行政院國家科學委員會專題研究計畫　成果報告

無母數隨機邊界模型之貝氏分析

計畫類別：個別型計畫
計畫編號：NSC94-2415-H-032-006-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：淡江大學財務金融學系

計畫主持人：黃河泉

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 9 月 5 日

Bayesian Inference of the Nonparametric Stochastic Frontier Models

Ho-Chuan (**River**) Huang [1]

Department of Banking and Finance
Tamkang University
Tamsui, Taipei County, **Taiwan** 251
TEL: +886-2-26215656 ext. 3337
FAX: +886-2-26214755
E-mail: river@mail.tku.edu.tw

September 4, 2006

[1]The author is grateful to NSC for financial support under grant number NSC 94-2415-H-032-006.

Abstract

Stochastic frontier models are often used to measure the extent of inefficiency of a firm. However, it is found that such a measure is sensitive to the specification of the functional form on the frontiers. As a result, misspecifications in the technology (frontier function) may lead to incorrect conclusions drawn from the resulting frontier even if the distributions of the composed-errors are correctly specified. This study considers a nonparametric stochastic frontier model in which the restrictive assumptions on the parametric specifications are relaxed. The inference is carried out via the Bayesian Markov chain Monte Carlo algorithm (the Gibbs sampler) which provides estimates exhibiting finite-sample properties. The full conditional distributions required in the implementation of the Gibbs sampler are derived. An empirical application to the real data is conducted to illustrate the practical use of our proposed model and estimation technique.

Keywords: stochastic frontier, nonparametric, Gibbs sampler, Metropolis-Hastings

1 Introduction

Stochastic frontier models, developed by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977), have been commonly used in the estimation of firms' technical (production or cost) inefficiencies. By definition, the production frontier denotes the maximum amount of output that can be produced by a certain technology with a given level of inputs. However, in practice, the actual output of a firm will typically fall below the maximum that is technically feasible. Thus, the deviation of actual from maximum output can be used as a measure of inefficiency and is the main focus of interest in many studies. For instance, the recent empirical applications include banking (Greene, 2005; Kumbhakar and Tsionas, 2005), health care (Griffin and Steel, 2004; Greene, 2004), life insurance (Greene and Segal, 2004), investment (Wang, 2003), sports (Koop, 2004; Amos, Beard and Caudill, 2005) and world production (Tsionas and Kumbhakar, 2004), to name a few.

In its basic form, the stochastic frontier model uses a parametric representation of technology along with a two-part composed-error term. Within this framework, the observed output or cost is decomposed into three components — the actual frontier, which depends on a set of explanatory variables; a symmetric disturbance, which captures other effects such as measurement error, and a particular one-sided disturbance which denotes deviations of the individual unit from the frontier, i.e., a measure of inefficiency. Existing extensions of the basic stochastic frontier approach include at least the following aspects. First, a more flexible distributional assumption of the one-sided disturbance is adopted for measuring inefficiencies. In contrast to the half-normal distribution of Aigner, Lovell and Schmidt (1977) and the exponential distribution of Meeusen and van den Broeck (1977), latter generalizations include the truncated-normal density of Stevenson (1980), the gamma density of Greene (1990) and the generalized gamma distributions and mixtures of generalized gamma distributions of Griffin and Steel (2003). In contrast, Park and Simar (1994) consider a parametric frontier and are nonparametric on the inefficiency distribution. Griffin and Steel (2004) propose a semiparametric Bayesian framework in which the dis-

tribution of inefficiencies is modeled nonparametrically through a Dirichlet process prior. Second, the distribution of technical inefficiency is allowed to depend on some exogenous variables. For example, Huang and Liu (1994) and Battese and Coelli (1995) allow the mean of the distribution to depend on firm-specific characteristics whereas Caudill, Ford and Gropper (1995) and Hadri (1999) parameterize the variance of the distribution as a function of appropriate explanatory variables. Recently, Wang (2002) and Hadri, Guermat and Whittaker (2003) provide a flexible parameterization to allow exogenous influences on both the mean and variance of the technical inefficiency distribution.

Third, alternative functional forms of the stochastic frontiers are examined. The most commonly-used specifications include a variant of the Cobb-Douglas or Translog models. Despite of the simplicity, it is well known that the primary objective of composed-error models, i.e., measurement of firms' inefficiencies, can be very sensitive to the choice of functional form of the frontier. Therefore, Koop, Osiewalski and Steel (1994) propose the asymptotically ideal model whereas Zhu, Ellinger and Shumway (1995) and Giannakas, Tran and Tzouvelekas (2003) consider a generalized quadratic Box-Cox transformation of the stochastic frontiers. However, most of the existing models explicitly or implicitly assume that all firms under investigation share exactly the same technology and differ only with respect to their degree of inefficiency. In practice, however, firms may adopt different technologies for a variety of reasons. As argued in Tsionas (2002), adoption of a new technology is costly, and firms adopt new technologies only with considerable lags. If costs related to installation and personnel training differ across firms, it follows that at any given point in time there will be some variability in the types of technology used by firms. Therefore, we might expect the production possibilities to be different in a cross-section of firms. Thus, Tsionas (2002) and Huang (2004) consider a random-coefficient stochastic frontier to separate technical inefficiency from technological differences across firms.

