

Bayesian Inference in a Sample Selection Model

Martijn van Hasselt *

Department of Economics

The University of Western Ontario

October 2008

Abstract

This paper develops methods of Bayesian inference in a sample selection model. The main feature of this model is that the outcome variable is only partially observed. We first construct a Gibbs sampling algorithm for the case that the likelihood of the selection and outcome equation is bivariate normal. Based on this, we extend the algorithm to allow for potential nonnormality of the likelihood. Specifically, we construct Gibbs samplers based on mixtures of normal distributions. These methods are appealing because they allow simultaneous testing for the presence of a selection effect and for departures from normality. We illustrate the various Gibbs samplers using simulated data and Mroz's (1987) labor supply data.

JEL Codes: C11, C14, C15, C34

Keywords: sample selection, Gibbs sampling, mixture distributions

*Social Science Centre, room 4033. 1151 Richmond St. N., London, Ontario, Canada, N6A 5C2. Tel: 1-519-661-2111, Fax: 1-519-661-3666, E-mail: mvanhass@uwo.ca

1 Introduction

In this paper we develop methods of Bayesian inference in a sample selection model. In general sample selection occurs when the data at hand is not a random sample from the population of interest. Instead, members of the population may have selected themselves into (or out of) the sample, based on a combination of observable quantities and unobserved heterogeneity. In this case inference based on the selected sample alone may suffer from selection bias.

A selection model typically consists of two components. The first is an equation that determines the level of the outcome variable. The second is an equation describing the selection mechanism: it determines whether we observe the outcome or not. Such a model can be given a structural interpretation, in which the dependent variable in the selection equation represents an agent's latent utility. If this utility crosses a certain threshold level, the agent behaves in such a way that his or her outcome is observed. On the other hand, if utility remains below the threshold, the agent behaves differently and the outcome is not observed. Thus, a selection model (labeled SSM hereafter) can be viewed as a model for *potential* outcomes which are only *partially* realized and observed. This interpretation applies most directly to the context of estimating wage offer distributions: here the wage offer is a potential outcome which is realized only when an individual actually participates in the labor force.

Sample selection models have been used in cases where the interpretation of potential outcomes is perhaps less appropriate. Consider the example of a model for an individual's medical expenditures. When we observe zero expenditures, it is true that we cannot observe what expenditures *would have been*, had the individual decided to seek medical help. Alternatively, rather than treating a zero outcome as missing data, the interpretation as a corner solution outcome may be more appealing. Duan et al. (1983) argue that in this case a so-called 'two-part model' should be used. This model, proposed by Cragg (1971), describes observed outcomes directly; it essentially assumes away the selection effect. Following Duan et al. (1983) there has been additional debate as to which model is more adequate, both in terms of interpretation and predictive ability; see, for example, Manning et al. (1987), Leung and Yu (1996) and Dow and Norton (2003). It is important to note here that a sample selection model does not generally nest a two-part model, since the parameters in each model have a different interpretation. A two-part model delivers marginal effects on observed

outcomes, whereas a selection model delivers marginal effects on potential outcomes. Whether we are interested in actual or potential outcomes depends on the application at hand. As a modeling device, however, a selection model may be used to analyze actual outcomes, since the latent model structure implies a model for the observed data.

Early contributions to the sample selection literature are, among others, Gronau (1974) and Heckman (1979). Gronau's paper analyzes the potential for selection bias in the labor market when observed wages are used to make inference about the distribution of wage offers. Heckman (1979) treats sample selection as a specification error and proposes a two-step estimator that corrects for omitted variable bias. Heckman's two-step procedure, together with full information maximum likelihood, are now widely used in applied work. An obvious problem with both methods is that they rely on parametric assumptions about the distribution of unobservables. When these assumptions are violated, the estimators may become inconsistent. More recently, flexible semiparametric methods have been proposed that aim to relax strong distributional assumptions in the SSM. Examples include Olsen (1982), Lee (1982, 1994), Cosslett (1991), Ichimura and Lee (1991) and Ahn and Powell (1993). Good surveys of this literature are given by Vella (1998) and Lee (2003).

Our paper develops a Bayesian approach to estimating a sample selection model. Despite the numerous contributions in classical econometrics mentioned above, the Bayesian literature on this topic is relatively sparse. Bayarri and DeGroot (1987), Bayarri and Berger (1998) and Lee and Berger (2001) focus on the idea that the likelihood of the latent model needs to be reweighted, with a potentially unknown weight function, in order to obtain the likelihood of the observed data. These papers propose methods that apply in the context of a univariate sample subject to a selection mechanism. More recently, Chib et al. (2008) consider Bayesian inference in a regression model, subject to sample selection and endogeneity.

The model used in this paper is sometimes referred to as a type-2 Tobit model (e.g., Amemiya, 1985). In essence this model has a binary selection rule: for each individual in the sample we only observe whether a latent variable (e.g. utility) crosses a threshold or not. In some cases the selection process can be more informative. In Lee (1994) the selection rule takes a Tobit form. The outcome of interest is then observed based on a selection variable, which itself is (partially) observed. For example, in a joint model for hours worked and wages, we observe hours and the wage offer if hours are positive. Though we do not treat this case explicitly in our paper, the methods presented here

could be modified in a straightforward manner.

Our main contribution is to provide Gibbs sampling algorithms for use in the sample selection model. These algorithms essentially produce an approximate sample from the posterior distribution of the model parameters. We start by considering inference in a fully parametric Bayesian model, based on the multivariate normal distribution. This basic model may, of course, be subject to misspecification of the likelihood. To address this shortcoming, we then extend the analysis to allow for a more flexible family of distributions. In particular, the focus is on mixtures of normal distributions with either a fixed or random number of mixture components. This will be referred to as a semiparametric Bayesian model. We believe the methods developed in this paper present a viable alternative to classical semiparametric techniques, for several reasons. First, posterior standard deviations are an automatic by-product of the sampling algorithm. In contrast, classical inference requires standard errors, which are often hard to calculate in semiparametric models. Second, our method allows immediate testing for both the presence of a selection effect and departures from normality. Finally, in case a researcher has prior information (in the form of parameter restrictions or likely regions in the parameter space, for example), this information can be incorporated into the analysis through the prior distribution.

Bayesian inference in limited dependent variable models can proceed by a combination of *Gibbs sampling* (e.g., Casella and George, 1992) and *data augmentation* (Tanner and Wong 1987). These methods are useful when either the joint posterior is analytically intractable or difficult to sample from. Gibbs sampling entails generating consecutive draws from the conditional posterior of each parameter, given the remaining ones. Data augmentation treats missing observations as additional parameters and samples new values as part of the algorithm. The combined algorithm yields a Markov chain that, under some regularity conditions, has the posterior as its invariant distribution. Parameter values generated by the chain are then an approximate sample from the posterior. An excellent treatment of Gibbs sampling and, more generally, Markov Chain Monte Carlo (MCMC) methods can be found in Gilks, Richardson, and Spiegelhalter (1996).

Applications of MCMC have become widespread in Bayesian econometrics. Discrete choice models are treated in Albert and Chib (1993), McCulloch and Rossi (1994) and McCulloch et al. (2000). The analysis of the parametric selection model in our paper is most closely related to Li (1998), Huang (2001) and Munkin and Trivedi (2003), the common element being a simultaneous

equations structure with limited dependent variables. Our extension to a semiparametric model draws on a large body of literature on nonparametric Bayesian methods, starting with Ferguson (1973), and incorporates it into a sample selection framework.¹ We reiterate here that the semiparametric model in our paper is based on introducing flexibility into the likelihood. In contrast, Chib, Greenberg, and Jeliazkov (2008) use a selectivity correction in the equation for the outcome variable, which is essentially a control function approach. We leave a comparison with this method to future work.

The remainder of this paper is organized as follows. Section 2 presents the selection model and two Gibbs sampling algorithms based on a bivariate normal likelihood. We briefly touch on the two-part model and ways to distinguish it from the SSM. Section 3 presents a semiparametric version of the selection model and constructs the corresponding Gibbs sampler. Section 4 illustrates the various methods with some simulated data, whereas section 5 contains an application to estimating a wage equation from Mroz’s (1987) labor supply data. Finally, section 6 concludes. Details regarding the construction of the Gibbs sampler are collected in the appendix.

2 A Sample Selection Model

2.1 The Model

We use the following version of a selection model, which Amemiya (1985, ch.10) calls a type-2 Tobit model:

$$\begin{aligned}
 I_i &= x'_{i1}\beta_1 + u_{i1}, \\
 s_i &= \mathbb{I}\{I_i > 0\}, \\
 y_i^* &= x'_{i2}\beta_2 + u_{i2}, \\
 y_i &= s_i y_i^*.
 \end{aligned} \tag{1}$$

The subscript i denotes the i^{th} observation in a sample of size n . The vectors x_{i1} and x_{i2} have k_1 and k_2 elements, respectively, and the data consists of $\{x'_{i1}, x'_{i2}, s_i, y_i\}_{i=1}^n$. The variable s_i is simply

¹To quote Müller and Quintana (2004), “Nonparametric Bayesian inference is an oxymoron and a misnomer”. The term nonparametric refers to the fact that these methods bear some resemblance to classical nonparametric techniques, such as kernel smoothing.

the indicator of observing y_i^* . Note the distinction between the partially latent outcome y_i^* and the observed outcome y_i . If $I_i > 0$ there is selection into the sample and we observe $y_i = y_i^*$. On the other hand, if $I_i \leq 0$ then y_i^* is not observed and we set $y_i = 0$. The zero here is simply an arbitrary label that indicates a missing outcome. In some cases this is quite natural: when y_i^* represents the wage offer, it is only observed when someone actually participates in the labor force. If someone is not employed, then y_i^* is not realized, and actual wages y_i are zero.

A fully parametric model is obtained when the joint distribution of u_{i1} and u_{i2} is bivariate normal:

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad (2)$$

where ρ is the correlation coefficient. The random variable s_i then has a Bernoulli distribution with

$$\Pr\{s_i = 1 | x_{i1}, \beta_1\} = \Phi(x'_{i1}\beta_1/\sigma_1).$$

Here $\Phi(\cdot)$ denotes the CDF of the standard normal distribution. Though β_1 and σ_1 are not separately identified, we retain both parameters for reasons explained shortly. Let $\beta = (\beta'_1, \beta'_2)'$.

The likelihood for the entire sample is

$$\begin{aligned} f(y, s | \beta, \Sigma) &= \prod_{i=1}^n [\Phi(x'_{i1}\beta_1/\sigma_1)]^{s_i} [1 - \Phi(x'_{i1}\beta_1/\sigma_1)]^{1-s_i} \times \\ &\quad \prod_{i: y_i \neq 0} f_{u_2 | I > 0}(y_i - x'_{i2}\beta_2), \end{aligned} \quad (3)$$

where $f_{u_2 | I > 0}$ is the density of u_{i2} conditional on $I_i > 0$. Here y and s denote the entire sample of y_i 's and s_i 's. Throughout this paper we will omit conditioning on the covariates (x'_{i1}, x'_{i2}) for notational simplicity.

Let $\phi(\cdot)$ denote the standard normal density function. It follows from (2) that the conditional distribution of I_i given u_{i2} is normal with mean $x'_{i1}\beta_1 + (\rho\sigma_1/\sigma_2)u_{i2}$ and variance $\sigma_1^2(1 - \rho^2)$. Then:

$$\begin{aligned} f_{u_2 | I > 0}(a) &= \frac{\int_0^\infty f_{u_2, I}(a, I) dI}{P(I > 0)} \\ &= \frac{f_{u_2}(a)}{\Phi(x'_{i1}\beta_1/\sigma_1)} \int_0^\infty f_{I | u_2}(I | u_2 = a) dI \end{aligned}$$

$$= \frac{\sigma_2^{-1} \phi(a/\sigma_2)}{\Phi(x'_{i1}\beta_1/\sigma_1)} \Phi\left(\frac{x'_{i1}\beta_1 + (\rho\sigma_1/\sigma_2)a}{\sqrt{\sigma_1^2(1-\rho^2)}}\right).$$

Plugging this back into (3) the likelihood becomes

$$f(y, s|\beta, \Sigma) = \prod_{i:y_i=0} [1 - \Phi(x'_{i1}\beta_1/\sigma_1)] \times \prod_{i:y_i \neq 0} \sigma_2^{-1} \phi\left(\frac{y_i - x'_{i2}\beta_2}{\sigma_2}\right) \Phi\left(\frac{x'_{i1}\beta_1}{\sigma_1\sqrt{1-\rho^2}} + \frac{\rho(y_i - x'_{i2}\beta_2)}{\sigma_2\sqrt{1-\rho^2}}\right). \quad (4)$$

2.2 Gibbs Sampling in the SSM

For a Bayesian analysis of the model in (1) the likelihood is combined with a prior distribution $f(\beta, \Sigma)$ to yield the posterior:

$$f(\beta, \Sigma|y, s) \propto f(\beta, \Sigma) \times f(y, s|\beta, \Sigma).$$

Given the likelihood in (4), however, there is no choice of prior for (β, Σ) will yield a tractable posterior distribution. We therefore develop a Gibbs sampling algorithm that simulates draws from the posterior distribution of (β, Σ) . The updating step for Σ generates new values for σ_1^2 , σ_2^2 and the covariance σ_{12} . The implied value of ρ is then computed as $\rho = \sigma_{12}/(\sigma_1\sigma_2)$.

Our Gibbs sampler involves the unidentified parameters β_1 and σ_1 . The sampled values of (β_1, σ_1) themselves are therefore not informative, in the sense that there is no updating of the prior.² The output from the algorithm, however, can be used to approximate the posterior of identified parameters such as β_1/σ_1 and ρ . We follow the approach of McCulloch and Rossi (1994), who apply this idea in the context of the multinomial Probit model. The main advantage of retaining the unidentified parameters is that it preserves the natural conjugacy structure in the model, and allows for an easier approximation of the posterior. We will elaborate on this shortly.

