have the following stochastic frontier cost model

$$y_i = x_i'\beta + v_i + u_i; \;\; v_i \sim IN(0, \sigma^2),\; u_i \sim IN(\mu, \omega^2),\; u_i \geq 0 \;\; for \;\; i = 1,..,n,\; \beta \in R^k \qquad (46)$$

where $y$ is log cost and $x_i$ is a vector of input prices and outputs[17]; $v_i$ 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 $x_i$.

To make the frontier model more flexible (non-parametric), we adopt the following strategy. Consider the usual parametric ML estimator for the normal ($v$) and truncated

---

[16] 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.

[17] The cost function specification is discussed in details in section 5.2.

normal ($u$) stochastic cost frontier model that solves the following problem (Stevenson, 1980):

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

where

$$l(y_i; x_i, \theta) = [\Phi(\psi)]^{-1} \Phi\left[\frac{\sigma^2 \psi + \omega(y_i - x_i'\beta)}{\sigma(\omega^2 + \sigma^2)^{1/2}}\right] \left[2\pi(\omega^2 + \sigma^2)\right]^{-1/2} \exp\left[-\frac{(y_i - x_i'\beta - \mu)^2}{2\left(\omega^2 + \sigma^2\right)}\right] \tag{47}$$

$\psi = \mu / \omega$, and $\Phi$ 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 non-parametric model involves the following steps. First, we choose a kernel function. A reasonable choice is

$$K_H(d) = (2\pi)^{-m/2} \, |H|^{-1/2} \exp\left(-\tfrac{1}{2} d' H^{-1} d\right), \; d \in R^m, \tag{48}$$

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, we choose a particular point $x \in X$, and solve the following problem:

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

A solution to this problem provides the LML parameter estimates $\tilde{\beta}(x), \tilde{\sigma}(x), \tilde{\omega}(x)$ and $\tilde{\psi}(x)$. Also notice that the 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.[18]

[18] 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

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

where $z$ are the given values for the vector of exogenous variables. The main feature of this formulation is that the $\beta$ parameters as well as $\sigma$, $\omega$, and $\psi$ will now be functions of $z$ instead of $x$.

### 4.1 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: Are large banks necessarily more efficient? Since the banking industry consists of a 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.[19] Moreover, the production technology among banks is likely to differ.[20] These problems are avoided in the non-parametric LML model that makes parameters bank-specific without using any ad hoc specification.

#### *4.1.1 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.

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

[19] 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, Chapter 3.4)). Consequently, estimates of economies of scale are likely to be wrong.

[20] 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.

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-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 (cost of each input divided by input quantity). Total cost is then defined as the sum of cost of these five inputs.

#### *4.1.2 Results*

We estimated the following Cobb-Douglas cost function

$$\ln(C / w_3) = \beta_0 + \sum_m \beta_{ym} \ln y_m + \sum_{j \neq 3} \beta_{wj} \ln(w_j / w_3) + v_i + u_i \tag{49}$$

where as before $v_i \sim IN(0,\sigma^2)$ and $u_i \sim IN(\mu,\omega^2)$, $u_i \geq 0$ $i=1,.n$, $\beta \in R^{k+m}$. Here $C$ is total cost (normalized by the price of $x_3$) and the independent variables contain $m$ (5) outputs and $k$ (4) normalized input prices.[21]

The estimates of $\sigma_v$ and $\psi$ show large variations while the opposite is true for $\sigma_u$. These large variations in estimated coefficients show why estimating a single set of parameters for all banks might not be a good idea.

We measure technical inefficiency using the following procedure. Suppose we localize with respect to observation $j$ and denote the resulting LML estimates of the frontier 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

$$m_{i,(j)} = \frac{\sigma_{(j)}\lambda_{(j)}}{1+\lambda_{(j)}^2}\left[\frac{\phi(z_{i,(j)})}{\Phi(z_{i,(j)})} - z_{i,(j)}\right], \tag{50}$$

[21] The normalization is done to impose linear homogeneity (in input prices) restrictions.

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 $i = 1, ..n$. Therefore, $m_{i,(j)}$ is the inefficiency measure[22] for observation $i$ when we localize with respect to observation $j$. A reasonable inefficiency measure for observation $i$ is provided by $m_i^* = \sum_{j=1}^{n} m_{i,(j)} W_{i,(j)}$ which is a weighted average of all $m_{i,(j)}$ based on the LML weights ($W_{i,(j)}$). Naturally, the dominating element in this average will be $m_{i,(i)}$, 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 $\beta, \mu, \sigma$ and $\omega$ 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 an 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 mergers of large banks may not be very high (Berger and Humphrey (1992)), especially from an efficiency point of view.

## 5. Conclusions

In this paper we presented three new techniques to estimate technical inefficiency using stochastic frontier technique. First, we presented a technique to estimate a non-homogeneous technology using the IO technical inefficiency. We then discussed the IO and OO controversy in the light of distance functions, and the dual cost and profit functions. The second part of the paper addressed the latent class modeling approach incorporating behavioral heterogeneity. The final part of the paper addressed LML method that can solve the functional form issue in parametric stochastic frontier.