Alternative modeling strategies and generalizations are the semiparametric or nonparametric analysis and inference. For example, Fan, Li and Weersink (1996) extend the linear stochastic frontier model to a semiparametric stochastic frontier model in

which the functional form of the frontier is left unspecified but the distributions of the composite error terms are of known form. They propose semiparametric pseudo-likelihood estimators based kernel estimation which are robust to possible misspecifications of the frontier as opposed to existing parametric estimators. Similarly, Huang and Fu (1999) also advocate a nonparametric specification of the frontier and adopt a parametric inefficiency distribution. In particular, they utilize the approach of average derivative to estimate slopes of a stochastic frontier function and the method of pseudolikelihood to infer inefficiency without making an assumption or approximation on the functional specification. In contrast, Park and Simar (1994) assume a parametric frontier and focus on the nonparametric inefficiency distribution. This setup is extended by Park, Sickles and Simar (1998) to allow for dependence between inefficiencies and regressors, and by Sickles, Good and Getachew (2002) to model the multiple output/multiple input technology.

In the same spirits, we propose a novel nonparametric stochastic frontier model to relax the restrictive assumption on the functional form of the frontier which represents the production technology. This can be very important since misspecifications in the technology (frontier function) may lead to incorrect conclusions drawn from the resulting frontier even if the distributions of the composed-errors are correctly specified. This study differs from the existing studies in some respects. First, the analysis and inference are from Bayesian point of view via the Markov chain Monte Carlo algorithm. The estimation and model comparison are straightforward to implement and intuitively feasible. Second, in contrast to classical approaches, we can obtain the whole density of the parameters of interests so that the uncertainty of parameters (or prediction) is taken into account. Third, it is well known that the classical nonparametric regression analysis relies heavily on large samples. The curse of dimensionality often makes the nonparametric estimators unreliable using sample of the regular size. In contrast, our Bayesian approach provides estimates which exhibit finite-sample properties.

2 The parametric framework

Since the introduction of Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977), the stochastic production frontier approach often assumes a parametric representation of technology along with a two-part composed-error term in the measurement of firm's (in)efficiency. Specifically, a standard linear specification takes the form as,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i - u_i \tag{1}$$

where y_i is the logarithm of the observed output and $x_{i1}, x_{i2}, \cdots, x_{ik}$ are the logarithms of k inputs for the 'i'th firm. The symmetric disturbance term ϵ_i, denoting either statistical noise or measurement error, is commonly assumed to be distributed as $iid\,\mathcal{N}(0, \sigma^2)$ for $i = 1, 2, \cdots, n$. Moreover, the one-sided (non-negative) error term u_i represents the extent of technical inefficiency. Obviously, the firm is fully efficient when $u_i = 0$.

In some cases we are interested in measuring cost rather than production inefficiencies. Then, equation (1) can be adapted to be a stochastic cost frontier which represents the minimum attainable cost of producing a given level of outputs. In a very similar way, a typical stochastic frontier model may be specified as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i + u_i \tag{2}$$

where so that, if the signs of y_i and $x_{i1}, x_{i2}, \cdots, x_{ik}$ are reversed, all results of the production frontier can be directly applied to the cost frontier as well.

Extensions of equation (1) include at least two main directions. The first one is relaxing the distributional assumptions of technical inefficiency. Originally, Aigner, Lovell and Schmidt (1977) consider a half-normal distribution while Meeusen and van den Broeck (1977) adopt an exponential distribution for u_i. Later, Stevenson (1980) extends the half-normal assumption to the truncated normal distribution and Greene (1990) generalizes the exponential distribution to the more flexible gamma density for measuring the technical inefficiency u_i. Recently, Griffin and Steel (2004) consider generalized gamma distributions and mixtures of generalized gamma distributions

and Griffin and Steel (2004) model the distribution of inefficiencies nonparametrically through a Dirichlet process prior.

3 The semi- and/or non-parametric specification

In order avoid possible model mis-specifications which might invalidate the estimation of technology and the measure of inefficiency in the parametric setup, we can consider a more flexible nonparametric inference of the stochastic frontier model. In contrast to conventional parametric models, the nonparametric approaches do not need to specify the functional forms between output and inputs *ex ante* and let the data determine what the relationship looks like.

3.1 The single-input case

For illustrative purpose, we first consider a simple case with only one input, i.e., [1]

$$y_i = f(z_i) + \epsilon_i - u_i \tag{3}$$

where, in contrast to equation (1), we have only one input z_i. More importantly, the relationship between y_i and z_i is characterized by the unknown (nonparametric) function $f(\cdot)$. The distributional assumptions of the error terms are $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $u_i \sim \mathcal{E}(\theta) = \theta \exp\{-\theta u_i\}$, respectively. The assumption of the exponentially distributed inefficiency u_i can be easily extended to the more general gamma distribution with some additional effort, e.g., Tsionas (2000) and Huang (2004).

Without loss of generality, the observations are ordered so that $z_1 \leq z_2 \leq \cdots \leq z_n$. By stacking the observations, we have,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_n) \end{bmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} - \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \tag{4}$$

By defining $y = (y_1, y_2, \cdots, y_n)'$, $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$, $u = (u_1, u_2, \cdots, u_n)'$, equation (4) can be re-written as,

$$y = \gamma + \epsilon - u \tag{5}$$

[1] We use x to denote variables entering the regression parametrically and z to represent inputs treated nonparametrically.

where $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_n)' = [f(z_1), f(z_2), \cdots, f(z_n)]'$ denotes the n points on the nonparametric regression line to be estimated. As noted by Koop and Poirier (2004), without imposing any additional structure to the above model, we are plagued by the problem of 'insufficient observations' in that we have more unknown parameters than available observations. However, the problem can be resolved through the use of prior information about the degree of smoothness of the nonparametric regression lines.

