Bayesian Inference in the Sample Selection and Two-Part Models

Martijn van Hasselt*
Department of Economics
The University of Western Ontario

September 2007

Abstract

This paper considers two models to deal with an outcome variable that contains a large fraction of zeros: the sample-selection model and the two-part model. Whereas the sample selection model allows correlation between the selection and outcome mechanisms, the two-part model assumes conditional independence. Using a fully parametric Bayesian approach, we present Markov Chain Monte Carlo (MCMC) algorithms for the model parameters that are based on data augmentation. With a Gaussian specification of the likelihood the models are, in some sense, nested. In order to determine which model is more appropriate, inference can focus on the correlation coefficient. Alternatively, a Bayes factor can be computed. The Bayesian semiparametric approach uses flexible families of distributions. In particular, we present MCMC schemes based on mixtures of normal distributions and Dirchlet process priors. The various methods are illustrated using simulated data.

1 Introduction

In many economic applications a nonnegative outcome variable of interest is typically characterized by a certain fraction of observations clustered at zero and a distribution of positive values that is highly skewed. Examples are consumer expenditures on durable goods and labor supply functions (hours worked). Our paper considers two models that are commonly used in the literature to analyze this type of data. The essential difference between these models is how they interpret a zero value of the outcome. In the first model, that we will refer to as a *sample selection model* (SSM), the data generating process is split up into two stages. The first stage describes the *selection* mechanism, which determines whether we see a positive outcome or not. Usually, the selection mechanism takes the form of a structural equation for an underlying latent variable such as utility. If the latent variable falls below a certain threshold, the outcome variable is zero; if it exceeds this threshold a positive outcome is observed. The second stage determines the *level* of the outcome.

*This is still a work in progress and feedback is appreciated. I am grateful to Tony Lancaster for sparking my interest in the topic and providing many helpful comments and suggestions. All errors are, of course, my own. Contact: mvanhass@uwo.ca

If the first stage dictates that the outcome should be positive we observe the level determined in the second stage. Otherwise, we observe a zero. Thus, the zeros represent missing data: we do not observe what the positive outcome *would have been*. As an example, consider a consumer who is thinking of purchasing a car. Depending on her income, preferences, the availability of public transportation and the travel distance to work, she will first decide on whether to buy a car at all. If not, we observe zero expenditures and the potential outcome is unknown. On the other hand, if she does decide to buy, the potential expenditures are realized (i.e. observed) and may vary depending on, say, income, travel distance and individual tastes.

Sample selection models, such as the one used in this paper, typically desribe *potential* outcomes which are only *partially* observed. In this case the observed positive values of the outcome variable follow a pattern that is derived from the latent structure. One could wonder whether the potential outcome is an interesting quantity to begin with. Regardless, as a modeling device the SSM can be used to describe actual outcomes as well. More generally, sample selection occurs when the data at hand is not a random sample from the population. When making inferences it is then important to understand the process that generated the sample. Individuals may select themselves into (or out of) the sample based on a combination of observable quantities and unobserved heterogeneity. If this heterogeneity also affects the outcome variable, inferences using the selected sample may be subject to selection bias.

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. Since our paper takes a Bayesian approach we will not discuss the frequentist literature any further[1].

The second model in this paper is a *two-part* model (2PM)[2]. One of the first discussions of the 2PM goes back to Cragg (1971). As in the sample selection model, two stages are distinguished in the outcome generating process: a selection mechanism and an outcome level function. The main difference with the SSM is that the 2PM models the *observed* outcomes directly, rather than specifying a latent structure. In this framework a zero does not represent missing data, but can be interpreted as a corner solution in a consumer optimization problem. An application can be found in Duan et al. (1983) who use the two-part model to analyze individuals' medical expenditures.

There has been some debate in the health literature as to which model is more appropriate for describing medical expenditures. Duan et al. (1983) argue that the 2PM is to be preferred since it models actual as opposed to potential outcomes. Whether we are interested in actual or potential outcomes depends on the particular application at hand. Regardless, the SSM can also be used to analyze actual outcomes because the latent structure implies a model for the observed data. Another claim, e.g. Hay and Olsen (1984), has been that in the parametric (Gaussian)

[1] Excellent frequentist surveys are Lee (2003) and Vella (1998). The latter focuses on semiparametric models.

[2] The term *hurdle model* is also used. Wooldridge (2002) and Cameron and Trivedi (2005) contain good discussions of the 2PM and related models.

case the 2PM is actually nested within the SSM. This is shown to be incorrect in Duan et al. (1984). Moreover, the model parameters are not directly comparable because they have a different interpretation in each model. Marginal effects on the observed outcomes are directly available in the 2PM, whereas in the SSM they involve some nonlinear transformation of the model parameters and are typically covariate dependent.

The aforementioned debate has prompted many authors to compare the SSM and 2PM. Manning, Duan, and Rogers (1987) conduct an extensive Monte Carlo study and find that overall the 2PM performs very well in terms of predictive ability, even if the SSM is the true model generating the data. Given a joint normality assumption the 2PM is observationally equivalent to the SSM when the cross-equation correlation is zero. In principle the null hypothesis that the 2PM is the true model can then be tested via a classical t-test. Leung and Yu (1996) present simulation evidence suggesting that such a test may perform poorly due to near multicollinearity. Dow and Norton (2003) propose a test based on the difference in empirical mean squared error between the two models[3].

Our paper takes a Bayesian approach to estimating the sample selection and two-part models. The main contribution is to provide sampling algorithms to approximate the posterior distributions of interest. We start by considering inference in a fully parametric Bayesian model based on multivariate normal distributions. This analysis is then extended to allow for more flexible families of distributions. In particular, the focus is on mixtures of normal distributions with either a fixed or random number of mixture components. We will refer to this case as a semiparametric Bayesian model.

Bayesian inference in limited dependent variable models can proceed by a combination of *Gibbs sampling* (e.g. Casella and George 1992) and *data augmentation* (e.g. 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[4]. The combined algorithm is 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 Markov Chain Monte Carlo (MCMC) methods can be found in Gilks, Richardson, and Spiegelhalter (1996).

Applications of MCMC are becoming widespread. Discrete choice models are treated in Albert and Chib (1993), McCulloch and Rossi (1994) and McCulloch, Polson, and Rossi (2000). The analysis of the parametric sample 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 combined with a limited range of the dependent variable. The extension to semiparametric

[3] There are some problems with this test as well. First, the choice of null hypothesis is arbitrary and second, the test suffers from the same power problems as the t-test.

[4] Thus, data augmentation can be viewed as stochastic imputation of missing values.

models draws on a large body of literature on nonparametric Bayesian methods[5] and incorporates it into a sample selection framework. The remainder of this paper is organized as follows. Section 2 presents the fully parametric versions of the SSM and 2PM and several Gibbs sampling algorithms. In section 3 a small simulation experiment is considered. Section 4 presents our semiparametric models and construct the corresponding MCMC algorithms. Finally, section 5 concludes.

2 Parametric Models

2.1 The Sample Selection Model

We use the following version of the SSM, which is sometimes referred to as a *type 2 Tobit* model (e.g. Amemiya 1985, ch. 10):

$$\begin{aligned} I_i &= x_{i1}'\beta_1 + u_{i1}, \\ s_i &= \mathbb{I}\{I_i > 0\}, \\ m_i &= x_{i2}'\beta_2 + u_{i2}, \\ \ln y_i &= \begin{cases} m_i & \text{if } s_i = 1 \\ -\infty & \text{if } s_i = 0 \end{cases}. \end{aligned} \tag{2.1}$$

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. The equation for I_i is a selection equation: if $I_i > 0$, then a positive outcome y_i is observed; $I_i \leq 0$ corresponds to $y_i = 0$. The variable s_i is simply the indicator of a positive outcome. The equation for m_i represents the logarithm of the *potential* outcomes. Potential outcomes are realized only when $s_i = 1$. Thus, m is a partially observed, partially latent variable. If the outcome y_i is zero, and hence $s_i = 0$, then m_i is unobserved. On the other hand, if y_i is positive and $s_i = 1$, the potential outcome m_i $(= \ln y_i)$ is realized[6,7]. To summarize, we observe the vectors (x_{i1}', x_{i2}', s_i) for all i and the values m_i that belong to the set $\{m_i : s_i = 1\}$. For the parametric Bayesian analysis of this model it is assumed that the joint distribution of u_{i1} and u_{i2} is bivariate normal:

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{2.2}$$

where ρ is the correlation coefficient.

