Centre for Efficiency and Productivity Analysis

Working Paper Series
No. 04/2003

Title
Semiparametric Estimation of Stochastic Frontiers
A Bayesian Penalized Approach

Authors
Gholamreza Hajargasht

Date:

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

Semiparametric Estimation of Stochastic Frontiers
A Bayesian Penalized Approach

Gholamreza Hajargasht*

School of Economics,
University of Queensland,
St Lucia QLD 4072, Australia

ABSRACT

Almost all previous approaches to estimating semiparametric frontier models, where the functional form for the production (cost) function is unknown, have been local nonparametric (ie. kernel) approaches. In this paper we use a penalized (ie. spline) approach. We show how this approach can be applied to a variety of frontier models, including panel models with fixed and random effects, within a Bayesian framework. We also apply our approach to different multivariate settings, including additive and additive with interaction models. The latter is a promising model because it is very flexible and does not suffer the severe curse of dimensionality problem common with fully nonparametric functions. We illustrate our method using a simulated example.

* I'm indebted to my supervisors, Prof. Prasada Rao and Dr. Chris O'Donnell, for their support and great suggestions.

1. Introduction

DEA and stochastic frontier analysis are the two approaches commonly used to estimate frontier functions and efficiency. DEA models are considered nonparametric, which means there is no need to specify a functional form for the production function, and they are usually non-stochastic, which means the effect of noise and random errors are ignored (or measured as inefficiency effects). Stochastic frontier models take noise and random errors into account but they are usually parametric – we have to specify functional forms for the production functions and inefficiency distributions.

There have been some efforts to relax parametric assumptions in stochastic frontier models. Some studies like, Park et. al. (1994, 1998), Horrace (2001) and Griffin & Steel (2002), have focused on the estimation of a frontier model with a parametric (linear) production function but unknown functional form for the inefficiency distribution. In other studies the focus has been on the estimation of a stochastic frontier model with unknown functional form for the production function. Fan et. al. (1995) provided a two stage pseudo maximum likelihood approach to estimate such a model. The approaches of Kneip & Simar (1996) and Henderson (2002) are only applicable to models with panel data. Adams et. al. (1999) estimate a stochastic panel frontier model relaxing the parametric assumption on the inefficiency distribution and on a subgroup of regressors in a multi-output distance function. Kumbhakar & Tsionas (2002) have used a local likelihood approach to estimate a frontier model. Their model could be called a fully nonparametric model because both the parameters of the production function and the inefficiency distribution remain fixed only in a local neighbourhood.

In this paper we focus on the estimation of a stochastic production (cost) frontier with unknown functional form (we make the usual parametric assumptions about the inefficiency distribution). In contrast to above studies we use a penalized (ie. spline) approach to nonparametric estimation rather than a local (ie. kernel) approach. Unlike the local approaches, it is straightforward to apply our approach to different stochastic frontier models, including models with panel data, fixed and random effects models, and even the "true fixed effect" model of Greene (2002). It is also easier to impose

economic regularity restrictions using our approach compared to kernel-based approaches.

Another novelty of this paper is the use of an additive function with interactions (Sperlich et. al. 2002) in the stochastic frontier. The additive model with interactions is a very flexible model; it is a generalization of some important functional forms like the translog; and the curse of dimensionality problem is not severe compared to a fully nonparametric function.

Although it is possible to estimate our model using a penalized maximum likelihood method, we use a fully Bayesian approach. There are some arguments in favour of the Bayesian approach to frontier model estimation. As Koop and Steel (2002, p.525) claim, the theoretical justification for point and interval estimates of inefficiencies based on maximum likelihood is not strong, but the Bayesian approach provides finite sample distributions for firm inefficiencies, and that allows us to obtain point and interval estimates easily. It is also possible to impose curvature restrictions using a Bayesian approach (eg. Cuesta et. al. 2001). Another reason for using a Bayesian approach here is that the Bayesian approach to nonparametric estimation yields the smoothing parameters automatically. It is also easier to implement a Gibbs sampling algorithm rather than maximize a complicated likelihood function with many parameters.

We begin the paper in Section 2 with an introduction to the Bayesian approach to parametric frontiers, and we show that the analysis can be extended to a univariate semiparametric model using an appropriate prior borrowed from the Bayesian nonparametric literature. In Section 3 we generalize the analysis to multivariate additive and partially linear additive models. We discuss semiparametric fixed and random effects models in Section 4. There are different approaches to Bayesian nonparametrics, and our methodology is independent of the approach used. We explain one of the popular approaches, called smoothing splines, in Section 5. In Section 6 we discuss the estimation of more elaborate multivariate models using P-splines. In Section 7 we apply our method to a simulated example. In section 8 the proposed method is applied to a real data example.

This paper is heavily based on the Bayesian approach to parametric frontier estimation, and also Bayesian semiparametric estimation with splines. We refer the reader to Koop & Steel (2001) and Tsionas (2001) for details on the Bayesian approach to estimating frontiers, and Hastie & Tibshirani (1990, 1998), Green & Silverman (1994), Eubank (1999), Fahrmeier (2000), Koop & Poirier (2001), Ruppert & Carroll (2000) and Berry et. al. (2002) for additional information on spline and Bayesian semiparametric estimation.

2. The Bayesian Approach to Estimating Univariate Semiparametric Frontiers

We start this section with a brief review of the Bayesian approach to parametric frontier estimation, and then we extend it to univariate nonparametric functions. The stochastic frontier model may be specified as follows:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} - z_i + \varepsilon_i \qquad (1)$$

where y represents the log of output, **x** is a vector of inputs in logs, z represents inefficiency effects and ε is a random error. The subscript i=1,2,…,n indexes firms. The following parametric assumptions are made in the specification of the above model:

1) The production function is linear in the parameters.
2) z_i has an known distribution i.e. exponential[1] with parameter γ^{-1}
3) u_i is distributed as $N(0,\sigma^2)$

In a fully Bayesian approach, our aim is to obtain the posterior $p(\beta,\sigma^{-2},\mathbf{z},\gamma^{-1}\,|\,\mathbf{y},\mathbf{x})$. According to Bayes's theorem we can write:

$$p(\boldsymbol{\beta},\sigma^{-2},\mathbf{z},\gamma^{-1}\,|\,\mathbf{y},\mathbf{x}) \propto p(\mathbf{y}\,|\,\boldsymbol{\beta},\mathbf{z},\sigma^{-2})\,p(\sigma^{-2})\,p(\mathbf{z}\,|\,\gamma^{-1})\,p(\gamma^{-1})\,p(\boldsymbol{\beta})$$

It can easily be shown that $p(\mathbf{y}_i \mid \beta,\mathbf{z},\sigma^{-2}) = N(\mathbf{x}_i\beta - z_i,\sigma^2)$ and $p(z_i \mid \gamma^{-1}) \propto \gamma^{-1}\exp(-\gamma^{-1}z_i)$. Fernandez et. al. (1997) have shown that the posterior distribution is not well defined when the usual non-informative priors are assumed for

[1] - Other distributions like half normal, truncated normal and gamma can also be used.

