Tilburg University

Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models

Osiewalski, J.; Steel, M.F.J.

Publication date:
1996

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):

Osiewalski, J., & Steel, M. F. J. (1996). *Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models*. (CentER Discussion Paper; Vol. 1996-03). *Econometrics*.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models¹

Jacek Osiewalski
Department of Econometrics
Academy of Economics
31-510 Kraków, Poland

Mark F.J. Steel
CentER and Department of Econometrics
Tilburg University, 5000 LE Tilburg
The Netherlands

ABSTRACT: In this paper we describe the use of modern numerical integration methods for making posterior inferences in composed error stochastic frontier models for panel data or individual cross-sections. Two Monte Carlo methods have been used in practical applications. We survey these two methods in some detail and argue that Gibbs sampling methods can greatly reduce the computational difficulties involved in analyzing such models.

Introduction

The stochastic frontier or composed error framework was first introduced in Meeusen and van den Broeck (1977) and Aigner, Lovell and Schmidt (1977) and has been used in many empirical applications. The reader is referred to Bauer (1990) for a survey of the literature.

In previous papers [van den Broeck, Koop, Osiewalski and Steel (1994), hereafter BKOS; Koop, Steel and Osiewalski (1995); Koop, Osiewalski and Steel, hereafter KOS (1994, 1995a, 1995b, 1996)] we used Bayesian methods to analyze stochastic frontier models and argued that such methods had several advantages over their classical counterparts in the treatment of these models. Most importantly, the Bayesian methods we outlined enabled us to provide exact finite sample results for any feature of interest and to take fully into account parameter uncertainty. In

¹We acknowledge many helpful discussions with Gary Koop in the course of our joint work on previous stochastic frontier papers.

addition, they made it relatively easy to treat uncertainty about which model to use since they recommended taking weighted averages over all models, where the weights were posterior model probabilities. We successfully applied the Bayesian approach in various empirical problems, ranging from hospital efficiencies [KOS (1996)] to analyses of the growth of countries in KOS (1995a,b). Throughout the paper, we shall label the individual units as "firms", but the same models can, of course, be used in other contexts.

In this paper we describe numerical integration methods which are used to perform a fully Bayesian analysis of stochastic frontier models. We present these methods in a much more systematic way than in previous papers and summarize our empirical experiences. For a more general survey of numerical methods in Bayesian econometrics see Geweke (1995).

The paper is organized as follows. Section 1 introduces the Bayesian model we consider in the paper. Section 2 discusses the necessity of modern (Monte Carlo type) numerical integration for making posterior inferences in this context. Section 3 is devoted to Monte Carlo with Importance Sampling and Section 4 presents Gibbs sampling.

Section 1: The Sampling Model and Prior Distribution

The basic stochastic frontier sampling model considered here is given by:

$$y_{it} = h(x_{it}, \beta) + v_{it} - z_{it}, \quad (1)$$

where y_{it} is the natural logarithm of output (or the negative of log cost) for firm i at time t ($i=1,...,N$; $t=1,...,T$); x_{it} is a row vector of exogenous variables; h , a known measurable function, and β , a vector of k unknown parameters, define the deterministic part of the frontier; v_{it} is a symmetric disturbance capturing the random nature of the frontier itself (due to, e.g., measurement error); z_{it} is a nonnegative disturbance capturing the level of inefficiency of firm i at time t ; and z_{it} and v_{it} are independent of each other and across firms and time. Thus, we do not allow for autocorrelated errors at this stage of our research. Generally, efficiency will be measured as $r_{it} = \exp(-z_{it})$, which is an easily interpretable quantity in $(0,1)$.

Note that this framework is suitable for panel data, but the case of just one cross-section, considered in many previous papers, is easily covered as it corresponds to $T=1$. Here we make the assumption that z_{it} is independent (conditionally upon whatever parameters are necessary to describe its sampling distribution) over both i and t , as in KOS (1995a,b); see also Pitt and Lee (1981, Model II). KOS (1996) follow an alternative modeling strategy and assume that the inefficiency level is an individual (firm) effect, i.e. $z_{it} = z_i$ ($t=1,...,T$); see also Pitt and Lee (1981, Model I) and Schmidt and Sickles (1984). In a Bayesian context, KOS (1996) define the difference between fixed and random effects models as a difference in the structure of the prior information.

Whenever the effects are marginally prior independent, they talk of fixed effects, whereas the situation with prior dependent effects, linked through a hierarchical prior, is termed "Bayesian random effects". They provide both theoretical and empirical support for the latter type of model. Thus, we shall focus on Bayesian random effects models in the sequel.

In general, a fully parametric Bayesian analysis requires specifying

- (i) a sampling distribution parameterized by a finite-dimensional vector (say, $\theta \in \Theta$),
- (ii) a prior distribution for that θ .