Since only the selection indicator s_i is observed, the variable I_i is latent and treated as an additional parameter. The same is true for the unobserved values of y_i^* . The full set of parameters is thus (β, Σ) combined with $I = (I_1, \dots, I_n)$ and $\mathcal{Y}^* \equiv \{y_i^* : s_i = 0\}$. Gibbs sampling now proceeds by consecutive sampling from the *conditional* posterior distributions $f(\beta|\Sigma, I, \mathcal{Y}^*, y, s)$,

²When the prior is improper, the generated values of (β_1, σ_1) are a random walk; see McCulloch and Rossi (1994).

$f(\Sigma|\beta, I, \mathcal{Y}^*, y, s)$ and $f(I, \mathcal{Y}^*|\beta, \Sigma, y, s)$. Sampling values from the latter distribution amounts to data-augmentation: at each iteration we create a ‘complete’ data set. The algorithm described here yields a sequence $\{\beta^{(m)}, \Sigma^{(m)}, I^{(m)}, \mathcal{Y}^{*(m)}\}_{m=1}^M$. The sampled values of (β, Σ) now form an approximate sample from the posterior $f(\beta, \Sigma|y, s)$.

It remains to find the various conditional posterior distributions. For the data-augmentation step we need to distinguish two cases: $s_i = 0$ and $s_i = 1$. Suppose first that $s_i = 1$ so that y_i^* is observed and $I_i > 0$. From (2) it follows that I_i has the following conditional distribution, truncated from below at zero:

$$f(I_i|s_i = 1, y_i^*, \beta, \Sigma) = N(x'_{i1}\beta_1 + \rho\sigma_1\sigma_2^{-1}(y_i^* - x'_{i2}\beta_2), \sigma_1^2(1 - \rho^2)) \mathbb{I}\{I_i > 0\}. \quad (5)$$

If $s_i = 0$ then it is known that $I_i \leq 0$ but the actual values (I_i, y_i^*) are not observed. A value of I_i can be generated from the $N(x'_{i1}\beta_1, \sigma_1^2)$ distribution truncated from above at zero. The value of y_i^* is a realization of its conditional distribution given I_i :³

$$f(I_i|s_i = 0, \beta, \Sigma) = N(x'_{i1}\beta_1, \sigma_1^2) \mathbb{I}\{I_i \leq 0\}, \quad (6)$$

$$f(y_i^*|I_i, \beta, \Sigma) = N(x'_{i2}\beta_2 + \rho\sigma_1^{-1}\sigma_2(I_i - x'_{i1}\beta_1), \sigma_2^2(1 - \rho^2)). \quad (7)$$

To find the conditional posterior of β given the completed data and Σ , it is convenient to write the model in (1) in a SUR form. To this end let y^* , u_1 and u_2 all be n -dimensional vectors that contain the individual observations i . Let X_1 and X_2 be matrices with i^{th} rows x'_{i1} and x'_{i2} , respectively, and define

$$W = \begin{bmatrix} I \\ y^* \end{bmatrix} : 2n \times 1, \quad X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} : 2n \times (k_1 + k_2), \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} : 2n \times 1.$$

It follows that $W = X\beta + u$, where $E(u) = 0$ and $V(u) = \Sigma \otimes I_n$. The completed-data likelihood based on (2) can be written as

$$f(W|\beta, \Sigma) \propto |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \left[e'S^{-1}e + (\beta - \hat{\beta})'X'S^{-1}X(\beta - \hat{\beta}) \right] \right\}, \quad (8)$$

³Of course, when $s_i = 0$ we could also generate y_i^* first and then I_i from its conditional distribution given y_i^* , right-truncated at zero.

where $S = \Sigma \otimes I_n$, $\hat{\beta}$ is the GLS estimator

$$\hat{\beta} = (X'S^{-1}X)^{-1}X'S^{-1}W,$$

and $e = W - X\hat{\beta}$ the residual. Note that the conditional posterior of β satisfies $f(\beta|W, \Sigma, y, s) = f(\beta|W, \Sigma)$ because (y, s) is a function of W . Combining the likelihood $f(W|\beta, \Sigma)$ with a normal $N(b_0, B_0)$ prior distribution for β , it is a standard result that the posterior is again normal with mean and variance given by

$$E(\beta|W, \Sigma) = [B_0^{-1} + X'S^{-1}X]^{-1} [B_0^{-1}b_0 + X'S^{-1}X\hat{\beta}], \quad (9)$$

$$V(\beta|W, \Sigma) = [B_0^{-1} + X'S^{-1}X]^{-1}. \quad (10)$$

Finally, it remains to find the conditional posterior of Σ . An equivalent form of the SUR likelihood in (8) is

$$\begin{aligned} f(W|\beta, \Sigma) &\propto |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2}(W - X\beta)'S^{-1}(W - X\beta) \right\} \\ &\propto |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2}\text{tr}(\Sigma^{-1}A) \right\}, \end{aligned} \quad (11)$$

where $\text{tr}(\cdot)$ is the trace and A is defined as

$$A = \begin{bmatrix} (I - X_1\beta_1)'(I - X_1\beta_1) & (I - X_1\beta_1)'(y^* - X_2\beta_2) \\ (y^* - X_2\beta_2)'(I - X_1\beta_1) & (y^* - X_2\beta_2)'(y^* - X_2\beta_2) \end{bmatrix}. \quad (12)$$

By inspection of (11) it can be seen that the inverse Wishart distribution is the natural conjugate prior. If Σ has an inverse Wishart distribution with parameter matrix H and degrees of freedom v , we will write $\Sigma \sim \mathcal{W}^{-1}(H, v)$. The density of Σ is given by

$$f(\Sigma|H, v) \propto |\Sigma|^{-(v+3)/2} \exp \left\{ -\frac{1}{2}\text{tr}(\Sigma^{-1}H) \right\}, \quad v \geq 2,$$

and has mean $(v - 3)^{-1}H$ (e.g., Muirhead 1982, page 97). Multiplying this density with (??) we

find

$$f(\Sigma|\beta, W) \propto |\Sigma|^{-(n+v+3)/2} \exp\left\{-\frac{1}{2}\text{tr}(\Sigma^{-1}(A+H))\right\}, \quad v \geq 2. \quad (13)$$

Thus, the conditional posterior of Σ is a $\mathcal{W}^{-1}(A+H, n+v)$ distribution. The Gibbs sampler can now be summarized as follows:

Algorithm 1 (Unidentified Parameters) *For given starting values of $(\beta, \Sigma, I, \mathcal{Y}^*)$:*

1. if $s_i = 1$, sample I_i from (5); if $s_i = 0$, sample I_i from (6) and y_i^* from (7);
2. sample β from a multivariate normal with mean (9) and variance (10);
3. sample Σ from (13);
4. return to step 1 and repeat.

In order to execute step 3 of the algorithm, note that by definition of the inverse Wishart distribution the precision matrix $P(= \Sigma^{-1})$ has a conditional posterior Wishart distribution with parameter matrix $(A+H)^{-1}$ and $(n+v)$ degrees of freedom. A draw of Σ can then be obtained by first generating $(n+v)$ vectors z_j from the $N_2(0, (A+H)^{-1})$ distribution and computing $\Sigma = (\sum_{j=1}^{n+v} z_j z_j')^{-1}$.

Algorithm 1 yields a realization of a Markov chain that is informative about the posterior distribution of $(\beta_1/\sigma_1, \beta_2, \sigma_2, \rho)$. A disadvantage of using unidentified parameters is that it may be difficult to choose appropriate priors for β_1/σ_1 and ρ .⁴ For that reason we can also use an idea proposed by Koop and Poirier (1997) in the context of a switching model: an alternative Gibbs sampler can be constructed by imposing the restriction $\sigma_1 = 1$, reparameterizing the model and placing priors on the identified parameters directly; see also Li (1998) and McCulloch et al. (2000) for additional applications.

In our model this approach can be implemented by writing the covariance matrix as

$$\Sigma = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \xi^2 + \sigma_{12}^2 \end{bmatrix},$$

where σ_{12} is the covariance between u_{i1} and u_{i2} and $\xi^2 \equiv \sigma_2^2 - \sigma_{12}^2$ is the conditional variance of u_{i2} given u_{i1} . In order to generate draws (σ_{12}, ξ^2) in the Gibbs sampler, we need the conditional

⁴In other words, the induced prior of $(\beta_1/\sigma_1, \rho)$ needs to be checked to ensure it is appropriately reflecting the researcher's beliefs.

posterior $f(\sigma_{12}, \xi | W, \beta)$. Note that given (W, β) , the errors u are known and (σ_{12}, ξ^2) are the parameters of a normal linear regression model:

$$u_{i2} = \sigma_{12}u_{i1} + \eta_i, \quad \eta_i \sim N(0, \xi^2).$$

Thus, the conditional posterior of interest satisfies

$$\begin{aligned} f(\sigma_{12}, \xi^2 | W, \beta) &= f(\sigma_{12}, \xi^2 | u, \beta) \\ &\propto f(u | \sigma_{12}, \xi^2, \beta) f(\sigma_{12}, \xi^2 | \beta) \\ &\propto f(u | \sigma_{12}, \xi^2) f(\sigma_{12}, \xi^2), \end{aligned}$$

where we take (σ_{12}, ξ^2) a priori independent of β .

The natural conjugate prior for (σ_{12}, ξ^2) is of the normal-inverse gamma form. By definition ξ^2 has an inverse gamma distribution with parameters (c_0, d_0) , written as $\xi^2 \sim IG(c_0, d_0)$, if ξ^{-2} has a prior gamma distribution $G(c_0, d_0)$. The prior density of ξ^2 is then

$$f(\xi^2 | c_0, d_0) = \frac{d_0^{c_0}}{\Gamma(c_0)} (\xi^2)^{-(c_0+1)} e^{-d_0/\xi^2}.$$

We set the conditional prior of σ_{12} equal to

$$f(\sigma_{12} | \xi^2, \tau, g_0) = N(g_0, \tau \xi^2).$$

The reason for prior dependence between σ_{12} and ξ^2 is that the induced prior for the correlation coefficient can be made roughly uniform by an appropriate choice of τ . Here (c_0, d_0, g_0, τ) is a set of hyperparameters. It is easy to show that now the posteriors take the following form:

$$\begin{aligned} \xi^2 | W, \beta, \sigma_{12} &\sim IG(\tilde{c}, \tilde{d}), \\ \tilde{c} &= c_0 + \frac{n+1}{2}, \\ \tilde{d} &= d_0 + \frac{(\sigma_{12} - g_0)^2}{2\tau} + \frac{1}{2} (u_2 - \sigma_{12}u_1)' (u_2 - \sigma_{12}u_1), \end{aligned} \tag{14}$$

and

$$\sigma_{12}|W, \beta, \xi^2 \sim N\left(\frac{g_0/\tau + u'_1 u_2}{\tau^{-1} + u'_1 u_1}, \frac{\xi^2}{\tau^{-1} + u'_1 u_1}\right). \quad (15)$$

The Gibbs sampler with identified parameters, which is similar to Li's (1998) algorithm, can now be summarized as follows:

Algorithm 2 (Identified Parameters) For given starting values of $\{\beta, \sigma_{12}, \xi^2, I, \mathcal{Y}^*\}$:

1. sample β from a multivariate normal with mean (9) and variance (10);
2. if $s_i = 1$, sample I_i from (5); if $s_i = 0$, sample I_i from (6) and y_i^* from (??);
3. sample ξ^2 from (14) and σ_{12} from (15);
4. return to step 1 and repeat.

2.3 Independence

An interesting special case arises when $\rho = 0$ or $\sigma_{12} = 0$. Then u_{i1} has no effect on y_i^* , whether we observe y_i^* or not, and inference on β_2 would only use the subsample where $y_i \neq 0$. Another way to see this is through the likelihood: (4) becomes

$$f(y, s|\beta, \Sigma) = \prod_{i=1}^n [\Phi(x'_{i1}\beta_1/\sigma_1)]^{s_i} [1 - \Phi(x'_{i1}\beta_1/\sigma_1)]^{1-s_i} \times \prod_{i:y_i \neq 0} \sigma_2^{-1} \phi\left(\frac{y_i - x'_{i2}\beta_2}{\sigma_2}\right). \quad (16)$$

This likelihood is identical to that of a two-part model, in which the nonzero outcomes are conditionally independent of the selection mechanism.⁵ Since the likelihood is separable in (β_1, σ_1^2) and (β_2, σ_2^2) , constructing a Gibbs sampler is now considerably easier. The function $f(y, s|\beta, \Sigma)$ factors into a Probit likelihood and a normal linear regression likelihood. If we impose independence between (β_1, σ_1^2) and (β_2, σ_2^2) in the prior, this is carried over to the posterior. As a consequence we can sample (β_1, σ_1^2) and (β_2, σ_2^2) each in their own 'mini Gibbs sampler'.

⁵The difference between a two-part model and a selection model concerns the interpretation of β_2 . In a selection model β_2 is the marginal effect on potential outcomes, whereas in a two-part model it is a marginal effect on realized non-zero outcomes.