σ^{-2} and γ^{-1}. They have proposed the following gamma prior[2] for these parameters: $p(\sigma^{-2}) \sim G(a_0, b_0), p(\gamma^{-1}) \sim G(a,b)$. It has also been shown that a proper or bounded prior in the form of $p(\boldsymbol{\beta}) \propto I(E)$, where $I(E)$ is the indicator function for the economic regularity conditions, is a good prior for $\boldsymbol{\beta}$. Here we assume that $I(E)$ is equal to one for all values of $\boldsymbol{\beta}$.[3]

With the above information, and using the assumption that z_i and ε_i are iid, we can obtain the following posterior:

$$p(\boldsymbol{\beta}, \sigma^{-2}, \mathbf{z}, \gamma^{-1} \mid \mathbf{y}, \mathbf{x}) \propto \sigma^{-2(n/2+a_0-1)} \exp\left\{-\frac{2b_0 + \sum_{i=1}^{n} \{y_i + z_i - \mathbf{x}_i\beta)\}^2}{2\sigma^2}\right\} \gamma^{-(a+n-1)} \exp\left\{-(b + \sum_{i=1}^{n} z_i)\gamma^{-1}\right\}$$

For further inference we must be able to draw from the above density. But this posterior is not a standard one and we can not draw from it directly. However, we can derive the following conditional distributions:

$$\boldsymbol{\beta} \mid \mathbf{z}, \sigma^{-2}, \gamma^{-1} \sim N\{\mathbf{S}(\mathbf{y} + \mathbf{z}), \sigma(\mathbf{x}'\mathbf{x})^{-1}\} \text{ where } \mathbf{S} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$$

$$\sigma^{-2} \mid \mathbf{z}, \boldsymbol{\beta}, \gamma^{-1} \sim G\left\{a_0 + n/2, b_0 + \frac{(\mathbf{y} + \mathbf{z} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} + \mathbf{z} - \mathbf{x}\boldsymbol{\beta})}{2}\right\}$$

$$z_i \mid \sigma^{-2}, \boldsymbol{\beta}, \gamma^{-1} \sim N\{\mathbf{x}_i\boldsymbol{\beta} - y_i - \gamma^{-1}\sigma^2, \sigma^2\}, z_i \geq 0$$

$$\gamma^{-1} \mid \mathbf{z}, \sigma^{-2}, \boldsymbol{\beta} \sim G(n + a, b + \sum_{i=1}^{n} z_i)$$

The above conditional distributions are all standard distributions, and drawing random numbers from them is fairly easy. Specifically, a Gibbs sampler with data augmentation can be set up to generate a sample of values for the parameters. The sample can be used to obtain expectations, standard errors and confidence intervals for the parameters.

Extending the above analysis to semiparametric univariate frontier models is fairly straightforward. Let the semiparametric univariate stochastic frontier model be defined as:

---

[2] - We define the Gamma distribution as $p(z \mid a, b) = \frac{b^a}{\Gamma(a)} z^{a-1} e^{-bz}$

[3] - We can impose economic conditions like curvature restrictions by letting I(E) be one for those values of $\boldsymbol{\beta}$ which satisfy the restriction and zero otherwise.

$$y_i = f(x_i) - z_i + \varepsilon_i \qquad (2)$$

The difference between (1) and (2) is that we now have an unknown form for regression function *f*. If we can specify an appropriate prior for $\mathbf{f} = \{f(x_1), f(x_2), \ldots, f(x_n)\}$, it will be possible to set up a Bayesian approach similar to the parametric case. If we choose a non-informative prior we a get a function which interpolate the data and that's not we expect from a regression estimation, we usually expect that the estimated function satisfies some degree of smoothness In the Bayesian semiparametric literature the following prior has been proposed for f:

$$\mathbf{f} \sim N(0, \mathbf{K}^{-1}\tau^2)$$

where **K** is an n by n matrix defined differently in different approaches to Bayesian nonparametrics[4]. The above prior penalizes the roughness of *f* – it reflects our prior belief that the estimated *f* must not be too rough or wiggly. We don't specify the matrix **K** here – we will come back to **K** when discussing smoothing splines and P-splines in the next sections. For the moment, we have **f** and an extra parameter τ. Then using Bayes's Theorem we can write the posterior as:

$$p(\mathbf{f},\sigma^{-2},\tau^{-2},\mathbf{z},\gamma^{-1} \mid \mathbf{y},\mathbf{x}) \propto p(\mathbf{y} \mid \mathbf{f},\mathbf{z},\sigma^{-2})\, p(\mathbf{f} \mid \tau^{-2})\, p(\sigma^{-2})\, p(\tau^{-2})\, p(\mathbf{z} \mid \gamma^{-1})\, p(\gamma^{-1})$$

The first term on the right hand side can be derived easily as $p(\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{z}, \sigma^{-2}) \sim N(f(x_i) - z_i, \sigma^2)$. For the second term, **f**, we use the prior discussed above. For σ^{-2}, τ^{-2} and γ^{-1} we again use gamma priors such that $p(\sigma^{-2}) \sim G(a_0, b_0)$, $p(\tau^{-2}) \sim G(a_1, b_1)$, and $p(\gamma^{-1}) \sim IG(a, b)$. Then our posterior can be written as

$$p(\mathbf{f},\sigma^{-2},\tau^{-2},\mathbf{z},\gamma^{-1} \mid \mathbf{y},\mathbf{x}) \propto \sigma^{-2(n/2+a_0-1)} \exp\left\{-\frac{2b_0 + \sum_{i=1}^{n}\{y_i + z_i - f(x_i)\}^2}{2\sigma^2}\right\} \tau^{-2(a_1-1+(n-2)/2)} \exp\left\{-\frac{2b_1 + \mathbf{K'fK}}{2\tau^2}\right\} \gamma^{-(a+n-1)} \exp\left\{-(b + \sum_{i=1}^{n} z_i)\gamma^{-1}\right\}$$

Again, the above density function is not a standard one and so we can't draw directly from it. However, we can easily obtain the following conditional distributions:

---

[4] There are a number of approaches to Bayesian nonparametric estimation: smoothing spline, regression spline, P-spline, Koop and Poirier's approach, Fahrmeier's approach. We will discuss smoothing spline and P-spline in more detail later in this paper.

