Bayesian Sampling Algorithms for the Sample Selection and Two-Part Models[1]

Martijn van Hasselt

Department of Economics

Brown University

January 30, 2005

[1]I am very grateful to Tony Lancaster for sparking my interest in the topic and providing helpful comments, suggestions and guidance along the way. Frank Kleibergen and participants at the Micro Lunch and the Econometrics Seminar at Brown provided valuable feedback. All remaining errors are my own. Comments are most welcome. Please do not circulate without permission. Contact: Martijn_van_Hasselt@Brown.edu

Abstract

This paper considers two models, namely a sample selection model and a two-part model, for an outcome variable that contains a large fraction of zeros, such as individual expenditures on health care. The sample selection model assumes two phases that determine the outcome: a decision process and an outcome process. Both of these processes may be correlated and, conditional on a favorable decision, the outcome is observed. The two-part model assumes that the decision and outcome processes are uncorrelated. The paper addresses the problem of selecting between these two models. Under a Gaussian specification of the likelihood the models are nested and inference can focus on the correlation coefficient. Using a fully parametric Bayesian approach, I present sampling algorithms for the model parameters that are based on data augmentation. In addition to the sampler output of the correlation coefficient, a Bayes factor can be computed to distinguish between models. The paper illustrates all methods and their potential pitfalls using simulated datasets.

1 Introduction

When modeling individual expenditures on durable goods or health care the data on the outcome variable is typically characterized by a certain fraction of observations clustered at zero and a distribution of positive values that is highly skewed. In a consumer optimization problem a zero, i.e. no demand or expenditures, can be viewed as a corner solution whereas a positive outcome indicates an interior solution.

The current paper considers two specific models that are commonly used in the literature to analyze this kind of data. One essential difference between these two models is how they interpret a zero in the data. In the first model, that we will refer to as a *sample selection model* or SSM, the decision process of each individual is split up into two stages. In the first stage the individual decides whether or not to spend. This stage is described by a structural equation for an underlying latent variable such as utility. If the latent variable falls below a certain threshold, expenditures are zero; if it exceeds this threshold positive expenditures are observed. In the second stage the individual makes a decision on the *level* of spending. If the first stage dictates that expenditures 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 an individual would have spent, had she decided to spend at all. Put differently, in a sample selection model *potential* expenditures are modeled, which are only *partially* observed. An important consequence is that the observed positive values of expenditures follow a pattern that is derived from the latent structure.

More generally sample selection occurs when the observed data is not obtained through randomly sampling the population but rather reflects the outcome of individuals' decision making processes. If the goal is to learn something about the entire population it is important to have an understanding of the process generating the sample. Individuals may select themselves into (or out of) the sample based on observable quantities or unobserved heterogeneity. When latent variables in the latter case also effect the outcome variable, inference using the selected sample may be subject to selection bias.

Early contributions to the sample selection literature are Gronau (1974) and Heckman (1979), among others. Gronau (1974) analyzes self-selection and the potential for selection bias in the labor market when actual observed wages are used to make inference on the distribution of wage

offers. Heckman (1979) treats sample selection as a specification error and proposes a by now very well known two-step estimator that corrects for omitted variable bias. As the current paper takes a Bayesian approach we do not further discuss the frequentist literature at this point. Good recent surveys are Lee (2003) and Vella (1998) who focuses on semiparametric estimation.

The second model is referred to as a *two-part* model or 2PM. One of the first discussions of this model goes back to Cragg (1971). As in the sample selection model two stages are distinguished in the decision making process: the decision whether to spend or not and the decision how much to spend. The level of observed positive expenditures is modeled directly, rather than potential expenditures. The two-part model therefore focuses on actual outcomes. In this framework a zero is truly a zero and does not represent missing data. Two-part (and more generally multi-part) models are described in Wooldridge (2002) and used in Duan, Manning, Morris, and Newhouse (1983) to analyze individuals' medical expenditures.

There has been some debate in the literature as to which model is more appropriate for describing health care expenditures. Duan, Manning, Morris, and Newhouse (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. Hay and Olsen (1984) claim that the 2PM is nested within the SSM and imposes error independence across equations. However, Duan, Manning, Morris, and Newhouse (1984) construct a 2PM counter example in which the errors are dependent. The main difference between the two approaches is that by construction the 2PM assumes away the selection effect. Cross-equation correlations therefore do not appear in the likelihood. The SSM offers a different perspective by initially modeling potential outcomes from which a model for the actual outcomes can be derived. As a consequence the parameters of the two models have a slightly different interpretation.

In their health expenditure application Duan et al. (1983, 1984) find that the 2PM outperforms the SSM in terms of mean squared forecast error (MSFE). Manning, Duan, and Rogers (1987) compare the models on the basis of MSFE and mean prediction error in an extensive Monte Carlo study. They find that the 2PM overall performs very well, even if the SSM is the true model. It is important to note that the goal of these studies is to predict outcomes rather than accurately estimate the model parameters.

Under certain distributional assumptions and given the specific versions of the SSM and 2PM we use in this paper, the 2PM is nested within the SSM.[1] The null hypothesis that the 2PM is the true model, or at least cannot be distinguished from the SSM, can then easily be tested through a classical t-test on the relevant parameter. However, Leung and Yu (1996) present simulation evidence suggesting that this test may perform poorly due to near multicollinearity. For that reason Dow and Norton (2003) propose a test based on the difference in empirical mean squared error (EMSE). A problem with this method is that the EMSE comparison is based on the null hypothesis that the sample selection model represents the truth. This choice of null hypothesis is arbitrary and in general it is not clear what would happen if the null and alternative hypotheses are reversed. Moreover, in the simulation design in which the t-test has very low power, the EMSE test fails to select the correct model.

This paper takes a Bayesian and fully parametric approach to the problem of distinguishing between the 2PM and SSM. Our goal is to make inference about the cross-equation correlation, rather than predicting outcomes. In the case of bivariate normal errors the relative support the data offers to either of the models can be assessed by simulating the posterior distribution of this correlation. To this end we present several Markov Chain Monte Carlo (MCMC) sampling algorithms. If selecting a single model is the ultimate goal of the analysis a posterior odds ratio or Bayes factor can be computed to guide the selection process. An interesting article that is worth mentioning at this point is Munkin and Trivedi (2003) who use a three-equation system to simultaneously model a count variable (visits to the doctor), a continuous nonnegative variable (expenditures) and a treatment indicator (choice of insurance scheme). They label the choice of treatment 'self-selection' whereas in our context the term refers to individuals displaying a positive outcome or not. Although their model is useful for analyzing nonnegative outcome variables, by construction it does not allow for zeros in the continuous outcome variable.[2] Therefore Munkin and Trivedi's (2003) model as it stands cannot be used when there are zeros in the data.