In the Probit part we impose the restriction $\sigma_1 = 1$, because it reduces the number of steps needed in the algorithm. The Probit algorithm described here is due to Ahn and Powell (1993). The parameters are (I, β_1) and it remains to find the conditional posteriors $p(I|\beta_1, s)$ and $p(\beta_1|I, s) = p(\beta_1|I)$.⁶ Since $I = X_1\beta_1 + u_1$ and $u_1 \sim N(0, I_n)$, it follows that

$$f(I|\beta_1) = (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \left[e'e + (\beta_1 - \hat{\beta}_1)' X_1' X_1 (\beta_1 - \hat{\beta}_1) \right] \right\},$$

where $\hat{\beta}_1 = (X_1' X_1)^{-1} X_1' I$ and $e = I - X_1 \hat{\beta}_1$. Combining a normal $N(b_1, B_1)$ prior distribution for β_1 with the likelihood of I given above, we get

$$\begin{aligned} \beta_1|I &\sim N(\bar{b}_1, \bar{B}_1), \\ \bar{B}_1 &= (B_1^{-1} + X_1' X_1)^{-1}, \\ \bar{b}_1 &= (B_1^{-1} + X_1' X_1)^{-1} (B_1^{-1} b_1 + X_1' X_1 \hat{\beta}_1). \end{aligned} \tag{17}$$

Since $I_i|\beta_1$ has a normal distribution with mean $x'_{i1}\beta_1$ and unit variance, the distribution of I_i given β_1 and s_i is truncated normal:

$$\begin{aligned} f(I_i|\beta_1, s_i = 0) &= N(x'_{i1}\beta_1, 1) \mathbb{I}\{I_i \leq 0\}, \\ f(I_i|\beta_1, s_i = 1) &= N(x'_{i1}\beta_1, 1) \mathbb{I}\{I_i > 0\}. \end{aligned} \tag{18}$$

Inference on (β_2, σ_2^2) uses only the subsample in which $y_i \neq 0$. Let \tilde{y} , \tilde{X}_2 and $\tilde{u}_2 = \tilde{y} - \tilde{X}_2\beta_2$ all refer to this subsample of size \tilde{n} . If the priors are $\sigma_2^2 \sim IG(c_0, d_0)$ and $\beta_2 \sim N(b_2, B_2)$, then standard results for the linear model with normal errors yield

$$\sigma_2^2|\beta_2, \tilde{y} \sim IG\left(c_0 + \frac{\tilde{n}}{2}, d_0 + \frac{1}{2}\tilde{u}_2'\tilde{u}_2\right), \tag{19}$$

$$\beta_2|\sigma_2^2, \tilde{y} \sim N(\bar{b}_2, \bar{B}_2), \tag{20}$$

$$\bar{B}_2 = \left(B_2^{-1} + \sigma_2^{-2} \tilde{X}_2' \tilde{X}_2 \right)^{-1},$$

$$\bar{b}_2 = \left(B_2^{-1} + \sigma_2^{-2} \tilde{X}_2' \tilde{X}_2 \right)^{-1} \left(B_2^{-1} b_2 + \sigma_2^{-2} \tilde{X}_2' \tilde{X}_2 \hat{\beta}_2 \right),$$

⁶The equality follows because s is a function of I .

where $\hat{\beta}_2$ is the OLS estimator from regressing \tilde{y} on \tilde{X} . The simplified Gibbs sampler can now be summarized as follows:

Algorithm 3 (Independence) For given starting values of $(I, \beta_1, \beta_2, \sigma_2^2)$:

1. Sample β_1 from (17) and I from (18);
2. sample σ_2^2 from (19) and β_2 from (20);
3. return to step 1 and repeat.

2.4 Testing for Independence

In the bivariate normal selection model a test for the presence of a selection effect is a test of the hypothesis $H_0 : \rho = 0$. In a classical setting this may be done via a t-test on the coefficient of the inverse Mills ratio. In the Bayesian approach we could simply use the output of algorithms 1 or 2 to look at the posterior of ρ (or σ_{12}) and then determine whether zero is a likely value or not.

An alternative is to compute the Bayes factor.⁷ Suppose that two competing models, or hypotheses, M_0 and M_1 are entertained to describe the outcome y . A model in this context consists of a prior distribution on the appropriate parameters and a likelihood for the data. Given prior probabilities $f(M_0)$ and $f(M_1)$ of the two models, the posterior odds ratio is computed as

$$\begin{aligned} \frac{f(M_0|y)}{f(M_1|y)} &= \frac{f(y|M_0)}{f(y|M_1)} \times \frac{f(M_0)}{f(M_1)} \\ &= B_{01} \times \frac{f(M_0)}{f(M_1)}, \end{aligned}$$

where B_{01} is the Bayes factor of model 0 versus model 1. In other words, the Bayes factor transforms the prior odds ratio into the posterior odds ratio. The Bayes factor itself is the ratio of the prior predictive distributions or *marginal likelihoods* $f(y|M_j)$, where

$$f(y|M_j) = \int f_j(y|\theta_j, M_j) f_j(\theta_j|M_j) d\theta_j, \quad j = 0, 1.$$

Here θ_j is the set of parameters under model M_j and $f_j(y|\theta_j, M_j)$ the corresponding likelihood. Bayes factors larger than 1 indicate support for model M_0 . Conversely, a value of B_{01} much smaller

⁷Bayes factors have been researched extensively. The article by Kass and Raftery (1995) is an excellent survey.

than 1 suggests that model M_1 is more likely. In our context let M_0 be the selection model with $\rho = 0$ imposed, whereas M_1 is the unrestricted model. The output of algorithms 2-3 can be used to approximate B_{01} in a relatively straightforward way. Here we briefly discuss two approaches.

In the parameterization involving σ_{12} and ξ^2 , suppose that β and (σ_{12}, ξ^2) are independent in the prior. It then follows from the arguments in Verdinelli and Wasserman (1995) that B_{01} is equal to

$$B_{01} = f(\sigma_{12} = 0|y)E \left[\frac{1}{f(\sigma_{12} = 0|\xi^2)} \right],$$

where the expectation is with respect to $f(\beta, \xi^2|y, s, \sigma_{12} = 0)$. Given a sequence $\{\beta^{(m)}, \xi^{2(m)}, W^{(m)}\}_{m=1}^M$, the first term can be estimated by

$$\hat{f}(\sigma_{12} = 0|y) = \frac{1}{M} \sum_{m=1}^M f(\sigma_{12} = 0|W^{(m)}, \beta^{(m)}, \xi^{2(m)}).$$

This simply requires M evaluations of the density in (15) at zero. To estimate the expected value, run algorithm 2 again with $\sigma_{12} = 0$ held fixed. This yields a sequence $\{\beta^{(l)}, \xi^{2(l)}\}_{l=1}^L$ from which

$$\hat{E} \left[\frac{1}{f(\sigma_{12} = 0|\xi^2)} \right] = \frac{1}{L} \sum_{l=1}^L \left[\frac{1}{f(\sigma_{12} = 0|\xi^{2(l)})} \right].$$

This only requires L evaluations of the conditional prior of σ_{12} given ξ^2 .

A second method to compute the Bayes factor is the one proposed by Chib (1995). Output from the Gibbs sampling algorithms 2 and 3 can be used to estimate $f(y|M_0)$ and $f(y|M_1)$ separately and compute their ratio.⁸ Starting with the unrestricted model, we have

$$f(y|M_1) = \frac{f(y|\beta, \sigma_{12}, \xi^2) f(\beta, \sigma_{12}, \xi^2)}{f(\beta, \sigma_{12}, \xi^2|y)}.$$

Note that this equation holds for all parameter values in the support of $f(\beta, \sigma_{12}, \xi^2|y)$. Now pick a specific value $(\beta^*, \sigma_{12}^*, \xi^{*2})$, for example the sample average from the Gibbs output. Then

$$\log f(y|M_1) = \log f(y|\beta^*, \sigma_{12}^*, \xi^{*2}) + \log f(\beta^*, \sigma_{12}^*, \xi^{*2}) - \log f(\beta^*, \sigma_{12}^*, \xi^{*2}|y).$$

⁸Estimating the marginal likelihood from algorithm 1 yields unstable results. Apparently the lack of full identification in that sampler causes problems.

The first two terms on the right-hand side can immediately be calculated. It remains to estimate the value of the posterior. To this end, write

$$\log f(\beta^*, \sigma_{12}^*, \xi^{*2} | y) = \log f(\xi^{*2} | y) + \log f(\sigma_{12}^* | \xi^{*2}, y) + \log p(\beta^* | \sigma_{12}^*, \xi^*, y).$$

Each term can be estimated using the conditional posterior and averaging over the draws from the Gibbs sampler. For the second and third terms, the sampler needs to be run with ξ^2 fixed at ξ^{*2} and (σ_{12}, ξ^2) fixed at $(\sigma_{12}^*, \xi^{*2})$, respectively. The argument for estimating $f(y | M_0)$ is analogous. For more detail, see Chib (1995).

3 Mixture Modeling

3.1 Finite Mixtures of Normals

There are several ways the assumption of bivariate normality in the SSM can be relaxed. We first consider using a mixture of normal distributions with a fixed number of mixture components. In the model in (1) suppose now that

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \sum_{j=1}^k \gamma_j N(\mu_j, c_u \Sigma_j), \quad \gamma_j \geq 0, \quad \sum_{j=1}^k \gamma_j = 1,$$

where $c_u > 0$ is a scaling factor and

$$\mu_j = \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \end{pmatrix}, \quad \Sigma_j = \begin{bmatrix} \sigma_{j,1}^2 & \sigma_{j,12} \\ \sigma_{j,12} & \sigma_{j,2}^2 \end{bmatrix}.$$

Thus, the distribution is a mixture of k components and with probability γ_j the errors are distributed according to $N(\mu_j, c_u \Sigma_j)$. Mixtures of normals are very flexible: even with a small number of mixture components (say, 2 or 3), they can display skewness, excess kurtosis and multimodality (e.g., Geweke 2005, chapter 6). Note that the mixture distribution trivially reduces to (2) when $k = 1$, $\mu_j = 0$ and $c_u = 1$.

For each observation let $\zeta_i \in \{1, \dots, k\}$ be a component indicator, so that

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} | \zeta_i = j \sim N(\mu_j, c_u \Sigma_j), \quad j = 1, \dots, k.$$

Since $\zeta = (\zeta_1, \dots, \zeta_n)$ is unknown, it is treated as a parameter vector. The complete set of parameters in the mixture model is now $\{\beta, I, \mathcal{Y}^*, \zeta, c_u\}$, combined with $\gamma = (\gamma_1, \dots, \gamma_k)$, $\mu = \{\mu_j\}_{j=1}^k$ and $\Sigma = \{\Sigma_j\}_{j=1}^k$.

We highlight some of the main features of the sampler and leave additional details to appendix 6. Recall that in the SUR formulation $W = X\beta + u$. Let $W_{(j)}, X_{(j)}, u_{(j)}$ be the submatrices corresponding to the j^{th} mixture component. Similar to (8) we can write

$$\begin{aligned} f(W|\beta, \zeta, \mu, \Sigma, c_u) &\propto \prod_{j=1}^k |c_u \Sigma_j|^{-n_j/2} \times \exp \left\{ -\frac{1}{2} \sum_{j=1}^k e'_{(j)} S_j^{-1} e_{(j)} \right\} \\ &\times \exp \left\{ -\frac{1}{2} (\beta - \hat{\beta})' \sum_{j=1}^k X'_{(j)} S_j^{-1} X_{(j)} (\beta - \hat{\beta}) \right\}, \end{aligned}$$

where $S_j = c_u \Sigma_j \otimes I_{n_j}$ and $n_j = \sum_{i=1}^n \mathbb{I}\{\zeta_i = j\}$. Here $\hat{\beta}$ is the GLS estimator

$$\hat{\beta} = \left[\sum_{j=1}^k X'_{(j)} S_j^{-1} X_{(j)} \right]^{-1} \sum_{j=1}^k X'_{(j)} S_j^{-1} (W_{(j)} - \mu_j \otimes \iota_{n_j}),$$

and $e_{(j)} = W_{(j)} - \mu_j \otimes \iota_{n_j} - X_{(j)} \hat{\beta}$ the residual. Note that once we condition on the component indicators ζ , the completed-data likelihood does not depend on γ . Provided β is a priori independent of γ , the same will be true for the conditional posterior of β . Combining the likelihood with a $N(b_0, B_0)$ prior, the conditional posterior is again normal with mean and variance given by

$$E(\beta|W, \zeta, \mu, \Sigma, c_u) = V(\beta|W, \zeta, \mu, \Sigma, c_u) \left[B_0^{-1} b_0 + \sum_{j=1}^k X'_{(j)} S_j^{-1} X_{(j)} \hat{\beta} \right] \quad (21)$$

$$V(\beta|W, \zeta, \mu, \Sigma, c_u) = \left[B_0^{-1} + \sum_{j=1}^k X'_{(j)} S_j^{-1} X_{(j)} \right]^{-1}. \quad (22)$$

Sampling I , \mathcal{Y}^* and Σ is similar as before, so we will be brief. For simplicity we do not list every

conditioning argument:

$$I_i | s_i = 1, \zeta_i = j \sim N \left(x'_{i1} \beta_1 + \mu_{j1} + \frac{\sigma_{j,12}}{\sigma_{j,2}^2} (y_i^* - x'_{i2} \beta_2 - \mu_{j2}), c_u \sigma_{j,1}^2 (1 - \rho_j^2) \right) \mathbb{I}\{I_i > 0\}, \quad (23)$$

$$I_i | s_i = 0, \zeta_i = j \sim N(x'_{i1} \beta_1 + \mu_{j1}, c_u \sigma_{j,1}^2) \mathbb{I}\{I_i \leq 0\}, \quad (24)$$

$$y_i^* | I_i, \zeta_i = j \sim N \left(x'_{i2} \beta_2 + \mu_{j2} + \frac{\sigma_{j,12}}{\sigma_{j,1}^2} (I_i - x'_{i1} \beta_1 - \mu_{j1}), c_u \sigma_{j,2}^2 (1 - \rho_j^2) \right). \quad (25)$$

The component indicator ζ_i has a prior multinomial distribution with $\Pr\{\zeta_i = j | \gamma\} = \gamma_j$, $j = 1, \dots, k$. Write the i^{th} observation as the bivariate linear model $w_i = X_i \beta + u_i$, where

$$w_i = (I_i, y_i^*)', \quad X_i = \begin{bmatrix} x'_{i1} & 0 \\ 0 & x'_{i2} \end{bmatrix}, \quad u_i = (u_{i1}, u_{i2})'.$$

The conditional posterior distribution of ζ_i follows from Bayes' rule:

$$\begin{aligned} \Pr\{\zeta_i = j | w_i, \beta, \mu, \Sigma, \gamma, c_u\} &\propto p(w_i | \beta, \mu, \Sigma, \gamma, c_u, \zeta_i = j) \times \Pr\{\zeta_i = j | \gamma\} \\ &\propto \gamma_j |c_u \Sigma_j|^{-1/2} \\ &\quad \times \exp \left\{ -\frac{1}{2c_u} (w_i - X_i \beta - \mu_j)' \Sigma_j^{-1} (w_i - X_i \beta - \mu_j) \right\}. \end{aligned} \quad (26)$$

The parameter γ is a set of multinomial probabilities, which suggests using a Dirichlet prior distribution. If $f(\gamma_1, \dots, \gamma_k) = \mathcal{D}(\gamma_0, \dots, \gamma_0)$, it is shown in the appendix that the posterior is given by

$$\gamma | \zeta \sim \mathcal{D}(\gamma_0 + n_1, \dots, \gamma_0 + n_k). \quad (27)$$

The use of a uniform Dirichlet prior, i.e. γ_0 is the parameter for each state $j = 1, \dots, k$, is appropriate here, because the state labels are not identified. This lack of knowledge is reflected by a prior in which the states are exchangeable.