In Bayesian analysis, we can treat u as additional parameters to be estimated. As a result, the (augmented) likelihood function becomes,

$$L(y|\gamma, \sigma^2, u) = (2\pi)^{-\frac{n}{2}}(\sigma^{-2})^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y + u - \gamma)'(y + u - \gamma)\right\} \tag{6}$$

All the priors are assumed to be independent. In particular, we follow Koop and Poirier (2004) to assume

$$D\gamma \sim \mathcal{N}(0, V(\eta)) \tag{7}$$

where

$$D_{(n-2)\times n} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}$$

so that $D\gamma$ represents the vector of second differences of points on the nonparametric regression line. [2]

For simplicity, we take $V(\eta) = \eta I_{n-2}$ where η^{-1} has a gamma prior, i.e.,

$$\eta^{-1} \sim \mathcal{G}(\nu_{\eta,0}, \delta_{\eta,0}) \tag{8}$$

Clearly, as $\eta \to \infty$, the prior becomes diffuse and the resulting estimates will be undersmoothed. In contrast, as $\eta \to 0$, prior information will dominate, and will restrict the second differences to be identically zero (potentially oversmoothing). In this

[2] As an alternative, we can consider the first differencing matrix,

$$D_{(n-1)\times n} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

sense, the scalar parameter η acts as a smoothing parameter in spirit to a bandwidth parameter in classical kernel-based methods.

The prior of σ^{-2} is also gamma distributed as,

$$\sigma^{-2} \sim \mathcal{G}(\nu_{\sigma^2,0}, \delta_{\sigma^2,0}) \tag{9}$$

Since we adopt the exponential distribution for the inefficiency term $u_i, i = 1, 2, \cdots, n$, the prior of u_i is,

$$u_i \sim \theta \exp(-\theta u_i) \tag{10}$$

and the prior of θ is assumed to be,

$$\theta \sim \mathcal{G}(\nu_{\theta,0}, \delta_{\theta,0}) \tag{11}$$

In order to implement the Gibbs sampler, we have to derive the relevant full conditional distributions for all parameters. As shown immediately, all the full conditionals are of standard forms and are easy to simulate from.

- The full conditional of γ:

  By combining (6) and (7), the full conditional of γ is,

  $$\gamma | y, \eta, \sigma^2, u, \theta \sim \mathcal{N}(\gamma_n, G_n) \tag{12}$$

  where

  $$\begin{aligned} \gamma_n &= G_n \left[(y+u)/\sigma^2 \right] \\ G_n &= \left(D'D/\eta + I_n' I_n/\sigma^2 \right)^{-1} \end{aligned}$$

- The full conditional of η^{-1}:

  By combining (6) and (8), the full conditional of η^{-1} is,

  $$\eta^{-1} | y, \gamma, \sigma^2, u, \theta \sim \mathcal{N}(\nu_{\eta,n}, \delta_{\eta,n}) \tag{13}$$

  where

  $$\begin{aligned} \nu_{\eta,n} &= \nu_{\eta,0} + \frac{n-2}{2} \\ \delta_{\eta,n} &= \delta_{\eta,0} + \frac{(D\gamma)'(D\gamma)}{2} \end{aligned}$$

- The full conditional of σ^{-2}:

  By combining (6) and (9), the full conditional of σ^{-2} is,

$$\sigma^{-2}|y,\gamma,\eta,u,\theta \sim \mathcal{N}(\nu_{\sigma^2,n},\delta_{\sigma^2,n}) \tag{14}$$

  where

$$\begin{aligned}\nu_{\sigma^2,n} &= \nu_{\sigma^2,0}+\frac{n}{2}\\ \delta_{\sigma^2,n} &= \delta_{\sigma^2,0}+\frac{(y+u-\gamma)'(y+u-\gamma)}{2}\end{aligned}$$

- The full conditional of $u_i, i=1,2,\cdots,n$:

  By combining (6) and (10), the full conditional of u_i for each i is,

$$u_i|y,\gamma,\eta,\sigma^2,\theta \sim \mathcal{N}_{[0,\infty]}\left(\gamma_i-y_i-\theta\sigma^2,\sigma^2\right) \tag{15}$$

- The full conditional of θ:

  By combining (6) and (11), the full conditional of θ is,

$$\theta|y,\gamma,\eta,\sigma^2,u \sim \mathcal{G}(\nu_{\theta,n},\delta_{\theta,n}) \tag{16}$$

  where

$$\begin{aligned}\nu_{\theta,n} &= \nu_{\theta,0}+\frac{n}{2}\\ \delta_{\theta,n} &= \delta_{\theta,0}+\frac{(y+u-\gamma)'(y+u-\gamma)}{2}\end{aligned}$$

Thus, posterior analysis can be carried out using the Gibbs sampler which sequentially draws from (12), (13), (14), (15) and (16), and all of these densities are of standard forms.

3.2 The multiple-input case

However, in reality, the production of output often requires multiple inputs. As a result, we consider a more general and flexible multiple-input model. In particular, we assume that a vector of k explanatory variables $x_i=(x_{i1},x_{i2},\cdots,x_{ik})'$ are treated

parametrically while there are p inputs $z_i = (z_{i1}, z_{i2}, \cdots, z_{ip})'$ entering the regression nonparametrically. The model can be written as,

$$y_i = x_i'\beta + f(z_{i1}, z_{i2}, \cdots, z_{ip}) + \epsilon_i - u_i$$

However, as the dimension of z_i increases, we will encounter the problem of "curse of dimensionality" which might make the preceding estimation approach work poorly. Thus, instead, our interest focuses on the additive models which do not suffer from this curse.

Specifically, consider the following partially linear stochastic frontier (PLSF) model with additive nonparametric components, [3]

$$y_i = x_i'\beta + f_1(z_{i1}) + f_2(z_{i2}) + \cdots + f_p(z_{ip}) + \epsilon_i - u_i \tag{17}$$

where the output y_i is affected by the $k \times 1$ vector of inputs x_i with magnitude measured by the corresponding coefficients β in a parametric and linear way. In contrast, the $p \times 1$ vector of inputs z_i influences the output y_i through the nonparametric and unknown function $f_1(z_{i1}), f_2(z_{i2}), \cdots, f_p(z_{ip})$, respectively.