The remainder of this paper is organized as follows: section 2 presents the particular versions of the 2PM and SSM we use and the distributional assumptions that enable a fully parametric

[1] Given the different interpretation of each model the word nested is slightly misleading. We take nested to mean there is a value of the parameter vector such that the two models are observationally equivalent.

[2] Specifically, the continuous outcome variable is modeled as having an exponential distribution where the logarithm of the mean is a linear function of covariates.

Bayesian analysis. Section 3 discusses three Gibbs sampling algorithms. Section 4 reviews some material on Bayes factors and two ways to compute them. Section 5 contains some simulation evidence, whereas section 6 assesses the performance of our methods in Leung and Yu's (1996) simulation designs. Section 7 discusses some extensions and modifications of the Gibbs sampler that may perform better if the likelihood has multiple local maxima. Finally, section 8 concludes and provides directions for future research.

2 The Sample Selection and Two-Part Models

Because it facilitates the discussion we will occasionally refer to the outcome variable y_i as expenditures. As in Leung and Yu (1996) we use the following version of the SSM:

$$\begin{aligned} I_i &= x_{i1}'\alpha + u_{i1}, \\ m_i &= x_{i2}'\beta + u_{i2}, \\ \ln y_i &= \begin{cases} m_i & \text{if } I_i > 0 \\ -\infty & \text{if } I_i \leq 0 \end{cases}. \end{aligned} \tag{2.1}$$

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: it determines whether an agent spends a positive amount or not, depending on whether I_i is positive or not. The equation for m_i represents the logarithm of *potential* expenditures. Potential expenditures are effected by a set of covariates x_{i2} and only observed when $I_i > 0$. Thus, m is a partially observed, partially latent variable. If expenditures y_i are zero we know that $I_i \leq 0$ and m_i is unobserved. On the other hand if y_i is positive we know that $I_i > 0$ and $m_i = \ln y_i$ is observed.[3]

For the fully parametric Bayesian analysis of this model it is assumed that the joint distribution of u_{i1} and u_{i2} is bivariate normal:

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

[3] This is a generic version of the sample selection model which appears in many places in the literature; see Lee (2003).

where ρ is the correlation coefficient. If s_i is the indicator of observing potential expenditures, i.e. $s_i = 1\{I_i > 0\}$, we observe (x'_{i1}, x'_{i2}, s_i) for all $i = 1, ..., n$. Moreover, m_i is observed only if $s_i = 1$. If $s_i = 0$ then m_i is not observed and $y_i = 0$. The random variable s_i has a Bernoulli distribution with

$$\begin{aligned}
\Pr\left\{s_i = 1 | x'_{i1}\alpha\right\} &= \Pr\left\{u_{i1} > -x'_{i1}\alpha\right\} \\
&= \Pr\left\{u_{i1}/\sigma_1 > -x'_{i1}\alpha/\sigma_1\right\} \\
&= \Phi\left(x'_{i1}\alpha/\sigma_1\right),
\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The likelihood of this model for the n observations can now be written as

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

where $p_{u_2|I>0}$ is the density of u_{i2} conditional on $I_i > 0$. To further simplify the above expression, let $f_N(a|b,c)$ and $F_N(a|b,c)$ denote the density and cumulative distribution functions, respectively, of a normal random variable with mean b, variance c, evaluated at a. If $\phi(\cdot)$ denotes the standard normal density function and $\bar{u}_i = \ln y_i - x'_{i2}\beta$, then

$$\begin{aligned}
p_{u_2|I>0}(\bar{u}_i) &= \frac{\int_0^\infty p_{u_2,I}(\bar{u}_i, I) dI}{P(I > 0)} \\
&= \frac{p_{u_2}(\bar{u}_i)}{\Phi\left(x'_{i1}\alpha/\sigma_1\right)} \int_0^\infty p_{I|u_2}(I|\bar{u}_i) dI \\
&= \frac{f_N\left(\bar{u}_i|0,\sigma_2^2\right)}{\Phi\left(x'_{i1}\alpha/\sigma_1\right)} \left[1 - F_N\left(0|x'_{i1}\alpha + (\rho\sigma_1/\sigma_2)\bar{u}_i, \sigma_1^2(1-\rho^2)\right)\right] \\
&= \frac{\sigma_{u2}^{-1}\phi\left(\bar{u}_i/\sigma_2\right)}{\Phi\left(x'_{i1}\alpha/\sigma_1\right)} \Phi\left(\frac{x'_{i1}\alpha + (\rho\sigma_1/\sigma_2)\bar{u}_i}{\sqrt{\sigma_1^2(1-\rho^2)}}\right).
\end{aligned}$$

Plugging this back into (2.3) the likelihood of the SSM becomes

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

From the last expression it is clear that α and σ_1 are not jointly identified through the likelihood. The identification problem that is standard in the Probit model is usually resolved by imposing the restriction $\sigma_1 = 1$. Because we will present a Gibbs sampling algorithm in the next section that involves nonidentified parameters to make inference on the identified parameters, we choose not to impose the variance restriction at this point.

The version of the 2PM we use is

$$\begin{aligned} I_i &= x_{1i}'\alpha + \varepsilon_{i1}, \\ \ln\left(y_i|I_i>0\right) &= x_{i2}'\beta + \varepsilon_{i2}. \end{aligned} \quad (2.5)$$

For this exposition we take $\varepsilon_{i1} \sim N\left(0,\sigma_1^2\right)$ and $\varepsilon_{i2} \sim N\left(0,\sigma_2^2\right)$. The selection equation is the same as in the SSM: if $I_i > 0$ then $y_i > 0$ and the logarithm is well-defined. If $I_i \leq 0$ then $y_i = 0$. The main difference with the SSM 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 $x_{i1}'\alpha$. In contrast ε_{i2} only effects the logarithm of positive values of expenditures and by construction $E\left(\varepsilon_{i2}|I_i>0\right) = 0$. The 2PM is silent about the joint distribution of $(\varepsilon_{i1},\varepsilon_{i2})$ and assumes that *conditional on* $\varepsilon_{i1} > -x_{i1}\alpha$ the errors ε_{i1} and ε_{i2} are independent[4].

[4] This does not imply that ε_{i1} and ε_{i2} are independent. See Duan, Manning, Morris, and Newhouse (1984) for an example.