In order to satisfy (i) and obtain the likelihood function we assume that v_{it} is $N(0, \sigma^2)$, i.e. Normal with zero mean and constant variance σ^2 , and z_{it} is Exponential with mean (and standard deviation) λ_{it} which can depend on some (say, $m-1$) exogenous variables explaining possible systematic differences in efficiency levels. In particular, we assume

$$\lambda_{it} = \prod_{j=1}^m \phi_j^{-w_{itj}} \quad (2)$$

where $\phi_j > 0$ are unknown parameters and $w_{it1} = 1$. If $m > 1$, the distribution of z_{it} can differ for different t or i or both and thus we call this case the Varying Efficiency Distribution (VED) model. If $m=1$, then $\lambda_{it} = \phi_1^{-1}$ and all inefficiency terms constitute independent draws from the same distribution. This case is called the Common Efficiency Distribution (CED) model. The choice of the Exponential distribution for z_{it} stems from the findings in BKOS, where a more general family of Erlang distributions and a truncated Normal were considered as possible specifications; their results clearly indicated that an Exponential distribution is least sensitive to the particular choice of the prior on λ_{it} . Note that the sampling density of the observable y_{it} given x_{it} , $w_{it} = (w_{it1}, \dots, w_{itm})'$ and $\theta = (\beta',$

$\sigma^2, \phi_1, \dots, \phi_m)'$ is a location mixture of Normals with the Exponential density of the inefficiency term as the mixing density:

$$p(y_{it} | x_{it}, w_{it}, \theta) = \int_0^\infty f_N(y_{it} | h(x_{it}, \beta) - z_{it}, \sigma^2) f_G(z_{it} | 1, \prod_{j=1}^m \phi_j^{w_{itj}}) dz_{it}, \quad (3)$$

where $f_G(\cdot | a, b)$ indicates the Gamma density with mean a/b and variance a/b^2 . Alternatively, the sampling density can be represented as

$$p(y_{it} | x_{it}, w_{it}, \theta) = \lambda_{it}^{-1} \exp[-\lambda_{it}^{-1} (m_{it} + \frac{1}{2} \sigma^2 \lambda_{it}^{-1})] \Phi(m_{it}/\sigma), \quad (4)$$

where

$$m_{it} = h(x_{it}, \beta) - y_{it} - \sigma^2 / \lambda_{it},$$

$\Phi(\cdot)$ denotes the distribution function of $N(0,1)$, and λ_{it} is given by (2). See Greene (1990) for a similar expression and BKOS for the generalization to the Erlang context. The likelihood function, $L(\theta|\text{data})$, is the product of the densities (4) over t and i . As a result of integrating out the inefficiency terms z_{it} , the form of (4) is quite complicated indeed, and even the numerical evaluation of the ensuing likelihood function is nontrivial, as shown in BKOS.

An important aspect of any efficiency analysis is making inferences on individual efficiencies of observed firms. It is easy to show that, conditionally on the parameters and the data, the unobserved inefficiency term z_{it} of an observed y_{it} has a truncated Normal distribution with density

$$p(z_{it} | y_{it}, x_{it}, w_{it}, \theta) = [\Phi(m_{it}/\sigma)]^{-1} f_N^1(z_{it} | m_{it}, \sigma^2) I(z_{it} \geq 0), \quad (5)$$

and the first two moments about 0 equal

$$E(z_{it} | y_{it}, x_{it}, w_{it}, \theta) = m_{it} + \sigma [\Phi(m_{it}/\sigma)]^{-1} f_N^1(m_{it}/\sigma | 0, 1) \quad (6)$$

and

$$E(z_{it}^2 | y_{it}, x_{it}, w_{it}, \theta) = \sigma^2 + m_{it}^2 + m_{it} \sigma [\Phi(m_{it}/\sigma)]^{-1} f_N^1(m_{it}/\sigma | 0, 1); \quad (7)$$

see BKOS and Greene (1990). Both the density value at a prespecified point z_{it} and the moments are complicated functions of the model parameters (given the data). If, instead of the efficiency of a particular firm i , we are interested in the efficiency of a hypothetical (unobserved) firm belonging to that industry, we should simply focus on the Exponential distribution with mean λ_{it} as in (2), given certain (representative) values for w_{it} if $m > 1$.

In principle, the prior distribution of θ can be any distribution, but it is usually preferred not to introduce too much subjective information about the parameters. Therefore, we propose the following prior structure

$$p(\theta) = p(\sigma^{-2}) p(\beta) p(\phi_1, \dots, \phi_m) \propto f_G(\sigma^{-2} | \frac{n_0}{2}, \frac{s_0}{2}) f(\beta) \prod_{j=1}^m f_G(\phi_j | a_j, g_j), \quad (8)$$

which reflects lack of prior knowledge about the frontier parameters β , possibly except for regularity conditions imposed by economic theory. That is, we assume $f(\beta) \equiv 1$ if there are no regularity constraints, and, if such conditions are imposed, $f(\beta) = 1$ for all β satisfying them and $f(\beta) = 0$ for all other β . Alternatively, we could use a proper prior distribution on β , possibly

truncated to the region of regularity. Typically, we shall choose the prior hyperparameters n_0 and s_0 so as to represent very weak prior information on the precision of the stochastic frontier. For the m parameters of the efficiency distribution we take proper, independent Gamma priors in order to avoid the pathology described by Ritter (1993). Following KOS (1996, 1995b), we suggest using $a_j = g_j = 1$ for $j > 1$, $a_1 = 1$ and $g_1 = -\ln(r^*)$, where $r^* \in (0, 1)$ is the hyperparameter to be elicited. In the CED model ($m=1$), r^* can be interpreted as the prior median efficiency, because it is exactly the median of the marginal prior distribution of firm efficiency $r_{it} = \exp(-z_{it})$; see BKOS. In the VED case ($m > 1$) our prior for $\phi = (\phi_1, \dots, \phi_m)'$ is quite noninformative and centered over the prior for the CED model. Note that the prior on ϕ , a parameter which is common to all firms, induces prior links between the firm-specific inefficiency terms. This becomes clear if we consider (2) and the fact that the conditional mean of z_{it} is given by λ_{it} . Thus, we are in a Bayesian random effects framework.