As in (5), equation (17) can be rewritten as,

$$y = X\beta + \gamma_1 + \gamma_2 + \cdots + \gamma_p + \epsilon - u \tag{18}$$

where X is a $n \times k$ matrix with ith row given by x_i', and $\gamma_j = (\gamma_{1j}, \gamma_{2j}, \cdots, \gamma_{np})' = [f_j(z_{1j}), f_j(z_{2j}), \cdots, f_j(z_{nj})]'$, $j = 1, 2, \cdots, p$. In the one-input case where z_i is a scalar for $i = 1, 2, \cdots, n$, we can sort the data points so that $z_1 \leq z_2 \leq \cdots \leq z_n$. In contrast, in the multiple-input case, we have a $p \times 1$ vector of explanatory variables which can be used to order the data, so there is not one simple ordering which can be adopted. However, as argued in Koop and Poirier (2004), Bayesian inference can still be carried out in the same manner as in the one-input case by setting up a Gibbs sampler which involves sequentially drawing from $\pi(\gamma_1|y, \gamma_2, \cdots, \gamma_p, \Theta)$, $\pi(\gamma_2|y, \gamma_1, \gamma_3, \cdots, \gamma_p, \Theta)$, $\cdots$, $\pi(\gamma_p|y, \gamma_1, \cdots, \gamma_{p-1}, \Theta)$ along with the full conditional densities of the remaining model parameters $\Theta = (\beta, \eta, \sigma^2, u, \theta)$.

[3] Please also see Fan, Li and Weersink (1996) for a similar specification.

The prior of β is chosen to be of natural conjugate form, i.e., $\beta \sim \mathcal{N}(\beta_0, B_0)$. The other priors are assumed to be independent and comparable to the ones used in the single-input case, i.e., $D\gamma_j^{(j)} \sim \mathcal{N}(0, \eta_j I_{n-2})$, $\eta_j^{-1} \sim \mathcal{G}(\nu_{\eta_j,0}, \delta_{\eta_j,0})$, and the priors of σ^{-2}, u_i and θ are the same as in (9), (10) and (11). Moreover, given u, the complete likelihood function is,

$$\begin{aligned} L(y|\gamma, \sigma^2, u) &= (2\pi)^{-\frac{n}{2}} (\sigma^{-2})^{\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}(y + u - X\beta - \gamma_1 - \cdots - \gamma_p)' \\ &\quad (y + u - X\beta - \gamma_1 - \cdots - \gamma_p)\} \end{aligned} \tag{19}$$

Let $\Theta = (u', \theta, \beta', \sigma^{-2}, \gamma', \eta')'$, where $u = (u_1, u_2, \cdots, u_n)'$, $\gamma = (\gamma_1', \gamma_2', \cdots, \gamma_p')'$, and $\eta = (\eta_1, \eta_2, \cdots, \eta_p)'$ denote the unknown parameters on which we are interested in drawing inferences. Moreover, let $\Theta_{\backslash u_1}$ denote all other the parameters in Θ by deleting u_1. Similar notations are applied to the other cases.

- The full conditional distribution of the latent inefficiency u_i for $i = 1, 2, \cdots, n$, can be shown to follow a truncated normal distribution. Specifically,

$$u_i |\, y, \Theta_{\backslash u_i} \sim \mathcal{N}_{[0,\infty]}(x_i'\beta + \gamma_{i1} + \cdots + \gamma_{ip} - y_i - \theta\sigma^2, \sigma^2) \tag{20}$$

  Note that the notation $\gamma_{\backslash j}$ denotes $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_k)$ by deleting the j^{th} element γ_j.

- The full conditional distribution of θ is gamma distributed as,

$$\theta |\, y, \Theta_{\backslash \theta} \sim \mathcal{G}(\nu_{\theta,n}, \delta_{\theta,n}) \tag{21}$$

  where

$$\begin{aligned} \nu_{\theta,n} &= \nu_{\theta,0} + n \\ \delta_{\theta,n} &= \delta_{\theta,0} + \sum_{i=1}^{n} u_i \end{aligned}$$

- The full conditional distribution of β is normally distributed as,

$$\beta |\, y, \Theta_{\backslash \beta} \sim \mathcal{N}(\beta_n, B_n) \tag{22}$$

where

$$\begin{aligned}\beta_n &= B_n\left[B_0^{-1}\beta_0 + X'(y+u-\gamma_1-\cdots-\gamma_p)/\sigma^2\right]\\ B_n &= \left(B_0^{-1}+X'X/\sigma^2\right)^{-1}\end{aligned}$$

- The full conditional distribution of σ^{-2} is gamma distributed as,

$$\sigma^{-2}|\, y, \Theta_{\backslash\sigma^{-2}} \sim \mathcal{N}(\nu_{\sigma^{-2},n}, \delta_{\sigma^{-2},n}) \tag{23}$$

where

$$\begin{aligned}\nu_{\sigma^{-2},n} &= \nu_{\sigma^{-2},0} + \frac{n}{2}\\ \delta_{\sigma^{-2},n} &= \delta_{\sigma^{-2},0} + \frac{(y+u-X\beta-\gamma_1-\cdots-\gamma_p)'(y+u-X\beta-\gamma_1-\cdots-\gamma_p)}{2}\end{aligned}$$