[5] To 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 methods, such as kernel smoothing.

[6] This version of the sample selection model is widely used in the literature; see, for example, Lee (2003).

[7] The logarithmic transform is very common in this type of models. A discussion of its rationales can be found in Manning (1998).

The random variable s_i has a Bernoulli distribution with

$$\Pr\{s_i = 1|x_{i1}, \beta_1\} = \Phi\left(x_{i1}'\beta_1/\sigma_1\right),$$

where $\Phi(\cdot)$ 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. The likelihood for the n observations can be written as

$$\begin{aligned} p_{SSM}(\ln y|\beta_1, \beta_2, \Sigma) &= \prod_{i=1}^{n} \left[\Phi\left(x_{i1}'\beta_1/\sigma_1\right)\right]^{s_i} \left[1 - \Phi\left(x_{i1}'\beta_1/\sigma_1\right)\right]^{1-s_i} \times \\ &\quad \prod_{i:y_i>0} p_{u_2|I>0}\left(\ln y_i - x_{i2}'\beta_2\right), \end{aligned} \tag{2.3}$$

where $p_{u_2|I>0}$ is the density of u_{i2} conditional on $I_i > 0$. Let $f_N(a|b,c)$ and $F_N(a|b,c)$ denote the density and CDF, respectively, of a normal random variable with mean b, variance c, evaluated at the point a. Let $\phi(\cdot)$ denote the standard normal density function and define $\tilde{u}_i = \ln y_i - x_{i2}'\beta_2$, so that

$$\begin{aligned} p_{u_2|I>0}(\tilde{u}_i) &= \frac{\int_0^\infty p_{u_2,I}(\tilde{u}_i, I)dI}{P(I>0)} \\ &= \frac{p_{u_2}(\tilde{u}_i)}{\Phi\left(x_{i1}'\beta_1/\sigma_1\right)} \int_0^\infty p_{I|u_2}(I|\tilde{u}_i)dI \\ &= \frac{f_N\left(\tilde{u}_i|0,\sigma_2^2\right)}{\Phi\left(x_{i1}'\beta_1/\sigma_1\right)} \left[1 - F_N\left(0|x_{i1}'\beta_1 + (\rho\sigma_1/\sigma_2)\tilde{u}_i, \sigma_1^2(1-\rho^2)\right)\right] \\ &= \frac{\sigma_2^{-1}\phi\left(\tilde{u}_i/\sigma_2\right)}{\Phi\left(x_{i1}'\beta_1/\sigma_1\right)} \Phi\left(\frac{x_{i1}'\beta_1 + (\rho\sigma_1/\sigma_2)\,\tilde{u}_i}{\sqrt{\sigma_1^2(1-\rho^2)}}\right). \end{aligned}$$

Plugging this back into (2.3) the likelihood becomes

$$\begin{aligned} p_{SSM}(\ln y|\beta_1, \beta_2, \Sigma) = &\prod_{i:y_i=0} \left[1 - \Phi\left(x_{i1}'\beta_1/\sigma_1\right)\right] \times \\ &\prod_{i:y_i>0} \sigma_2^{-1}\phi\left(\frac{\ln y_i - x_{i2}'\beta_2}{\sigma_2}\right) \Phi\left(\frac{x_{i1}'\beta_1}{\sigma_1\sqrt{1-\rho^2}} + \frac{\rho\left(\ln y_i - x_{i2}'\beta_2\right)}{\sigma_2\sqrt{1-\rho^2}}\right). \end{aligned} \tag{2.4}$$

2.2 Gibbs Sampling in the SSM

By inspection of the likelihood (2.4) it appears that no choice of prior for $(\beta_1, \beta_2, \Sigma)$ will yield a tractable posterior distribution. We therefore develop a Gibbs sampling algorithm that simulates draws from the posterior distribution of (α, 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) are therefore not informative, in the sense that there is no updating of the prior[8]. 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 now turn to the Gibbs samplers for the parametric SSM.

Since only the selection indicator s_i is observed, the variable I_i is latent and treated as an additional parameter in the algorithm. The same can be said about the unobserved values of m_i. Data-augmentation 'completes' the data and generates a sequence of $(\beta_1, \beta_2, \Sigma, I, m)$ values that are approximately drawn from their joint posterior. Discarding the values of (I, m) we then obtain a sample (again, approximately) from the posterior distribution of $(\beta_1, \beta_2, \Sigma)$. In what follows all conditional distributions are to be understood as conditional on the data as well. This conditioning is omitted for notational simplicity.

The SSM in (2.1) can be written as a SUR model. Let $I = (I_1, \ldots, I_n)'$, $m = (m_1, \ldots, m_n)'$, $u_1 = (u_{11}, \ldots, u_{n1})'$ and $u_2 = (u_{12}, \ldots, u_{n2})'$ be $n \times 1$ vectors. Define the following matrices:

$$\begin{aligned}
W &= \begin{bmatrix} I \\ m \end{bmatrix} : 2n \times 1, \quad X_1 = \begin{bmatrix} x_{11}' \\ \vdots \\ x_{n1}' \end{bmatrix} : n \times k_1, \\
X_2 &= \begin{bmatrix} x_{12}' \\ \vdots \\ x_{n2}' \end{bmatrix} : n \times k_2, \quad X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} : 2n \times (k_1 + k_2), \\
\delta &= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} : (k_1 + k_2) \times 1, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} : 2n \times 1.
\end{aligned}$$

Then $W = X\delta + u$, where $E(u) = 0$ and $V(u) = \Sigma \otimes I_n$. The likelihood of the normal SUR model is

$$\begin{aligned}
p(W|\delta, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}(W - X\delta)'(\Sigma^{-1} \otimes I_n)(W - X\delta)\right\} \\
&\propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} tr\left(B\Sigma^{-1}\right)\right\}, \qquad (2.5)
\end{aligned}$$

where $tr(\cdot)$ is the trace of a square matrix and B is defined as

$$B = \begin{bmatrix} (I - X_1\beta_1)'(I - X_1\beta_1) & (I - X_1\beta_1)'(m - X_2\beta_2) \\ (m - X_2\beta_2)'(I - X_1\beta_1) & (m - X_2\beta_2)'(m - X_2\beta_2) \end{bmatrix}. \qquad (2.6)$$

Starting with the conditional posterior of δ, note that $p(\delta|I, m, \Sigma, s) = p(\delta|I, m, \Sigma)$ because s is a

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

function of I. The likelihood in (2.5) can be rewritten as

$$\begin{aligned} p(W|\delta,\Sigma) &\propto |\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\left[e'S^{-1}e+(\delta-\hat{\delta})'X'S^{-1}X(\delta-\hat{\delta})\right]\right\}, \\ e &= W-X\hat{\delta}, \\ \hat{\delta} &= (X'S^{-1}X)^{-1}X'S^{-1}W, \\ S^{-1} &= \Sigma^{-1}\otimes I_n. \end{aligned}$$

Combining this with a normal $N(\delta_0, D_0)$ prior for δ, the posterior is again normal with mean and variance given by

$$E(\delta|W,\Sigma) = \left[D_0^{-1}+X'S^{-1}X\right]^{-1}\left[D_0^{-1}\delta_0+X'S^{-1}X\hat{\delta}\right], \tag{2.7}$$

$$V(\delta|W,\Sigma) = \left[D_0^{-1}+X'S^{-1}X\right]^{-1}. \tag{2.8}$$

To sample (I_i, m_i) we need to distinguish two cases: $s_i = 0$ and $s_i = 1$. Suppose first that $s_i = 1$ so that m_i is observed and $I_i > 0$. From (2.2) it follows that I_i, conditional on m_i and $I_i > 0$, has a normal distribution with mean $x_{i1}'\beta_1 + \rho\sigma_1\sigma_2^{-1}(m_i - x_{i2}'\beta_2)$ and variance $\sigma_1\sqrt{1-\rho^2}$, truncated from below at zero:

$$p(I_i|s_i=1,m_i,\beta_1,\beta_2,\Sigma) = N\left(x_{i1}'\beta_1+\rho\sigma_1\sigma_2^{-1}\left(m_i-x_{i2}'\beta_2\right),\sigma_1\sqrt{1-\rho^2}\right)\mathbb{I}\{I_i>0\}. \tag{2.9}$$