In order to find the conditional posterior of (μ_j, Σ_j) we take its prior to be

$$f(\Sigma_j) = \mathcal{W}^{-1}(H, v), \quad f(\mu_j | c_u, \Sigma_j) = N(0, \tau c_u \Sigma_j).$$

The advantage of this choice is that the posterior of (μ_j, Σ_j) conditional on (β, ζ, c_u) is again of the

normal-inverse Wishart form. This allows us to sample (μ_j, Σ_j) jointly, which should improve the convergence behavior of the Markov chain.⁹ It follows that

$$f(\Sigma_j|W, \beta, \zeta, c_u) = \mathcal{W}^{-1}(H_j, v + n_j), \quad (28)$$

$$H_j = H + \frac{1}{c_u} \frac{n_j \bar{u}_j \bar{u}_j'}{1 + \tau n_j} + \frac{1}{c_u} \sum_{i: \zeta_i=j} (u_i - \bar{u}_j)(u_i - \bar{u}_j)',$$

and $\bar{u}_{(j)} = n_j^{-1} \sum_{\zeta_i=j} u_i$. Also:

$$f(\mu_j|\Sigma_j, W, \beta, \zeta, c_u) = N\left(\frac{\tau n_j}{1 + \tau n_j} \bar{u}_{(j)}, \frac{\tau c_u}{1 + \tau n_j} \Sigma_j\right). \quad (29)$$

Allowing for prior dependence between μ_j and Σ_j is often reasonable: if $c_u \Sigma_j$ is large the errors are highly variable, making it harder to pin down their location. The parameter τ can be thought of as a tuning parameter that can be used to control the potential multimodality in the distribution of u_i , independently of the variance.

The conditionally conjugate prior distribution for c_u is the inverse gamma distribution. In particular, we take $f(c_u) = IG(v/2, d)$. Since c_u enters the prior of μ_j and the likelihood, it is shown in the appendix that

$$f(c_u|W, \delta, \mu, \Sigma, \zeta, \gamma) = IG\left(\frac{v}{2} + n + k, \tilde{d}\right), \quad (30)$$

$$\tilde{d} = d + \frac{1}{2\tau} \sum_{j=1}^k \mu_j' \Sigma_j^{-1} \mu_j + \frac{1}{2} \sum_{i=1}^n (u_i - \mu_{j(i)})' \Sigma_{j(i)}^{-1} (u_i - \mu_{j(i)}),$$

where $j(i)$ is the value of $j \in \{1, \dots, k\}$ such that $\zeta_i = j$. The Gibbs sampler for the mixture of normals sample selection model can now be summarized as follows:

Algorithm 4 (Mixture SSM) For given starting values of $\{\beta, \mu, \Sigma, \zeta, \gamma, c_u, I, \mathcal{Y}^*\}$:

1. Sample β from a normal distribution with mean (21) and variance (22);
2. if $s_i = 1$ sample I_i from (23); if $s_i = 0$ sample I_i from (24) and y_i^* from (25);
3. sample ζ_i from (26) for $i = 1, \dots, n$;

⁹Alternatively, if μ_j is a priori independent of Σ_j , it would be necessary to sample from $p(\mu_j|\Sigma_j, W, \beta, \zeta, c_u)$ and $p(\Sigma_j|\mu_j, W, \beta, \zeta, c_u)$ consecutively. This adds a 'block' to the Gibbs sampler, which slows down convergence.

4. sample γ from (27);
5. sample Σ_j from (28 and μ_j from (29) for $j = 1, \dots, k$;
6. sample c_u from (30);
7. return to step 1 and repeat.

The priors used throughout this section are proper. It is important to note that algorithm 4 cannot be modified to allow for improper priors. As shown by Roeder and Wasserman (1997), the use of improper priors in a mixture model leads to improper posterior distributions that are meaningless. Also, the component labels j are not identified without further prior information. If the components have a structural interpretation, such as representing different states of the world or distinct subpopulations, and are of primary importance, then the algorithm above is not appropriate.¹⁰ In our case, however, we merely use mixtures as a modeling device, and labeling issues are not a concern.

3.2 Dirichlet Mixtures of Normals

The method discussed above requires one to choose the number of mixture components beforehand. If the econometrician is uncomfortable doing this, he could use various choices of k and compare models on the basis of their posterior probabilities. An alternative which does not require model selection is to use the so-called Dirichlet process prior, developed by Ferguson (1973) and Antoniak (1974). In this section we incorporate this methodology into a sample selection framework. The main appeal of using a Dirichlet process prior lies in the fact that the error distribution is modeled as a mixture of normals with a *random* number of components. This number is an additional parameter and through Bayesian updating we can obtain its posterior. Also, the prior allows us to center in some sense the semiparametric model around the parametric one. Our approach here is based on results from Escobar (1994), Escobar and West (1995) and is closely related to Conley, Hansen, McCulloch, and Rossi (2007), who consider the use of Dirichlet process priors in an instrumental variables model.¹¹

¹⁰For a discussion, see Geweke (2005, chapter 6) and the references cited therein.

¹¹Escobar and West (1998), MacEachern (1998) and Müller and Quintana (2004) are excellent reviews of semi-parametric modeling with Dirichlet processes.

The basic setup can be described as follows. In (1) let

$$\begin{aligned} \begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} | \theta_i &\sim N(\mu_i, c_u \Sigma_i), \quad \theta_i = (\mu_i, \Sigma_i), \\ \theta_i | G &\sim G, \quad G | \alpha, G_0 \sim \mathcal{DP}(\alpha, G_0). \end{aligned} \tag{31}$$

Here θ_i is simply the set of parameters for the normal distribution (apart from the common scale factor c_u). Our discussion of the SSM in section 2 involved setting $\theta_i = (0, \Sigma)$ for all i and specifying a prior on the elements of Σ . The semiparametric model in (31) allows each pair (u_{i1}, u_{i2}) to have a distinct normal distribution, conditional on θ_i . The parameters $\{\theta_i\}_{i=1}^n$ are i.i.d. draws from a distribution G . If G is chosen to be a $N(\mu_0, S_0)$ distribution, possibly augmented with a hyperprior on (μ_0, S_0) , then we have specified a hierarchical normal model, which still imposes a lot of structure on, say, the marginal distribution of the errors. In particular, it would not allow multimodality or skewness. Instead, in (31) the distribution G itself is treated as unknown and given a Dirichlet process (DP) prior¹². Thus, G can be viewed as a *random* probability measure. G_0 is a distribution that in some sense is a prior guess about G . Specifically, the marginal prior distribution of θ_i is exactly G_0 (e.g., Ferguson, 1973, Antoniak, 1974). The parameter α reflects the prior belief that G_0 is the actual distribution of θ_i . This belief becomes stronger as $\alpha \rightarrow \infty$.

We present here a Gibbs sampler while keeping α fixed. In the simulations and application in sections 4 and 5 α is random and updated in the Gibbs sampler. Additional details on the prior of α and the form of G_0 are discussed in appendix 6. Throughout this section the conditioning on (α, G_0) is implicit. The parameters are now $\{\theta_i\}_{i=1}^n$, G and c_u , in addition to β , I and \mathcal{Y}^* . A convenient simplification occurs because G can be integrated out of the posterior (e.g., Escobar, 1994), so that the Gibbs sampler does not involve G .

The likelihood of W , conditional on $\theta = \{\theta_i\}_{i=1}^n$, is

$$p(W | \beta, \theta, c_u) \propto \prod_{i=1}^n |c_u \Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2c_u} (w_i - X_i \delta - \mu_i)' \Sigma_i^{-1} (w_i - X_i \delta - \mu_i) \right\}.$$

Combining this with a $N(b_0, B_0)$ prior for β and collecting terms, it follows that the posterior of β

¹²Suppose Ω is the sample space and $\{A_j\}_{j=1}^k$ is any measurable partition. If $G \sim \mathcal{DP}(\alpha, G_0)$, then the collection of random probabilities $\{G(A_j)\}_{j=1}^k$ follows a Dirichlet distribution.

is again normal with mean and variance given by

$$E(\beta|W, \theta, c_u) = V(\beta|W, \theta, c_u) \left[B_0^{-1} b_0 + c_u^{-1} \sum_{i=1}^n X_i' \Sigma_i^{-1} X_i \hat{\beta} \right], \quad (32)$$

$$V(\beta|W, \theta, c_u) = \left[B_0^{-1} + c_u^{-1} \sum_{i=1}^n X_i' \Sigma_i^{-1} X_i \right]^{-1}. \quad (33)$$

As before $\hat{\beta}$ is the GLS estimator, which in this case equals

$$\hat{\beta} = \left[\sum_{i=1}^n X_i' \Sigma_i^{-1} X_i \right]^{-1} \sum_{i=1}^n X_i' \Sigma_i^{-1} (w_i - \mu_i).$$

Sampling I_i when $s_i = 1$ and (I_i, y_i^*) when $s_i = 0$ is done by generating draws from the distributions in (23), (24) and (25), where we now condition on $\theta_i = (\mu_i, \Sigma_i)$, instead of ζ_i .

Let $\theta_{-i} = \{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n\}$. Blackwell and MacQueen (1973) show that if $\theta_i|G \sim G$ and $G \sim \mathcal{DP}(\alpha, G_0)$, then the distribution of θ_i given θ_{-i} with G integrated out is given by

$$\theta_i|\theta_{-i} \begin{cases} = \theta_j & \text{w. prob. } \frac{1}{\alpha+n-1}, \quad j \neq i \\ \sim G_0 & \text{w. prob. } \frac{\alpha}{\alpha+n-1} \end{cases}. \quad (34)$$

That is, θ_i equals one of the other θ_j 's with nonzero probability, or is distributed according to G_0 . This property is known as the Pólya urn representation of a sample from the Dirichlet process. Using Bayes' rule the posterior distribution takes a similar form:

$$\theta_i|\theta_{-i}, W, \beta \begin{cases} = \theta_j & \text{w. prob. } c^{-1} p(w_i|X_i, \beta, \theta_j, c_u), \quad j \neq i \\ \sim p(\theta_i|w_i, X_i, \beta, c_u) & \text{w. prob. } c^{-1} \alpha p(w_i|X_i, \beta, c_u) \end{cases}. \quad (35)$$

Here c^{-1} is a normalizing constant, $p(\theta_i|w_i, \beta, c_u)$ is the posterior obtained from the prior $dG_0(\theta_i)$, and $p(w_i|X_i, \beta, c_u)$ is the marginal likelihood after integrating out θ_i with respect to $dG_0(\theta_i)$:

$$\begin{aligned} p(\theta_i|w_i, X_i, \beta, c_u) &\propto p(w_i|X_i, \beta, \theta_i, c_u) dG_0(\theta_i), \\ p(w_i|X_i, \beta, c_u) &= \int p(w_i|X_i, \beta, \theta_i, c_u) dG_0(\theta_i). \end{aligned}$$

More details are given in appendix 6.

Finally, updating c_u conditional on the remaining parameters is exactly the same as in the finite mixture case discussed earlier. The Gibbs sampler for the Dirichlet process selection model can now be summarized as follows:

Algorithm 5 (Dirichlet process) *For given starting values of $\{\beta, \theta, I, \mathcal{Y}^*, c_u\}$:*

1. *Sample β from a normal distribution with mean (32) and variance (33);*
2. *if $s_i = 1$ sample I_i from (23); if $s_i = 0$ sample I_i from (24) and y_i^* from (25); all draws here are conditional on $\theta_i = (\mu_i, \Sigma_i)$;*
3. *sample θ_i from (35) for $i = 1, \dots, n$;*
4. *sample c_u from (30);*
5. *return to step 1 and repeat.*

Algorithm 5 can be extended in several ways. As noted before it is possible to place a prior distribution on α . Recall that α represents the prior belief that G_0 is the distribution of θ_i . If α is large, then we will see many unique values in θ , which yields a model with a large number of mixture components. Alternatively, if α is small, then θ will likely see few unique values. In fact, Antoniak (1974) shows that k_n , the number of unique θ values in a sample of size n , satisfies $E(k_n|\alpha) \approx \alpha \log((\alpha + n)/\alpha)$. The limit case $\alpha = 0$ results in $\theta_i = \bar{\theta}$ for all i , which leads to the parametric model in section 2. By placing a prior on α it is possible to learn about the number of mixture components, after seeing the data. Escobar and West (1995) use a convenient gamma prior which yields a posterior mixture of two gamma distributions. See appendix 6 for details. Alternatively, it is possible to specify a prior for k_n explicitly (rather than implicitly through α) as in Escobar (1994).