[6]-The additive regression model has been discussed in detail in Hastie and Tibshirani (1990). The standard frequentist method for estimation of additive regression is backfitting.

$$\mathbf{f} \mid \mathbf{z}, \sigma^{-2}, \tau^{-2}, \gamma^{-1} \sim N\{\mathbf{S}(\mathbf{y}+\mathbf{z}), \sigma\mathbf{S}^{1/2}\} \text{ where } \mathbf{S} = \{\mathbf{I} + (\sigma/\tau)\mathbf{K}\}^{-1}$$

$$\sigma^2 \mid \mathbf{f}, \mathbf{z}, \tau^{-2}, \gamma^{-1} \sim G\left\{a_0 + n/2, b_0 + \frac{(\mathbf{y}+\mathbf{z}-\mathbf{f})'(\mathbf{y}+\mathbf{z}-\mathbf{f})}{2}\right\}$$

$$\tau^{-2} \mid \mathbf{f}, \mathbf{z}, \sigma^{-2}, \gamma^{-1} \sim G\left\{a_1 + (n-2)/2, b_1 + \frac{\mathbf{f}'\mathbf{K}\mathbf{f}}{2}\right\}$$

$$z_i \mid \mathbf{f}, \sigma^{-2}, \tau^{-2}, \gamma^{-1} \sim N\{f(x_i) - y_i - \gamma^{-1}\sigma^2, \sigma^2\}, z_i \geq 0$$

$$\gamma^{-1} \mid \mathbf{f}, \mathbf{z}, \sigma^{-2}, \tau^{-2} \sim G\left(n + a, b + \sum_{i=1}^{n} z_i\right)$$

A Gibbs sampler with data augmentation can be set up by sequentially drawing from the above conditional distributions. Notice that we don't need to use anything different from the parametric case; all that we need to do is generating random numbers from truncated normal and gamma distributions.

3. Extension to Additive and Partially Linear Additive Models

In this section we extend the previous analysis to two special multivariate functions: the additive and partially linear additive functions. We will discuss less restrictive multivariate models later in Section 6.

The stochastic frontier model with an additive production structure can be written as[6]

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) - z_i + \varepsilon_i$$

A tailor-made Gibbs sampling algorithm, which can be interpreted as a "Bayesian backfitting" approach (Hastie and Tibshirani, 1998), can be setup for Bayesian estimation of additive models. Assuming $\mathbf{f}_j \sim N(0, \mathbf{K}_j^{-}\tau_j^2)$ as a prior distribution for $\mathbf{f_j}$ we can setup our Gibbs sampling procedure by drawing from the following conditional distributions.

$$\mathbf{f}_j \mid \mathbf{y}, \mathbf{z}, \sigma^{-2}, \tau_j^{-2} \sim N\{\mathbf{S}_j(\mathbf{y} - \sum_{i \neq j} \mathbf{f}_i(\mathbf{x}_i) + \mathbf{z}), \sigma \mathbf{S}_j^{1/2}\}$$

$$\sigma^{-2} \mid \mathbf{y}, \mathbf{f}, \mathbf{z} \sim G\left\{a_0 + n/2, b_0 + \frac{(\mathbf{y} + \mathbf{z} - \sum_{j=1}^{p} \mathbf{f}_j)'(\mathbf{y} + \mathbf{z} - \sum_{j=1}^{p} \mathbf{f}_j)}{2}\right\}$$

$$\tau^{-2} \mid \mathbf{f}, \mathbf{z}, \sigma^{-2}, \gamma^{-1} \sim G\left\{a_j + (n-2)/2, b_j + \frac{\mathbf{f}_j' \mathbf{K}_j \mathbf{f}_j}{2}\right\}$$

$$z_i \mid f, y, \sigma^{-2}, \gamma^{-1} \sim N\{f(x_i) - y_i - \gamma^{-1}\sigma^2, \sigma^2\}, z_i \geq 0$$

$$\gamma^{-1} \mid z \sim G(n + a, b + \sum_{i=1}^{n} z_i)$$

The above Bayesian analysis can be easily extended to the following partially linear additive model where $\mathbf{w}_i$ is vector of variables that is related to y_i in a linear fashion.

$$y_i = \mathbf{w}_i \boldsymbol{\beta} + \sum_{j=1}^{p} f_j(x_{ij}) - z_i + \varepsilon_i$$

4. Bayesian Semiparametric Frontiers with Random and Fixed Effects

It is increasingly common to use panel data in stochastic frontier analysis. Two different models – the fixed effects and random effects models – have been proposed for stochastic frontier models with panel data. The Bayesian approach to parametric fixed and random effects estimation has been discussed in Koop and Steel (2001). Here we describe the semiparametric Bayesian alternative to random and fixed effect models.