The likelihood of the 2PM can be written as[5]

$$p_{2PM}\left(\ln y|\alpha,\beta,\sigma_1,\sigma_2\right) = \prod_{i=1}^{n}\left[\Phi\left(x_{i1}'\alpha/\sigma_1\right)\right]^{s_i}\left[1-\Phi\left(x_{i1}'\alpha/\sigma_1\right)\right]^{1-s_i}\times \prod_{i:y_i>0}\left(2\pi\sigma_2^2\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma_2^2}\left(\ln y_i - x_{i2}'\beta\right)^2\right\}. \quad (2.6)$$

By comparing (2.4) and (2.6) 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 in this distributional framework we can consider inference on the correlation coefficient.

3 Posterior Analysis Via Gibbs Sampling

In this section we will present several Gibbs samplers that will aid in distinguishing between the 2PM and SSM. Two approaches are considered. First we develop a Gibbs sampler for the SSM. The output from this algorithm can be used to make inference about the cross-equation correlation and to compute a Bayes factor for the hypothesis that $\rho = 0$. Second, a Gibbs sampler for the 2PM is given whose output can be used to compute a Bayes factor in a different way. Discussion of the Bayes factor and its computation is postponed until the next section. The following two subsections contain the algorithms.

3.1 The Sample Selection Model

By inspection of the likelihood (2.1) it appears that no choice of prior for ρ will yield a tractable posterior distribution. We therefore first develop a Gibbs sampling algorithm that simulates draws from the posterior distribution of (α,β,Σ) and then use these realizations to approximate the posterior of ρ. Since only the selection indicator s_i is observed, the variable I_i is latent and hence treated as an additional parameter in the algorithm. The same can be said about m_i which is partially observed. Hence, through *data-augmentation* we are able to complete the algorithm and generate a sequence of realizations of $(\alpha,\beta,\Sigma,I,m)$ from the posterior[6]. In what follows all

[5] Although α and σ_1 are not jointly identified we do not impose the restriction $\sigma_1 = 1$ at this point for reasons explained earlier.

[6] Albert and Chib (1993) provide an application of data-augmentation to binary and polychotomous response data.

conditional distributions are to be understood as also being conditional on the data. For convenience of the exposition we will not denote this dependency explicitly in our notation.

The most convenient way to analyze the model is to first write it as a 'seemingly unrelated regressions' (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:

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

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

$$\begin{aligned}
p(W|\delta, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}(W - X\delta)'(\Sigma^{-1} \otimes I_n)(W - X\delta)\right\} \\
&\propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} tr\left(B\Sigma^{-1}\right)\right\}, \qquad (3.1)
\end{aligned}$$

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

$$B = \begin{bmatrix} (I - X_1\alpha)'(I - X_1\alpha) & (I - X_1\alpha)'(m - X_2\beta) \\ (m - X_2\beta)'(I - X_1\alpha) & (m - X_2\beta)'(m - X_2\beta) \end{bmatrix}. \qquad (3.2)$$

Starting with the conditional posterior of (α, β), note that $p(\alpha, \beta|I, m, \Sigma, s) = p(\alpha, \beta|I, m, \Sigma)$ be-

cause s is a function of I. The likelihood in (3.1) 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 $\delta=(\alpha,\beta)$, the posterior is again normal:

$$E(\delta|W,\Sigma) = \left[D_0^{-1}+X'S^{-1}X\right]^{-1}\left[D_0^{-1}\delta_0+X'S^{-1}X\hat{\delta}\right], \tag{3.3}$$

$$V(\delta|W,\Sigma) = \left[D_0^{-1}+X'S^{-1}X\right]^{-1}. \tag{3.4}$$

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 $x_{i1}'\alpha+\rho\sigma_1\sigma_2^{-1}(m_i-x_{i2}'\beta)$ and variance $\sigma_1\sqrt{1-\rho^2}$, truncated from below at zero:

$$p(I_i|m_i,I_i>0,\alpha,\beta,\Sigma)=\begin{cases} \left[\Phi\left(\frac{x_{i1}'\alpha+\rho\sigma_1\sigma_2^{-1}(m_i-x_{i2}'\beta)}{\sigma_1\sqrt{1-\rho^2}}\right)\right]^{-1}\times & \\ \sigma_1^{-1}\left(1-\rho^2\right)^{-1/2}\phi\left(\frac{I_i-\left(x_{i1}'\alpha+\rho\sigma_1\sigma_2^{-1}(m_i-x_{i2}'\beta)\right)}{\sigma_1\sqrt{1-\rho^2}}\right) & \text{if } I_i>0 \\ 0 & \text{if } I_i\le 0 \end{cases}. \tag{3.5}$$

If $s_i=0$ then it is known that $I_i\le 0$ but the actual values (I_i, m_i) are not observed. A value of I_i can be generated from the $N\left(x_{i1}'\alpha,\sigma_1^2\right)$ distribution truncated from above at zero[7]. The value of m_i is a realization of its conditional distribution given I_i that follows from (2.2):

$$p(I_i|I_i\le 0,\alpha,\beta,\Sigma)=\begin{cases} \left[1-\Phi\left(x_{i1}'\alpha/\sigma_1\right)\right]^{-1}\sigma_1^{-1}\phi\left(\frac{I_i-x_{i1}'\alpha}{\sigma_1}\right) & \text{if } I_i\le 0 \\ 0 & \text{if } I_i>0 \end{cases}, \tag{3.6}$$

$$p(m_i|I_i,\alpha,\beta,\Sigma)=\sigma_2^{-1}\left(1-\rho^2\right)^{-1/2}\phi\left(\frac{m_i-\left(x_{i2}'\beta+\rho\sigma_1^{-1}\sigma_2\left(I_i-x_{i1}'\alpha\right)\right)}{\sigma_2\sqrt{1-\rho^2}}\right). \tag{3.7}$$

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

Finally it remains to find the conditional posterior of Σ. By inspection of the SUR likelihood (3.1) it can be seen that the inverse Wishart distribution is the natural conjugate prior. If an $m \times m$ matrix Σ has an inverse Wishart distribution with parameter matix H and degrees of freedom v we will write $\Sigma \sim \mathcal{W}^{-1}(H, v, m)$ and its density is given by

$$p(\Sigma|H, v, m) \propto |\Sigma|^{-(v+m+1)/2} \exp\left\{-\frac{1}{2} tr\left(\Sigma^{-1} H\right)\right\}, \quad v \geq m \tag{3.8}$$