The Markov chain constructed in algorithm 5 may converge slowly if the Gibbs sampler ‘gets stuck’ at a few fixed values of θ_i . From (35) this could happen when the sum (over $j \neq i$) of $c^{-1}p(w_i|X_i, \beta, \theta_j, c_u)$ gets large relative to $c^{-1}\alpha p(w_i|X_i, \beta, c_u)$. It is possible to slightly reparameterize the model and associated Gibbs sampler, such that at each iteration the distinct values in θ are ‘remixed’. The Gibbs samplers in sections 4 and 5 use remixing. Details are given in appendix 6; see also West et al. (1994), MacEachern (1998) and MacEachern and Müller (1998).

3.3 Independence and Departures from Normality

Algorithm 5 provides information on the number of mixture components that are used to fit the likelihood. At each iteration of the sampler we can determine and store the number k of unique elements in $\{\theta_i\}_{i=1}^n$. The sampled values of k then allow us to calculate the posterior distribution of k . In particular, if $k = 1$ has a low posterior probability, then the normal likelihood based on (2) is not appropriate. Thus, an immediate by-product of the Dirichlet process Gibbs sampler is a test for (departures from) normality.

In the case of a finite mixture model with k fixed at a value greater than 1, it is less obvious how to construct a test for normality. Suppose for example that we use the Gibbs sampler in algorithm 4 with $k = 2$. If the true model has bivariate normal errors, then we are overfitting the number of components. The posterior should then reveal $\mu_1 \approx \mu_2$ and $\Sigma_1 \approx \Sigma_2$. Constructing a test statistic would involve comparing a relatively large number of parameters. This approach is therefore not as appealing as using the Dirichlet mixture directly.

A potential complication in the mixture model is that the dependence between u_{i1} and u_{i2} is not easily captured by a single parameter. Recall that when the errors are bivariate normal, a test for the presence of a selection effect can be based on the correlation coefficient ρ . In a mixture model this is no longer true. Moreover, it is not clear that correlation is a meaningful measure of dependence.

The degree of dependence between u_{i1} and u_{i2} can be assessed by drawing a set of values from the posterior predictive distribution, and calculating a dependence measure from these draws. In the mixture model the sampled values $\{\gamma, \mu, \Sigma, c_u\}$ are (approximately) draws from the posterior, and these are the only parameters that determine the error distribution. A predicted realization u_{n+1} can be then be generated according to

$$u_{n+1} | \gamma, \mu, \Sigma, c_u \sim \sum_{j=1}^k \gamma_j N(\mu_j, c_u \Sigma_j).$$

In the Dirichlet mixture model a predicted value is generated as follows: after updating $\{\mu_i, \Sigma_i\}_{i=1}^n$, sample a new value $\{\mu_{n+1}, \Sigma_{n+1}\}$ from (35) and sample u_{n+1} from $N(\mu_{n+1}, c_u \Sigma_{n+1})$. Given a set of draws $\{u_{n+1}^{(m)}\}_{m=1}^M$ the correlation between $u_{n+1,1}$ and $u_{n+1,2}$ can be computed. Of course, with

a mixture distribution a zero correlation does not imply the absence of a selection effect. It is therefore useful to also report alternative measures such as Kendall’s tau and Spearman’s rho (e.g., Nelsen, 2006).

4 Simulations

We now illustrate the use of the Dirichlet mixture model in algorithm 5 using some simulated data. A sample of size $n = 1,000$ is generated from the model in (1):

$$\begin{aligned} I_i &= \beta_0 + x_{i1}\beta_1 + u_{i1}, \\ \log y_i^* &= \beta_0 + x_{i2}\beta_2 + u_{i2}, \\ y_i &= \mathbb{I}\{I_i > 0\}y_i^*. \end{aligned}$$

In the first design the errors (u_{i1}, u_{i2}) are bivariate normal with $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and correlation $\rho = 0.5$. The coefficients are $\beta_1 = \beta_2 = 1$, and x_{i1} and x_{i2} are independently $U(0, 6)$ distributed. Note that the intercept β_0 controls the fraction of the sample that has $y_i = 0$. We set $\beta_0 = -1.50$ which implies that roughly 25% of the outcomes $\log y_i^*$ are missing (and labeled zero). The algorithm is run three times from different starting values. The number of iterations is 8,000, the first 4,000 of which are discarded. Thus, the figures in this section are estimated posterior distributions based on 12,000 random draws. The posterior estimates from algorithm 5 are given in figures 1 and 2. The posterior distributions of β_2 and ρ are centered around their true values. The sampler overwhelmingly fits a single mixture component to the data, with $\Pr\{k = 1|y\}$ close to 90%. The contour plot of $f(u_{i1}, u_{i2})$ shows that the normal model describes the error distribution well near the center. In the tail areas it is likely that the sampler is fitting more than one mixture component, thus creating some nonnormality.

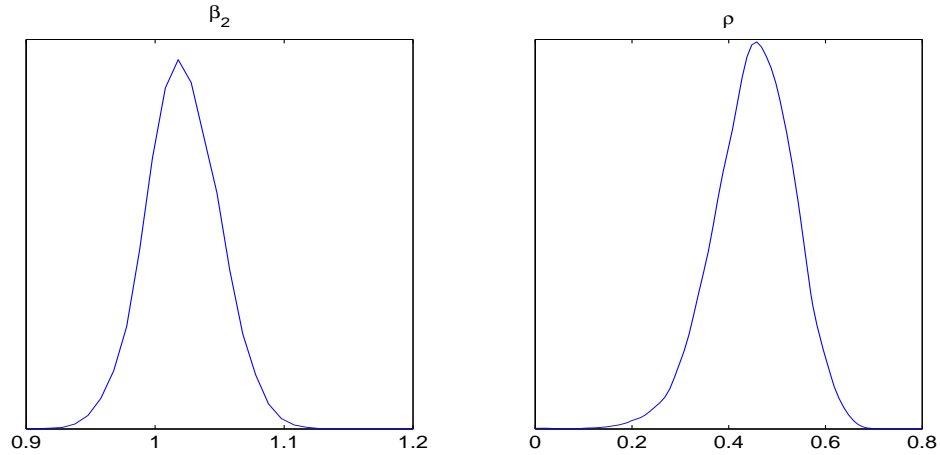


Figure 1: Dirichlet posterior of β_2 and ρ (normal errors)

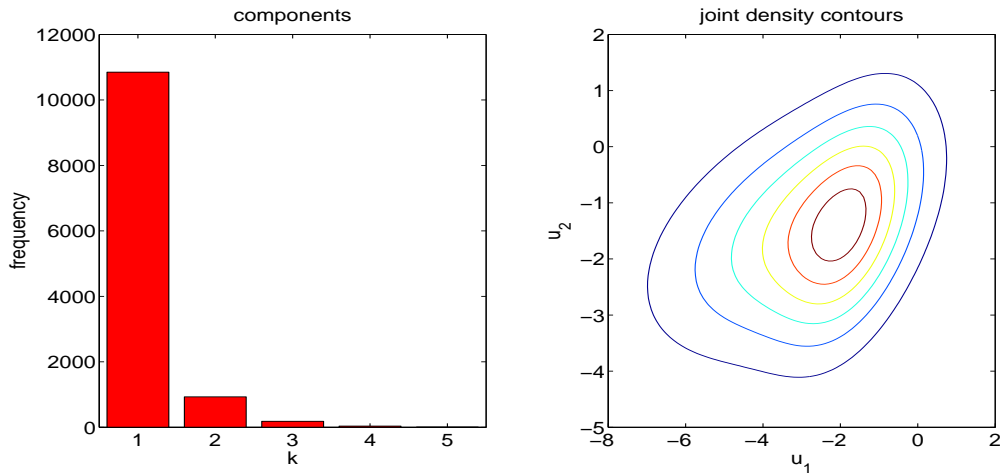


Figure 2: number of components and error density contours (normal errors)

The second design is largely similar, except that (u_{i1}, u_{i2}) is now distributed according to a mixture of two normals:

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \gamma_1 N(\mu_1, \Sigma) + (1 - \gamma_1) N(\mu_2, \Sigma),$$

where $\gamma_1 = 0.3$, $\mu_1 = (0, -2.1)$ and $\mu_2 = (0, 0.9)$. The covariance matrix is the same as before. The intercept is set to $\beta_0 = -1.50$, so that around 25% of the sample has a missing outcome. From figures 3 and 4 we see that the posterior of β_2 is slightly upward biased. The posterior of ρ strongly indicates a positive relation between u_{i1} and u_{i2} . The Dirichlet model correctly fits two mixtures

components in 55% of the time, with the posterior probability of three components dropping down to 26%.

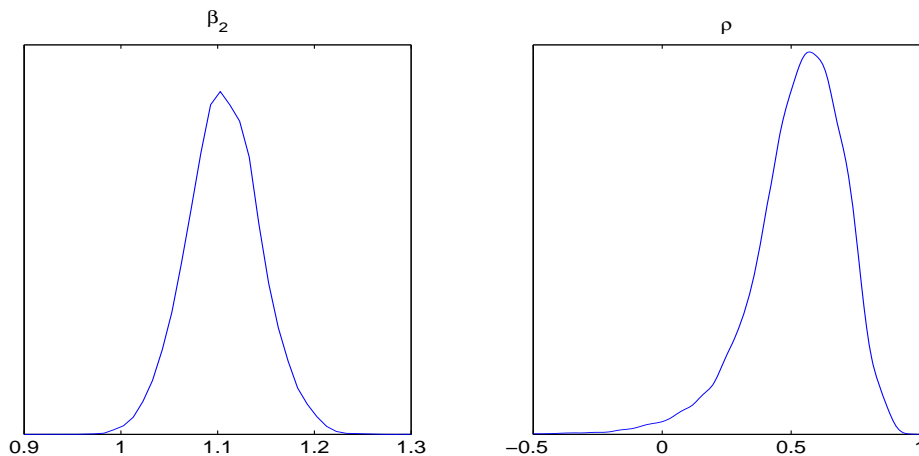


Figure 3: Dirichlet posterior of β_2 and ρ (mixture errors)

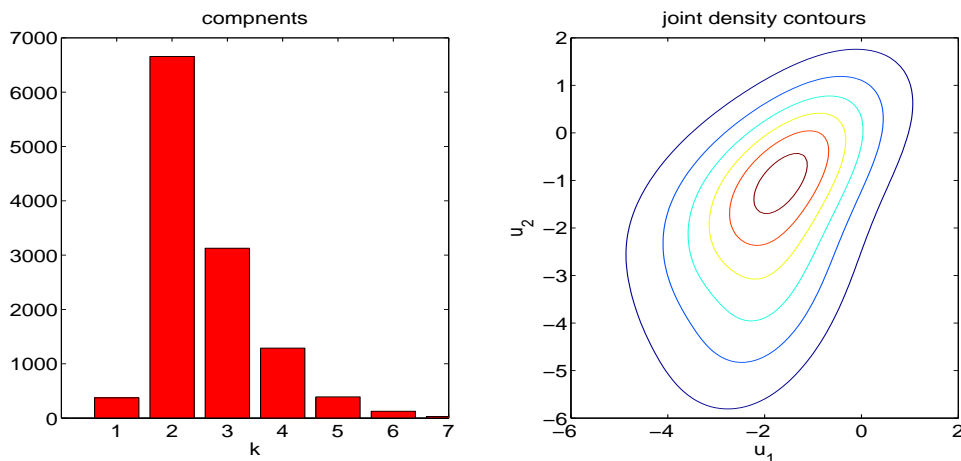


Figure 4: number of components and error density contours (mixture errors)

In the final design we generate v_{i1} and v_{i2} as independent $\log(\chi_1^2)$ errors and set $u_{i1} = v_{i1} + 0.25v_{i2}$ and $u_{i2} = 0.25v_{i1} + v_{i2}$. This creates heavily left-skewed error with a correlation of 0.47. The intercept is set to $\beta_0 = 0.25$, which results in roughly 25% missing outcomes. The posterior of β_2 in figure 5 is quite accurate, but the correlation is harder to pin down. On the other hand, figure 6 shows that there is substantial evidence for nonnormality: the posterior probability of either three or four mixture components is 46%.