Consider the following random effects model for a (possibly unbalanced) panel data set

$$y_{it} = f(x_{it}) - z_i + \varepsilon_{it}$$
$$t = 1,2,....T_i, i = 1,2,...n$$
$$z_i \sim Gamma(1, \gamma^{-1}), \varepsilon_{it} = N(0, \sigma_\varepsilon^2)$$

Assuming an inverse gamma prior for σ^2, τ^2 and γ as before, it is not difficult to show that our posterior will be

$$p(\mathbf{f},\sigma^{-2},\tau^{-2},\mathbf{z},\gamma^{-1}\,|\,\mathbf{Y},\mathbf{x}) \propto \sigma^{-2(nT/2+a_0-1)} \exp\left\{-\frac{2b_0+\sum_{t=1}^{T_i}\sum_{i=1}^{n}\{y_{it}+z_i-f(x_{it})\}^2}{2\sigma^2}\right\} \tau^{-2(a_1-1+(n-2)T/2)} \exp\left\{-\frac{2b_1+\mathbf{f}'\mathbf{K}\mathbf{f}}{2\tau^2}\right\} \gamma^{-(a+n-1)} \exp\left\{-\gamma^{-1}(b+\sum_{i=1}^{n} z_i)\right\}$$

where $T = \sum_{i=1}^{n} T_i$. This posterior distribution can be used for further inference using a Gibbs sampler. The conditional distributions are:

$$\mathbf{f} \mid \mathbf{z}, \sigma^{-2}, \tau^{-2}, \gamma^{-1} \sim N\{\mathbf{S}(\mathbf{Y}+\mathbf{z}), \sigma\mathbf{S}^{1/2}\}$$

$$\sigma^{-2} \mid \mathbf{z}, \tau^{-2}, \mathbf{f}, \gamma^{-1} \sim G\left\{a_0 + nT/2, b_0 + \frac{\sum_{t=1}^{T_i}\sum_{i=1}^{n}\{y_{it}+z_i-f(x_{it})\}^2}{2}\right\}$$

$$\tau^{-2} \mid \mathbf{z}, \sigma^{-2}, \mathbf{f}, \gamma^{-1} \sim G\left\{a_1 + (n-2)T/2, b_1 + \frac{\mathbf{f}'\mathbf{K}\mathbf{f}}{2}\right\}$$

$$z_i \mid \sigma^{-2}, \tau^{-2}, \mathbf{f}, \gamma^{-1} \sim N\left(\frac{\sum_{t=1}^{T_i}[f(x_{it}) - y_{it} - \gamma^{-1}\sigma^2]}{T_i} - T_i\gamma^{-1}\sigma^2, T_i\sigma^2\right), z_i \geq 0$$

$$\gamma^{-1} \mid \mathbf{z}, \sigma^{-2}, \tau^{-2}, \mathbf{f}, \sim G(n+a, b+\sum_{i=1}^{n} z_i)$$

Note that for posterior analysis we only need to draw random numbers from truncated normal and Gamma distributions. The extension of the above panel data model to the multivariate additive case within a Gibbs sampler is straightforward.

The fixed effects model can be written as the following partially linear model:

$$y_{it} = \sum_{i=1}^{n-1} \alpha_i D_i + f(x_{it}) + \varepsilon_{it}$$

where D_i represents the dummy variable associated with the i-th firm. This model is nothing more than a partially linear regression model and can be estimated using Gibbs sampling easily.

5. Smoothing splines and Bayesian semiparametrics

Consider the following univariate estimation problem:

$$y_i = f(x_i) + \varepsilon_i \ ; \ \varepsilon_i \sim N(0, \sigma^2)$$

where the functional form for f is unknown. One method of estimating such a model is by minimizing of following penalized sum of square criterion

$$J(f) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(x)\}^2 dx$$

over all functions *f(x)* such that the integral exists. The integral represents a roughness penalty and λ is the smoothing parameter. Larger values of λ result in a smoother curve. The solution function *f* to the above minimization problem has been shown to be a natural cubic spline with knots at each of the unique values of x_i. Let $\mathbf{f} = \{f(x_1), f(x_2), \ldots, f(x_n)\}$. Then using the cubic spline nature of *f* it can be shown that the penalty term can be written as $\lambda \int \{f''(x)\}^2 dx = \lambda \mathbf{f}' \mathbf{K} \mathbf{f}$, where **K** is an n by n matrix of rank n-2 and is defined in Green and Silverman (1994, pp13). Then using matrix algebra it is easy to show that the smoothing spline minimizer *J*(*f*) is $\mathbf{f} = \mathbf{S}(\lambda)\mathbf{y}$ where $\mathbf{S}(\lambda) = (\mathbf{I} + \lambda \mathbf{K})^{-1}$ and $\mathbf{y} = \{y_1, y_2, \ldots . y_n\}$.

In the Bayesian approach to smoothing splines, the following partially improper Gaussian prior is given to f:

$$\mathbf{f} \sim N(0, \mathbf{K}^{-} \tau^2)$$

where $\mathbf{K}^{-}$ is the generalized inverse of **K** and $\tau^2 = \sigma^2 / \lambda$. Knowing **K**, we can do Bayesian analysis of the stochastic frontier model as discussed in Sections 2 and 3. Generalizations to additive and partially linear additive multivariate production functions are straightforward within the framework discussed in the previous sections, but for less restrictive multivariate production functions we use a P-spline approach.

6. P-splines and multivariate stochastic frontiers

So far we have proposed a Bayesian estimation method for a frontier model with additive production structure, which is not very flexible. The purpose of this section is to extend our analysis to less restrictive forms of multivariate frontier models.

We can consider a fully nonparametric multivariate function, but as is well known in the nonparametric econometrics literature, there is a curse of dimensionality with a multivariate nonparametric function. However, there is a form of multivariate production function called an additive with interaction model, which is very flexible and is a generalization of some important functions like the translog, generalized Leontief and quadratic functional forms. This form can be written as[7]:

$$f(x_1, x_2, ...x_p) = \alpha + \sum_{i=1}^{p} f_i(x_i) + \sum_{i<j}^{p} f_{ij}(x_i, x_j)$$

The rest of this section considers Bayesian methods to estimate a multivariate frontier model with the above structure. We use a P-spline estimator, which is simpler than smoothing splines when estimating these models.

There are two general approaches to spline fitting – smoothing spline and regression splines. Smoothing splines use all the observations as knots. Consequently, when the number of observations is large they become computationally impractical, and generalizing them to multivariate function estimation (except for the additive and partially linear model) is not straightforward. We can fit regression splines using ordinary least squares once the knots have been selected, but knot selection procedures are complicated and computationally intensive (Smith and Kohn 1996). P-splines, introduced by Eilers & Max (1996) and Ruppert & Carroll (2000), combine features of smoothing splines and regression splines in such a way that, unlike regression splines, the locations of knots are not crucial, and they have far fewer parameters than smoothing splines. The following discussion is based on Ruppert and Carroll's introduction to P-splines, and we refer the reader to their papers for more information.

---

[7] - For identification purposes we need to put some restriction on components of the interaction models . This has been discussed in Sperlich & et al. (2002) and Chen (1993). Here because we are mostly interested in estimation of efficiencies we don't discuss them in more detail.