Multiplication of this density with the SUR likelihood and substituting $m = 2$ it can be seen that

$$p(\Sigma|\alpha, \beta, I, m) \propto |\Sigma|^{-(n+v+3)/2} \exp\left\{-\frac{1}{2} tr\left(\Sigma^{-1}(B + H)\right)\right\}, \quad v \geq 2 \tag{3.9}$$

where B was defined in (3.2). Thus the conditional posterior of Σ is $\mathcal{W}^{-1}(B + H, n + v, 2)$. The Gibbs sampler can now be summarized as follows:

Algorithm 1 (Unidentified Parameters) *For given starting values of* $(\alpha, \beta, \Sigma, I, m)$*:*

1. *Sample* (α, β) *from a normal distribution with mean (3.3) and variance (3.4);*
2. *If* $s_i = 1$ *sample* I_i *from (3.5). If* $s_i = 0$ *sample* I_i *from (3.6) and* m_i *from (3.7);*
3. *Sample* Σ *from (3.9);*
4. *Return to* 1 *and repeat* T *times.*

Note that this Gibbs sampler involves the unidentified parameters α and σ_1. Thus we expect the posterior distribution of (α, σ_1) to be quite uninformative. However, the output from the algorithm can be used to approximate the posterior of an identified parameter such as ρ. McCulloch and Rossi (1994) employ this technique in the context of a multinomial Probit model. Their finding is that the algorithm typically converges very rapidly. Of course it remains to be seen whether this happens in the SSM. From the discussion so far it is clear that the main advantage of working with unidentified parameters is that standard normal and inverse Wishart priors can be used.

The Gibbs sampler with unidentified parameters cannot be trivially modified[8] to satisfy the restriction $\sigma_1 = 1$. Although Σ has an inverse Wishart distribution, Σ *conditional* on $\sigma_1 = 1$ does

[8]The naive solution of simply replacing the $(1, 1)$ element of Σ by 1 may yield a matrix which is not positive semi-definite.

not. A reparameterization of the covariance matrix, however, will allow us to impose the restriction $\sigma_1 = 1$ and still work with easily tractable priors and posteriors. This idea is used by Koop and Poirier (1997) to analyze the correlation in a regime-switching model. McCulloch, Polson, and Rossi (2000) develop an algorithm for the multinomial Probit model.

When we impose the restriction $\sigma_1 = 1$ and use (2.2) we know that $\text{Var}(u_{i2}|u_{i1}) = \sigma_2^2\left(1-\rho^2\right) = \sigma_2^2 - \sigma_{12}^2$, where σ_{12} is the covariance between u_{i1} and u_{i2}. Now define $\xi^2 \equiv \text{Var}(u_{i2}|u_{i1})$ and write Σ as

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

The likelihood in this new parameterization is

$$\begin{aligned} p_{SSM}\left(\ln y|\alpha,\beta,\Sigma\right) &= \prod_{i:y_i=0}\left[1-\Phi\left(x_{i1}'\alpha\right)\right] \times \prod_{i:y_i>0}\left(\xi^2+\sigma_{12}^2\right)^{-1/2}\phi\left(\frac{\ln y_i - x_{i2}'\beta}{\sqrt{\xi^2+\sigma_{12}^2}}\right) \times \\ &\quad \prod_{i:y_i>0}\Phi\left(\frac{x_{i1}'\alpha\left(\xi^2+\sigma_{12}^2\right)+\sigma_{12}\left(\ln y_i - x_{i2}'\beta\right)}{\xi\sqrt{\xi^2+\sigma_{12}^2}}\right). \end{aligned} \tag{3.10}$$

In order to generate draws (σ_{12},ξ) in the Gibbs sampler we need the conditional posterior $p\left(\sigma_{12},\xi|I,m,\alpha,\beta\right)$. Note that as a result of bivariate normality of (u_{i1},u_{i2}) we can write

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

Thus, if the vectors u_1 and u_2 were known (data) inference about (σ_{12},ξ) can be made using standard Bayesian techniques for the normal linear model. The parameters (α,β) do not effect the posterior. Going back to the posterior of interest we can see that conditional on (I,m,α,β) the values of u_1 and u_2 are known, so that

$$\begin{aligned} p\left(\sigma_{12},\xi|I,m,\alpha,\beta\right) &= p\left(\sigma_{12},\xi|u_1,u_2,\alpha,\beta\right) \\ &= p\left(\sigma_{12},\xi|u_1,u_2\right) \\ &\propto p\left(u_1,u_2|\sigma_{12},\xi\right)\pi\left(\sigma_{12},\xi\right). \end{aligned}$$

The natural conjugate prior on (σ_{12}, ξ) for this model is of the normal-inverse gamma form[9]:

$$\begin{aligned} p(\xi|c_0, d_0) &= \frac{2d_0^{c_0}}{\Gamma(c_0)} \xi^{-(2c_0+1)} \exp\left\{-d_0/\xi^2\right\}, \\ p(\sigma_{12}|g, \tau, \xi) &= \left(2\pi\tau\xi^2\right)^{-1/2} \exp\left\{-\frac{1}{2\tau\xi^2}(\sigma_{12}-g)^2\right\}, \end{aligned} \tag{3.11}$$

where (c_0, d_0, g, τ) is a set of hyperparameters. As we shall see later, this prior specification induces a prior for the correlation coefficient that can be made roughly uniform by an appropriate choice of τ. It is easy to show that

$$\begin{aligned} (\xi|\alpha, \beta, I, m, \sigma_{12}) &\sim \Gamma^{-1}\left(\tilde{c}_0, \tilde{d}_0\right), \\ \tilde{c}_0 &= c_0 + \frac{n+1}{2}, \\ \tilde{d}_0 &= d_0 + \frac{1}{2\tau}(\sigma_{12}-g)^2 + \frac{1}{2}(u_2 - \sigma_{12}u_1)'(u_2 - \sigma_{12}u_1), \end{aligned} \tag{3.12}$$

and

$$(\sigma_{12}|\alpha, \beta, I, m, \xi) \sim N\left(\frac{g/\tau + u_1'u_2}{1/\tau + u_1'u_1}, \frac{\xi^2}{1/\tau + u_1'u_1}\right) \tag{3.13}$$

The Gibbs sampler with identified parameters can now be summarized as

Algorithm 2 (Fully Identified Parameters) *For given starting values of* $(\alpha, \beta, I, m, \xi, \sigma_{12})$*:*

1. *Sample* (α, β) *from a normal distribution with mean (3.3) and variance (3.4);*
2. *If* $s_i = 1$ *sample* I_i *from (3.5). If* $s_i = 0$ *sample* I_i *from (3.6) and* m_i *from (3.7);*
3. *Sample* ξ *from (3.12) and* σ_{12} *from (3.13);*
4. *Return to* 1 *and repeat* T *times.*

[9] In what follows $\Gamma^{-1}(c_0, d_0)$ will denote the inverse-gamma distribution with density function (3.11).