Section 2: Bayesian Inference and Numerical Integration

The Bayesian approach combines all the information about the model parameters in their posterior density

$$p(\theta | data) \propto p(\theta) L(\theta | data).$$

As this is a multivariate density, the crucial task of any applied Bayesian study is "to calculate relevant summaries of the posterior distribution, to express the posterior information in a usable form, and to serve as formal inferences if appropriate. It is in the task of summarizing that computation is typically needed." [O'Hagan (1994), p.205].

Often, there is a variety of potential models that we could use for the analysis of a particular phenomenon, without any compelling indication as to which model to choose in a particular empirical context. The Bayesian paradigm provides a very straightforward and easily understood solution in such a situation. If we denote the models by M_j , $j=1, \dots, p$, and assign prior probabilities $P(M_j)$ to each model, then the posterior probability of each model (within the set of models considered) is given by

$$P(M_j | data) = \frac{K_j P(M_j)}{\sum_{i=1}^p K_i P(M_i)}, \quad (9)$$

where, for each model, K_j denotes the integrating constant of the posterior density, i.e.

$$K_j = \int_{\Theta} p(\theta) L_j(\theta | data) d\theta,$$

with $L_j(\theta | data)$ the likelihood corresponding to M_j . The ratio K_j/K_i is called the Bayes factor of model j against model i and summarizes the relative evidence in the data supporting M_j as opposed to M_i . Posterior model probabilities can guide us in choosing a single model, but are more naturally used as weights in a pooled analysis (over the whole set of p models), thus formally taking model uncertainty into account.

Using $g(\theta; data)$ as a generic notation for any function of interest, we can represent our Bayesian inference problem as the ratio of integrals

$$E[g(\theta; data) | data] = \frac{\int_{\Theta} g(\theta; data) p(\theta) L(\theta | data) d\theta}{\int_{\Theta} p(\theta) L(\theta | data) d\theta}. \quad (10)$$

The main numerical difficulty amounts to evaluating this ratio. In the case of the stochastic frontier model, the likelihood is too complex to analytically calculate any such posterior summary. Numerical integration methods are unavoidable. Most quantities of interest such as moments of the parameters or of functions of the parameters, probabilities of certain regions for the parameters, etc. can be expressed as expectations of some $g(.,.)$ in (10). The integrating constant of the posterior, which is crucial for the calculation of Bayes factors, as mentioned above, exactly corresponds to the integral in the denominator of (10).

Note that traditional quadratures, like Cartesian product rules, are not helpful because they are feasible only when the dimension of θ , equal $k+m+1$, is very small. Iterative quadrature, as discussed in Naylor and Smith (1982), combined with a judicious choice of parameterization, has been known to handle problems up to 5 or 6 dimensions. However, our model is inherently "non-standard" and the dimension of θ is already 5 in the simplest case of the CED model ($m=1$) with a Cobb-Douglas frontier depending on $\beta=(\beta_1 \ \beta_2 \ \beta_3)'$ ($k=3$), and will be much higher for more complicated models. This effectively renders these type of numerical integration procedures quite useless.

Another interesting approach is based on Laplace's method to approximate the integrals in (10). Tierney and Kadane (1986) show how the ratio form of (10) can be exploited in applying this method. This method is quite inexpensive to implement as it requires no generation of drawings from the posterior, but relies crucially upon properties of the integrand (such as unimodality) and quickly becomes problematic in multidimensional settings. As a consequence, our stochastic frontier problems do not lend themselves to analysis through such approximations.

A straightforward Monte Carlo method would be to generate i.i.d. drawings from the posterior distribution and use this to calculate empirical sampling counterparts of the integrals in (10). Such a Direct Monte Carlo integration would be fairly easy to implement, but requires the ability to draw from the posterior distribution. The latter is often not a trivial matter, and certainly not feasible for the class of models treated here. Ley and Steel (1992) use Rejection Sampling [see e.g. Devroye (1986) or Smith and Gelfand (1992)] to draw directly from the posterior for a version of our model with no measurement error v_{it} in (1). However, when measurement error is allowed for, these rejection methods can no longer be used.

In the next Sections, we will discuss in more detail two numerical integration procedures that have successfully been used for a Bayesian analysis of the type of stochastic frontier models introduced in Section 1. In particular, these methods are Monte Carlo Importance Sampling integration and Markov Chain Monte Carlo methods.