If $s_i = 0$ then it is known that $I_i \leq 0$ but the actual values (I_i, m_i) are not observed. A value of I_i can be generated from the $N\left(x_{i1}'\beta_1, \sigma_1^2\right)$ distribution truncated from above at zero[9]. The value of m_i is a realization of its conditional distribution given I_i:[10]

$$p(I_i|s_i=0,\beta_1,\beta_2,\Sigma) = N\left(x_{i1}'\beta_1,\sigma_1^2\right)\mathbb{I}\{I_i\leq 0\}, \tag{2.10}$$

$$p(m_i|I_i,\beta_1,\beta_2,\Sigma) = N\left(x_{i2}'\beta_2+\rho\sigma_1^{-1}\sigma_2\left(I_i-x_{i1}'\beta_1\right),\sigma_2^2\left(1-\rho^2\right)\right). \tag{2.11}$$

Finally it remains to find the conditional posterior of Σ. By inspection of the SUR likelihood (2.5) it can be seen that the inverse Wishart distribution is the natural conjugate prior. If Σ has an inverse Wishart distribution with parameter matix H and degrees of freedom v, we will write $\Sigma \sim \mathcal{W}^{-1}(H, v)$ and its density is given by

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

[9] All draws from truncated normal distributions can easily be obtained through the inverse c.d.f. method, e.g. Lancaster (2004, p.190-191).

[10] When $s_i = 0$ we could also generate m_i first and then I_i from its conditional distribution given m_i, right-truncated at zero.

Multiplying this density with the SUR likelihood we get

$$p\left(\Sigma|\alpha,\beta,I,m\right) \propto |\Sigma|^{-(n+v+3)/2} \exp\left\{-\frac{1}{2} tr\left(\Sigma^{-1}\left(B+H\right)\right)\right\}, \quad v \geq 2, \tag{2.12}$$

where B was defined in (2.6). Thus the conditional posterior of Σ is $\mathcal{W}^{-1}\left(B+H, n+v\right)$. The Gibbs sampler can now be summarized as follows:

Algorithm 1 (Unidentified Parameters) *For given starting values of* $(\beta_1, \beta_2, \Sigma, I\}$ *and* $\{m_i : y_i = 0\}$*:*

1. *Sample* (β_1, β_2) *from a multivariate normal with mean (2.7) and variance (2.8);*
2. *if* $s_i = 1$*, sample* I_i *from (2.9); if* $s_i = 0$*, sample* I_i *from (2.10) and* m_i *from (2.11);*
3. *sample* Σ *from (2.12);*
4. *return to step* 1 *and repeat.*

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[11]. Our algorithm cannot be trivially modified to satisfy the restriction $\sigma_1 = 1$[12]. For that reason we also adopt the approach of McCulloch, Polson, and Rossi (2000). In essence this entails a reparameterization of the model and placing priors on the identified parameters directly; see also Koop and Poirier (1997) and Li (1998) for applications of this idea to a regime switching model and a Tobit model with endogeneity, respectively.

Given the bivariate normality of (u_{i1}, u_{i2}) and imposing the restriction $\sigma_1 = 1$, it follows that

$$\xi^2 \equiv \text{var}(u_{i2}|u_{i1}) = \sigma_2^2 - \sigma_{12}^2.$$

The covariance matrix can now be written as

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

and the likelihood of the SSM becomes

$$p_{SSM}\left(\ln y|\beta_1,\beta_2,\Sigma\right) = \prod_{i:y_i=0}\left[1-\Phi\left(x_{i1}'\beta_1\right)\right] \times \prod_{i:y_i>0}\left(\xi^2+\sigma_{12}^2\right)^{-1/2}\phi\left(\frac{\ln y_i - x_{i2}'\beta_2}{\sqrt{\xi^2+\sigma_{12}^2}}\right) \times$$

[11] In 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.

[12] Although Σ has an inverse Wishart distribution, Σ conditional on $\sigma_1 = 1$ does not. Nobile (2000) proposes a methods to sample the remaining elements of Σ, conditional on the value of a diagonal element. The algorithm in our paper is based on McCulloch, Polson, and Rossi (2000).

$$\prod_{i:y_i>0} \Phi\left(\frac{x_{i1}'\beta_1\left(\xi^2+\sigma_{12}^2\right)+\sigma_{12}\left(\ln y_i - x_{i2}'\beta_2\right)}{\xi\sqrt{\xi^2+\sigma_{12}^2}}\right). \tag{2.13}$$

In order to generate draws (σ_{12}, ξ) in the Gibbs sampler, we need the conditional posterior $p\left(\sigma_{12}, \xi | I, m, \beta_1, \beta_2\right)$. Given (I, m, β_1, β_2), however, the errors u_1 and u_2 are known and (σ_{12}, ξ^2) are the parameters of a normal linear regression model:

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

Thus, the conditional posterior of interest satisfies

$$\begin{aligned} p\left(\sigma_{12}, \xi^2 | I, m, \beta_1, \beta_2\right) &= p\left(\sigma_{12}, \xi^2 | u_1, u_2, \beta_1, \beta_2\right) \\ &\propto p(u_1, u_2 | \sigma_{12}, \xi^2, \beta_1, \beta_2)\pi(\sigma_{12}, \xi^2 | \beta_1, \beta_2) \\ &\propto p\left(u_1, u_2 | \sigma_{12}, \xi^2\right)\pi\left(\sigma_{12}, \xi^2\right), \end{aligned}$$

where we take (σ_{12}, ξ^2) a priori independent of (β_1, β_2). The natural conjugate prior for (σ_{12}, ξ^2) is of the normal-inverse gamma form. We say ξ^2 has a prior 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 $G(c_0, d_0)$. Then

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

The conditional prior of σ_{12} is given by

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

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, τ) is a set of hyperparameters. It is easy to show that the posteriors take the following form:

$$\begin{aligned} \xi^2 | I, m, \beta_1, \beta_2, \sigma_{12} &\sim IG(\bar{c}, \bar{d}), \\ \bar{c} &= c_0 + \frac{n+1}{2}, \\ \bar{d} &= d_0 + \frac{(\sigma_{12}-g)^2}{2\tau} + \frac{1}{2}\left(u_2 - \sigma_{12}u_1\right)'\left(u_2 - \sigma_{12}u_1\right), \end{aligned} \tag{2.14}$$

and

$$\sigma_{12} | I, m, \beta_1, \beta_2, \xi^2 \sim N\left(\frac{g/\tau + u_1'u_2}{\tau^{-1} + u_1'u_1}, \frac{\xi^2}{\tau^{-1} + u_1'u_1}\right). \tag{2.15}$$

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* $(\beta_1, \beta_2, \sigma_{12}, \xi^2, I\}$ *and* $\{m_i : y_i = 0\}$*:*

1. *Sample* (β_1, β_2) *from a multivariate normal with mean (2.7) and variance (2.8);*
2. *if* $s_i = 1$*, sample* I_i *from (2.9); if* $s_i = 0$*, sample* I_i *from (2.10) and* m_i *from (2.11);*
3. *sample* ξ^2 *from (2.14) and* σ_{12} *from (2.15);*
4. *return to step* 1 *and repeat.*

2.3 The Two-Part Model

The version of the 2PM we use is

$$\begin{aligned} I_i &= x_{1i}'\beta_1 + \varepsilon_{i1}, \\ s_i &= \mathbb{I}\{I_i > 0\}, \\ \ln(y_i|s_i) &= \begin{cases} x_{i2}'\beta_2 + \varepsilon_{i2} & \text{if } s_i = 1 \\ -\infty & \text{if } s_i = 0 \end{cases}. \end{aligned} \tag{2.16}$$

For the parametric model we take

$$\varepsilon_{i1} \sim N\left(0, \sigma_1^2\right), \quad \varepsilon_{i2} \sim N\left(0, \sigma_2^2\right).$$

The selection equation is the same as in the SSM: if $I_i > 0$, then $y_i > 0$ and the logarithm of y_i is well-defined. If $I_i \leq 0$ then $y_i = 0$. The main difference between the SSM and the 2PM concerns the errors ε_{i2} and u_{i2}. In the sample selection model u_{i2} is an error that corresponds to potential outcomes. Conditional on $I_i > 0$ the error then has a nonzero mean that depends on Σ and $x_{i1}'\beta_1$. In contrast, ε_{i2} only affects the logarithm of positive values of expenditures and by construction $E(\varepsilon_{i2}|I_i > 0) = 0$. The 2PM is silent about the joint distribution of $(\varepsilon_{i1}, \varepsilon_{i2})$ and assumes that *conditional on* $\varepsilon_{i1} > -x_{i1}\beta_1$, the errors ε_{i1} and ε_{i2} are independent[13].