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

## References


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

Alvarez, A., C. Arias and S.C. Kumbhakar (2003), "Empirical consequences of direction choice in technical efficiency analysis", manuscript, SUNY Binghamton, NY.

Arias, C. and A. Alvarez (1998), "A Note On Dual Estimation of Technical Efficiency", Paper presented at the First Oviedo Efficiency Workshop, University of Oviedo.

Atkinson, S.E. and C. Cornwell (1994), "Estimation of Output and Input Technical Efficiency Using a Flexible Functional Form and Panel Data", *International Economic Review*, 35, 245-256.

Baltagi, B.H., J.M. Griffin and S.R. Vadali (1998), "Excess capacity: A permanent characteristic of US airlines?", *Journal of Applied Econometrics* 13, 645-657.

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

Beard, T., S. Caudill and D. Gropper (1991), "Finite Mixture Estimation of Multiproduct Cost Functions", *Review of Economics and Statistics* 73, 654-664.

Beard, T., S. Caudill and D. Gropper (1997), "The Diffusion of Production Processes in the U.S. Banking Industry: A Finite Mixture Approach," *Journal of Banking and Finance* 21, 721-740.

Berger, A. N. and D. B. Humphrey (1992), "Measurement and Efficiency Issues in Commercial Banking," in Z. Griliches, ed., *Output Measurement in the Service Sectors*. National Bureau of Economic Research Studies in Income and Wealth Volume 56. Chicago: University of Chicago Press.

Berger, A. N. and L. J. Mester (1997), "Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?," *Journal of Banking and Finance* 21:7 (July), 895-947.

Caudill, S. (2003), "Estimating a Mixture of Stochastic Frontier Regression Models via the EM Algorithm: A Multiproduct Cost Function Application", *Empirical Economics* 28,581-598.

Cuesta, R. A. and J.L. Zofío (1999), "Hyperbolic Efficiency and Distance Functions: A Parametric Approach", Seminar presented at the Department of Econometrics. School of Economics, University of New England, Armidale, Australia.

Färe, R., S. Grosskopf and C. A. K. Lovell (1985), *The Measurement of Efficiency of Production*, Boston: Kluwer Nijhoff Publishing.

Greene, W. (2001), "Fixed and Random Effects in Nonlinear Models", Working Paper, Department of Economics, Stern School of Business, NYU.

Greene, W. (2002), "Alternative Panel Data Estimators for Stochastic Frontier Models", Working Paper, Department of Economics, Stern School of Business, NYU.

Greene, W. (2003), "Simulated Likelihood Estimation of the Normal-Gamma Stochastic Frontier Function", *Journal of Productivity Analysis*, 19, 179–190, 2003.

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

Hagenaars, J.A. and A.L. McCutcheon (2002), *Applied Latent Class Analysis*, Cambridge University Analysis, New York, NY.

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.

Kaparakis, E.I., S.M. Miller and A. Noulas (1994), “Short-Run Cost-Inefficiency of Commercial Banks: A flexible Stochastic Frontier Approach,” *Journal of Money, Credit, and Banking* 26, 875-893.

Kumbhakar, S.C., L. Orea, A. Rodríguez and E. Tsionas (2003), “Estimation of a Mixture of Input and Output Distance Functions and Efficiency Indices”, Efficiency Series Paper 08/2003, Department of Economics, University of Oviedo.

Kumbhakar, S.C. and E. Tsionas (2003), “Estimation of Stochastic Frontier Production Functions with Input-Oriented Technical Efficiency”, mimeo, Binghamton University, New York.

Kumbhakar, S.C. and E. Tsionas (2003), “Do Firms Maximize Profit? A Latent Class Approach,” mimeo, Binghamton University, New York.

Kumbhakar, S.C. and E. Tsionas (2002), “Non-parametric stochastic frontier models,” mimeo, Binghamton University, New York.

Kumbhakar S.C. and C.A.K. Lovell (2000), *Stochastic Frontier Analysis*, Cambridge University Press, New York.

McLachlan, G.J. and D. Peel (2000), *Finite Mixture Models*, John Wiley & Sons, New York.

Meeusen, W. and J. van den Broeck (1977), “Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error," *International Economic Review* 18:2 (June), 435-44.

Orea, L. and S.C. Kumbhakar (2003), “Efficiency Measurement using a Latent Class Stochastic Frontier Model,” *Empirical Economics (forthcoming).*

Orea, L., D. Roibás and A. Wall (2003), “Choosing the Technical Efficiency Orientation to Analyze Firms’ Technology: A Model Selection Test Approach,” *Journal of Productivity Analysis* (*forthcoming*).