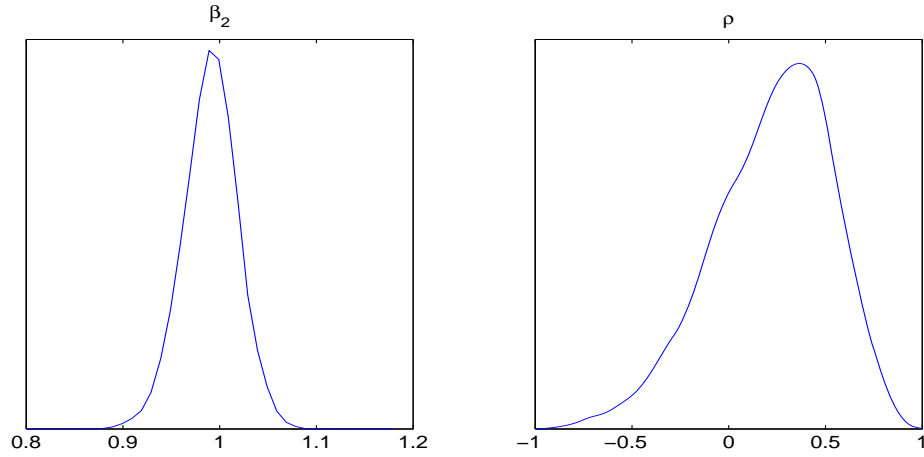


Figure 5: Dirichlet posterior of β_2 and ρ ($\log(\chi_1^2)$ errors)

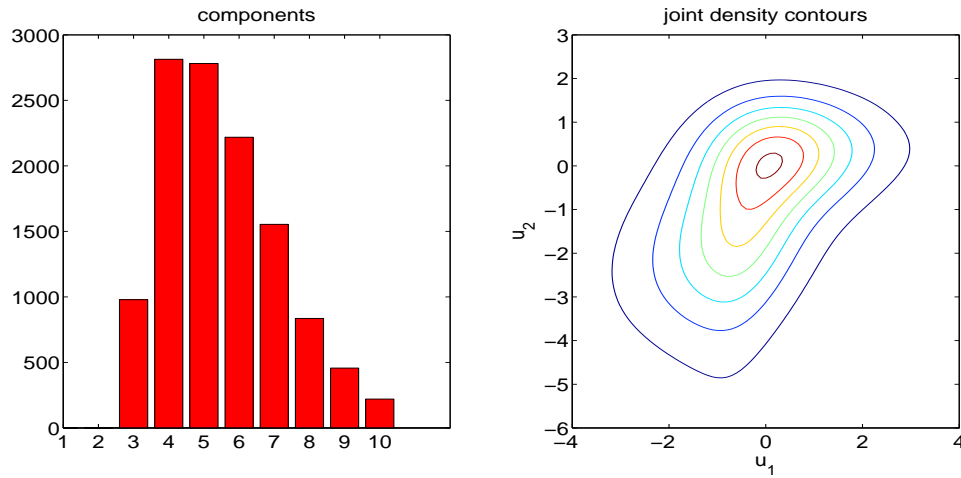


Figure 6: number of components and error density contours ($\log(\chi_1^2)$ errors)

5 Mroz's (1987) Labor Supply Data

In this section we apply the various Gibbs samplers to estimate a wage equation, using the labor supply data of Mroz (1987). This data is a PSID subsample from 1975 and contains information on 753 married women, 428 of whom worked at some point during the year. We estimate the

parameters of the following wage function:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exp}_i + \beta_3 \text{exp}_i^2 + u_{i2}, \quad (36)$$

where educ_i and exp_i denote education and experience levels, respectively. Wage is only observed if the individual selects into the work force. The selection equation contains the covariates in (36), plus age, family income (excluding the woman’s income) and the number of children in the household. Since the results for algorithms 1 and 2 are largely similar, we only report results for the identified sampler. We first assess whether wage outcomes are conditionally independent of selection into the work force.

Model	Parameter	Mean	Std. Dev.	95% HPD
SSM	β_1	0.108	0.015	[0.078, 0.137]
	β_2	0.042	0.015	[0.010, 0.070]
	β_3	-0.001	0.000	[-0.002, 0.000]
SSM ($\rho = 0$)	β_1	0.107	0.014	[0.079, 0.135]
	β_2	0.042	0.013	[0.016, 0.068]
	β_3	-0.001	0.000	[-0.001, 0.000]

Table 1: wage equation parameters, algorithms 2 and 3

Table 1 shows that both algorithms 2 and 3 yield very similar posteriors for the coefficients of the wage equation. This is true because there is no evidence that the correlation between u_{i1} and u_{i2} is nonzero. The Bayes factor of $\rho = 0$ versus the alternative is 12.01. To allow for potential nonnormality, we also estimate the selection model with a mixture of two normals and the Dirichlet mixture of normals.

Model	Parameter	Mean	Std. Dev.	95% HPD
Mixture ($k = 2$)	β_1	0.105	0.011	[0.083, 0.127]
	β_2	0.025	0.013	[0.000, 0.049]
	β_3	-0.000	0.000	[-0.001, 0.000]
Dirichlet	β_1	0.104	0.010	[0.082, 0.123]
	β_2	0.024	0.012	[0.001, 0.048]
	β_3	-0.000	0.000	[-0.001, 0.000]

Table 2: wage equation parameters, algorithms 4 and 5

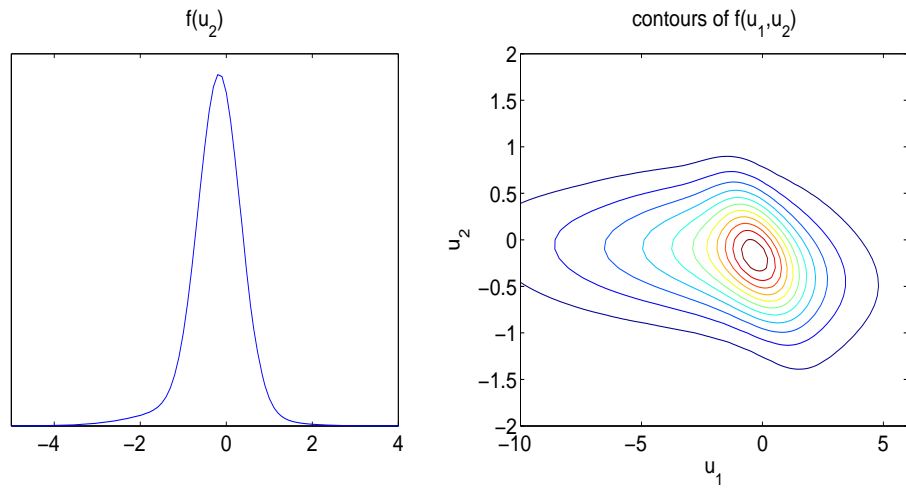


Figure 7: estimated error densities from mixture model ($k = 2$)

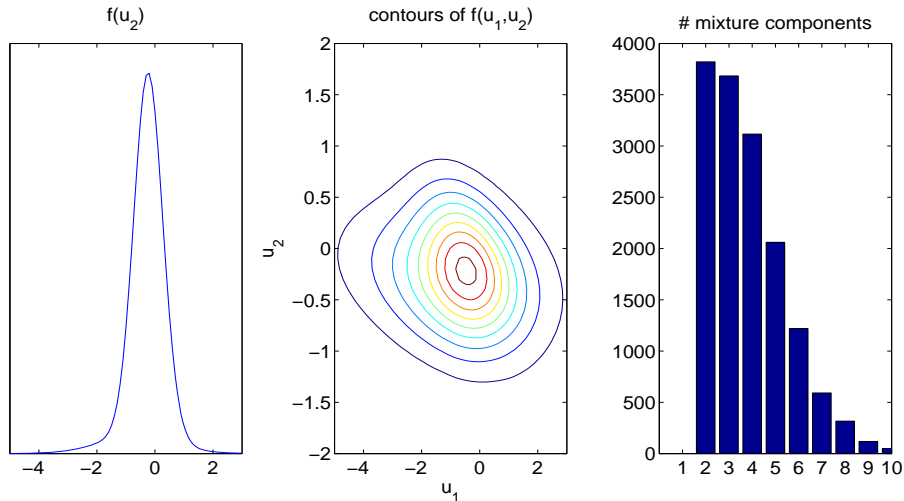


Figure 8: estimated error densities from Dirichlet model

The (unconditional) correlation between the selection and wage equations in the mixture model has posterior mean -0.201 and 95% HPD $[-0.49, 0.12]$. Thus, at least in terms of correlation there is no evidence of selection on unobservables. The mixture model shows some evidence of nonnormality in the error distribution in figure 7, but the posteriors summary statistics for β_1 and β_3 are not very different from those given in table 1. Interestingly, the return to experience is lower in the mixture and Dirichlet models, while the posterior of the return to education is less dispersed. From figure 8 we see that the Dirichlet sampler fits only a small number of mixture components to the error distribution. The posterior probability of 2, 3 or 4 components is 25%, 25% and 21%, respectively.

6 Conclusion

In this paper we have developed Gibbs sampling algorithms for use in a sample selection models. The model essentially consists of a latent structure which is only partially observed. Under the strong parametric assumption of normal errors, the Gibbs sampler is easy to construct in combination with data augmentation. The sampling scheme can be formulated with or without an identification restriction, but we have found there to be hardly any difference in the posterior approximation.¹³

¹³In contrast, McCulloch et al. (2000) find large differences in the autocorrelation and convergence behavior of the chains, in case of the multinomial probit model.

A Bayesian semiparametric procedure can be based on introducing more flexibility in the (joint) error distribution. One option is the use of a mixture of normal distributions. We have shown how to construct a Gibbs sampler by augmenting the parameter space with a set of latent state variables. Within this sampling algorithm the number of mixture components is fixed. In principle different specifications can be compared on the basis of a Bayes factor.

A second option and an attractive alternative to comparing many different models, is the use of Dirichlet process mixtures. We have modeled the errors as having a bivariate normal distribution whose parameters may or may not differ across observations. Thus, the Dirichlet process mixture amounts to specifying a mixture distribution with an unknown number of components. The data is then used to make inference about this number. Our paper also provides a Gibbs sampler for this case. The only requirement for tractability is the choice of conjugate priors.

The use of Dirichlet mixtures in modeling sample selection is appealing for several reasons. First, the likelihood is based on a flexible mixture distribution, rather than imposing normality. Second, the Gibbs sampler includes an automatic fitting of the number of mixture components. The posterior distribution of this number can then be used to test for normality. Finally, the sampler may be used to check for the presence of a selection effect. It should be noted here that we have not compared our methods to the control function approach, which is more common in classical econometrics. Recently, Chib et al. (2008) have proposed a Gibbs sampler based on a flexible form of the regression function. We leave a thorough comparison with their method to future work.

Appendix A: Finite Mixtures

Here we provide some additional details on implementing the Gibbs sampler in algorithm 4.

Sampling ζ_i

To sample ζ_i from its posterior distribution (26) in the mixture model, we can use the following steps:

1. Calculate

$$\Pr\{\zeta_i = j | w_i, \beta, \mu, \Sigma, \gamma, c_u\} = \frac{\gamma_j |c_u \Sigma_j|^{-1/2} \exp\left\{-\frac{1}{2c_u}(w_i - X_i\beta - \mu_j)' \Sigma_j^{-1}(w_i - X_i\beta - \mu_j)\right\}}{\sum_{l=1}^k \gamma_l |c_u \Sigma_l|^{-1/2} \exp\left\{-\frac{1}{2c_u}(w_i - X_i\beta - \mu_l)' \Sigma_l^{-1}(w_i - X_i\beta - \mu_l)\right\}}.$$

2. Calculate the CDF $\Pr\{\zeta_i \leq j | w_i, \beta, \mu, \Sigma, \gamma\}$ for $j = 1, \dots, k-1$;

3. Generate a random number u from the $U(0, 1)$ distribution;

4. Find j^* , such that

$$\Pr\{\zeta_i \leq j^* - 1 | w_i, \beta, \mu, \Sigma, \gamma\} < u \leq \Pr\{\zeta_i \leq j^* | w_i, \beta, \mu, \Sigma, \gamma\},$$

and set $\zeta_i = j^*$.

Sampling γ_j

If γ has a Dirichlet prior distribution with parameters $(\gamma_0, \dots, \gamma_0)$, its density is given by

$$f(\gamma) = \frac{\Gamma(k\gamma_0)}{[\Gamma(\gamma_0)]^k} \gamma_1^{\gamma_0-1} \dots \gamma_k^{\gamma_0-1} \mathbb{I}\left\{\sum_{j=1}^k \gamma_j = 1\right\}.$$

The distribution of the component indicators is multinomial:

$$f(\zeta_1, \dots, \zeta_n | \gamma) = \gamma_1^{n_1} \dots \gamma_k^{n_k},$$

where $n_j = \sum_{i=1}^n \mathbb{I}\{\zeta_i = j\}$ and $\sum_{j=1}^k n_j = n$. If $(\beta, \mu, \Sigma, c_u)$ is a priori independent of (ζ, γ) , the posterior of γ conditional on the completed data W and the remaining parameters satisfies

$$f(\gamma | W, \beta, \mu, \Sigma, \zeta, c_u) \propto f(W | \beta, \mu, \Sigma, \zeta, \gamma, c_u) \times f(\zeta | \gamma) \times f(\gamma).$$

Conditional on ζ the likelihood $f(W | \beta, \mu, \Sigma, \zeta, \gamma, c_u)$ does not depend on γ , so that

$$\begin{aligned} f(\gamma | W, \beta, \mu, \Sigma, \zeta, c_u) &= f(\gamma | \zeta) \\ &\propto f(\zeta | \gamma) f(\gamma), \end{aligned}$$

from which (27) follows. To generate draws from this distribution, perform the following two steps (e.g., (1973), 1973): (1) for $j = 1, \dots, k$, generate $Z_j \sim G(\gamma_0 + n_j, 1)$ and (2) set $\gamma_j = Z_j / \sum_{l=1}^k Z_l$.

Sampling (μ_j, Σ_j)

Multiplying the completed-data mixture likelihood with the normal-inverse Wishart prior, we have

$$\begin{aligned} f(\mu_j, \Sigma_j | W, \beta, \zeta, c_u) &\propto |\Sigma_j|^{-(v+3)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma_j^{-1} H) \right\} \\ &\times |\tau c_u \Sigma_j|^{-1/2} \exp \left\{ -\frac{1}{2\tau c_u} \mu_j' \Sigma_j^{-1} \mu_j \right\} \\ &\times |c_u \Sigma_j|^{-n_j/2} \exp \left\{ -\frac{1}{2c_u} \sum_{i:\zeta_i=j} (u_i - \mu_j)' \Sigma_j^{-1} (u_i - \mu_j) \right\}, \end{aligned}$$

where $w_i = X_i \beta + u_i$. The exponent involving μ_j can be rewritten as

$$\exp \left\{ -\frac{1}{2c_u} \left[(\mu_j - b_j)' M_j^{-1} (\mu_j - b_j) - b_j' M_j^{-1} b_j + \sum_{i:\zeta_i=j} u_i' \Sigma_j^{-1} u_i \right] \right\},$$

with b_j and M_j are the mean and variance in (29). The posterior of μ_j given Σ_j then follows. As for Σ_j , we can write

$$\begin{aligned} f(\mu_j, \Sigma_j | W, \beta, \zeta, c_u) &\propto f(\mu_j | \Sigma_j, W, \beta, \zeta, c_u) \\ &\times |\Sigma_j|^{-(v+n_j+3)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma_j^{-1} H) \right\} \\ &\times \exp \left\{ -\frac{1}{2c_u} \left[\sum_{i:\zeta_i=j} u_i' \Sigma_j^{-1} u_i - b_j' M_j^{-1} b_j \right] \right\} \\ &\propto f(\mu_j | \Sigma_j, W, \beta, \zeta, c_u) |\Sigma_j|^{-(v+n_j+3)/2} \\ &\times \exp \left\{ -\frac{1}{2} \text{tr} \Sigma_j^{-1} \left(H + c_u^{-1} \sum_{i:\zeta_i=j} u_i u_i' - c_u^{-1} \frac{\tau n_j^2}{1 + \tau n_j} \bar{u}_j \bar{u}_j' \right) \right\}, \end{aligned}$$

from which (28) follows.