The likelihood of the 2PM is

$$\begin{aligned} p_{2PM}(\ln y|\beta_1, \beta_2, \sigma_1, \sigma_2) &= \prod_{i=1}^{n} \left[\Phi\left(x_{i1}'\beta_1/\sigma_1\right)\right]^{s_i} \left[1 - \Phi\left(x_{i1}'\beta_1/\sigma_1\right)\right]^{1-s_i} \times \\ &\quad \prod_{i:y_i>0} \sigma_2^{-1}\phi\left(\frac{\ln y_i - x_{i2}'\beta_2}{\sigma_2}\right). \end{aligned} \tag{2.17}$$

By comparing (2.4) and (2.17) it is clear that the former reduces to the latter when $\rho = 0$. This suggests that in order to discriminate between the SSM and 2PM we can consider inference on the correlation coefficient.

[13] This does not imply that ε_{i1} and ε_{i2} are independent; see Duan et al. (1984) for an example.

2.4 Gibbs Sampling in the 2PM

Approximating the posterior of $(\beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ in the two-part model is considerably easier because ρ (or σ_{12}) is absent from the model. Moreover, the likelihood in (2.17) factors into a Probit term involving (β_1, σ_1^2) and a Gaussian term involving (β_2, σ_2^2). 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'.

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 Albert and Chib (1993). The parameters are (I, β_1) and it remains to find the conditional posteriors $p(I|\beta_1, s)$ and $p(\beta_1|I, s) = p(\beta_1|I)$[14]. Since $I = X_1\beta_1 + u_1$ and $u_1 \sim N(0, I_n)$, it follows that

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

Combining a normal $N(b_1, B_1)$ prior distribution for β_1 with the likelihood of I given above, we get

$$\begin{aligned} \beta_1|I &\sim N\left(\bar{b}_1, \bar{B}_1\right), \\ \bar{B}_1 &= \left(B_1^{-1} + X_1' X_1\right)^{-1}, \\ \bar{b}_1 &= \left(B_1^{-1} + X_1' X_1\right)^{-1}\left(B_1^{-1} b_1 + X_1' X_1 \hat{\beta}_1\right). \end{aligned} \tag{2.18}$$

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:

$$\begin{aligned} p(I_i|\alpha, s_i = 0) &= N\left(x_{i1}'\alpha, 1\right)\mathbb{I}\left\{I_i \leq 0\right\}, \\ p(I_i|\alpha, s_i = 1) &= N\left(x_{i1}'\alpha, 1\right)\mathbb{I}\left\{I_i > 0\right\}. \end{aligned} \tag{2.19}$$

Inference on $\left(\beta, \sigma_2^2\right)$ uses only the subsample in which $y_i > 0$. Let $\ln y^+$, X_2^+ and $u_2^+ = \ln y^+ - X_2^+\beta$ all refer to this subsample of size 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

$$\sigma_2^2|\beta_2, \ln y^+ \sim IG(\bar{c}, \bar{d}), \tag{2.20}$$

$$\bar{c} = c_0 + \frac{n^+}{2}, \quad \bar{d} = d_0 + \frac{1}{2}u_2^{+\prime} u_2^+,$$

$$\beta_2|\sigma_2^2, \ln y^+ \sim N\left(\bar{b}_2, \bar{B}_2\right), \tag{2.21}$$

$$\bar{B}_2 = \left(B_2^{-1} + \sigma_2^{-2} X_2^{+\prime} X_2^+\right)^{-1},$$

[14] The equality follows because s is a function of I.

$$\bar{b}_2 = \left(B_2^{-1} + \sigma_2^{-2} X_2^{+\prime} X_2^{+}\right)^{-1} \left(B_2^{-1} b_2 + \sigma_2^{-2} X_2^{+\prime} X_2^{+} \hat{\beta}_2\right),$$
$$\hat{\beta}_2 = \left(X_2^{+\prime} X_2^{+}\right)^{-1} X_2^{+\prime} \ln y^{+}.$$

The Gibbs sampler in the 2PM can now be summarized as follows:

Algorithm 3 (Two-Part Model) *For given starting values of* $\left(I, \beta_1, \beta_2, \sigma_2^2\right)$*:*

1. *Sample* β_1 *from (2.18) and* I *from (2.19);*
2. *sample* σ_2^2 *from (2.20) and* β *from (2.21);*
3. *return to step* 1 *and repeat.*

3 A Simulation Experiment

We now consider a small simulation experiment to compare algorithms 1, 2 and 3. A sample of size $n = 1,000$ is generated from the sample selection model in 2.1:

$$\begin{aligned}
I_i &= \beta_{10} + x_{i1}\beta_{11} + u_{i1}, \\
m_i &= \beta_{20} + x_{i2}\beta_{21} + u_{i2}, \\
y_i &= \mathbb{I}\{I_i > 0\} e^{m_i}, \\
\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} &\sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \\
\alpha_0 &= \beta_0, \quad \alpha_1 = \beta_1 = 1.
\end{aligned}$$

Note that β_{10} controls the probability p_0 of observing a zero outcome. We let $x_{i1}, x_{i2} \sim U(0, 10)$, so that $\beta_{10} = -2.50$ and $\beta_{10} = -5.00$ correspond to $p_0 = 0.25$ and $p_0 = 0.50$, respectively (cf. Leung and Yu 1996). With the exclusion restriction $x_{i1} \neq x_{i2}$, the parameter β_{21} measures the marginal effect of x_{i2} on the observed *positive* outcomes in both the SSM and 2PM. Of course, the correlation ρ is absent in the 2PM. We start each algorithm at three different points. The number of iterations is $6,000$ with a burn-in period of $3,000$. The figures in this section are therefore estimated posterior distributions based on $9,000$ random draws.

As appears from figures 3.1-3.3, all three Gibbs samplers produce posterior approximations centered around the true parameter value. It is interesting to note that Gibbs sampling in the two-part model does locate the coefficient of the selection equation, β_{12}, and the marginal effect of x_{i2} on positive outcomes, namely β_{22}.