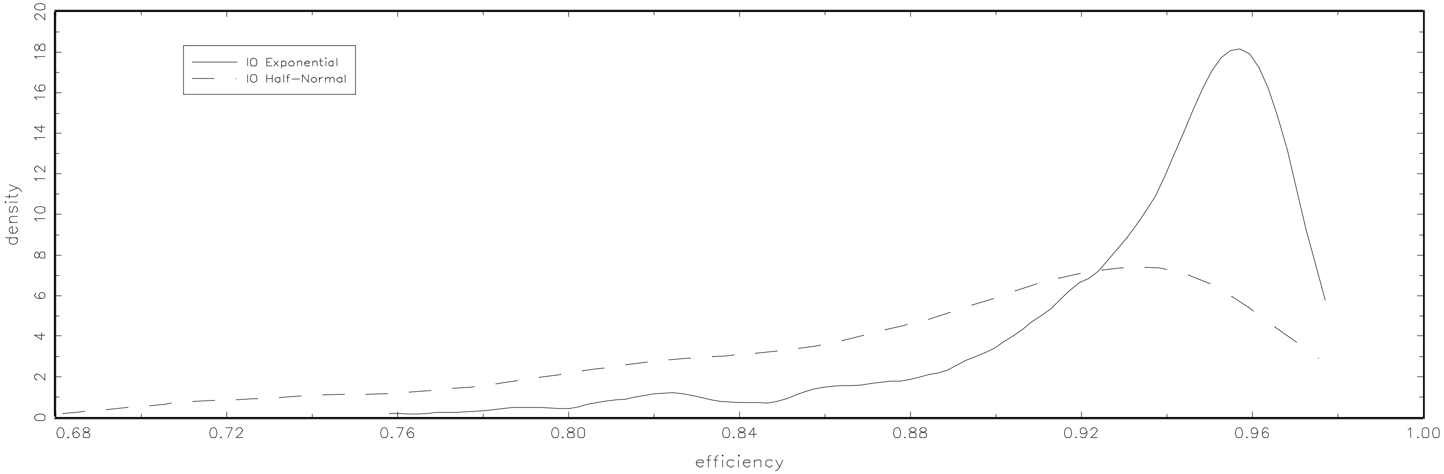
Ray, S.C. (2003), *Data Envelopment Analysis*, Cambridge University Press, New York (*forthcoming*).

Schankerman, M., and M. I. Nadiri (1986), “A Test of Static Equilibrium Models and Rates of Return to Quasi-Fixed Factors, with an Application to the Bell System,” *Journal of Econometrics* 33:1/2 (October/November), 97-118.

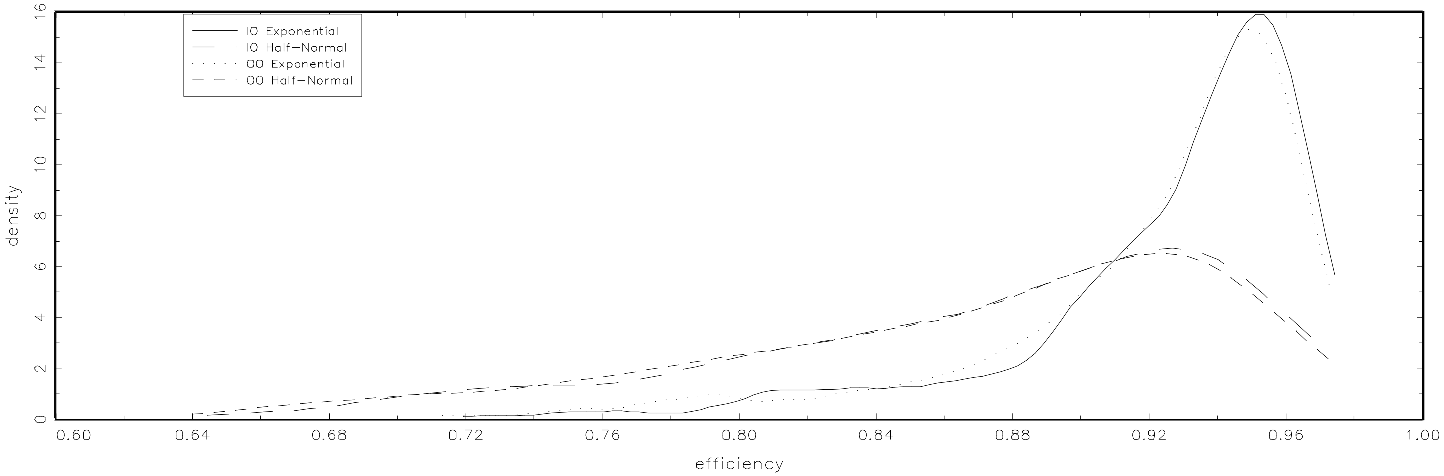
Stevenson, R. E. (1980), "Likelihood Functions for Generalized Stochastic Frontier Estimation," *Journal of Econometrics* 13:1 (May), 57-66.