Section 3: Monte Carlo with Importance Sampling

The Monte Carlo - Importance Sampling (MC-IS) approach to calculating the generic integral

$$I = \int_{\Theta} A(\theta) d\theta = \int_{\Theta} g(\theta; data) p(\theta) L(\theta | data) d\theta \quad (11)$$

evaluates the integrand, $A(\theta)$, at random points drawn independently from some known density $s(\theta)$, called the importance function. Since

$$I = \int_{\Theta} \frac{A(\theta)}{s(\theta)} s(\theta) d\theta = E_s \left[\frac{A(\theta)}{s(\theta)} \right], \quad (12)$$

where E_s denotes expectation with respect to s , we estimate I by the sample mean

based on M independent drawings $\theta^{(1)}, \dots, \theta^{(M)}$ from $s(\theta)$. The ratio of integrals in (10) is estimated

$$\hat{I} = \frac{1}{M} \sum_{l=1}^M A(\theta^{(l)})/s(\theta^{(l)}) \quad (13)$$

by the appropriate ratio of sample means. The asymptotic variance of the latter ratio (as M tends to infinity) can be estimated from the same sample $(\theta^{(1)}, \dots, \theta^{(M)})$. For theoretical details see e.g. O'Hagan (1994, p.223-224), Geweke (1989). An alternative way of interpreting I in (11) is the following:

$$I = \int_{\Theta} g(\theta; data) \frac{p(\theta)L(\theta|data)}{s(\theta)} s(\theta) d\theta = E_s[g(\theta; data) \frac{p(\theta)L(\theta|data)}{s(\theta)}], \quad (14)$$

which immediately shows us that the sample estimate of I in (13) can also be written as

$$\hat{I} = \frac{1}{M} \sum_{l=1}^M g(\theta^{(l)}; data) w(\theta^{(l)}), \quad (15)$$

where $w(\theta)=p(\theta)L(\theta|data)/s(\theta)$ indicates the "weight" attached to each drawing. Note that through choosing $g(\theta; data)=1$, we approximate K_j in (9), i.e. the integrating constant, by the sample average of the weights. Typically, this integrating constant of the posterior will not be known, so that we only have the posterior kernel at our disposal. This essentially means that we will need to evaluate the ratio of integrals in (10) in order to obtain a correctly scaled Monte Carlo estimate of the particular $g(\theta; data)$ we have in mind.

To achieve convergence in practice, it is imperative that we avoid having wildly varying weights $w(\theta^{(l)})$; in particular, we want to avoid having one or two drawings that completely dominate the sample average in (15). Thus, we choose the importance function with thick enough tails, so that the ratio of the posterior to the importance function does not blow up if we happen to draw far out in the tails. Usually, a multivariate Student t density with a small number of degrees of freedom (like a Cauchy density), centered around the maximum likelihood estimate, will do.

BKOS presented an application of MC-IS to simple 7- and 8-parameter stochastic frontier cost models for cross-section data ($T=1$) and found this numerical method feasible, but very time-consuming and requiring a lot of effort on the part of the researcher. First, finding (and fine-tuning!) a good importance function was not a trivial task. Second, even about 150,000 drawings, obtained in a typical overnight run on a PC, did not lead to smooth density plots, as a result of some very large weights. Those experiences were not promising for highly dimensional problems. Using the same models and data, Koop, Steel and Osiewalski (1995) showed that Gibbs sampling, a chief competitor to MC-IS, is superior to the latter method both in terms of the total computing

time (required to obtain reliable final results) and the effort of the researcher. We will discuss Gibbs sampling in the next section.

It is worth noting that the actual implementation of MC-IS, presented by BKOS, did not rely on analytical formulae for features of the conditional distribution of the z_{it} s given the data and parameters [the strategy of using knowledge of the conditional properties is often referred to as an application of the Rao-Blackwell Theorem, see e.g. Gelfand and Smith (1990)]. Instead of evaluating expressions like (5) for the density plot or (6) and (7) for moments at each draw $\theta^{(l)}$, $z_{it}^{(l)}$ was randomly drawn from the conditional density $p(z_{it}|y_{it}, x_{it}, w_{it}, \theta = \theta^{(l)})$, and the posterior summaries were calculated on the basis of the sample $(z_{it}^{(1)}, \dots, z_{it}^{(M)})$. However, the increase in dimensionality of the Monte Carlo due to sampling z_{it} s for a few selected firms is not that important because each z_{it} is drawn from its actual conditional posterior distribution. In the joint posterior density

$$p(z_{it}, \theta | data) = p(z_{it} | y_{it}, x_{it}, w_{it}, \theta) p(\theta | data),$$

only the marginal density of θ has to be mimicked by the importance function $s(\theta)$ which creates the main challenge for MC-IS. The conditional density of z_{it} is a univariate Normal density truncated to be nonnegative. Simulation from this density is simple using the algorithm described in Geweke (1991). This algorithm is a combination of several different algorithms depending on where the truncation point is. For example, if the truncation point is far in the left tail the algorithm simply draws from the unrestricted Normal, discarding the rare draws which are beyond the truncation point. For other truncation points the algorithm uses rejection methods based on densities which tend to work well in the relevant region.

Section 4: Gibbs Sampling

Gibbs sampling is a technique for obtaining a sample from a joint distribution of a random vector α by taking random draws from only full conditional distributions. A detailed description of the technique can be found in e.g. Casella and George (1992), Gelfand and Smith (1990), and Tierney (1994).