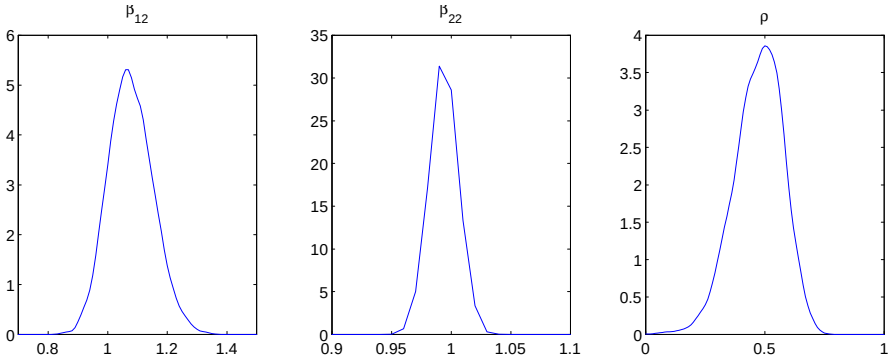

Figure 3.1: Algorithm 1; $p_0 = 0.25$

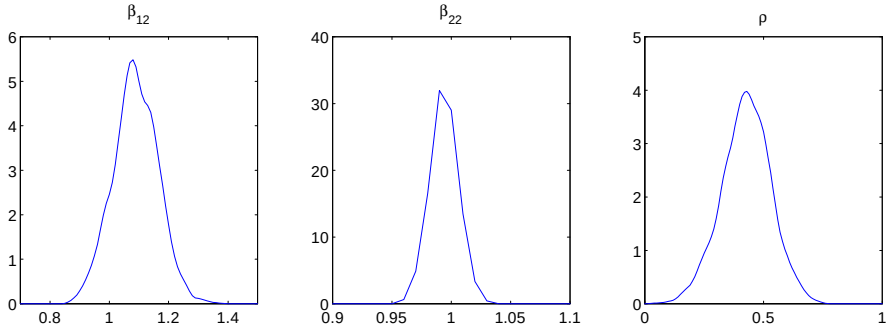

Figure 3.2: Algorithm 2; $p_0 = 0.25$

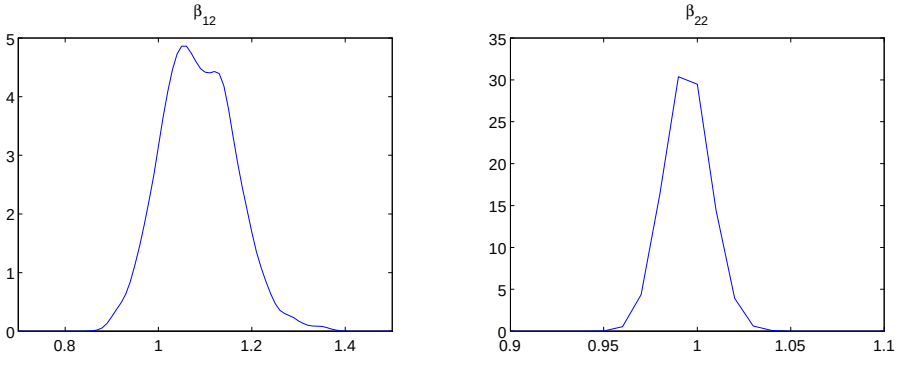

Figure 3.3: Algorithm 3; $p_0 = 0.25$

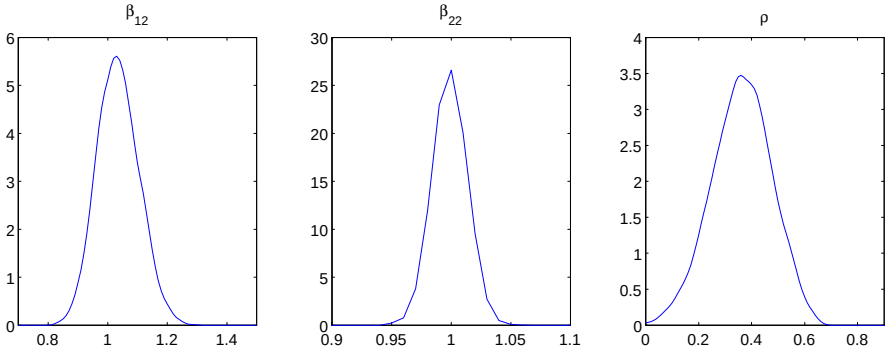

Figure 3.4: Algorithm 1; $p_0 = 0.50$

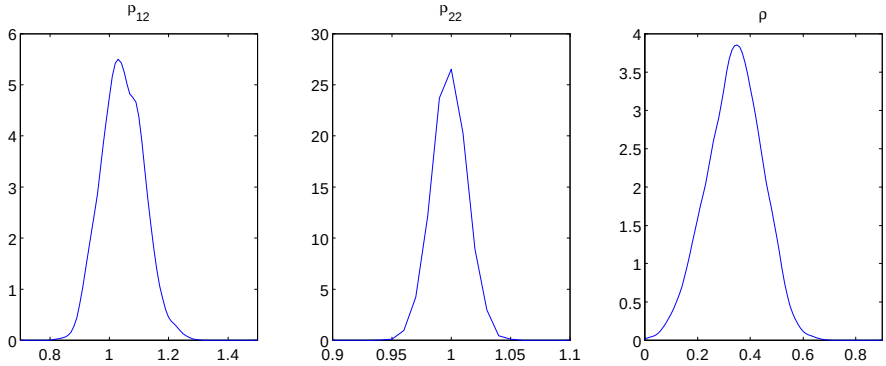

Figure 3.5: Algorithm 2; $p_0 = 0.50$

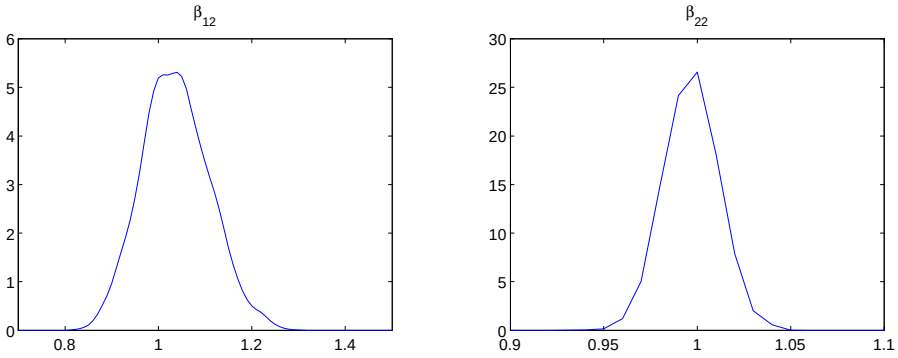

Figure 3.6: Algorithm 3; $p_0 = 0.50$

The results for more severe selection, $p_0 = 0.50$, are given in figures 3.4-3.6. Perhaps surprisingly, even with a larger fraction of zero outcomes all posteriors for β_{12} and β_{22} are centered around the true value. It does become slightly harder to identify the cross-equation correlation in the SSM: the posterior of ρ in both the underidentified and identified models are 'biased' towards zero. In results not reported here, we have found that all algorithms appear to converge in terms of the Gelman-Rubin R-statistic (e.g. Gelman, Carlin, Stern, and Rubin 1995, chapter 11). In addition, the Bayes factor for testing whether $\rho = 0$ decisively rejects the 2PM in favor of the SSM.

4 Bayesian Semiparametric Models

4.1 Mixtures of Normals

There are several ways the assumption of bivariate normality in the SSM can be relaxed. Here we first consider using a mixture of normal distributions with a fixed number of mixture components. The sample selection model is given in (2.1) but now

$$\begin{aligned}\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} &\sim \sum_{j=1}^{k} \gamma_j N(\mu_j, \Sigma_j), \quad \gamma_j \geq 0, \quad \sum_{j=1}^{k} \gamma_j = 1, \\ \mu_j &= \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \end{pmatrix}, \quad \Sigma_j = \begin{bmatrix} \sigma_{j1}^2 & \sigma_{j12} \\ \sigma_{j12} & \sigma_{j2}^2 \end{bmatrix}.\end{aligned}$$

Mixtures of normals are very flexible distributions and even with small values of k (say, 2 or 3) can display skewness, excess kurtosis and bimodality (e.g. Geweke 2005).