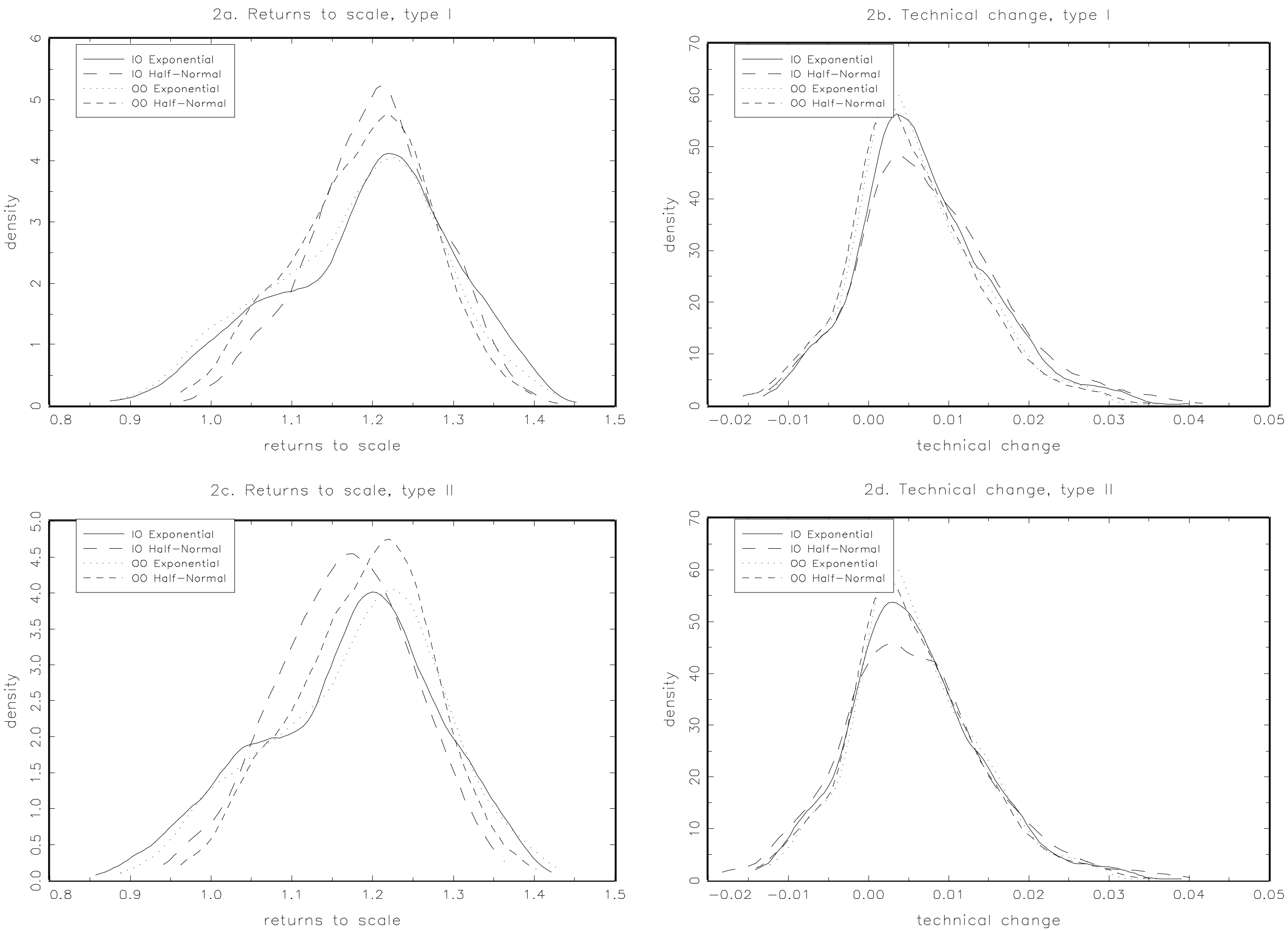
Tibshirani, R. (1984), *Local likelihood estimation*, Ph.D. thesis, Stanford University.