3.2 The Two-Part Model

Sampling from the posterior distribution of $(\alpha, \beta, \sigma_2)$[10] in the 2PM is considerably easier than in the SSM because the likelihood is separable in α and (β, σ_2):

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

Thus, the likelihood consists of a probit part and a log-normal part. The posterior of α can be sampled using data-augmentation as in Albert and Chib (1993): recall that $s_i = 1\{I_i > 0\} = 1\{y_i > 0\}$ is observed but the actual value of I_i is not. Again the vector $I = (I_1, \ldots, I_n)$ is treated as a parameter. The goal is now to find $p(I|\alpha, s)$ and $p(\alpha|I, s) = p(\alpha|I)$[11]. We write $I = X_1\alpha + u_1$ where $u_1 \sim N_n(0, I_n)$. Then

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

Combining a normal $N(\alpha_0, A_0)$ prior distribution for α with the likelihood of I given above, it follows that

$$\begin{aligned} \alpha|I &\sim N\left(\bar{\alpha}, \bar{A}\right), \\ \bar{A} &= \left(A_0^{-1} + X_1'X_1\right)^{-1}, \\ \bar{\alpha} &= \left(A_0^{-1} + X_1'X_1\right)^{-1}\left(A_0^{-1}\alpha_0 + X_1'X_1\hat{\alpha}\right). \end{aligned} \tag{3.14}$$

To compute $p(I|\alpha, s)$ we need to consider the case $s_i = 0$ and $s_i = 1$. Since $I_i|\alpha$ has a normal distribution with mean $x_{i1}'\alpha$ and unit variance, the distribution of I_i given α *and* s_i is truncated

[10] We have chosen to impose $\sigma_1 = 1$ at this point because the algorithm of this section will be used in conjunction with algorithm 2 (which imposes the same restriction) to compute Bayes factors.

[11] This follows because s is a function of I.

normal:

$$p(I_i|\alpha, s_i = 0) = \begin{cases} \frac{(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}\left(I_i - x_{i1}'\alpha\right)^2\right\}}{1-\Phi\left(x_{i1}'\alpha\right)} & \text{if } I_i \leq 0 \\ 0 & \text{if } I_i > 0 \end{cases}, \tag{3.15}$$

$$p(I_i|\alpha, s_i = 1) = \begin{cases} \frac{(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}\left(I_i - x_{i1}'\alpha\right)^2\right\}}{\Phi\left(x_{i1}'\alpha\right)} & \text{if } I_i > 0 \\ 0 & \text{if } I_i \leq 0 \end{cases}.$$

Inference on (β, σ_2) only uses 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 $\pi(\sigma) = \Gamma^{-1}(c_0, d_0)$ and $\pi(\beta) = N(\beta_0, B_0)$ it follows that

$$\sigma_2|\beta, \ln y^+ \sim \Gamma^{-1}\left(\tilde{c}_0, \tilde{d}_0\right), \tag{3.16}$$

$$\begin{aligned} \tilde{c}_0 &= c_0 + \frac{n^+}{2}, \\ \tilde{d}_0 &= d_0 + \frac{1}{2}u_2^{+\prime}u_2^+, \end{aligned}$$

$$\beta|\sigma_2, \ln y^+ \sim N\left(\bar{\beta}, \bar{B}\right), \tag{3.17}$$

$$\begin{aligned} \bar{B} &= \left(B_0^{-1} + \sigma_2^{-2}X_2^{+\prime}X_2^+\right)^{-1}, \\ \bar{\beta} &= \left(B_0^{-1} + \sigma_2^{-2}X_2^{+\prime}X_2^+\right)^{-1}\left(B_0^{-1}\beta_0 + \sigma_2^{-2}X_2^{+\prime}X_2^+\hat{\beta}\right), \\ \hat{\beta} &= \left(X_2^{+\prime}X_2^+\right)^{-1}X_2^{+\prime}\ln y^+. \end{aligned}$$

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

Algorithm 3 (Two-Part Model) *For given starting values of* $(\alpha, I, \beta, \sigma_2)$*:*

1. *Sample* α *from (3.14) and* I *from (3.15);*
2. *Sample* β *from (3.17) and* σ_2 *from (3.16);*
3. *Return to* 1 *and repeat* T *times.*

The next section discusses how these algorithms can be used to compute Bayes factors.

4 Bayes Factors

The Bayes factor provides a way to compare different models on the basis of their prior predictive distribution for the outcome variable. Suppose that two competing models, M_1 and M_2, are entertained to describe the outcome $\ln y$. A model in this context consists of a prior distribution on the appropriate parameters and a likelihood for the data. Given prior probabilities $\pi(M_1)$ and $\pi(M_2)$ on the two models the posterior odds ratio is computed as

$$\begin{aligned} \frac{p(M_1|\ln y)}{p(M_2|\ln y)} &= \frac{p(\ln y|M_1)}{p(\ln y|M_2)} \times \frac{\pi(M_1)}{\pi(M_2)} \\ &= B_{12} \times \text{prior odds ratio.} \end{aligned}$$

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. In this context Kass and Raftery (1995) is a good survey article. In what follows we consider two ways to compute the Bayes factor.

4.1 The Savage Density Ratio

Let M_1 denote the SSM with the restriction $\rho = \sigma_{12} = 0$ imposed and M_2 the unrestricted SSM. The marginal likelihood $m_j(\ln y) \equiv p(\ln y|M_j)$ is given by

$$m_j(\ln y) = \int p_j(\ln y|\alpha, \beta, \sigma_{12}, \xi)\, \pi_j(\alpha, \beta, \sigma_{12}, \xi)\, d\alpha \cdots d\xi,$$

where $p_j(\cdot|\cdot)$ and $\pi_j(\cdot)$ are the likelihood and prior under model $j = 1, 2$, respectively. The prior $\pi_2(\alpha, \beta, \sigma_{12}, \xi)$ for the unrestricted model follows from the ones used to arrive at algorithm 2 in section 3.1:

$$\begin{aligned} \delta &\equiv \left(\alpha', \beta'\right)' \sim N(\delta_0, D_0), \\ \xi &\sim \Gamma^{-1}(c_0, d_0), \quad \sigma_{12}|\xi \sim N\left(g, \tau\xi^2\right). \end{aligned}$$