Suppose we want to estimate the following nonparametric model:

$$y_i = f(x_i) + \varepsilon_i$$

Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_{k+2})'$ and consider the following regression spline model[8]

$$f(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \beta_{1+k} (x - \kappa_k)_+$$

where $(u)_+ = uI(u \geq 0)$ and $\kappa_1 < ..... < \kappa_K$ are fixed knots. In the P-spline approach we allow K to be large and fixed, but we put a penalty on the $\{\beta_{k+1}\}_{k=1}^{K}$ (the set of jumps in the derivative of $f(x, \boldsymbol{\beta})$) such that our penalized least square criterion will be

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}' \mathbf{K} \boldsymbol{\beta}$$

where $\mathbf{X}$ is a matrix with $\mathbf{x}_i = (1, x_i, (x_i - \kappa_1)_+, (x_i - \kappa_k)_+)$ it's i-th row, and $\mathbf{K}$ is a diagonal matrix whose first two diagonal elements are 0 and the remaining diagonal elements are 1. Simple calculation shows that the penalized least square minimizer $\boldsymbol{\beta}$ will be

$$\boldsymbol{\beta}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{K})^{-1} \mathbf{X}'\mathbf{y}$$

The extension to additive functions is straightforward. Suppose we have a bivariate additive function of the following form:

$$f(x_1, x_2) = a + f_1(x_1) + f_2(x_2)$$

Then we can write the regression spline function as

$$f(x_1, x_2, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sum_{k=1}^{K} \beta_{2+k} (x_1 - k_k)_+ \sum_{m=1}^{M} \beta_{2+K+m} (x_2 - \mu_m)_+$$

Define $\mathbf{X} = \mathbf{(X(1), X(2), X(3))}$ where $\mathbf{X(1)} = \mathbf{(1, x_1, x_2)}$, $\mathbf{X(2)} = \mathbf{((x_1 - k_1)_+, ...(x_1 - k_K)_+}$

[8] - Here for ease of illustration we use a linear basis. Other basis like polynomial and B-spline should be used in real practice.

and $\mathbf{X(3)} = ((\mathbf{x}_2 - \boldsymbol{\mu}_1)_+, ...(\mathbf{x}_2 - \boldsymbol{\mu}_M)_+)$. Then our penalized least squares criterion will be

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}$$

where K is a block diagonal matrix with blocks $0I_3, \lambda_1 I_K, \lambda_2 I_M$, and I_M is an identity matrix of dimension M (note that we have used different smoothing parameters, λ, for different variables).

The above analysis can be extended to multivariate nonparametric models using tensor product splines. We discuss a bivariate model here, but generalization to the multivariate case is straightforward. Suppose $B(\mathbf{1}) = \{1\} \cup B_p(1) \cup B_{pp}(1)$ is the set of our basis functions where $\mathbf{1}$ is a vector of ones, $B_p(1) = \{\mathbf{x}_1\}$ and $B_{pl}(1) = \{(\mathbf{x}_1 - \kappa_{11})_+, ...(\mathbf{x}_1 - \kappa_{1k_1})_+\}$. The subscripts "l" and "pl" denote "linear" and "piecewise linear". $B(2)$ can also be defined in the same way for variable x_2. The tensor product regression spline basis is defined by $B(1,2) \equiv B(1) \otimes B(2)$, which is the set of all products $b(1)$ and $b(2)$ where $b(1) \in B(1), b(2) \in B(2)$. Let

$$B_{pl}(1,2) \equiv [B_{pl}(1) \cup B(2)] \otimes [B(1) \cup B_{pl}(2)]$$

Then we can write our regression spline as follows

$$f(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{X}(1)\boldsymbol{\beta}_1 + \mathbf{X}(2)\boldsymbol{\beta}_2 + \mathbf{X}(3)\boldsymbol{\beta}_3 + \mathbf{X}(4)\boldsymbol{\beta}_4$$

where $\mathbf{X}(1)$, $\mathbf{X}(2)$, $\mathbf{X}(3)$ and $\mathbf{X}(4)$ are equal to $\mathbf{X}(1) = (1, \mathbf{x}_1, \mathbf{x}_2)$, $\mathbf{X}(2) = B_{pl}(1)$, $\mathbf{X}(3) = B_{pl}(2)$ and $\mathbf{X}(4) = B_{pl}(1,2)$. We can define the penalized least squares criterion by using different smoothing parameters for $\boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$

It is clear that the dimension of the basis grows geometrically with the increase in the number of variables, illustrating the curse of dimensionality. So it might not be practical to estimate a fully nonparametric model for more than two variables. Instead we propose using an interaction model, which is very flexible and where the curse of

dimensionality is not that severe. Without loss of generality, we can write an additive model with interactions as an additive model of bivariate functions as follows

$$f(x_1,....,x_n) = \sum_{i<j} f_{ij}(x_i, x_j)$$

Combining our analysis of additive and bivariate models, we can obtain the regression spline for the interaction model. We see that with an appropriate definition of $\mathbf{X}(p)$, all the above models can be written in following regression spline form

$$f(x_1,....,x_n,\boldsymbol{\beta}) = \sum_{p=1}^{P} \mathbf{X}(p)\boldsymbol{\beta}(p) = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X} = \mathbf{(X(1),...,X}(P))$, $\boldsymbol{\beta}' = (\boldsymbol{\beta}(1),.....,\boldsymbol{\beta}(P))'$.