1a. Input–oriented technical efficiency

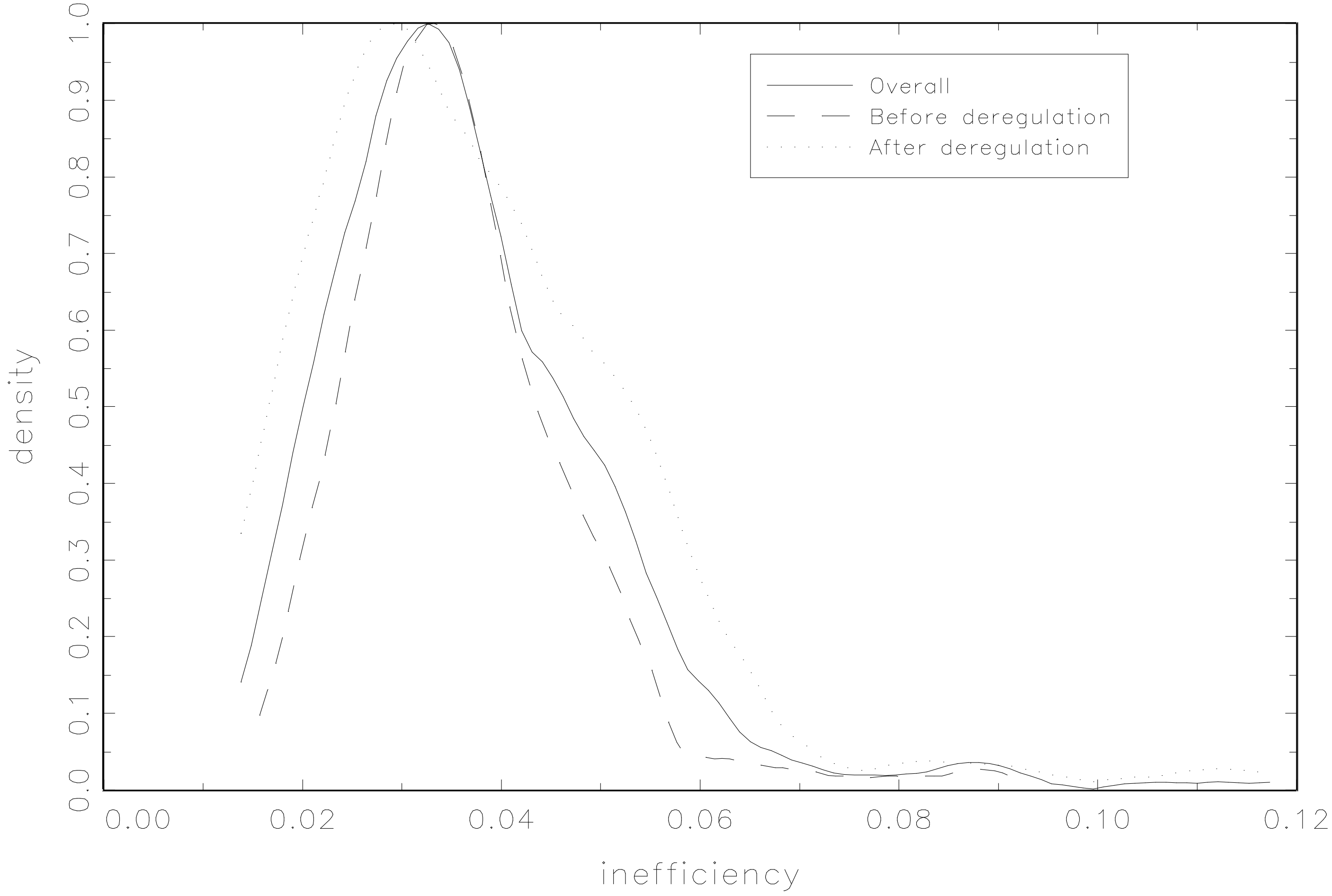



1b. Output–oriented technical efficiency

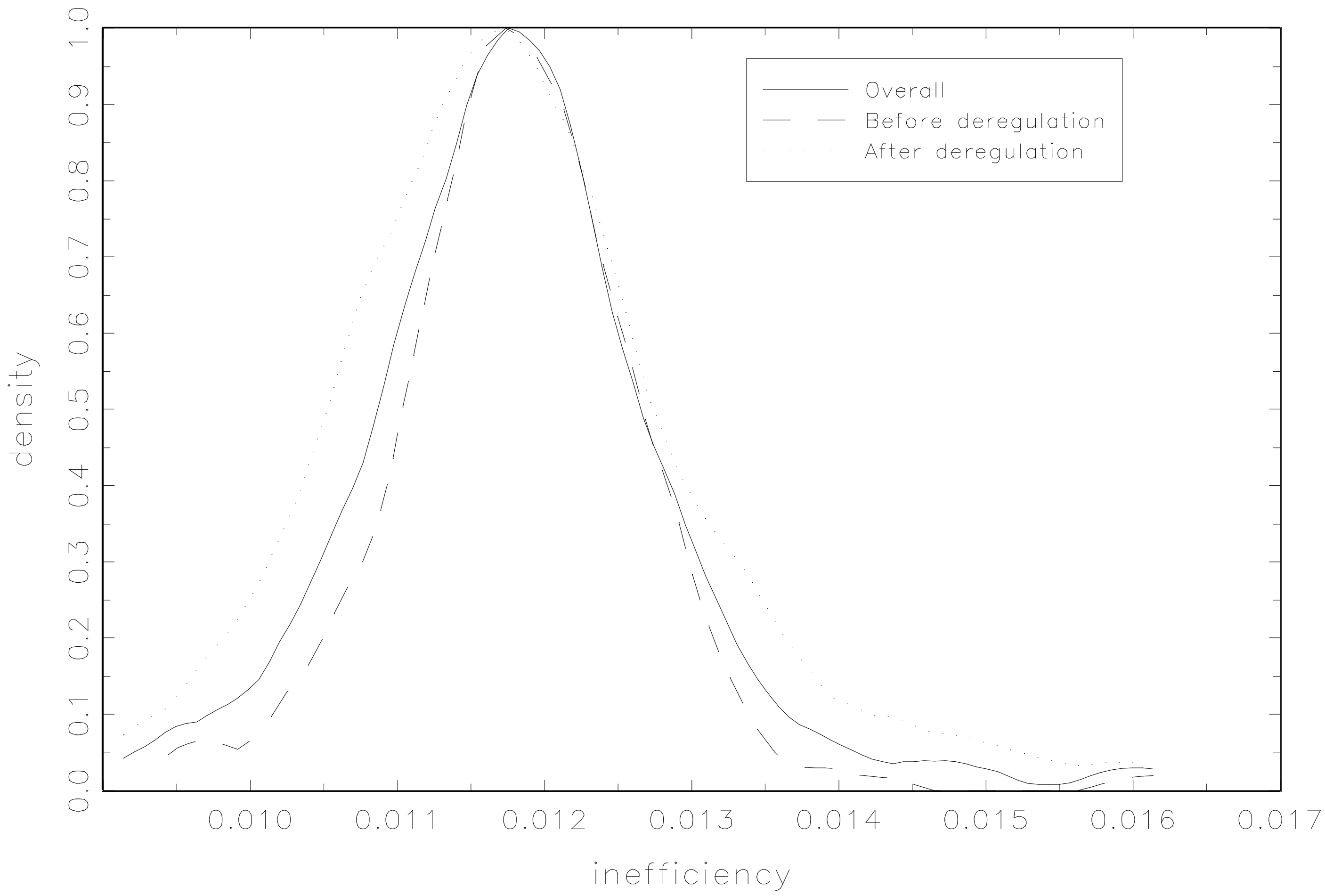



2a. Returns to scale, type I
IO Exponential
IO Half-Normal
OO Exponential
OO Half-Normal
density
returns to scale
2b. Technical change, type I
IO Exponential
IO Half-Normal
OO Exponential