The Savage density ratio method lets the restricted prior, in which σ_{12} no longer appears, follow from the unrestricted one:

$$\begin{aligned}\pi_1(\alpha,\beta,\xi) &= \pi_2(\alpha,\beta,\xi|\sigma_{12}=0)\\ &= \left[\frac{\pi_2(\alpha,\beta,\sigma_{12},\xi)}{\pi_2(\sigma_{12})}\right]_{\sigma_{12}=0},\end{aligned}$$

where $\pi_2(\sigma_{12})$ is the marginal prior of σ_{12} in the unrestricted SSM. In addition let $p_1(\ln y|\alpha,\beta,\xi) = [p_2(\ln y|\alpha,\beta,\sigma_{12},\xi)]_{\sigma_{12}=0}$. This greatly simplifies computations:

$$\begin{aligned}m_1(\ln y) &= \int p_1(\ln y|\alpha,\beta,\xi)\,\pi_1(\alpha,\beta,\xi)\,d\alpha d\beta d\xi\\ &= \left[\int p_2(\ln y|\alpha,\beta,\sigma_{12},\xi)\,\frac{\pi_2(\alpha,\beta,\sigma_{12},\xi)}{\pi_2(\sigma_{12})}d\alpha d\beta d\xi\right]_{\sigma_{12}=0}\\ &= \left[\frac{p_2(\ln y,\sigma_{12})}{\pi_2(\sigma_{12})}\right]_{\sigma_{12}=0},\\ m_2(\ln y) &= \int p_2(\ln y|\alpha,\beta,\sigma_{12},\xi)\,\pi_2(\alpha,\beta,\sigma_{12},\xi)\,d\alpha d\beta d\sigma_{12} d\xi\\ &= p_2(\ln y).\end{aligned}$$

It then follows that

$$\begin{aligned}B_{12} &= \frac{m_1(\ln y)}{m_2(\ln y)} = \left[\frac{p_2(\ln y,\sigma_{12})}{p_2(\ln y)\,\pi_2(\sigma_{12})}\right]_{\sigma_{12}=0}\\ &= \left[\frac{p_2(\sigma_{12}|\ln y)}{\pi_2(\sigma_{12})}\right]_{\sigma_{12}=0}.\end{aligned} \tag{4.1}$$

The Bayes factor is simply the ratio of the marginal posterior of σ_{12} and the marginal prior, evaluated at the point of interest. The denominator of (4.1) requires a single evaluation of the $t(2c_0, g, \tau d_0/c_0)$ density[12] at the point zero.

To calculate the numerator of (4.1) note that

$$p_2(\sigma_{12}|\ln y) = E\left[p_2(\sigma_{12}|\alpha,\beta,\xi,I,m,\ln y)\right],$$

where the expectation is taken with respect to $p_2(\alpha,\beta,\xi,I,m|\ln y)$. Given a sample

[12] This is the t distribution with $2c_0$ degrees of freedom, mean g and scale $\tau d_0/c_0$. The result follows from observing that $\xi \sim \Gamma^{-1}(c_0,d_0)$, $\sigma_{12}|\xi \sim N(g,\tau\xi^2)$ and $\pi_2(\sigma_{12}) = \int \pi_2(\sigma_{12}|\xi)\,\pi_2(\xi)\,d\xi$.

$\left\{\alpha_{(t)}, \beta_{(t)}, \xi_{(t)}, I_{(t)}, m_{(t)}\right\}_{t=1}^{T}$ generated by algorithm 2 the value of the posterior can be estimated through

$$\hat{p}_2\left(\sigma_{12}=0|\ln y\right)=\frac{1}{T}\sum_{t=1}^{T}p_2\left(\sigma_{12}=0|\alpha_{(t)},\beta_{(t)},\xi_{(t)},I_{(t)},m_{(t)}\right),$$

which requires T evaluations of the density (3.13) at the point zero. Finally, the estimated Bayes factor of the 2PM versus the SSM is $\hat{B}_{12}=\hat{p}_2\left(\sigma_{12}|\ln y\right)/\pi_2\left(\sigma_{12}\right)$ evaluated at $\sigma_{12}=0$. A value smaller than 1 indicates that the data favors the SSM. Similarly, a value greater than one suggests that the 2PM cannot be rejected and that there is little evidence of a selection effect.

4.2 Estimating The Marginal Likelihood from Gibbs Output

The second method we consider to compute the Bayes factor is the one proposed by Chib (1995). Output from the Gibbs sampling algorithms 3 and 2 can be used to estimate $m_1(\ln y)$ and $m_2(\ln y)$ separately and then report their ratio. We apply Chib's (1995) method first to the sample selection model. The marginal likelihood $m_2\left(\ln y\right)$ can be written as

$$m_2\left(\ln y\right)=\frac{p_2\left(\ln y|\alpha,\beta,\sigma_{12},\xi\right)\pi_2\left(\alpha,\beta,\sigma_{12},\xi\right)}{p_2\left(\alpha,\beta,\sigma_{12},\xi|\ln y\right)}.$$

Note that this equation holds for all parameter values in the support of $p_2\left(\alpha,\beta,\sigma_{12},\xi|\ln y\right)$. Now pick a specific value $(\alpha^*,\beta^*,\sigma_{12}^*,\xi^*)$, say the sample mean from the Gibbs output. Taking logarithms we get

$$\begin{aligned}\log m_2\left(\ln y\right) &= \log p_2\left(\ln y|\alpha^*,\beta^*,\sigma_{12}^*,\xi^*\right)+\log\pi_2\left(\alpha^*,\beta^*,\sigma_{12}^*,\xi^*\right)\\ &\quad -\log p_2\left(\alpha^*,\beta^*,\sigma_{12}^*,\xi^*|\ln y\right).\end{aligned}$$

Using (3.10) and the priors on $(\alpha,\beta,\sigma_{12},\xi)$ in section 3.1 the first two terms on the right-hand side can easily be calculated. It remains to estimate the value of the posterior. To this end, write

$$\begin{aligned}\log p_2\left(\alpha^*,\beta^*,\sigma_{12}^*,\xi^*|\ln y\right) &= \log p_2\left(\xi^*|\ln y\right)+\log p_2\left(\sigma_{12}^*|\xi^*,\ln y\right)\\ &\quad +\log p_2\left(\alpha^*,\beta^*|\sigma_{12}^*,\xi^*,\ln y\right).\end{aligned}\tag{4.2}$$