To construct a Gibbs sampler for this mixture model, let $\zeta = (\zeta_1, \ldots, \zeta_n)$ be the vector of component selectors; if $\zeta_i = j$ then the errors of observation i are drawn from the j^{th} mixture component:

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

Since the values of ζ_i are not observed, the parameters in the Gibbs sampler are now

$$\begin{aligned}&\beta_1, \beta_2, \zeta, I, \{m_i : y_i = 0\}, \\ \gamma &= \{\gamma_j\}_{j=1}^{k}, \quad \mu = \{\mu_j\}_{j=1}^{k}, \quad \Sigma = \{\Sigma_j\}_{j=1}^{k}\end{aligned}$$

We highlight some of the main features of the sampler and leave additional details to appendix A.1. Recall that in the SUR formulation $W = X\delta + u$. Let $W_{(j)}, X_{(j)}, u_{(j)}$ be the submatrices corresponding to the j^{th} mixture component. The SSM likelihood in (2.5) can be written as

$$\begin{aligned}p(W|\delta, \zeta, \mu, \Sigma) \quad &\propto \\ &\prod_{j=1}^{k} |\Sigma_j|^{-n_j/2} \exp\left\{-\frac{1}{2}(W_{(j)} - X_{(j)}\delta - \mu_j \otimes \iota_{n_j})' A_j (W_{(j)} - X_{(j)}\delta - \mu_j \otimes \iota_{n_j})\right\},\end{aligned}$$

where $A_j = \Sigma_j^{-1} \otimes I_{n_j}$ and $n_j = |\{i : \zeta_i = j\}|$. Using straightforward algebra, it can be shown that

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

where $\hat{\delta}$ is the GLS estimator and $e_{(j)}$ the residual:

$$\begin{aligned} \hat{\delta} &= \left[\sum_{j=1}^{k} X'_{(j)} A_j X_{(j)} \right]^{-1} \sum_{j=1}^{k} X'_{(j)} A_j (W_{(j)} - \mu_j \otimes \iota_{n_j}), \\ e_{(j)} &= W_{(j)} - \mu_j \otimes \iota_{n_j} - X_{(j)} \hat{\delta}. \end{aligned}$$

Combining this likelihood with a $N(d_0, D_0)$ prior, it follows that

$$p(\delta|W, \zeta, \mu, \Sigma) = N(\bar{d}, \bar{D}), \tag{4.1}$$

$$\bar{d} = \left[D_0^{-1} + \sum_{j=1}^{k} X'_{(j)} A_j X_{(j)} \right]^{-1} \left[D_0^{-1} d_0 + \sum_{j=1}^{k} X'_{(j)} A_j X_{(j)} \hat{\delta} \right],$$

$$\bar{D} = \left[D_0^{-1} + \sum_{j=1}^{k} X'_{(j)} A_j X_{(j)} \right]^{-1}.$$

Note that the likelihood of W does not depend on γ, once we condition on the component selectors ζ. The same is then true for the conditional posterior distribution of δ.

Sampling I, $\{m_i : y_i = 0\}$ and Σ is similar as before, so we will be brief. For simplicity we do not list every conditioning argument:

$$I_i|s_i = 1, \zeta_i = j \sim N\left(x'_{i1}\beta_1 + \mu_{j1} + \frac{\sigma_{j12}}{\sigma^2_{j2}}(m_i - x'_{i2}\beta_2 - \mu_{j2}), \sigma^2_{j1}(1 - \rho_j^2) \right) \mathbb{I}\{I_i > 0\}, \tag{4.2}$$

$$I_i|s_i = 0, \zeta_i = j \sim N(x'_{i1}\beta_1 + \mu_{j1}, \sigma^2_{j1}) \mathbb{I}\{I_i \leq 0\}, \tag{4.3}$$

$$m_i|I_i, \zeta_i = j \sim N\left(x'_{i2}\beta_2 + \mu_{j2} + \frac{\sigma_{j12}}{\sigma^2_{j1}}(I_i - x'_{i1}\beta_1 - \mu_{j1}), \sigma^2_{j2}(1 - \rho_j^2) \right). \tag{4.4}$$

Using a $\mathcal{W}^{-1}(H, v)$ prior for Σ_j, the posterior is again of the inverse Wishart form:

$$\begin{aligned} \Sigma_j|\zeta &\sim \mathcal{W}^{-1}(H + B_j, v + n_j), \\ B_j &= \begin{bmatrix} B_{j,11} & B_{j,12} \\ B_{j,12}' & B_{j,22} \end{bmatrix}, \\ B_{j,11} &= (I_{(j)} - X_{1(j)}\beta_1 - \mu_{j1} \otimes \iota_{n_j})'(I_{(j)} - X_{1(j)}\beta_1 - \mu_{j1} \otimes \iota_{n_j}), \end{aligned} \tag{4.5}$$

$$
\begin{aligned}
B_{j,12} &= (I_{(j)} - X_{1(j)}\beta_1 - \mu_{j1} \otimes \iota_{n_j})'(m_{(j)} - X_{2j)}\beta_2 - \mu_{j2} \otimes \iota_{n_j}), \\
B_{j,22} &= (m_{(j)} - X_{2j)}\beta_2 - \mu_{j2} \otimes \iota_{n_j})'(m_{(j)} - X_{2j)}\beta_2 - \mu_{j2} \otimes \iota_{n_j}).
\end{aligned}
$$

It remains to find the posteriors of ζ, γ and μ. The component selector ζ_i has a prior multinomial distribution with $\Pr\{\zeta_i = j|\gamma\} = \gamma_j$, $j = 1, \ldots, k$. Let the i^{th} observation in the SUR formulation be $w_i = X_i\delta + u_i$, where

$$
w_i = \begin{pmatrix} I_i \\ m_i \end{pmatrix}, \quad X_i = \begin{bmatrix} x'_{i1} & 0 \\ 0 & x'_{i2} \end{bmatrix}, \quad u_i = \begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix}.
$$

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

$$
\begin{aligned}
\Pr\{\zeta_i = j|w_i, \delta, \mu, \Sigma, \gamma\} &\propto p(I_i, m_i|\delta, \mu, \Sigma, \gamma, \zeta_i = j\} \times \Pr\{\zeta_i = j|\gamma\} \\
&\propto \gamma_j \left|\Sigma_j\right|^{-1/2} \exp\left\{-\frac{1}{2}(w_i - X_i\delta - \mu_j)'\Sigma_j^{-1}(w_i - X_i\delta - \mu_j)\right\}. \quad (4.6)
\end{aligned}
$$

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

$$
\gamma|\zeta \sim \mathcal{D}(\gamma_0 + n_1, \ldots, \gamma_0 + n_k). \quad (4.7)
$$

Only the likelihood contributions of those observations with $\zeta_i = j$ are informative about μ_j. The natural conjugate prior for μ_j for this partial likelihood is the normal distribution. If $\pi(\mu_j) = N(m_{j0}, M_{j0})$ and the μ_j's are a priori independent, then

$$
\begin{aligned}
&\mu_j|W, \delta, \Sigma, \zeta, \gamma \sim N(\bar{m}_j, \bar{M}_j), \quad (4.8) \\
&\bar{m}_j = \left[M_{j0}^{-1} + (\Sigma_j/n_j)^{-1}\right]^{-1} \left[M_{j0}^{-1}m_{j0} + \Sigma_j^{-1} \sum_{i:\zeta_i=j} (w_i - X_i\delta)\right], \\
&\bar{M}_j = \left[M_{j0}^{-1} + (\Sigma_j/n_j)^{-1}\right]^{-1}.
\end{aligned}
$$

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* $(\delta, \mu, \Sigma, \zeta, \gamma, I)$ *and* $\{m_i : y_i = 0\}$*:*

1. *Sample* $\delta = (\beta'_1, \beta'_2)'$ *from (4.1);*
2. *if* $s_i = 1$ *sample* I_i *from (4.2); if* $s_i = 0$ *sample* I_i *from (4.3) and* m_i *from (4.4);*
3. *sample* Σ_j *from (4.5) for* $j = 1, \ldots, k$*;*
4. *sample* ζ_i *from (4.6) for* $i = 1, \ldots, n$*;*
5. *sample* $\gamma = (\gamma_1, \ldots, \gamma_k)$ *from (4.7);*

6. *sample* μ_j *from (4.8) for* $j = 1, \ldots, k$*;*

7. *return to step* 1 *and repeat.*

A similar algorithm can be constructed for the two-part model. Some simplications will occur because of the assumed conditional independence between the selection and outcome equations. In (2.16) the distributions of both ε_{i1} and ε_{i2} could be modeled as a mixture of normals. One important point to note is that the use of improper priors in a mixture model leads to improper posterior distributions (e.g. Roeder and Wasserman 1997). Second, the state labels j are not identified without further prior information. If the states themselves are the primary focus, for example a state might represent a certain regime or subpopulation, then the algorithm above is not appropriate[15]. In our case, however, we merely use mixtures as a modeling device, and labeling issues are not a concern.