OO Half-Normal
density
technical change
2c. Returns to scale, type II
IO Exponential
IO Half-Normal
OO Exponential
OO Half-Normal
density
returns to scale
2d. Technical change, type II
IO Exponential
IO Half-Normal
OO Exponential
OO Half-Normal
density
technical change

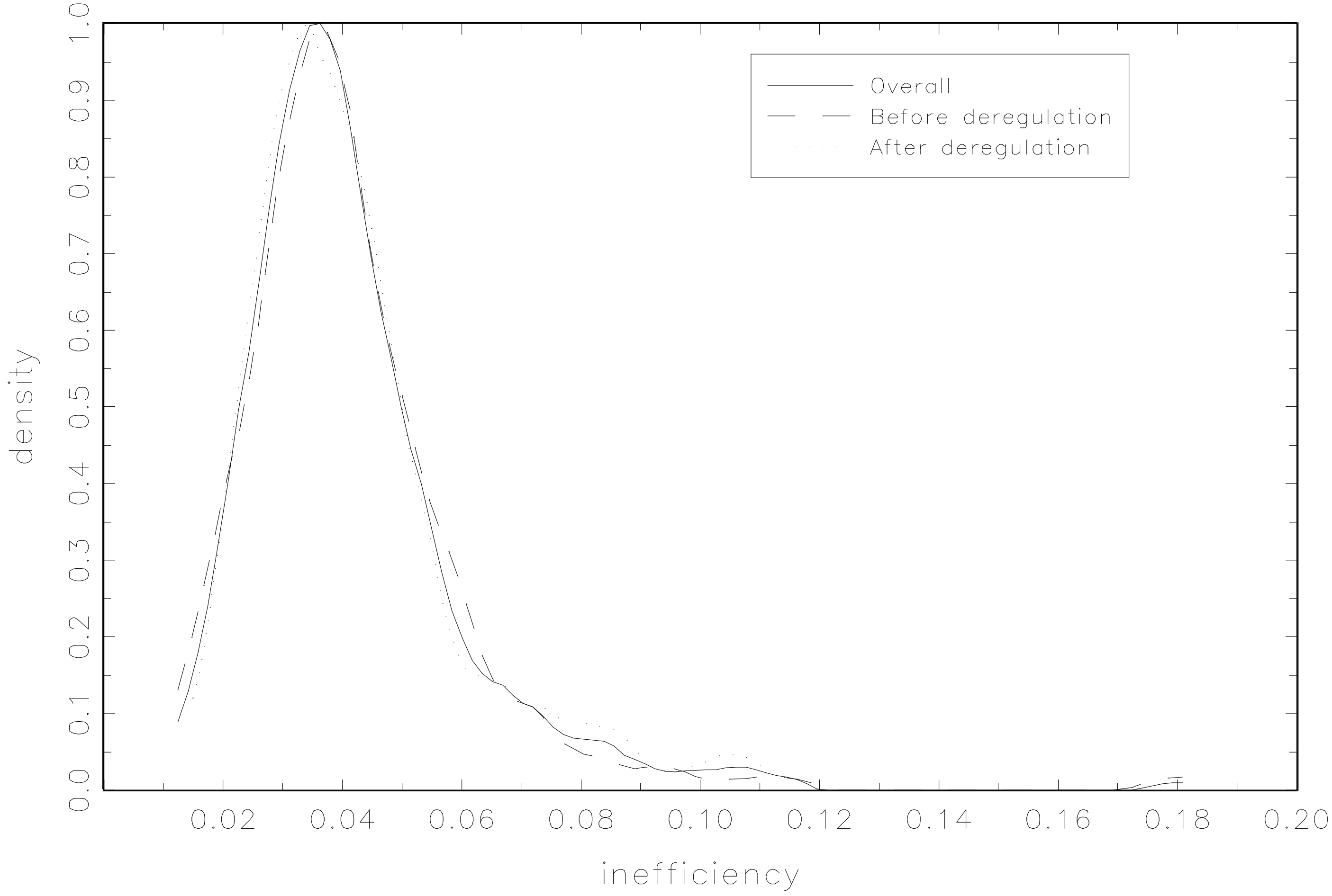
3a. Technical inefficiency from cost-share system