- In order to derive the full conditional distributions of γ_j as well as η_j for $j = 1, 2, \cdots, p$, we let $y^{(j)}$ denote the dependent variable ordered according to the j^{th} input, i.e., $z_{1j} \leq z_{2j} \leq \cdots \leq z_{nj}$, and define $X^{(j)}$, $\gamma_\ell^{(j)}, \ell = 1, 2, \cdots, p$ and $u^{(j)}$ in the same way. In addition, let

$$\tilde{y}^{(j)} = y^{(j)} - \left(\gamma_1^{(j)} + \cdots + \gamma_{j-1}^{(j)} + \gamma_{j+1}^{(j)} + \cdots + \gamma_p^{(j)}\right)$$

The full conditional distribution of γ_j can be shown to be normally distributed as,

$$\gamma_j|\, y, \Theta_{\backslash\gamma_j} \sim \mathcal{N}(\gamma_{j,n}, G_{j,n}) \tag{24}$$

where

$$\begin{aligned}\gamma_{j,n} &= G_{j,n}\left[(\tilde{y}^{(j)} + u^{(j)} - x_i^{(j)'}\beta)/\sigma^2\right]\\ G_{j,n} &= \left(D'D/\eta_j + I_n'I_n/\sigma^2\right)^{-1}\end{aligned}$$

- The full conditional distribution of η_j^{-1}, $j = 1, 2, \cdots, p$, is gamma distributed as,

$$\eta_j^{-1}|\, y, \Theta_{\backslash\eta_j} \sim \mathcal{G}(\nu_{\eta_j,n}, \delta_{\eta_j,n}) \tag{25}$$

where

$$\begin{aligned}\nu_{\eta_j,n} &= \nu_{\eta_j,0} + \frac{n-2}{2} \\ \delta_{\eta_j,n} &= \delta_{\eta_j,0} + \frac{\left(D\gamma_j^{(j)}\right)'\left(D\gamma_j^{(j)}\right)}{2}\end{aligned}$$

Thus, posterior analysis can be carried out via the Gibbs sampling algorithm which sequentially draws from (20), (21), (22), (23), (24), and (25), and all of these densities are of standard forms and are easy to simulate from.

4 Empirical applications

In order to illustrate the practicality of our model, we consider the estimation of a stochastic cost frontier. The theory of firm implies that a firm's costs should depend on the quantity of each output produced as well as the input prices faced by the firm. The data set used to illustrate the technique is collected by Christensen and Greene (1976) for a total of 123 electric utility companies in the United States in 1970. The same data set has been previously analyzed by Greene (1990), van den Broeck, Koop, Osiewalski and Steel (1994), Koop, Steel and Osiewalski (1995) and Tsionas (2002).

For comparison purpose, we first estimate the parametric Cobb-Douglas cost function which is specified as,

$$\ln\left(\frac{c}{p_f}\right)_i = \beta_0 + \beta_1 \ln q_i + \beta_2 (\ln q_i)^2 + \beta_3 \ln\left(\frac{p_l}{p_f}\right)_i + \beta_4 \ln\left(\frac{p_k}{p_f}\right)_i + \epsilon_i - u_i \qquad (26)$$

where c is total cost, q is output, and p_l, p_k and p_f are the three unit prices of labor, capital and fuel, respectively. As above, we assume that the symmetric disturbance term $\epsilon_i \sim iid\,\mathcal{N}(0, \sigma^2)$ and the non-negative error $u_i \sim iid\,\mathcal{E}(\theta)$.

As an alternative to the parametric setup, we now consider a semiparametric partially linear stochastic frontier model. In particular, we assume that the inputs prices, $\ln(p_l/p_f)$ and $\ln(p_k/p_f)$, enter the PLSF regression parametrically as in (26). In contrast to the quadratic specification of output, $\ln q$, we do not impose any functional assumption between cost and output. Instead, we let the data speak for themselves

by estimating a nonparametric component $f(\ln q)$ as,

$$\ln\left(\frac{c}{p_f}\right)_i = f(\ln q_i) + \beta_1 \ln\left(\frac{p_l}{p_f}\right)_i + \beta_2 \ln\left(\frac{p_k}{p_f}\right)_i + \epsilon_i - u_i \tag{27}$$

or, in term of previous notations,

$$y_i = f(z_i) + x_i'\beta + \epsilon_i - u_i \tag{28}$$

where $y_i = \ln\left(\frac{c}{p_f}\right)_i$, $z_i = \ln q_i$, $x_i = \left[\ln\left(\frac{p_l}{p_f}\right)_i, \ln\left(\frac{p_k}{p_f}\right)_i\right]'$, and $\beta = (\beta_1, \beta_2)'$.

Both models are estimated via the Gibbs sampler with data augmentation algorithm by assuming relatively diffuse priors. The Markov chain is then run for $20,000$ iterations. We collect the last $10,000$ sample variates after discarding the first $10,000$ draws. As a result, the following results are based on $10,000$ Gibbs output for making posterior inference.

The top panel of Table 1 reports the posterior moments of the parametric stochastic frontier model as specified in (26). First, we find that the posterior means of β coefficients are all positive as expected. Except for the coefficient of $\ln(p_k/p_f)$, all the other β coefficients are also highly significant according to either 95% or 90% Bayesian confidence intervals. Second, both the parameters on $\ln q_i$ and $(\ln q_i)^2$ are estimated to be significantly positive, indicating that, other things being equal, the cost is a convex function of the output produced. In other words, linear specification of the relationship between cost and output appears to be inadequate. These results are comparable to those found in Koop, Steel and Osiewalski (1995) and Tsionas (2002). As discussed earlier, our main concern is on the measurement of firm-specific efficiency. Figure 1 presents the kernel density of the (mean) efficiency measures of all firms. It is apparent that the efficiency distribution is highly left-skewed and exhibits large variation over firms.