Suppose we are able to partition α into $(\alpha_1', \dots, \alpha_p')$ in such a way that sampling from each of the conditional distributions (of α_i given the remaining subvectors; $i=1, \dots, p$) is relatively easy. Then the Gibbs sampler consists of drawing from these distributions in a cyclical way. That is, given the l th draw, $\alpha^{(l)}$, the next draw, $\alpha^{(l+1)}$, is obtained in the following pass through the sampler:

$\alpha_1^{(l+1)}$ is drawn from

$$p(\alpha_1 | \alpha_2 = \alpha_2^{(l)}, \dots, \alpha_p = \alpha_p^{(l)}),$$

$\alpha_2^{(l+1)}$ is drawn from

$$p(\alpha_2 | \alpha_1 = \alpha_1^{(l+1)}, \alpha_3 = \alpha_3^{(l)}, \dots, \alpha_p = \alpha_p^{(l)}),$$

... $\alpha_p^{(l+1)}$ is drawn from

$$p(\alpha_p | \alpha_1 = \alpha_1^{(l+1)}, \dots, \alpha_{p-1} = \alpha_{p-1}^{(l+1)}).$$

Note that each pass consists of p steps, i.e. drawings of the p subvectors of α . The starting point, $\alpha^{(0)}$, is arbitrary. Under certain general conditions [irreducibility and aperiodicity as described in e.g. Tierney (1994)], the distribution of $\alpha^{(l)}$ converges to the joint distribution, $p(\alpha)$, as l tends to infinity. Thus, we draw a sample directly from the joint distribution in an asymptotic sense. In practical applications we have to discard a (large) number of passes before convergence is reached.

In contrast to the MC-IS approach, we now no longer have independent drawings from the posterior, but the Markovian structure of the sampler induces serial correlation between the drawings. This will, in itself, tend to reduce the numerical efficiency of the Gibbs method [see Geweke (1992) for a formal evaluation of numerical efficiency based on spectral analysis]. However, variation of the weights has a similar negative effect on the efficiency of the MC-IS method, and the Gibbs sampler does not require a judicious choice of an importance function $s(\theta)$. Gibbs sampling is, in a sense, more "automatic" and less dependent upon critical input from the user. As the drawings in Gibbs sampling are (asymptotically) from the actual posterior distribution, which is properly normalized, there is no need to evaluate the integrating constant K_j separately. This has two consequences: we do not require the evaluation of a ratio of integrals as in (10) in order to "normalize" our estimates of $g(\theta; \text{data})$, but, on the other hand, the Gibbs sampler does not easily lend itself to approximating K_j which is naturally obtained by integrating the likelihood with the prior, as in (9). This makes the calculation of Bayes factors, used in assessing posterior model probabilities (see Section 2) less immediate than in the MC-IS context. However, a number of solutions based on the harmonic mean of the likelihood function is presented in Newton and Raftery (1994), and in Gelfand and Dey (1994). Chib (1995) proposes an entirely different approach based on a simple identity for K_j .

In order to efficiently use Gibbs sampling to make posterior inferences on both the parameters and firm efficiencies, we have to consider the joint posterior density of θ and z ,

$p(\theta, z | \text{data})$, where z is a $TN \times 1$ vector of all the z_{it} s. Now, instead of integrating out z , which was shown to lead to a very nonstandard likelihood function in Section 1, we shall consider θ given z and the data, which is quite easy to deal with. On the other hand, this implies that we also need to include z in the Gibbs sampler. The underlying idea of adding random variables to the sampler in order to obtain simple conditional distributions is often called Data Augmentation, after Tanner and Wong (1987). Note that the dimension is then $NT+k+m+1$, greater than the number of observations. Despite this high dimensionality, the steps involved in the Gibbs are very easy to implement, and the resulting sampler is found in Koop, Steel and Osiewalski (1995) to have very good numerical properties, and to be far preferable to Monte Carlo with Importance Sampling in the particular application used. The conditional posterior density $p(z | \text{data}, \theta)$ is the product of the TN independent truncated Normal densities given by (5), so we can very easily draw z_{it} s given the data and the parameters. These draws are immediately transformed into efficiency indicators defined as $r_{it} = \exp(-z_{it})$. Thus, this NT -dimensional step of each pass of our Gibbs sampler is quite simple. It is worth stressing at this stage that our Gibbs sampler, unlike the Monte Carlo approach outlined in BKOS, yields a draw of the whole vector z at each pass, and that for this reason, the efficiency measures for all N firms and all T periods are obtained as a by-product of our Gibbs sampling methodology.

However, the main difference with respect to MC-IS is in drawing θ . Now, the unwieldy form of the marginal posterior, $p(\theta | \text{data})$, is not used at all, and we focus instead on the conditional posterior densities of subvectors of θ given the remaining parameters and z . Given z , the frontier parameters (β, σ^2) are independent of ϕ and can be treated as the parameters of the (linear or nonlinear) Normal regression model in (1). Thus we obtain the following full conditionals for β and σ^2 :

$$p(\sigma^{-2} | \text{data}, z, \beta) = f_G(\sigma^{-2} | \frac{n_0 + TN}{2}, \frac{1}{2} \{s_0 + \sum_{t,i} [y_{it} + z_{it} - h(x_{it}, \beta)]^2\}), \quad (16)$$

$$p(\beta | \text{data}, z, \sigma^{-2}) \propto f(\beta) \exp[-\frac{1}{2} \sigma^{-2} \sum_{t,i} (y_{it} + z_{it} - h(x_{it}, \beta))^2]. \quad (17)$$