3b. Technical inefficiency from profix maximization system



3c. Technical inefficiency from the mixing model



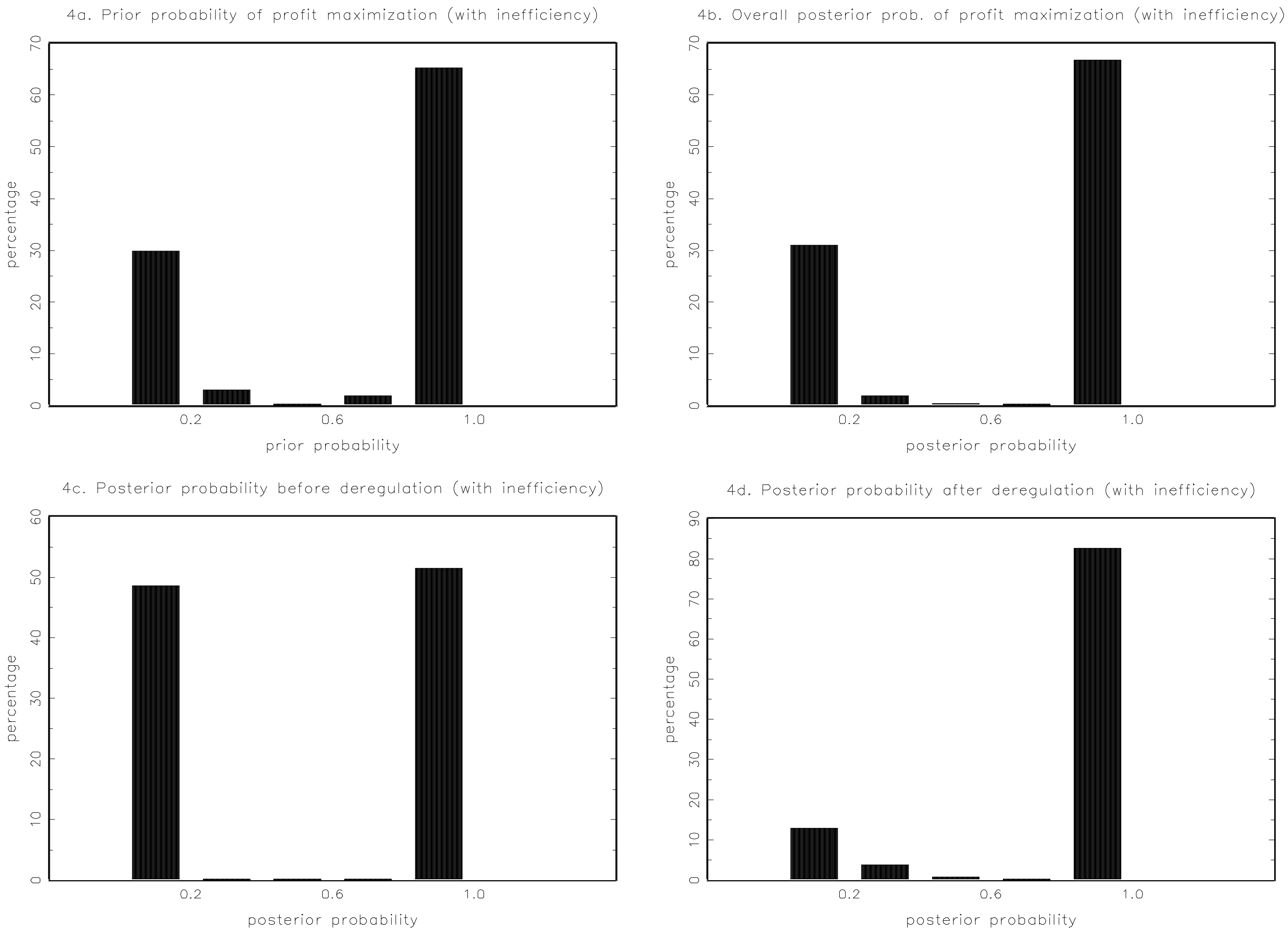

4a. Prior probability of profit maximization (with inefficiency)
percentage
0
10
20
30
40
50
60
70
0.2
0.6
1.0
prior probability
4b. Overall posterior prob. of profit maximization (with inefficiency)
percentage
0
10
20
30
40
50
60
70
0.2
0.6
1.0
posterior probability