Since

$$p_2\left(\xi^*|\ln y\right) = \int p_2\left(\xi^*|\alpha,\beta,\sigma_{12},I,m,\ln y\right) p_2\left(\alpha,\beta,\sigma_{12},I,m|\ln y\right) d\alpha\cdots dm$$

and a sample $\left\{\alpha_{(t)},\beta_{(t)},\xi_{(t)},I_{(t)},m_{(t)}\right\}_{t=1}^{T}$ is available from the posterior, we estimate this term by

$$\hat{p}_2\left(\xi^*|\ln y\right) = \frac{1}{T}\sum_{t=1}^{T} p_2\left(\xi^*|\alpha_{(t)},\beta_{(t)},\xi_{(t)},I_{(t)},m_{(t)},\ln y\right),$$

where each term in the sum requires evaluating the inverse-gamma density in (3.12)[13]. As for the second term in (4.2) we have

$$p_2\left(\sigma_{12}^*|\xi^*,\ln y\right) = \int p_2\left(\sigma_{12}^*|\alpha,\beta,\xi^*,I,m,\ln y\right) p_2\left(\alpha,\beta,I,m|\xi^*,\ln y\right) d\alpha d\beta dI dm.$$

In order to estimate this term we need a sample from the posterior distribution of (α,β,I,m), given $\ln y$ *and* ξ^*. The current sample does not satisfy this condition. Therefore the algorithm needs to implemented again, this time with ξ fixed at the value ξ^*. This yields a sequence $\left\{\alpha_{(r)},\beta_{(r)},I_{(r)},m_{(r)}\right\}_{r=1}^{R}$ that (approximately) comes from $p_2\left(\alpha,\beta,I,m|\xi^*,\ln y\right)$. The estimate is constructed as

$$\hat{p}_2\left(\sigma_{12}^*|\xi^*,\ln y\right) = \frac{1}{R}\sum_{r=1}^{R} p_2\left(\sigma_{12}^*|\alpha_{(r)},\beta_{(r)},\xi^*,I_{(r)},m_{(r)},\ln y\right),$$

where each term involves evaluating (3.13)[14]. Using similar logic fix $\sigma_{12}=\sigma_{12}^*$ and $\xi=\xi^*$ and run algorithm 2 again to yield a sample $\left\{I_{(q)},m_{(q)}\right\}_{q=1}^{Q}$ from $p_2\left(I,m|\sigma_{12}^*,\xi^*,\ln y\right)$. The third term in (4.2) is then estimated as

$$\hat{p}_2\left(\alpha^*,\beta^*|\sigma_{12}^*,\xi^*,\ln y\right) = \frac{1}{Q}\sum_{q=1}^{Q} p_2\left(\alpha^*,\beta^*|\sigma_{12}^*,\xi^*,I_{(q)},m_{(q)},\ln y\right).$$

Each term in the sum is the value of the multivariate normal density with mean (3.3) and variance (3.4) at the point (α^*,β^*). Only the mean varies with q.

Computations in the 2PM are largely similar so we will be brief. Again the logarithm of the

[13] Note that the parameter $\tilde{d}_0$ depends on $t=1,\ldots,T$.

[14] The mean and variance of the normal distribution in this case both depend on r.

marginal likelihood $m_1(\ln y)$ is split up into the logarithms of the likelihood, the prior and the posterior, evaluated at $(\alpha^*, \beta^*, \sigma_2^*)$. The first two terms are easy to compute. The value of the posterior is estimated using output from algorithm 3. Unlike the SSM the Gibbs sampler needs to be run only once. Suppose a sample $\left\{\alpha_{(t)}, I_{(t)}, \beta_{(t)}, \sigma_{(t)}\right\}_{t=1}^{T}$ is available. The estimate of $p_1(\alpha^*, \beta^*, \sigma_2^* | \ln y)$ uses

$$\begin{aligned}
\hat{p}_1(\alpha^* | \ln y) &= \frac{1}{T}\sum_{t=1}^{T} p_1\left(\alpha^* | I_{(t)}, \ln y\right), \\
\hat{p}_1(\sigma_2^* | \ln y) &= \frac{1}{T}\sum_{t=1}^{T} p_1\left(\sigma_2^* | \beta_{(t)}, \ln y^+\right),
\end{aligned}$$

where we evaluate (3.14) and (3.16) T times. Note that $p_1(\beta^* | \sigma_2^*, \ln y) = p_1(\beta^* | \sigma_2^*, \ln y^+)$ does not have to be estimated an only requires a single function evaluation of (3.17). This concludes our discussion of marginal likelihoods.

5 Simulation Results

In this section we examine each algorithm separately. The parameter of main interest in the sample selection model is the correlation ρ between the selection and outcome equations. In algorithm 1 σ_1 is unrestricted and we compute $\rho = \sigma_{12}/(\sigma_1\sigma_2)$ whereas in algorithm 2 $\sigma_1 = 1$ and $\rho = \sigma_{12}/\sqrt{\xi^2 + \sigma_{12}^2}$. The algorithm for the two-part model does not contain a correlation coefficient. We will investigate to what extent ignoring a nonzero correlation effects our ability to make inference about the remaining parameters.

Unless noted otherwise a data set is generated according to:

$$\begin{aligned}
x_{i1}, x_{i2} &\sim U(-5, 5), \quad \alpha = \beta = 0, \quad n = 100, \\
\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} &\sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.
\end{aligned} \tag{5.1}$$

Thus, the regressors in both the selection and outcome equations are uniformly distributed scalars. The true model is assumed to be the SSM where the outcome variable is generated according to (2.1). The Gibbs samplers are run for $20,000$ iterations after which the first $5,000$ draws, the

so-called burn-in period, are discarded.

5.1 SSM with Unidentified Parameters

The prior of $\delta = \left(\alpha', \beta'\right)'$ is taken to be reasonably vague: $\delta_0 = (0,0)'$ and $D_0 = 1000 * I_2$ where I_2 denotes the 2×2 identity matrix. Recall that in algorithm 1 the prior of Σ is of the inverse-Wishart form, denoted by $\Sigma \sim \mathcal{W}^{-1}(H, v, 2)$, where H is a symmetric positive definite parameter matrix. Since this prior induces a prior on the correlation coefficient, the parameter that we are mainly interested in, it is important to consider the choice of H and v carefully. We find that for $v = 2$, the smallest possible value, the induced prior of ρ has large modes at ± 1.[15] For $v = 3$ the prior is nearly uniform between -1 and $+1$ whereas for larger values it clusters around zero. This pattern emerges more or less independently of the choice of H. Therefore we take $v = 3$ in what follows. If h_{22} is the $(2,2)$ element of H it follows from the properties of the inverse-Wishart distribution (e.g. Zellner 1971, pp.395-396) that $\sigma_2 \sim \Gamma^{-1}(v/2, h_{22}/2)$[16]. By looking at the moments of this distribution we decide to take $h_{22} = 8$ which yields $E(\sigma_2) = 2.2568$ and $V(\sigma_2) = 2.9070$. The value of h_{11} is largely irrelevant for the shape of the induced priors. Finally we set $h_{12} = h_{21} = 0$ because a nonzero value induces a positive or negative slope in the prior of ρ.