The above analysis shows that we can write our multivariate frontier model as

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} - \mathbf{z} + \boldsymbol{\varepsilon}$$

which is very similar to the parametric case, the only difference being that here we have to specify a special prior for $\boldsymbol{\beta}$. We use the following prior used in Berry et. al. (2002):

$$\boldsymbol{\beta} \sim N(0, \mathbf{K}^{-})$$

where K is a block diagonal matrix with blocks $0_{p_1}, \tau_1 I_{p_2}, ..., \tau_{p-1} I_{p_p}$. Using this prior we can obtain the associated posterior in the same way as we did in previous sections. Again, the posterior is not a standard one, but we can derive the following conditional distributions and apply a Gibbs sampling for further analysis:

$$\boldsymbol{\beta} \mid \mathbf{z}, \sigma^{-2}, \boldsymbol{\tau}^{-2}, \gamma^{-1} = N\{\mathbf{S}_j(\mathbf{y}) + \mathbf{z}), \sigma \mathbf{S}_j^{1/2}\}$$

$$\sigma^{-2} \mid \mathbf{z}, \boldsymbol{\tau}^{-2}, \boldsymbol{\beta}, \gamma^{-1} \sim G\left\{a_0 + n/2, b_0 + \frac{(\mathbf{y} + \mathbf{z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} + \mathbf{z} - \mathbf{X}\boldsymbol{\beta})}{2}\right\}$$

$$\tau_j^{-2} \mid \mathbf{z}, \sigma^{-2}, \boldsymbol{\beta}, \gamma^{-1} \sim IG\left\{a_j + (m_j - 2)/2, b_j + \frac{\beta_j' \mathbf{K}_j \beta_j}{2}\right\}$$ [9]

$$z_i \mid \sigma^{-2}, \boldsymbol{\tau}^{-2}, \boldsymbol{\beta}, \gamma^{-1} \sim N\{\mathbf{X}_i\boldsymbol{\beta} - y_i - \gamma^{-1}\sigma^2, \sigma^2\}, z_i \geq 0$$

$$\gamma^{-1} \mid \mathbf{z}, \sigma^{-2}, \boldsymbol{\tau}^{-2}, \boldsymbol{\beta} \sim G(n + a, b + \sum_{i=1}^{n} z_i)$$

7. A Simulated Example

In this section we estimate an additive frontier model using simulated data. We generated data using the following model:

$$y_i = \frac{1}{5} x_{i1}^5 e^{-x_{i1}} + x_{2i} - z_i + \varepsilon_i$$

where i =1,2,……..…,100, $x_1 = seq\{1,.04,100\}$, $x_2 \sim uniform(2,4)$, $z_i \sim Gamma(1,5)$, $\varepsilon_i \sim Normal(0,.04)$. We have chosen a very nonlinear functional form for f_1 so that it shows the three well-known stages of production; for f_2 a linear function has been specified.

We used a smoothing spline to estimate above additive model as discussed in Section 3. For Bayesian analysis we need to specify priors for the parameters $a_0, b_0, a_1, b_1, a_2, b_2, a, b$. We used: $a_0 = 1, b_0 = .1, a_1 = 1, b_1 = .1, \; a_2 = 1, b_2 = .1, a = 1, b = .25$. The results are not very sensitive to moderate changes in these priors. Starting values for the parameters were obtained using a simple COLS estimator.

Posterior analysis was based on 10000 realizations; the first 2000 were excluded from final analysis as burn-in period. The means of the posterior samples were used as point estimates of the parameters. We have summarized the results in several graphs. First, the fitted values of f_1 and f_2 have been compared with simulated data on f_1 and f_2 in

[9] - m_j is the dimension of vector B_j

Figure 1. As we see, both seem to fit the real data very well. The dashed line represents the original simulated data and the thick line represents the estimated function.

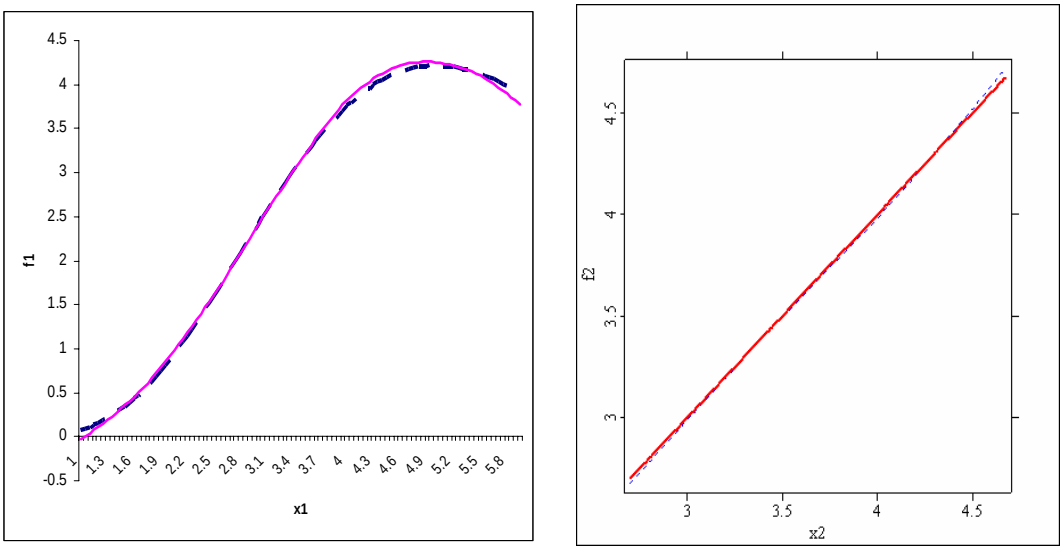


Figure 1. Fitted and simulated values of f_1 and f_2 plotted against x_1 and x_2

Figure 3 compares fitted values of output with original data on output. The estimated output seems to be a proper production frontier insofar as most data are below it.

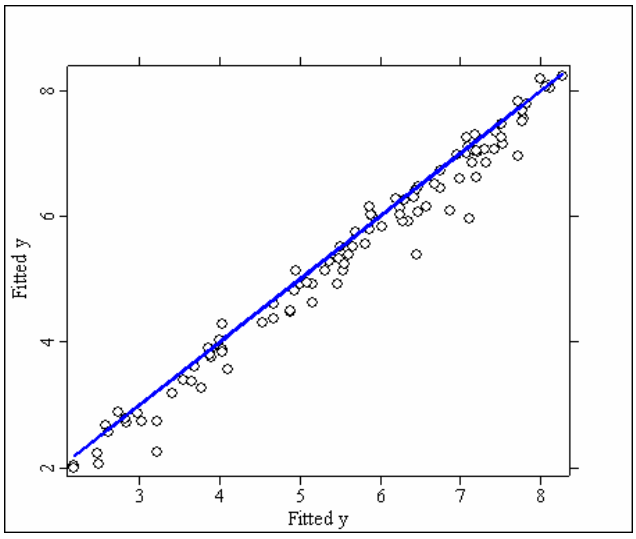


Figure 3. Fitted values compared with real data on output

In Figure 4 we have compared the estimated values of inefficiencies (z_i) with the original data on z_i. As we see, our estimates follow the real data but there are some significant differences. These differences are not unexpected in frontier analysis. If we

estimate a parametric frontier using standard methods we see similar differences between estimated inefficiencies and true inefficiencies.