4c. Posterior probability before deregulation (with inefficiency)
percentage
0
10
20
30
40
50
60
0.2
0.6
1.0
posterior probability
4d. Posterior probability after deregulation (with inefficiency)
percentage
0
10
20
30
40
50
60
70
80
90
0.2
0.6
1.0
posterior probability

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

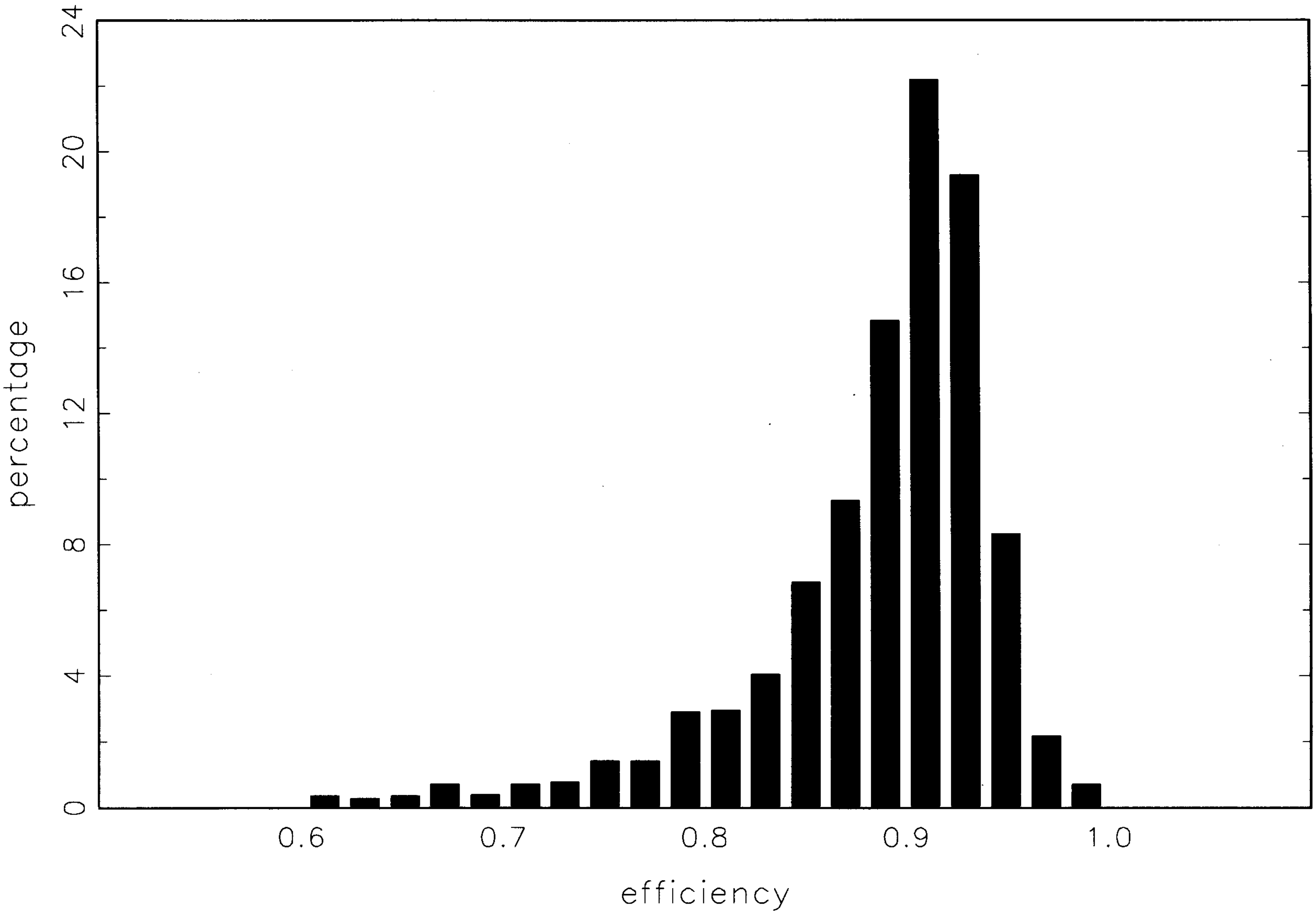


Figure 6: Plot of cost-inefficiency against log(assets) (local ML)