4.2 Dirichlet Mixtures of Normals

The approach 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 different models on the basis of their posterior probabilities. Here we explore the use of Dirichlet process priors in modeling sample selection. Building on a large literature that started with Ferguson ((1973, 1974)) and Antoniak (1974) we show that some of the existing methodology can be readily adapted to the models in this paper. The main appeal of using a Dirichlet process prior lies in the fact that the errors are modeled as a mixture of normals with a random number of mixture components. Through Bayesian updating we can directly make inference about this number. Also, the prior allows us to center in some sense the semiparametric model around the parametric one. The work in this section uses results from Escobar (1994) and Escobar and West (1995)[16], and is closely related to Conley, Hansen, McCulloch, and Rossi (2007), who consider the use of Dirchlet process priors in an instrumental variable model.

The basic setup can be described as follows. In (2.1) let

$$\begin{aligned} \begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} | \theta_i &\sim N(\mu_i, \Sigma_i), \quad \theta_i = (\mu_i, \Sigma_i), \\ \theta_i | G &\sim G, \\ G | \alpha, G_0 &\sim \mathcal{DP}(\alpha, G_0). \end{aligned} \tag{4.9}$$

Here θ_i is simply the set of parameters for the normal distribution. Our discussion of the SSM in section 2 involved $\theta_i = \theta = (\mu, \Sigma)$ and specifying a prior for θ[17]. The semiparametric model in (4.9) allows each pair (u_{i1}, u_{i2}) to have a distinct normal distribution, conditional on θ_i. The parameters

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

[16] Escobar and West (1998), MacEachern (1998) and Müller and Quintana (2004) are excellent reviews of semiparametric modeling with Dirichlet processes.

[17] To be precise, the prior of μ was degenerate at zero and $\pi(\Sigma) \sim \mathcal{W}^{-1}(H, v)$.

$\{\theta_i\}_{i=1}^n$ 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 (4.9) the distribution G itself is treated as unknown and given a Dirichlet process (DP) prior[18]. 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. Sepcifically, 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 $\alpha \to \infty$.

We construct a Gibbs sampler based on fixed values of (α, G_0). Additional details and prior distributions for (α, G_0) are discussed in appendix A.2. Throughout this section the conditioning on (α, G_0) is implicit. The parameters are now

$$\delta, I, \{m_i : y_i = 0\}, \{\theta_i\}_{i=1}^n, G.$$

A simplification occurs because G can be integrated out of the posterior (e.g. Escobar 1994), so that the Gibbs sampler only involves updating the remaining parameters.

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

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

Combining this with a $N(d_0, D_0)$ prior for δ and collecting terms, it can be shown that

$$\begin{aligned} \delta|W, \theta &\sim N(\bar{d}, \bar{D}), \\ \bar{D} &= \left[D_0^{-1} + \sum_{i=1}^n X_i'\Sigma_i^{-1}X_i\right]^{-1}, \\ \bar{d} &= \bar{D}\left[D_0^{-1}d_0 + \sum_{i=1}^n X_i'\Sigma_i^{-1}X_i\hat{\delta}\right], \end{aligned} \tag{4.10}$$

and $\hat{\delta}$ is the weighted least squares estimator:

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

Sampling I_i when $s_i = 1$ and (I_i, m_i) when $s_i = 0$ is done by generating draws from the distributions in (4.2), (4.3) and (4.4), where we now condition on $\theta_i = (\mu_i, \Sigma_i)$, instead of ζ_i.

Let $\theta_{-i} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\}$. Blackwell and MacQueen (1973) show that if $\theta_i|G \sim G$

[18] Suppose 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.

and $G \sim \mathcal{DP}(\alpha, G_0)$, then the distribution of θ_i given θ_{-i} with G integrated out is given by

$$\theta_i|\theta_{-i} \begin{cases} = \theta_j & \text{w. prob. } \frac{1}{\alpha+n-1}, \quad j \neq i \\ \sim G_0 & \text{w. prob. } \frac{\alpha}{\alpha+n-1} \end{cases} . \tag{4.11}$$

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

$$\theta_i|\theta_{-i}, W, \delta \begin{cases} = \theta_j & \text{w. prob. } c^{-1}p(w_i|X_i, \delta, \theta_j), \quad j \neq i \\ \sim p(\theta_i|w_i, X_i, \delta) & \text{w. prob. } c^{-1}\alpha p(w_i|X_i, \delta) \end{cases} . \tag{4.12}$$

Here c^{-1} is a normalizing constant, $p(\theta_i|w_i, \delta)$ is the posterior given prior $dG_0(\theta_i)$ and $p(w_i|X_i, \delta)$ is the likelihood with θ_i integrated out with respect to $dG_0(\theta_i)$:

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

More details are given in appendix A.2. The Gibbs sampler for the Dirichlet process SSM can now be summarized as follows:

Algorithm 5 (Dirichlet SSM) *For given starting values of* $(\delta, \theta, I\}$ *and* $\{m_i : y_i = 0\}$*:*

1. *Sample* δ *from (4.10);*
2. *if* $s_i = 1$ *sample* I_i *from (4.2); if* $s_i = 0$ *sample* I_i *from (4.3) and* m_i *from (4.4); all draws here are conditional on* $\theta_i = (\mu_i, \Sigma_i)$*;*
3. *sample* θ_i *from (4.12) for* $i = 1, \ldots, n$*;*
4. *return to step* 1 *and repeat.*

Algorithm 5 can be extended in several ways. First of all, 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 $E(k_n|\alpha) \approx \alpha \log((\alpha + n)/\alpha)$. By placing a prior on α it is possible to learn about the number of mixture components, after seeing the data; see Escobar (1994) and Escobar and West (1995) who discuss several tractable priors.

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 (4.12) this could happen when the sum (over $j \neq i$) of $c^{-1}p(w_i|X_i, \delta, \theta_j)$ gets large relative to $c^{-1}\alpha p(w_i|X_i, \delta)$. It is possible to slightly reparameterize

the model and associated Gibbs sampler, such that at each iteration the distinct values in θ are 'remixed'; see West, Müller, and Escobar (1994), MacEachern (1998) and the appendix.

5 Conclusion

In this paper we have considered models that can be used to describe nonnegative outcome variables. The sample selection model essentially treats the outcome as a censored quantity, and specifies a structure for the latent process. The two-part model focuses on the observed outcomes directly. Given the strong parametric assumption of normal errors, Bayesian inference in both models can proceed straightforwardly, using a combination of data augmentation and Gibbs sampling. The MCMC schemes can be formulated with or without an identification restriction. We have found there to be hardly any difference in the posterior approximation[19].

A possible Bayesian semiparametric treatment introduces more flexibility in the (joint) error distribution. When mixtures of normal distributions are used, together with natural conjugate priors, we can construct a Gibbs sampler by augmenting the parameter space with a set of latent state variables. Within this MCMC algorithm the number of mixture components is fixed. In principle different specifications can be compared on the basis of a Bayes factor.

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 differ across observations. In this case the Dirichlet process essentially amounts to using a mixture of normals with an unknown number of mixture components. The data may then be used to make inference about this number. Our paper has constructed an MCMC algorithm for use in the sample selection model. The only requirement for tractability is then the choice of conjugate priors.

Much work remains that we shall address in future work. First of all, we aim to provide a thorough comparison between the mixture of normals and Dirichlet process models. In our experience the mixture model shows convergence failure if the true distribution is (close to) a normal. It is not clear what would happen with a Dirichlet process. Finding analytical bounds on the convergence rates of each MCMC algorithm should be useful in this context. Second, imposing identification restrictions in the mixture model will typically destroy any natural conjugacy. For that case, reparameterizing the model may be helpful and we expect that then the more general Metropolis-Hastings algorithm is needed to approximate the posteriors. Finally, Richardson and Green (1997) have proposed a method for using mixtures with an unknown number of components. It remains to contrast their approach with the Dirichlet process model.

[19] This is not always the case: McCulloch, Polson, and Rossi (2000) find large differences in the autocorrelation and convergence behavior of the chains, in case of the multinomial probit model.

References

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 Nonparametric Problems," *The Annals of Statistics*, 2(6), 1152–1174.

Blackwell, D., and J. B. MacQueen (1973): "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1(2), 353–355.