The full conditional posterior densities of ϕ_j ($j=1, \dots, m$) have the general form

$$p(\phi_j | \text{data}, z, \phi_{(-j)}) \propto f_G(\phi_j | a_j + \sum_{t,i} w_{itj} g_j) \exp(-\phi_j \sum_{t,i} z_{it} D_{itj}), \quad (18)$$

where

$$D_{ir} = \prod_{j \neq r} \phi_j^{w_{ij}}$$

for $r=1, \dots, m$ ($D_{i1}=1$ when $m=1$) and $\phi_{(j)}$ denotes ϕ without its j th element. Since $w_{i1}=1$, the full conditional of ϕ_1 is just Gamma with parameters a_1+NT and $g_1+z_{11}D_{111}+\dots+z_{TN}D_{TN1}$.

Depending on the form of the frontier and on the values of w_{ij} s for $j>2$, the full conditionals for β and for ϕ_j ($j=2, \dots, m$) can be quite easy or very difficult to draw from. Drawing from nonstandard conditional densities within the Gibbs sampler requires special techniques, like rejection methods or the Metropolis-Hastings algorithm [see e.g. Tierney (1994) or O'Hagan (1994)]. In the context of stochastic frontiers, Koop, Steel and Osiewalski (1995) use rejection sampling to deal with non-Exponential Erlang inefficiency distributions for z_{ii} , and KOS (1994) apply an independence chain Metropolis-Hastings algorithm [see Tierney (1994)] to handle a complicated flexible form of $h(x_{ii}, \beta)$. Especially the latter procedure implies a substantial added complexity in the numerical integration and requires additional input from the user, not unlike the choice of an importance function in MC-IS. Therefore, we stress two important special cases where considerable simplifications are possible:

- (i) linearity of the frontier,
- (ii) 0-1 dummies for w_{ij} ($j=2, \dots, m$).

If $h(x_{ii}, \beta) = x_{ii}\beta$ then (17) is a k -variate Normal density, possibly truncated due to regularity conditions. That is, we have

$$p(\beta \mid \text{data}, z, \sigma^{-2}) \propto f(\beta) f_N^k(\beta \mid \hat{\beta}, \sigma^2(X'X)^{-1}), \quad (19)$$

where

$$\hat{\beta} = (X'X)^{-1}X'(y+z),$$

and y and X denote a $NT \times 1$ vector of y_{it} s and a $NT \times k$ matrix with x_{it} s as rows, respectively.

Cobb-Douglas or translog frontiers serve as examples of linearity in β ; see Koop, Steel and Osiewalski (1995) and KOS (1995a, 1996). It is worth stressing that linearity in all elements of β is not essential for obtaining tractable conditionals. Suppose that $\beta' = (\beta_1' \beta_2')$ and the frontier is linear in β_1 given β_2 . Then the full conditional of β_1 alone is multivariate (truncated) Normal. Even if the conditional of β_2 is not standard, its dimension is smaller than that of β . Of course, a bilinear frontier (linear in β_1 given β_2 and in β_2 given β_1) leads to (truncated) Normal conditional densities for both subvectors. Bilinear frontiers are thus quite easy to handle, as evidenced by KOS (1995b) who consider a Cobb-Douglas production frontier with simple parametric effective factor corrections.

The dichotomous (0-1) character of the variables explaining efficiency differences (w_{ij} ; $j=2,\dots,m$) greatly simplifies (18) which simply becomes a Gamma density:

$$p(\phi_j | data, z, \phi_{(-j)}) = f_G(\phi_j | a_j + \sum_{t,i} w_{tij}, g_j + \sum_{t,i} w_{tij} z_{it} D_{tij}), \quad (20)$$

From the purely numerical perspective, it pays to dichotomize original variables which are not 0-1 dummies.

The above discussion confirms that the stochastic frontier model considered in this paper can be analyzed using Gibbs sampling methods. That is, even though the marginal posteriors of θ and the z_{it} s are unwieldy, the conditionals for a suitable partition of the set of parameters should be much easier to work with. By taking a long enough sequence of successive draws from the conditional posterior densities, each conditional on previous draws from the other conditional densities, we can create a sample which can be treated as coming from the joint posterior distribution.

The posterior expectation of any arbitrary function of interest, $g(\theta, z; data)$, can be approximated by its sample mean, g^* , based on M passes. Moreover, numerical standard errors (NSEs) and relative numerical efficiency (RNE) can also be calculated in order to quantify the approximation error and the efficiency relative to i.i.d. drawings from the posterior, respectively; see Geweke (1992) and Koop, Steel and Osiewalski (1995). However, Geweke (1995) remarks that these measures lack formal theoretical foundations. Geweke (1992) also provides a formal convergence criterion which, however, is not suitable for detecting whether the sampler has stayed trapped in one part of the parameter space. For instance, if the posterior is bimodal and the Gibbs sampler remains in the region of one of the modes it is likely that Geweke's formal diagnostic could indicate convergence, even though an important part of the posterior is being ignored. Hence, we suggest an informal diagnostic which draws on the work of Gelman and Rubin (1992). This diagnostic involves the following steps:

- i) Simulate L , each of length $2S$, sequences from randomly selected, overly dispersed, starting values. Keep only the last S passes from each sequence.
- ii) Calculate the posterior mean of a feature of interest for each sequence. If the variance of the Gibbs passes of this feature within one of the m sequences is much larger than the variance of the calculated posterior means between sequences, then convergence has been achieved.