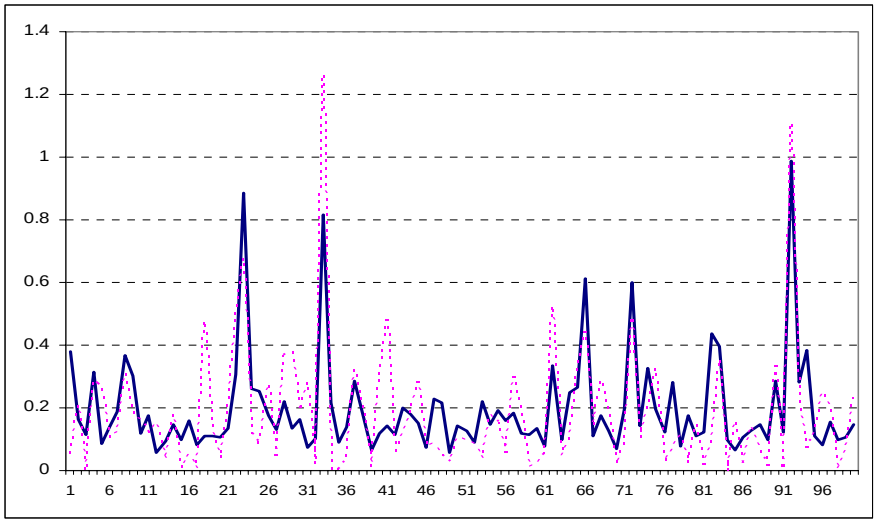


Figure 4. Estimated values of z compared to real data.
The dashed lines show the real data

A Real Data Example

In this section we apply our method to real data[10]. The data set consists of observations on 10 major privately-owned Texas electric utilities observed annually over 18 years from 1966 to 1985, and includes information on labour, capital and fuel (inputs) for electrical power generation (output). This data set has been already used in Kumbhakar (1996), Schmidt and etal. (1996, 1999, 2000).

We assume following random effect additive stochastic frontier

$$y_{it} = f_1(L_{it}) + f_2(K_{it}) + f_3(\mathrm{F}_{it}) - z_i + \varepsilon_{it}$$

where L, K and F represent labour, capital and fuel respectively. A Bayesian P-spline approach is used to estimate the above model. 20 equi-distance points were chosen as knots for each variable. The results of estimation of f_1, f_2, f_3 can be seen in figures 5 to 7. The shape of f_1 is not what we usually expect but pervious studies with the same data set confirm this negative relationship, in other parametric studies (i.e. Schmidt and etal) they have found a negative coefficients for L. for two other inputs the production function has a regular shape and it is not far from being linear.

[10] - The data has been downloaded from Journal of Applied Econometric Data Archive