In contrast, we also report the posterior results of the semiparametric model in the bottom panel of Table 1. Similar to the results obtained in the parametric model, the posterior mean of the coefficient on $\ln(p_l/p_f)$ remains positive and highly significant while the posterior mean of the coefficient on $\ln(p_k/p_f)$ turns out to be negative but is still insignificantly different from zero. Most notably, the unknown (nonparametric)

relationship between cost and output is estimated and displayed in Figure 2. It seems that the estimated nonparametric line is close to an approximately linear line with positive slope. This is in contrast with the convex function predicted by the parametric quadratic regression. Finally, we summarize the efficiency distribution from the semiparametric stochastic frontier model in Figure 3. Clearly, the density is different from that shown in Figure 1 derived from the parametric stochastic frontier function.

5 Conclusions

This paper considers the measurement of firm's specific (in)efficiency while allows for the possible heterogeneous technologies adopted by different firms. A very flexible stochastic frontier model with nonparametric specification is proposed to distinguish technical inefficiency from technological differences across firms. Posterior inference of the model is made possible via the simulation-based approach, namely, Markov chain Monte Carlo method.

The full conditionals of the parameters are all in standard forms and can be easily and directly simulated from using the Gibbs sampler with data augmentation algorithm. The model is applied to a real data set which has also been considered in Christensen and Greene (1976), Greene (1990), Tsionas (2002), among others. Empirical results show that the parametric quadratic specification does not seem tom be the best representation compared to our estimated nonparametric (approximately linear) relationship. As a result, we believe that the novel techniques proposed in this paper might allow for better understanding of firm efficiency than do traditional methods.

References

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

Amos, D., Beard, T. R. and Caudill, S. (2005), "A Statistical Analysis of the Handling Characteristics of Certain Sporting Arms: Frontier Regression, the Moment of Inertia, and the Radius of Gyration." *Journal of Applied Statistics* 32, 3-16.

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

Caudill,S. B., Ford, J. M. and Gropper, D. M. (1995), "Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroscedasticity." *Journal of Business & Economic Statistics* 13, 105-111.

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

Hadri, K. (1999), "Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function." *Journal of Business & Economic Statistics* 17, 359-363.

Hadri, K., Guermat, C. and Whittaker, J. (2003), "Estimation of Technical Inefficiency Effects using Panel Data and Doubly Heteroscedastic Stochastic Production Frontiers." *Empirical Economics* 28, 203-222.

Huang, C. J. and Fu, T. T. (1999), "An Average Derivative Estimation of Stochastic Frontiers." *Journal of Productivity Analysis* 12, 45-53.

Huang, C. J. and Liu, J. T. (1994), "Estimation of a Non-Neutral Stochastic Frontier Production Function." *Journal of Productivity Analysis* 5, 171-180.

Huang, River H. C. (2004), "Estimation of Technical Inefficiencies with Heterogeneous Technologies." *Journal of Productivity Analysis* 21, 277-296

Giannakas, K., Tran, K. C. and Tzouvelekas, V. (2003), "On the Choice of Functional Form in Stochastic Frontier Modeling." *Empirical Economics* 28, 75-100.

Greene, W. H. (1990), "A Gamma-Distributed Stochastic Frontier Model." *Journal of Econometrics* 46, 141-163.

Greene, W. H. (2004), "Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems." *Health Economics* 13, 959-980.

Greene, W. H. (2005), "Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model." *Journal of Econometrics* 126, 269-303.

Greene, W. H. and Segal, D. (2004), "Profitability and Efficiency in the U.S. Life Insurance Industry." *Journal of Productivity Analysis* 21, 229-247.

Griffin, J. E. and Steel, M. F. J. (2003), "Flexible Mixture Modelling of Stochastic Frontiers." Working paper.

Griffin, J. E. and Steel, M. F. J. (2004), "Semiparametric Bayesian Inference for Stochastic Frontier Models." *Journal of Econometrics* 123, 121-152.

Koop, G. (2004), "Modelling the Evolution of Distributions: An Application to Major League Baseball." *Journal of the Royal Statistical Society A* 167, 639-656.

Koop, G., Osiewalski, J. and Steel, M. F. J. (1994), "Bayesian Efficiency Analysis with a Flexible Form: The AIM Cost Function." *Journal of Business & Economic Statistics* 12, 339-346.

Koop, G. and Poirier, D. (2004), "Bayesian Variants of Some Classical Semiparametric Regression Techniques." *Journal of Econometrics* 123, 259-282.

Kumbhakar, S. C. and Tsionas, E. G. (2005), "Measuring Technical and Allocative Inefficiency in the Translog Cost System: A Bayesian Approach." *Journal of Econometrics* 126, 355-384.

Meeusen, W. and van den Broeck, J. (1977), "Efficiency Estimation from Cobb-Douglas Productions with Composed Errors." *International Economic Review* 8, 435-444.

Park, B. U., Sickles, R. C. and Simar, L. (1998), "Stochastic Panel Frontiers: A Semiparametric Approach." *Journal of Econometrics* 84, 273-301.

Park, B. U. and Simar, L. (1994), "Efficient Semiparametric Estimation in a Stochastic Frontier Model." *Journal of the American Statistical Association* 89, 929-936.

Sickles, R. C., Good, D. H. and Getachew, L. (2002), "Specification of Distance Functions using Semi- and Nonparametric Methods with an Application to the Dynamic Performance of Eastern and Western European Air Carriers." *Journal of Productivity Analysis* 17, 133-155.

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

Tsionas, E. G. (2002), "Stochastic Frontier Models with Random Coefficients." *Journal of Applied Econometrics* 17, 127-147.

Tsionas, E. G. and Kumbhakar, S. C. (2004), "Markov Switching Stochastic Frontier Model." *Econometrics Journal* 7, 398-425.