Sampling c_u

Given that $f(c_u) = IG(v/2, d)$ and c_u enters the prior of μ_j and the completed-data likelihood, we have

$$\begin{aligned}
f(c_u|W, \beta, \mu, \Sigma, \zeta, \gamma) &\propto c_u^{-(v/2+1)} e^{-d/c_u} \prod_{j=1}^k |\tau c_u \Sigma_j|^{-1/2} \exp \left\{ -\frac{1}{2\tau c_u} \mu_j' \Sigma_j^{-1} \mu_j \right\} \\
&\quad \times \prod_{i=1}^n |c_u \Sigma_{j(i)}|^{-1/2} \exp \left\{ -\frac{1}{2c_u} (w_i - X_i \delta - \mu_{j(i)})' \Sigma_{j(i)}^{-1} (w_i - X_i \delta - \mu_{j(i)}) \right\} \\
&\propto c_u^{-(v/2+k+n+1)} \exp \left\{ -\frac{1}{c_u} \left[d + \frac{1}{2\tau} \sum_{j=1}^k \mu_j' \Sigma_j^{-1} \mu_j \right] \right\} \\
&\quad \times \exp \left\{ -\frac{1}{c_u} \left[\frac{1}{2} \sum_{i=1}^n (u_i - \mu_{j(i)})' \Sigma_{j(i)}^{-1} (u_i - \mu_{j(i)}) \right] \right\},
\end{aligned}$$

from which (30) follows.

Appendix B: Dirichlet Mixtures

Posterior of θ

We will use the following shorthand notation for the Polya urn prior in (34):

$$f(\theta_i|\theta_{-i}) = \frac{\alpha}{\alpha + n - 1} dG_0(\theta_i) + \sum_{j \neq i} \frac{1}{\alpha + n - 1} \delta_{\theta_j}(\theta_i),$$

where $\delta_{\theta_j}(\cdot)$ is a measure with unit mass at θ_j . Then

$$\begin{aligned}
f(\theta_i|\theta_{-i}, W, \beta, c_u) &\propto \prod_{i=1}^n f(w_i|\beta, \theta_i, c_u) f(\theta_i|\theta_{-i}) \\
&\propto \frac{\alpha}{\alpha + n - 1} dG_0(\theta_i) f(w_i|\beta, \theta_i, c_u) + \sum_{j \neq i} \frac{f(w_i|\beta, \theta_j, c_u) \delta_{\theta_j}(\theta_i)}{\alpha + n - 1} \\
&\propto \alpha f(w_i|\beta, \theta_i, c_u) dG_0(\theta_i) + \sum_{j \neq i} f(w_i|\beta, \theta_j, c_u) \delta_{\theta_j}(\theta_i). \tag{37}
\end{aligned}$$

The normalizing constant c satisfies

$$\begin{aligned} c &= \int \left[\alpha f(w_i|\beta, \theta_i, c_u) dG_0(\theta_i) + \sum_{j \neq i} f(w_i|\beta, \theta_j, c_u) \delta_{\theta_j}(\theta_i) \right] d\theta_i \\ &= \alpha f(w_i|\beta, c_u) + \sum_{j \neq i} f(w_i|\beta, \theta_j, c_u). \end{aligned}$$

The posterior (37) then becomes

$$\begin{aligned} f(\theta_i|\theta_{-i}, W, \delta, c_u) &= \frac{1}{c} \left\{ \alpha f(w_i|\beta, c_u) \frac{f(w_i|\beta, \theta_i, c_u) dG_0(\theta_i)}{f(w_i|\beta, c_u)} + \sum_{j \neq i} f(w_i|\beta, \theta_j, c_u) \delta_{\theta_j}(\theta_i) \right\} \\ &= (c^{-1} \alpha f(w_i|\beta, c_u)) f(\theta_i|w_i, \beta) + \sum_{j \neq i} (c^{-1} f(w_i|\beta, \theta_j, c_u)) \delta_{\theta_j}(\theta_i), \end{aligned}$$

which yields (35).

In order to sample from this distribution three elements are needed: $f(w_i|\beta, \theta_j, c_u)$, $f(w_i|\beta, c_u)$ and $f(\theta_i|w_i, \beta, c_u)$. The first is simply the completed data likelihood contribution of observation i :

$$f(w_i|\beta, \theta_j, c_u) = (2\pi)^{-1} |c_u \Sigma_j|^{-1/2} \exp \left\{ -\frac{1}{2c_u} (w_i - X_i \beta - \mu_j)' \Sigma_j^{-1} (w_i - X_i \beta - \mu_j) \right\}. \quad (38)$$

For the second component, integrating out θ_j from the equation above yields $f(w_i|\beta, c_u)$. In order to do so analytically, we take

$$dG_0(\theta_i) = N(\mu_i; 0, \tau c_u \Sigma_i) \times \mathcal{W}^{-1}(\Sigma_i; H, v). \quad (39)$$

Thus, the base distribution is the product of an inverse Wishart distribution for Σ_i (with parameters H and v) and a conditional normal distribution for μ_i (with mean zero and variance $\tau c_u \Sigma_i$). This choice of G_0 allows us to center prior beliefs around the parametric SSM in section 2 by taking $\tau = 0$ and $\alpha = 0$.

The conditional joint distribution of w_i and θ_i is now

$$\begin{aligned} f(w_i|\beta, \theta_i, c_u) dG_0(\theta_i) &= k(v) |H|^{v/2} |\Sigma_i|^{-(v+3)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma_i^{-1} H) \right\} \\ &\quad \times c_u^{-1} (2\pi)^{-1} \tau^{-1} |\Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2\tau c_u} \mu_i' \Sigma_i^{-1} \mu_i \right\} \end{aligned}$$

$$\times c_u^{-1} (2\pi)^{-1} |\Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2c_u} (w_i - X_i\beta - \mu_i)' \Sigma_i^{-1} (w_i - X_i\beta - \mu_i) \right\},$$

where $k(v)$ is a normalizing constant:

$$k(v) = \frac{2^{-v} \pi^{-1/2}}{\Gamma\left(\frac{v}{2}\right) \Gamma\left(\frac{v-1}{2}\right)}. \quad (40)$$

Collecting all constants and rewriting the exponent as a quadratic function in μ_i , it follows that

$$\begin{aligned} f(w_i|\beta, \theta_i, c_u) dG_0(\theta_i) &= c_u^{-2} \tilde{k}(v) |\Sigma_i|^{-(v+5)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma_i^{-1} H) \right\} \\ &\quad \times \exp \left\{ -\frac{1}{2c_u} [u_i' \Sigma_i^{-1} u_i - b_i' M_i^{-1} b_i] \right\} \\ &\quad \times \exp \left\{ -\frac{1}{2c_u} (\mu_i - b_i)' M_i^{-1} (\mu_i - b_i) \right\}, \\ u_i &= w_i - X_i \delta, \\ b_i &= \frac{\tau}{1 + \tau} u_i, \\ M_i &= \frac{\tau}{1 + \tau} \Sigma_i, \\ \tilde{k}(v) &= (2\pi)^{-2} \tau^{-1} k(v) |H|^{v/2}. \end{aligned} \quad (41)$$

Integrating out μ_i we find

$$\begin{aligned} f(w_i, \Sigma_i|\beta, c_u) &= \int f(w_i|\beta, \theta_i, c_u) dG_0(\theta_i) d\mu_i \\ &= c_u^{-2} \tilde{k}(v) |\Sigma_i|^{-(v+5)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma_i^{-1} H) \right\} \\ &\quad \times \exp \left\{ -\frac{1}{2c_u} [u_i' \Sigma_i^{-1} u_i - b_i' M_i^{-1} b_i] \frac{2\pi\tau c_u}{(1 + \tau)} |\Sigma_i|^{1/2} \right\} \\ &= (2\pi)^{-1} \frac{k(v) |H|^{v/2}}{c_u(1 + \tau)} |\Sigma_i|^{-(v+4)/2} \\ &\quad \times \exp \left\{ -\frac{1}{2} \left[\text{tr}(\Sigma_i^{-1} H) + c_u^{-1} u_i' \Sigma_i^{-1} u_i - c_u^{-1} b_i' M_i^{-1} b_i \right] \right\} \\ &= (2\pi)^{-1} \frac{k(v) |H|^{v/2}}{c_u(1 + \tau)} |\Sigma_i|^{-(v+4)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma_i^{-1} H_i) \right\}, \end{aligned}$$

where

$$H_i = H + \frac{1}{c_u(1 + \tau)} (w_i - X_i\beta)(w_i - X_i\beta)'$$

Using the density of the inverse Wishart distribution Σ_i^{-1} can be integrated out, so that

$$\begin{aligned} f(w_i|\beta, c_u) &= \int p(w_i, \Sigma_i|\beta, c_u)d\Sigma_i \\ &= \frac{(2\pi)^{-1} k(v) |H|^{v/2}}{c_u(1+\tau) k(\tilde{v}) |H_i|^{\tilde{v}/2}} \\ &= \frac{(v-1) |H|^{v/2}}{2\pi c_u(1+\tau) |H_i|^{\tilde{v}/2}}, \end{aligned}$$

where $\tilde{v} = v + 1$. A typical choice of vague prior uses $H = \varepsilon_0 I_2$ for some large $\varepsilon_0 > 0$. In that case the determinants can be explicitly calculated. Using the result that for any $|I_p + aa'| = 1 + a'a$ for any $a \in \mathbb{R}^p$, it follows that

$$\begin{aligned} f(w_i|\beta, c_u) &= \frac{(v-1) \varepsilon_0^v}{2\pi c_u(1+\tau) \varepsilon_0^{v+1} (1 + [\varepsilon_0 c_u(1+\tau)]^{-1} u_i' u_i)^{(v+1)/2}} \\ &= \frac{(v-1)}{2\pi c_u \varepsilon_0(1+\tau)} \left[1 + \frac{1}{c_u \varepsilon_0(1+\tau)} u_i' u_i \right]^{-(v+1)/2}. \end{aligned}$$

Finally, from (35), if the value of θ_i is not equal to any other θ_j it is distributed according to the posterior arising from $G_0(\theta_i)$. The chosen form of G_0 now allows us to generate values of μ_i and Σ_i in two simple steps: first generate Σ_i from its posterior (conditional on $\{w_i, \beta, c_u\}$) and then generate μ_i from its posterior conditional on $\{\Sigma_i, w_i, \beta, c_u\}$.

Multiplying the likelihood (38) and prior (39) and rearranging the exponent, it follows that

$$\begin{aligned} f(\mu_i, \Sigma_i|w_i, \beta, c_u) &\propto |\Sigma_i|^{-(v+4)/2} \exp \left\{ -\frac{1}{2} \text{tr} (\Sigma_i^{-1} H_i) \right\} \\ &\quad \times |\Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2c_u} (\mu_i - b_i)' M_i^{-1} (\mu_i - b_i) \right\}, \end{aligned}$$

where b_i and M_i were defined in (41). Thus:

$$\begin{aligned} \Sigma_i|w_i, \beta, c_u &\sim \mathcal{W}^{-1}(H_i, v+1), \\ \mu_i|\Sigma_i, w_i, \beta, c_u &\sim N \left(\frac{\tau}{1+\tau} u_i, \frac{c_u \tau}{1+\tau} \Sigma_i \right). \end{aligned}$$

Remixing Unique Values of θ

To describe the remixing step for θ , let k be the number of unique values in θ , denoted by $\theta^* = \{\theta_j^*\}_{j=1}^k$. Define the component indicators $\zeta = \{\zeta_i\}_{i=1}^n$ as before:

$$\zeta_i = j \quad \Leftrightarrow \quad \theta_i = \theta_j^*, \quad j = 1, \dots, k.$$

Let k_{-i} be the number of distinct θ values in θ_{-i} and $n_{j,-i} = \sum_{l:l \neq i} \mathbb{I}\{\zeta_l = j\}$ for $j = 1, \dots, k_{-i}$.