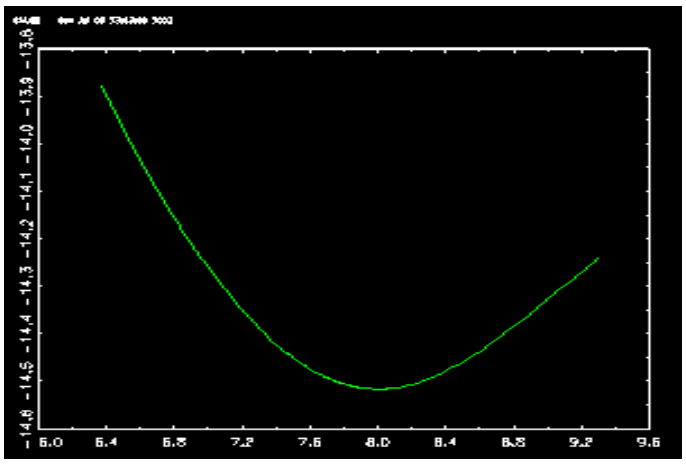


Figure 5 – the graph of f_1 versus L

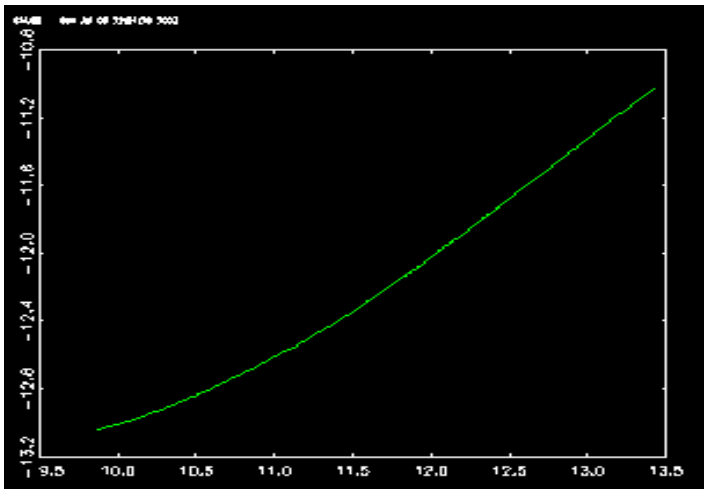


Figure 6 – the graph of f_3 versus F

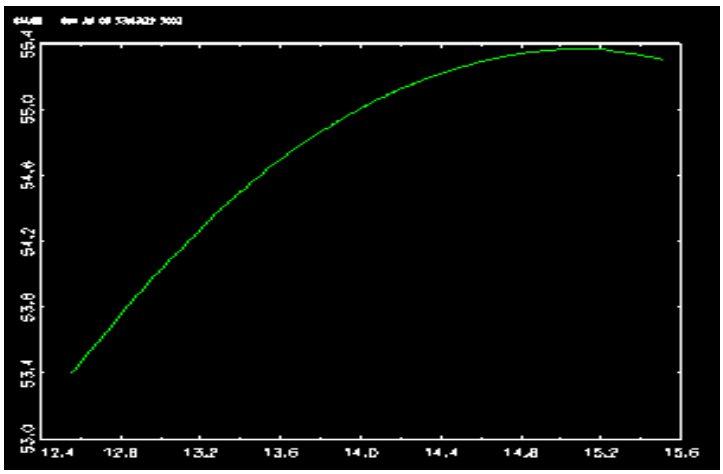
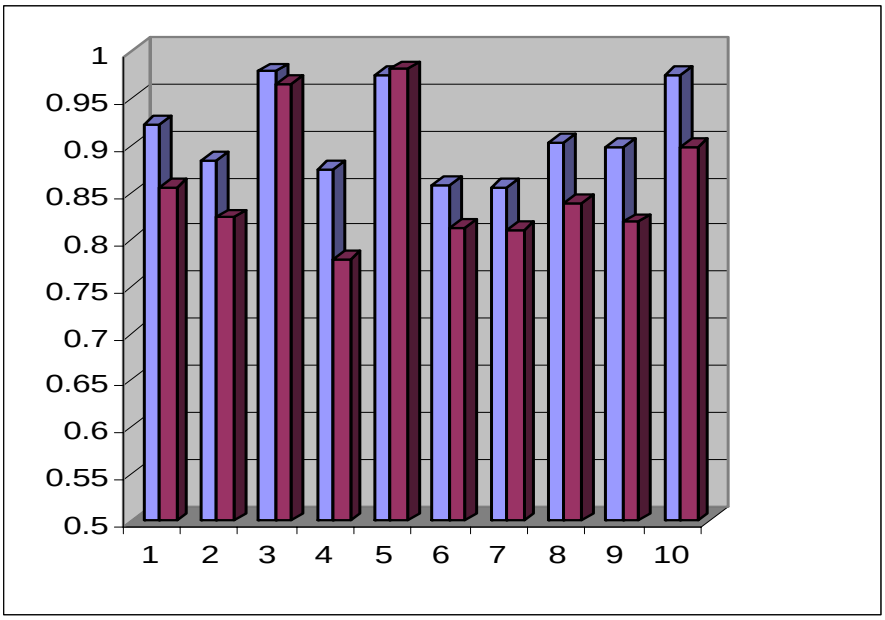


Figure 7 – the graph of f_2 versus K

In figure (8) we compare the nonparametric (calculated based on the method we have proposed) and parametric estimates (drawn from Schmidt and etal. 1999)of efficiencies of ten firms. The blue and red rectangles represent nonparametric and parametric estimates respectively. As we see the pattern is the same for both cases but the nonparametric estimates in this particular example are higher.



8. Conclusion

In this paper we propose a penalized approach to estimating semiparametric stochastic frontiers where the functional form for production (cost) function is assumed unknown. We use a Bayesian approach and show that by borrowing an appropriate prior from the Bayesian nonparametric literature we can easily generalize the Bayesian parametric stochastic frontier to the semiparametric case, we also see that we don't need any special econometric tools other than techniques for drawing from truncated and gamma distributions within a Gibbs sampling set-up. Throughout the paper it is shown that our approach is applicable to different stochastic frontier models including models with cross sections, fixed and random effects and different form of semiparametric multivariate production functions including the interesting case of additive with interaction functions.

References

Adams, R., A. Berger and R. Sickles (1999) Semiparametric approaches to stochastic panel frontiers with applications in the banking industry, *Journal of Business and Economic Statistics*, 17, 349-358.

Aigner, D.J., C.A.K. Lovell and P. Schmidt (1977), Formulation and estimation of stochastic frontier production functions. *Journal of Econometrics*, 6, 21-37.

Chen, Z. (1993), Fitting multivariate regression functions by interaction spline models, *Journal of Royal Statistical Society B*, 55, 473-491

Coelli, T.J., D.S. Rao,., and G.E. Battese (1998), *An introduction to efficiency and productivity analysis*, Kluwer Academic Publishers, Boston.

Cuesta R.A., C. J. O'Donnell, T. J. Coelli and S. Singh (2001), Imposing curvature restriction on production frontier, *CEPA working paper*, University of New England.

Eilers, P. H. C. and Marx, B. D. (1996), Flexible Smoothing With B-splines and Penalties (with discussion). *Statistical Science*, 11, 89-121.

Eubank, R. L. (1999), *Nonparametric regression and spline smoothing* (Second Edition), Marcel Dekker, Inc. New York.

Fernandez, C., J. Osiewalski, and M.F.J. Steel, 1997, On the use of panel data in stochastic frontier models, *Journal of Econometrics* 79: 169-193

Fan, Y., Q. Li and A.Weersink (1996), Semiparametric estimation of Stochastic Production Frontier, *Journal of Business and Economic Statistics*, 14, 460-477.

Green, P.J. and Silverman, B. (1994): *Nonparametric regression and generalized linear models.* Chapman and Hall, London

Griffin, J.E. and M. Steel (2002), Semiparametric Bayesian Inference for Stochastic Frontier Models, www.ukc.ac.uk/IMS/publications/documents/paper_408.pdf

Hastie, T. and Tibshirani, R. (1990): *Generalized additive models.* Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (2000): Bayesian backfitting. *Statistical Science*, 15, 193-223.

Henderson D.J. (2002), Nonparametric kernel measurement of technical efficiency, http://www.geocities.com/djh_ucr/pdffiles/henderson_v8.pdf

Koop, G. and D. Poirier (2002), Bayesian variants of some classical semiparametric regression techniques,

Koop, G., M.F.J. Steel, and J. Osiewalski, (1995), Posterior analysis of stochastic frontier models using Gibbs sampling, *Computational Statistics* 10, 353-373.

Koop G., and M.F.J. Steel, (2001), Bayesian analysis of stochastic frontier models. In *A Companion to Theoretical Econometrics*, Baltagi B. (ed). Blackwell, 520-573.

Kumbhakar S.C. and G. Tsionas (2002), Nonparametric Stochastic Frontier Models, http://www.sinica.edu.tw/~teps/A1-2.pdf

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

Park, B. and L. Simar (1994) Efficient semiparametric estimation in a stochastic frontier model, *Journal of the American Statistical Association*, 89,929-936.

Ruppert, D. and R.J Carroll, (2000), Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, **42**, 205-224.

Smith, M. and Kohn, R. (1996), Nonparametric regression via Bayesian variable selection, *Journal of Econometrics*, 75, 317-344.

Sperlich, S. D. Tjøstheim and L. Yang (2002), *Nonparametric estimation and testing of interaction in additive models, Econometric* Theory, 18, 197-251

Tsionas G. (2002), An Introduction To Efficiency Measurement Using Bayesian Stochastic Frontier Models, *Global Business & Economics Review* 4, 287-311