Wang, H. J. (2003), "A Stochastic Frontier Analysis of Financing Constraints on Investment: The Case of Financial Liberalization in Taiwan." *Journal of Business & Economic Statistics* 21, 406-419.

Wang, H. J. (2002), "Heteroscedasticity and Non-Monotonic Efficiency Effects of a Stochastic Frontier Model." *Journal of Productivity Analysis* 18, 241-253.

Zhu, S., Ellinger, P. N. and Shumway, C. R. (1995), "The Choice of Functional Form and Estimation of Banking Inefficiency." *Applied Economics Letters* 2, 375-379.

Table 1: Parametric vs Semiparametric (Partially Linear) Models

	The Parametric Results						
	Mean	Std	Median	2.5%	5%	95%	97.5%
`constant`	−7.4485	0.3454	−7.4542	−8.1217	−8.0099	−6.8629	−6.7462
$\ln q_i$	0.4210	0.0441	0.4211	0.3350	0.3482	0.4926	0.5063
$(\ln q_i)^2$	0.0298	0.0029	0.0299	0.0242	0.0251	0.0347	0.0355
$\ln(p_l/p_f)$	0.2498	0.0646	0.2506	0.1187	0.1426	0.3560	0.3783
$\ln(p_k/p_f)$	0.0503	0.0624	0.0491	−0.0708	−0.0506	0.1535	0.1770
σ^2	0.0140	0.0043	0.0133	0.0072	0.0080	0.0217	0.0233
θ	13.7278	6.9475	11.3832	7.1908	7.6161	29.8576	34.9799
	The Semiparametric Results						
	Mean	Std	Median	2.5%	5%	95%	97.5%
$f(\ln q_i)$	Figure 2						
$\ln(p_l/p_f)$	0.2603	0.0121	0.2645	0.2261	0.2355	0.2728	0.2755
$\ln(p_k/p_f)$	−0.0109	0.0279	−0.0118	−0.0652	−0.0550	0.0333	0.0467
σ^2	0.0003	0.0007	0.0001	0.0000	0.0000	0.0015	0.0022
θ	6.2341	0.7328	6.1916	4.9394	5.1188	7.5080	7.8363

* The posterior means and posterior standard deviations are obtained using 10,000 simulated draws after discarding the first 10,000 variates to mitigate the start-up effect.

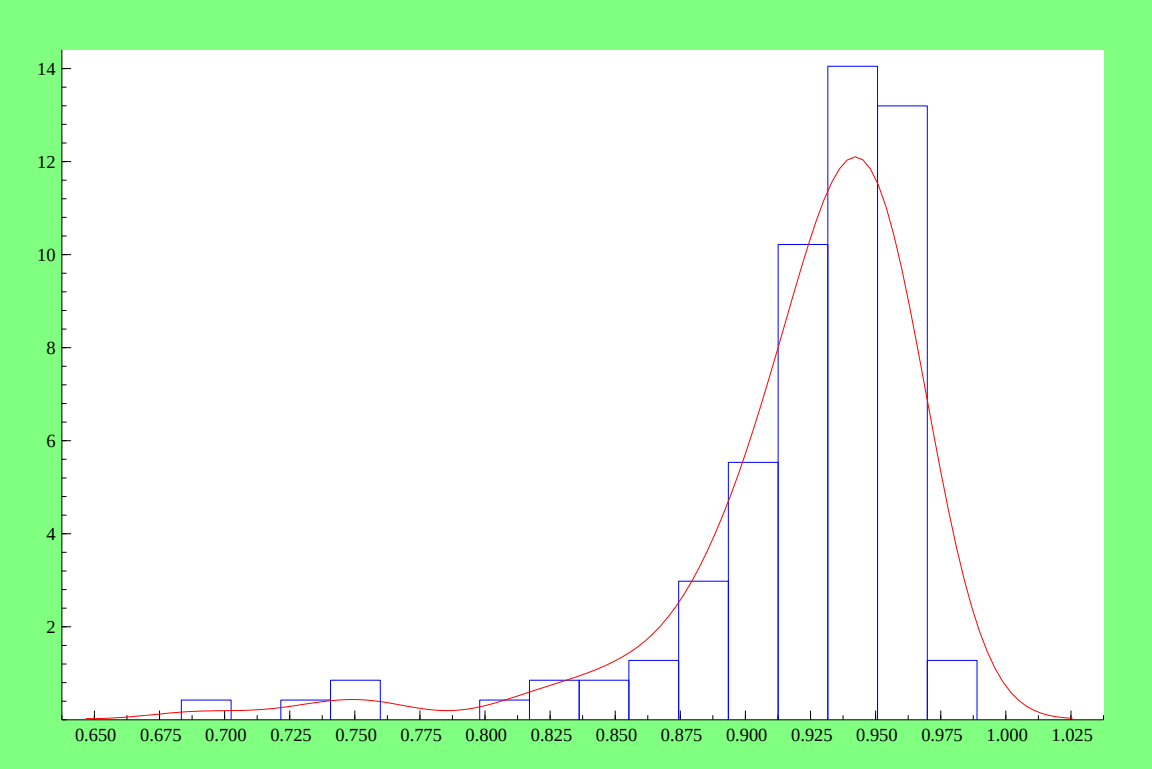

Figure 1: The efficiency distribution from the parametric specification.