The posterior of $\theta_i | \theta_{-i}$ in (35) then becomes

$$\theta_i | \theta_{-i}, W, \beta, c_u \begin{cases} = \theta_j^* & \text{w. prob. } c^{-1} n_{j,-i} f(w_i | \beta, \theta_j^*, c_u), \quad j = 1, \dots, k_{-i} \\ \sim f(\theta_i | w_i, \beta, c_u) & \text{w. prob. } c^{-1} \alpha f(w_i | \beta, c_u) \end{cases}. \quad (42)$$

Note that knowledge of θ is equivalent to knowing (θ^*, ζ, k) . The remixing algorithm is based on sampling (ζ, k) from its conditional distribution given θ^* , and θ^* from its conditional distribution given (ζ, k) . From (42) it follows immediately that

$$\Pr\{\zeta_i = j | \zeta_{-i}, W, \beta, \theta^*, c_u\} = c^{-1} n_{j,-i} f(w_i | \beta, \theta_j^*, c_u), \quad j = 1, \dots, k_{-i}. \quad (43)$$

Also, with probability

$$\Pr\{\zeta_i = 0 | \zeta_{-i}, W, \beta, \theta^*, c_u\} = 1 - c^{-1} \sum_{j=1}^{k_{-i}} n_{j,-i} f(w_i | \beta, \theta_j^*, c_u), \quad (44)$$

set ζ_i equal to zero and generate θ_i from $f(\theta_i | w_i, \beta, c_u)$. After cycling through for $i = 1, \dots, n$, and potentially relabeling ζ , we obtain a new value of (ζ, k) . In the prior θ^* represents k i.i.d. draws from G_0 (Antoniak, 1974). Then:

$$f(\theta_1^*, \dots, \theta_k^* | W, \beta, \zeta, k, c_u) \propto \prod_{j=1}^k \left\{ \prod_{i:\zeta_i=j} f(w_i | \beta, \theta_j^*, c_u) dG_0(\theta_j^*) \right\},$$

so that the θ_j^* 's are conditionally independent and

$$f(\theta_j^* | W, \beta, \zeta, k, c_u) \propto \prod_{i:\zeta_i=j} f(w_i | \beta, \theta_j^*, c_u) dG_0(\theta_j^*), \quad j = 1, \dots, k. \quad (45)$$

Applying this to (μ_j^*, Σ_j^*) yields exactly the posterior given by (28) and (29). Thus, the posterior for (μ_j, Σ_j) in the finite mixture model can be used in the Dirichlet model as a remixing distribution: first update $\{(\mu_i, \Sigma_i)\}_{i=1}^n$ and determine the number of unique components, then regenerate the unique components alone and assign the k values across the sample, according to the component indicators ζ_i . Step 3 in algorithm 5 can now be replaced by

- 3a. Sample ζ_i from the distribution in (43) and (44) for $i = 1, \dots, n$, and determine the number of unique components k ;
- 3b. sample (μ_j^*, Σ_j^*) from (28) and (29) for $j = 1, \dots, k$.

Uncertainty about α

It is possible to incorporate uncertainty about α into the Gibbs sampler, thereby allowing the data to determine how many mixture components are appropriate. A convenient approach, due to Escobar and West (1995), is described here. As before let $\theta^* = \{\theta_j^*\}_{j=1}^k$ be the collection of unique values of θ_i . Let α be a priori independent of $\{\beta, c_u\}$ with prior $f(\alpha)$. Note that the likelihood $f(w_i|\beta, \theta^*, c_u, \zeta, k, \alpha)$ does not depend on α . The posterior of α then satisfies

$$\begin{aligned} f(\alpha|W, \beta, \theta^*, c_u, \zeta, k) &\propto f(W|\beta, \theta^*, c_u, \zeta, k, \alpha) f(\theta^*, \zeta, k|\alpha, \beta, c_u) f(\alpha) \\ &\propto f(\theta^*|\zeta, k, \alpha, \beta, c_u) f(\zeta|k, \alpha, \beta, c_u) f(k|\alpha, \beta, c_u) f(\alpha). \end{aligned}$$

Given (34), the prior distribution of ζ given $\{k, \alpha, \beta, c_u\}$ only depends on the sample size and k . Also, the random number of mixture components only depends on α , so that $f(k|\alpha, \beta, c_u) = f(k|\alpha)$. Finally, conditional on $\{k, \zeta, c_u\}$ the unique values θ^* follow the prior distribution G_0 , which does not depend on α or β . With these conditional independence relations, the posterior simplifies to

$$\begin{aligned} f(\alpha|W, \beta, \theta^*, \zeta, k) &= f(\alpha|k) \\ &\propto f(k|\alpha) f(\alpha) \\ &\propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha + n)} f(\alpha). \end{aligned}$$

Using a property of the beta function $B(b_1, b_2)$:

$$\begin{aligned} B(b_1, b_2) &= \int_0^1 x^{b_1-1}(1-x)^{b_2-1} dx \\ &= \frac{\Gamma(b_1)\Gamma(b_2)}{\Gamma(b_1+b_2)}, \end{aligned}$$

the posterior of α can be written as

$$f(\alpha|k) \propto \frac{f(\alpha)\alpha^{k-1}(\alpha+n)}{\Gamma(n)} \int_0^1 \eta^\alpha(1-\eta)^{n-1} d\eta.$$

The posterior corresponds to a joint posterior of α and a latent variable $\eta \in (0, 1)$, given by

$$p(\alpha, \eta|k) \propto \frac{f(\alpha)\alpha^{k-1}(\alpha+n)}{\Gamma(n)} \eta^\alpha(1-\eta)^{n-1},$$

from which it is clear that $p(\eta|\alpha, k)$ is the beta $B(\alpha+1, n)$ distribution. The joint posterior suggests using a $G(c_1, c_2)$ distribution as prior for α . Then:

$$\begin{aligned} p(\alpha|\eta, k) &\propto \alpha^{c_1+k-1} e^{-\alpha(c_2-\log \eta)} + n\alpha^{c_1+k-2} e^{-\alpha(c_2-\log \eta)} \\ &\propto \left(\frac{[c_2 - \log \eta]^{c_1+k}}{\Gamma(c_1+k)} \right)^{-1} G(c_1+k, c_2 - \log \eta) \\ &\quad + n \left(\frac{[c_2 - \log \eta]^{c_1+k-1}}{\Gamma(c_1+k-1)} \right)^{-1} G(c_1+k-1, c_2 - \log \eta), \end{aligned}$$

which is a mixture of two gamma distributions. The mixing probability p_η satisfies

$$\begin{aligned} \frac{p_\eta}{1-p_\eta} &= \left(\frac{[c_2 - \log \eta]^{c_1+k}}{\Gamma(c_1+k)} \right)^{-1} n^{-1} \left(\frac{[c_2 - \log \eta]^{c_1+k-1}}{\Gamma(c_1+k-1)} \right) \\ &= \frac{c_1+k-1}{n(c_2 - \log \eta)}. \end{aligned} \tag{46}$$

We can now augment algorithm 5 with the following steps:

- 4a. sample η from $B(\alpha+1, n)$;
- 4b. calculate p_η according to (46);
- 4c. with probability p_η , sample α from $G(c_1+k, c_2 - \log \eta)$; with probability $1-p_\eta$, sample α

from $G(c_1 + k - 1, c_2 - \log \eta)$;

5. return to step 1 in algorithm 5 and repeat.

Acknowledgements

I would like to thank Tony Lancaster, Frank Kleibergen, William McCausland and Chris Bollinger for their many comments and suggestions. I am responsible for all remaining errors.

References

- AHN, H., AND J. L. POWELL (1993): “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism,” *Journal of Econometrics*, 58, 3–29.
- ALBERT, J. H., AND S. CHIB (1993): “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, 88(422), 669–679.
- AMEMIYA, T. (1985): *Advanced Econometrics*. Harvard University Press.
- ANTONIAK, C. E. (1974): “Mixtures of Dirichlet Processes with Applications to Bayesian Non-parametric Problems,” *The Annals of Statistics*, 2(6), 1152–1174.
- BAYARRI, M., AND J. BERGER (1998): “Robust Bayesian Analysis of Selection Models,” *Annals of Statistics*, 26, 645–659.
- BAYARRI, M., AND M. DEGROOT (1987): “Bayesian Analysis of Selection Models,” *The Statistician*, 36, 137–146.
- BLACKWELL, D., AND J. B. MACQUEEN (1973): “Ferguson Distributions via Pólya Urn Schemes,” *The Annals of Statistics*, 1(2), 353–355.
- CASELLA, G., AND E. I. GEORGE (1992): “Explaining the Gibbs Sampler,” *The American Statistician*, 46(3), 167–174.
- CHIB, S. (1995): “Marginal Likelihood from the Gibbs Output,” *Journal of the American Statistical Association*, 90(432), 1313–1321.

- CHIB, S., E. GREENBERG, AND I. JELIAZKOV (2008): “Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection,” working paper.
- CONLEY, T., C. HANSEN, R. MCCULLOCH, AND P. E. ROSSI (2007): “A Semiparametric Bayesian Approach to the Instrumental Variable Problem,” Graduate School of Business, University of Chicago Working Paper.
- COSSLETT, S. R. (1991): “Semiparametric Estimation of a Regression Model with Sample Selectivity,” in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 175–197. Cambridge.
- CRAGG, J. G. (1971): “Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods,” *Econometrica*, 39(5), 829–844.
- DOW, W. H., AND E. C. NORTON (2003): “Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions,” *Health Services and Outcomes Research Methodology*, 4, 5–18.
- DUAN, N., W. G. MANNING, C. N. MORRIS, AND J. P. NEWHOUSE (1983): “A Comparison of Alternative Models for the Demand for Medical Care,” *Journal of Business and Economic Statistics*, 1(2), 115–126.
- ESCOBAR, M. D. (1994): “Estimating Normal Means With a Dirichlet Process Prior,” *Journal of the American Statistical Association*, 89(425), 268–277.
- ESCOBAR, M. D., AND M. WEST (1995): “Bayesian Density Estimation and Inference Using Mixtures,” *Journal of the American Statistical Association*, 90(430), 577–588.
- (1998): “Computing Nonparametric Hierarchical Models,” in *Practical Nonparametric and Semiparametric Bayesian Statistics*, ed. by D. Dey, P. Müller, and D. Sinha, pp. 1–22. Springer.
- FERGUSON, T. S. (1973): “A Bayesian Analysis of Some Nonparametric Problems,” *The Annals of Statistics*, 1(2), 209–230.
- GEWEKE, J. (2005): *Contemporary Bayesian Econometrics and Statistics*. Wiley.

- GILKS, W., S. RICHARDSON, AND D. SPIEGELHALTER (1996): *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- GRONAU, R. (1974): “Wage Comparisons – A Selectivity Bias,” *The Journal of Political Economy*, 82(6), 1119–1143.
- HECKMAN, J. J. (1979): “Sample Selection as a Specification Error,” *Econometrica*, 47(1), 153–162.
- HUANG, H.-C. (2001): “Bayesian Analysis of the SUR Tobit Model,” *Applied Economics Letters*, 8, 617–622.
- ICHIMURA, H., AND L.-F. LEE (1991): “Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation,” in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 3–49. Cambridge.
- KASS, R. E., AND A. E. RAFTERY (1995): “Bayes Factors,” *Journal of the American Statistical Association*, 90(430), 773–795.
- KOOP, G., AND D. J. POIRIER (1997): “Learning About the Across-Regime Correlation in Switching Regression Models,” *Journal of Econometrics*, 78, 217–227.
- LEE, J., AND J. BERGER (2001): “Semiparametric Bayesian Analysis of Selection Models,” *Journal of the American Statistical Association*, 96, 1269–1276.
- LEE, L.-F. (1982): “Some Approaches to the Correction of Selectivity Bias,” *Review of Economic Studies*, 49, 355–372.
- (1994): “Semiparametric two-stage Estimation of Sample Selection Models Subject to Tobit-type Selection Rules,” *Journal of Econometrics*, 61, 305–344.
- (2003): “Self-Selection,” in *A Companion to Theoretical Econometrics*, ed. by B. H. Baltagi, chap. 18. Blackwell Publishing.
- LEUNG, S. F., AND S. YU (1996): “On the Choice Between Sample Selection and Two-Part Models,” *Journal of Econometrics*, 72, 197–229.

- LI, K. (1998): “Bayesian Inference in a Simultaneous Equation Model with Limited Dependent-Variables,” *Journal of Econometrics*, 85, 387–400.
- MACEACHERN, S. N. (1998): “Computational Methods for Mixture of Dirichlet Process Models,” in *Practical Nonparametric and Semiparametric Bayesian Statistics*, ed. by D. Dey, P. Müller, and D. Sinha, pp. 23–44. Springer.
- MACEACHERN, S. N., AND P. MÜLLER (1998): “Estimating Mixture of Dirichlet Process Models,” *Journal of Computational and Graphical Statistics*, 7(2), 223–238.
- MANNING, W., N. DUAN, AND W. ROGERS (1987): “Monte Carlo Evidence on the Choice Between Sample Selection and Two-Part Models,” *Journal of Econometrics*, 35, 59–82.
- MCCULLOCH, R. E., N. G. POLSON, AND P. E. ROSSI (2000): “A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters,” *Journal of Econometrics*, 99, 173–193.
- MCCULLOCH, R. E., AND P. E. ROSSI (1994): “An Exact Likelihood Analysis of the Multinomial Probit Model,” *Journal of Econometrics*, 64, 207–240.
- MROZ, T. A. (1987): “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economical and Statistical Assumptions,” *Econometrica*, 55, 765–799.
- MUIRHEAD, R. (1982): *Aspects of Multivariate Statistical Theory*. Wiley.
- MÜLLER, P., AND F. A. QUINTANA (2004): “Nonparametric Bayesian Data Analysis,” *Statistical Science*, 19(1), 95–110.
- MUNKIN, M. K., AND P. K. TRIVEDI (2003): “Bayesian Analysis of a Self-Selection Model with Multiple Outcomes Using Simulation-Based Estimation: An Application to the Demand for Healthcare,” *Journal of Econometrics*, 114, 197–220.
- NELSEN, R. (2006): *An Introduction to Copulas*. Springer, 2nd edn.
- OLSEN, R. J. (1982): “Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator,” *International Economic Review*, 23(1), 223–240.

- ROEDER, K., AND L. WASSERMAN (1997): “Practical Bayesian Density Estimation Using Mixtures of Normals,” *Journal of the American Statistical Association*, 92(439), 894–902.
- TANNER, M. A., AND W. H. WONG (1987): “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82, 528–550.
- VELLA, F. (1998): “Estimating Models with Sample Selection Bias: a Survey,” *Journal of Human Resources*, 33, 127–169.
- VERDINELLI, I., AND L. WASSERMAN (1995): “Computing Bayes Factors Using a Generalization of the Savage-Dickey DensityRatio,” *Journal of the American Statistical Association*, 90(430), 614–618.
- WEST, M., P. MÜLLER, AND M. D. ESCOBAR (1994): “Hierarchical Priors and Mixture Models, With Applications in Regression and Density Estimation,” in *Aspects of Uncertainty: a Tribute to D.V. Lindley*, ed. by P. Freeman, and A. Smith, pp. 363–386. Wiley.