Loosely speaking, if widely disparate starting values yield highly similar Gibbs estimates, then convergence has occurred and sensitivity to starting values is not an issue. Recently, Zellner and Min (1995) have proposed a number of new operational convergence criteria for the Gibbs sampler. In general, it pays to be very careful regarding convergence issues in a Gibbs exercise,

and for stochastic frontiers in particular the results of Ritter (1993) and Ritter and Simar (1994) indicate that posteriors can be badly behaved.

There are two main implementations of the Gibbs sampler. One version (sequential Gibbs sampling) takes one long run of M passes (after discarding an initial set of burn-in passes), the other (parallel Gibbs sampling) takes several shorter runs each starting at the same initial value (ie. at the same $\theta^{(0)}$) as in Gelfand and Smith (1990). For this latter strategy we carry out S runs, each containing L passes, and keep only the L th pass out of these runs. In both cases, we proceed in an iterative fashion, starting from an arbitrary initial value and using small Gibbs runs, typically with $M=500$ or $L=10$ and $S=500$, to choose a starting value, $\theta^{(0)}$.

The issue of whether to use one long run from the Gibbs sampler or to restart every L^{th} pass has been discussed in the literature [see, for example, Tanner (1991), Carlin, Polson and Stoffer (1992), Casella and George (1992), Gelman and Rubin (1992), Raftery and Lewis (1992) and Tierney (1994)]. Although the theoretical properties of these two variants of the Gibbs sampler are basically the same, they can, in practice, be quite different. Advocates for taking one long run note that the restarting method typically wastes a lot of information. Advocates for restarting the Gibbs sampler note that this technique can be very effective in breaking serial correlation between retained draws and prevents the path from becoming "temporarily trapped in a nonoptimal subspace" (Tanner (1991), p. 91; see also Zeger and Karim (1991) and Gelman and Rubin (1992)). The question of which variant is preferable is no doubt a problem-specific one, but in the application in Koop, Steel and Osiewalski (1995) the two methods yielded almost identical results. Since the sequential sampler was less computationally demanding, it seemed the superior method for the presented example. The RNEs with parallel Gibbs were typically around 1, suggesting that parallel Gibbs sampling was roughly as efficient as is sampling directly from the posterior. However, the sequential Gibbs sampler yielded RNE's which were typically around .5 to .8, only modestly worse than those with parallel Gibbs. Since, roughly speaking, parallel Gibbs was $L=5$ times as computationally demanding as sequential Gibbs, the reduction in RNE found by going to parallel Gibbs did not seem worth the cost. Note that, as a result of large variation in the weights, MC-IS on the same application lead to RNE's of the order of 0.001, clearly illustrating the numerical problems encountered by the latter method.

A variant of the sequential Gibbs sampler is to use a single run, but once it has converged to keep only the drawings which are several (say, t) passes apart. The separation t should be enough to produce essentially independent draws. This approach, like the parallel sampler, wastes a lot of information. Moreover, O'Hagan (1994, p.236) proves that if we simply average $M=rt$ consecutive draws after convergence, the result will be at least as good in terms of variance of g^*

as taking r draws t apart. Thus it seems preferable to use a single long run and accept that correlation exists between the passes. Of course, whenever correlation is present, we need to take this into account in estimating the asymptotic variance of our Monte Carlo estimates (and, thus, e.g. RNE). Sometimes, a simple approximation through batch means is used [see e.g. Ripley (1987)], whereas Geweke (1992) advocates spectral methods, also used in KOS (1994) and Koop, Steel and Osiewalski (1995).

Conclusion

In this paper, we have examined a Bayesian analysis of stochastic frontier models with composed error, and illustrated the considerable numerical requirements. We review the Bayesian paradigm and the basic inferential problems that it poses. We typically need to solve integrals, for which a number of candidate numerical methods can be considered in general. Based on our experience with two numerical strategies that show some promise for this complicated problem, we argue that Gibbs sampling methods can be used to greatly reduce the computational burden inherent to this analysis. We show how the posterior conditional densities can be used to set up a Gibbs sampler. In important special cases all conditionals are either truncated Normal, Normal or Gamma distributions which leads to enormous computational gains. We refer the reader to a number of papers [KOS (1994, 1995a, 1995b, 1996) and Koop, Steel and Osiewalski (1995)] where the Gibbs sampling strategy was successfully applied in a variety of empirical contexts.

References

- Aigner, D., C.A.K. Lovell, and P. Schmidt (1977). "Formulation and estimation of stochastic frontier productions function models," **Journal of Econometrics**, 6, 21-37.
- Bauer, P.W. (1990). "Recent developments in the econometric estimation of frontiers," **Journal of Econometrics**, 46, 39-56.
- van den Broeck, J., G. Koop, J. Osiewalski, and M. Steel (1994). "Stochastic frontier models: A Bayesian perspective," **Journal of Econometrics**, 61, 273-303.
- Carlin, B., N. Polson and D. Stoffer (1992). "A Monte Carlo approach to nonnormal and nonlinear state-space modelling," **Journal of the American Statistical Association**, 87, 493-500.
- Casella, G. and E. George (1992). "Explaining the Gibbs sampler," **The American Statistician**, 46, 167-174.
- Chib, S. (1995). "Marginal likelihood from the Gibbs output," **Journal of the American Statistical Association**, forthcoming.