Cameron, A. C., and P. K. Trivedi (2005): *Microeconometrics: Methods and Applications*. Cambridge.

Casella, G., and E. I. George (1992): "Explaining the Gibbs Sampler," *The American Statistician*, 46(3), 167–174.

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.

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.

——— (1984): "Choosing Between the Sample-Selection Model and the Multi-Part Model," *Journal of Business and Economic Statistics*, 2(3), 283–289.

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.

——— (1974): "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2(4), 615–629.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995): *Bayesian Data Analysis*. Chapman & Hall.

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.

Hay, J. W., and R. J. Olsen (1984): "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care," *Journal of Business and Economic Statistics*, 2(3), 279–289.

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.

Koop, G., and D. J. Poirier (1997): "Learning About the Across-Regime Correlation in Switching Regression Models," *Journal of Econometrics*, 78, 217–227.

Lancaster, T. (2004): *An Introduction to Modern Bayesian Econometrics*. Blackwell.

Lee, L. F. (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.

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.

MANNING, W. G. (1998): "The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem," *Journal of Health Economics*, 17, 283–295.

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.

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.

NOBILE, A. (2000): "Comment: Bayesian Multinomial Probit Models with a Normalization Constraint," *Journal of Econometrics*, 99, 335–345.

RICHARDSON, S., AND P. J. GREEN (1997): "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)," *Journal of the Royal Statistical Society B*, 59(4), 731–792.

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.

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. Freedman et al., pp. 363–386.

WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

A Additional Details for Section 4

A.1 Mixtures of Normals

To sample ζ_i from its posterior distribution (4.6) in the mixture model, we can use the following steps:

1. Calculate

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

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

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

4. Find j^*, such that

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

   and set $\zeta_i = j^*$.

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

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

The distribution of the component selectors is multinomial:

$$p(\zeta_1, \ldots, \zeta_n|\gamma) = \gamma_1^{n_1} \cdots \gamma_k^{n_k},$$

where $n_j = |\{i : \zeta_i = j\}|$ and $\sum_{j=1}^{k} n_j = n$. If (δ, μ, Σ) is a priori independent of (ζ, γ), the posterior of γ conditional on the completed data W and the remaining parameters satisfies

$$p(\gamma|W, \delta, \mu, \Sigma, \zeta) \propto p(W|\delta, \mu, \Sigma, \zeta, \gamma) \times p(\zeta|\gamma) \times \pi(\gamma).$$

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

$$\begin{aligned} p(\gamma|W, \delta, \mu, \Sigma, \zeta) &= p(\gamma|\zeta) \\ &\propto p(\zeta|\gamma)\pi(\gamma), \end{aligned}$$

from which (4.7) follows. To generate draws from this distribution, perform the following two steps (e.g. Ferguson 1973): (1) for $j = 1, \ldots, k$, generate $Z_j \sim G(\gamma_0 + n_j, 1)$ and (2) set $\gamma_j = Z_j / \sum_{l=1}^{k} Z_l$.

The posterior of μ_j is given by

$$\begin{aligned} p(\mu_j|W,\delta,\Sigma,\zeta,\gamma) &\propto \exp\left\{-\frac{1}{2}(\mu_j - m_{j0})' M_{j0}^{-1}(\mu_j - m_{j0})\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\sum_{i:\zeta_i=j}(w_i - X_i\delta - \mu_j)'\Sigma_j^{-1}(w_i - X_i\delta - \mu_j)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\mu_j'\bar{M}_j^{-1}\mu_j - 2\mu_j'\bar{M}_j^{-1}\bar{M}_j\bar{m}_j\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\mu_j - \bar{m}_j)'\bar{M}_j^{-1}(\mu_j - \bar{m}_j)\right\}, \end{aligned}$$

which is the kernel of the $N(\bar{m}_j, \bar{M}_j)$ distribution. Note that the posterior mean takes the usual form of a precision-weighted average of the prior mean and the OLS estimator:

$$\bar{m}_j = \left[M_{j0}^{-1} + (\Sigma_j/n_j)^{-1}\right]\left[M_{j0}^{-1}m_{j0} + (\Sigma_j/n_j)^{-1} n_j^{-1}\sum_{i:\zeta_i=j}(w_i - X_i\delta)\right].$$

Here the (system) OLS estimator is obtained by regressing $w_i - X_i\delta$ on $(1,1)'$ in the subsample for which $\zeta_i = j$, which yields a sample average. The variance of that estimator is simply $n_j^{-1}\Sigma_j$.

A.2 Dirichlet Mixtures of Normals

Posterior of θ

We will use the following shorthand notation for the Polya urn prior in (4.11):

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

where δ_{θ_j} is the measure with unit mass at θ_j. Then

$$\begin{aligned} p(\theta_i|\theta_{-i},W,\delta) &\propto \prod_{i=1}^{n} p(w_i|X_i,\delta,\theta_i)\pi(\theta_i|\theta_{-i}) \\ &\propto \frac{\alpha}{\alpha + n - 1}dG_0(\theta_i)p(w_i|X_i,\delta,\theta_i) + \sum_{j\neq i}\frac{p(w_i|X_i,\delta,\theta_i)\delta_{\theta_j}(\theta_i)}{\alpha + n - 1} \\ &\propto \alpha p(w_i|X_i,\delta,\theta_i)dG_0(\theta_i) + \sum_{j\neq i} p(w_i|X_i,\delta,\theta_j)\delta_{\theta_j}(\theta_i). \end{aligned} \tag{A.1}$$

The normalizing constant c satisfies

$$c = \int\left[\alpha p(w_i|X_i,\delta,\theta_i)dG_0(\theta_i) + \sum_{j\neq i} p(w_i|X_i,\delta,\theta_j)\delta_{\theta_j}(\theta_i)\right]d\theta_i$$

$$= \quad \alpha p(w_i|X_i,\delta) + \sum_{j\neq i} p(w_i|X_i,\delta,\theta_j).$$

The posterior (A.1) then becomes

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

which yields (4.12).

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 selectors $\zeta = \{\zeta_i\}_{i=1}^n$ as before:

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

Let k_{-i} be the number of distinct θ values in θ_{-i} and $n_{j,-i} = |\{l : l \neq i, \zeta_l = j\}|$ for $j = 1,\ldots,k_{-i}$. The posterior of $\theta_i|\theta_{-i}$ in (4.12) then becomes

$$\theta_i|\theta_{-i},W,\delta \begin{cases} = \theta_j^* & \text{w. prob. } c^{-1}n_{j,-i}p(w_i|X_i,\delta,\theta_j^*), \quad j = 1,\ldots,k_{-i} \\ \sim p(\theta_i|w_i,X_i,\delta) & \text{w. prob. } c^{-1}\alpha p(w_i|X_i,\delta) \end{cases}. \tag{A.2}$$

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 (A.2) it follows immediately that

$$\Pr\{\zeta_i = j|\zeta_{-i},W,\delta,\theta^*\} = c^{-1}n_{j,-i}p(w_i|X_i,\delta,\theta_j^*), \quad j = 1,\ldots,k_{-i}. \tag{A.3}$$

Also, with probability

$$\Pr\{\zeta_i = 0|\zeta_{-i},W,\delta,\theta^*\} = 1 - c^{-1}\sum_{j=1}^{k_{-i}} n_{j,-i}p(w_i|X_i,\delta,\theta_j^*), \tag{A.4}$$

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

$$p(\theta_1^*,\ldots,\theta_k^*|W,\delta,\zeta,k) \propto \prod_{j=1}^k \left\{\prod_{i:\zeta_i=j} p(w_i|X_i,\delta,\theta_j^*)dG_0(\theta_j^*)\right\},$$

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

$$p(\theta_j^*|W,\delta,\zeta,k) \propto \prod_{i:\zeta_i=j} p(w_i|X_i,\delta,\theta_j^*) dG_0(\theta_j^*), \quad j=1,\ldots,k. \tag{A.5}$$

Thus, step 3 in algorithm 5 can be replaced by

3a. Sample ζ_i from the distribution in (A.3) and (A.4) for $i=1,\ldots,n$;

3b. sample θ_j^* from (A.5) for $j=1,\ldots,k$.