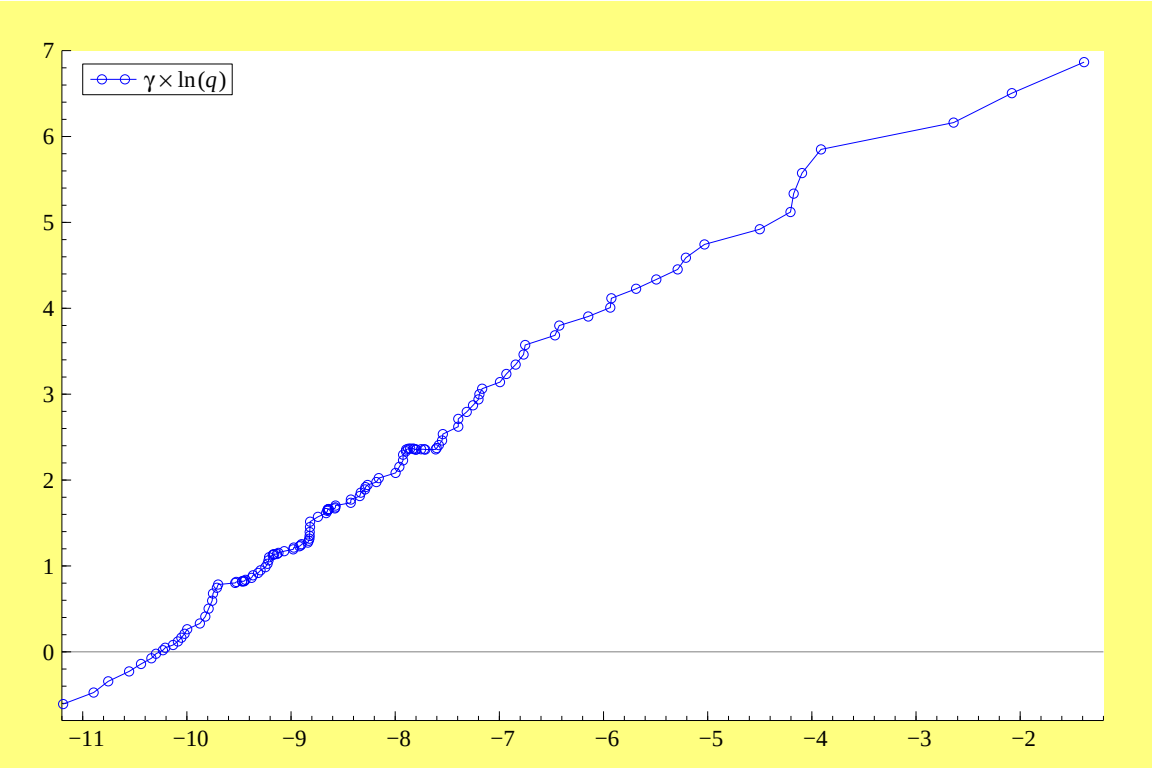

Figure 2: The estimated nonparametric relationship between $\ln(c/p_f)$ and $\ln(q)$.

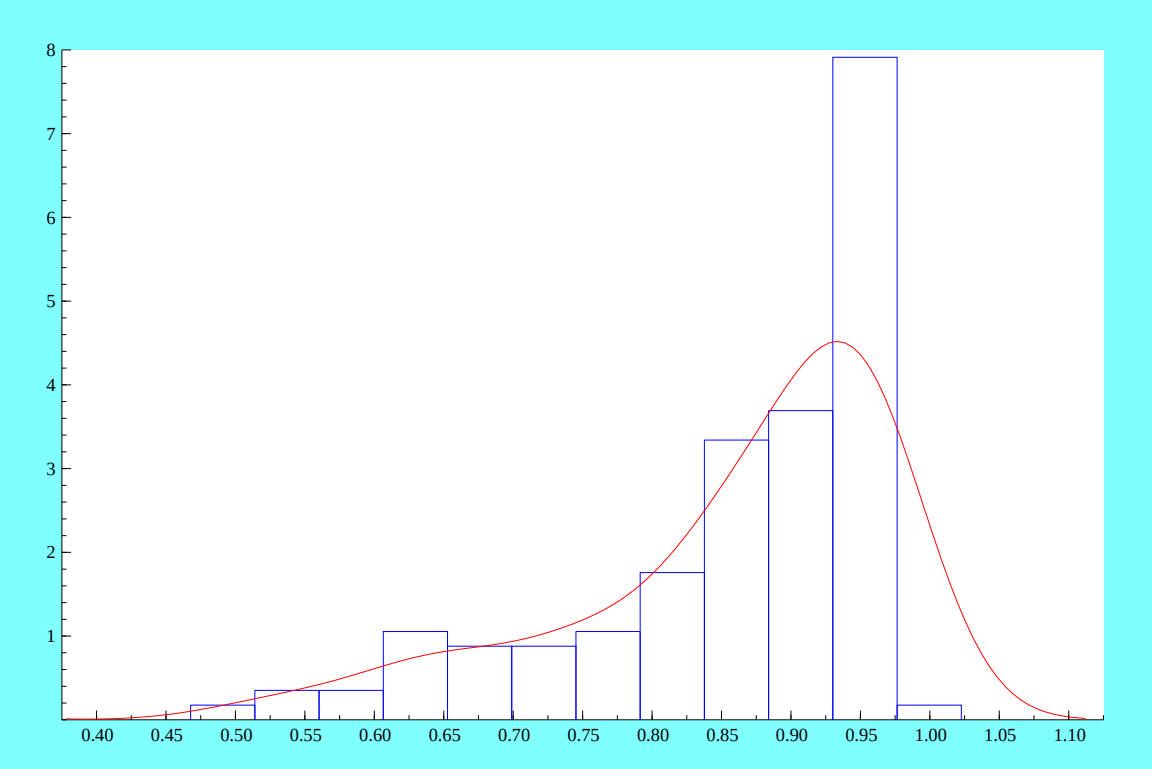

Figure 3: The efficiency distribution from the semi-parametric specification.