- Devroye, L. (1986). **Non-Uniform Random Variate Generation**, Springer-Verlag, New York.
- Gelfand, A.E. and D.K. Dey (1994). "Bayesian model choice: Asymptotics and exact calculations," **Journal of the Royal Statistical Society, Ser. B**, 56, 501-514.
- Gelfand, A.E. and A.F.M. Smith (1990). "Sampling-based approaches to calculating marginal densities," **Journal of the American Statistical Association**, 85, 398-409.
- Gelman, A. and D. Rubin (1992). "A Single series from the Gibbs sampler provides a false sense of security," in **Bayesian Statistics 4**, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), Oxford University Press, Oxford.
- Geweke, J. (1989). "Bayesian inference in econometric models using Monte Carlo integration," **Econometrica**, 57, 1317-1339.
- Geweke, J. (1991). "Efficient simulation from the multivariate Normal and Student-t distributions subject to linear constraints," in **Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface**, E.M. Keramidas and S.M. Kaufman (eds.), Interface Foundation of North America.
- Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in **Bayesian Statistics 4**, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), Oxford University Press, Oxford.
- Geweke, J. (1995). "Posterior simulators in econometrics," **Federal Reserve Bank of Minneapolis Working Paper 555**.
- Greene, W.H. (1990). "A gamma-distributed stochastic frontier model," **Journal of Econometrics**, 46, 141-163.
- Koop, G., J. Osiewalski and M.F.J. Steel (1994). "Bayesian efficiency analysis with a flexible form: The AIM cost function," **Journal of Business and Economic Statistics**, 12, 339-346.
- Koop, G., J. Osiewalski and M.F.J. Steel (1995a). "The components of output growth: A cross-country analysis," **CORE Discussion Paper 9503**.
- Koop, G., J. Osiewalski and M.F.J. Steel (1995b). "Measuring the sources of output growth in a panel of countries," **CORE Discussion Paper 9542**.
- Koop, G., J. Osiewalski and M.F.J. Steel (1996). "Bayesian efficiency analysis through individual effects: Hospital cost frontiers," **Journal of Econometrics**, forthcoming.
- Koop, G., M.F.J. Steel and J. Osiewalski (1995). "Posterior analysis of stochastic frontier models using Gibbs Sampling," **Computational Statistics**, forthcoming.
- Ley, E. and M.F.J. Steel (1992). "Bayesian econometrics: Conjugate analysis and rejection sampling using *Mathematica*," Chapter 15 in **Economic and Financial Modeling Using Mathematica**, H. Varian (ed.), Springer-Verlag, New York.
- Meeusen, W. and J. van den Broeck (1977). "Efficiency estimation from Cobb-Douglas production

functions with composed error," **International Economic Review**, 8, 435-444.

Naylor, J.C. and A.F.M. Smith (1982). "Applications of a method for the efficient computation of posterior distributions," **Applied Statistics**, 31, 214-225.

Newton, M. and A. Raftery (1994). "Approximate Bayesian inference by the weighted likelihood bootstrap," (with discussion), **Journal of the Royal Statistical Society, Series B**, 56, 3-48.

O'Hagan, A. (1994). **Bayesian Inference**, Edward Arnold, London.

Raftery, A. and S. Lewis (1992). "How many iterations in the Gibbs sampler?" in **Bayesian Statistics 4**, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), Oxford University Press, Oxford.

Ripley, B.D. (1987). **Stochastic Simulation**, Wiley, New York.

Ritter, C. (1993). "The Normal-Gamma frontier model under a common vague prior does not produce a proper posterior," manuscript.

Ritter, C. and L. Simar (1994). "Pitfalls of Normal-Gamma stochastic frontier models," manuscript.

Schmidt, P. and R.C. Sickles (1984). "Production frontiers and panel data," **Journal of Business and Economic Statistics**, 2, 367-374.

Smith, A.F.M. and A.E. Gelfand (1992). "Bayesian statistics without tears: A sampling-resampling perspective," **The American Statistician**, 46, 84-88.

Tanner, M.A. (1991). **Tools for Statistical Inference**, Springer-Verlag, New York.

Tanner, M.A. and Wong, W.H. (1987). "The calculation of posterior distributions by data augmentation," **Journal of the American Statistical Association**, 82, 528-540.

Tierney, L. (1994). "Markov chains for exploring posterior distributions," (with discussion), **Annals of Statistics**, 22, 1701-1762.

Tierney, L. and J.B. Kadane (1986). "Accurate approximations for posterior moments and marginal densities," **Journal of the American Statistical Association**, 81, 82-86.

Zeger, S. and M. Karim (1991). "Generalized linear models with random effects: A Gibbs sampling approach," **Journal of the American Statistical Association**, 86, 79-86.

Zellner, A. and C. Min (1995). "Gibbs sampler convergence criteria," **Journal of the American Statistical Association**, 90, 921-927.