Figure 5.1: Prior (dashed) and posterior (solid) of σ_2 and ρ in algorithm 1.

[15] The prior is approximated by generating a large number of Wishart distributed matrices and computing ρ from them.

[16] A similar result holds for σ_1 but it is of less importance because σ_1 is not identified.

Figure 5.2: Autocorrelation function of σ_2 and ρ in algorithm 1.

Parameter	Mean	Median	St.Dev.	2.5%	97.5%
α	-0.0125	-0.0125	0.0432	-0.0987	0.0716
β	0.0027	0.0029	0.0440	-0.0835	0.0897
σ_2	0.9422	0.9353	0.0913	0.7829	1.1422
ρ	0.3243	0.3311	0.1421	0.0292	0.5847

Table 5.1: Output summary for algorithm 1

From figure 5.1 it is clear that the likelihood for this data set is very informative about (σ_2, ρ). The graphs of α and β, not depicted here, are very similar. Table 5.1 contains some summary statistics from the Gibbs output. The 95% highest posterior density intervals for α and β are tightly centered around zero. The autocorrelation in the sampler output of these two parameters drops almost to zero for lags greater than 1. For σ_2 and ρ the autocorrelation is more persistent, see figure 5.2.

5.2 SSM with Identified Parameters

We now use algorithm 2 in which σ_{12} =Cov(u_{i1}, u_{i2}) and $\xi = \left(\text{Var}\left(u_{i2}|u_{i1}\right)\right)^{1/2}$ are sampled separately. To ensure that the prior of ξ has a negligible effect on the posterior, see (3.12), we set $c_0 = d_0 = 1$. When σ_{12} has zero prior mean ($g = 0$) it remains to choose a value of τ. To this end we simulated ξ from the $\Gamma^{-1}(1, 1)$ distribution and σ_{12} from $N\left(0, \tau\xi^2\right)$ for different values of τ. The correlation coefficient is calculated as $\rho = \sigma_{12}/\sqrt{\xi^2 + \sigma_{12}^2}$. The value $\tau = 0.5$ yields a prior for

ρ that is roughly uniform.[17]

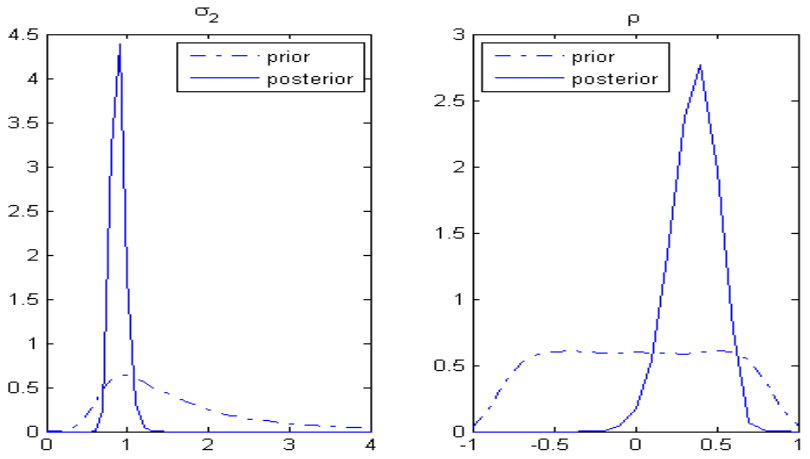

Figure 5.3: Prior (dashed) and posterior (solid) of σ_2 and ρ in algorithm 2.

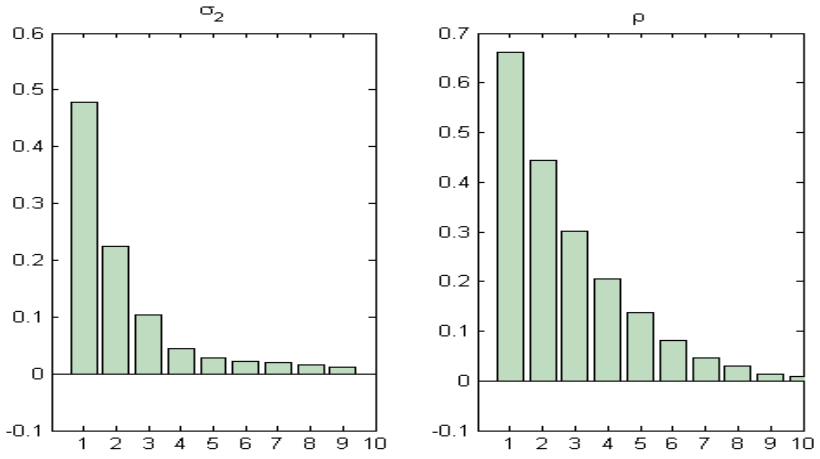

Figure 5.4: Autocorrelation function of σ_2 and ρ in algorithm 2.

The results are very similar to the ones presented in the previous section. One difference between tables 5.1 and 5.2 is that in the identified algorithm the distribution of σ_2 is shifted slightly to the left. It appears by comparing figures 5.4 and 5.2 that in terms of autocorrelation the two algorithms behave similarly. McCulloch, Polson, and Rossi (2000) find that in the context of the multinomial

[17] Larger values of τ cause bimodality at the extremes whereas smaller values of τ put almost zero mass beyond ± 0.5.

Parameter	Mean	Median	St.Dev.	2.5%	97.5%
α	-0.0127	-0.0128	0.0427	-0.0958	0.0705
β	0.0038	0.0038	0.0404	-0.0761	0.0827
σ_2	0.8885	0.8813	0.086	0.741	1.0757
ρ	0.3598	0.3682	0.1418	0.059	0.6087
σ_{12}	0.3217	0.3226	0.1362	0.0503	0.5846

Table 5.2: Output summary for algorithm 2

probit model a sampling scheme with unidentified parameters displays much less autocorrelation than one with identified parameters. Given our results so far, this is not the case in the sample selection model.

5.3 2PM

We will now look at a sample obtained from running algorithm 3 from section 3.2. Because the true correlation coefficient is nonzero it will be interesting to see how ignoring this correlation, as the two-part model does, effects inference on the remaining parameters.

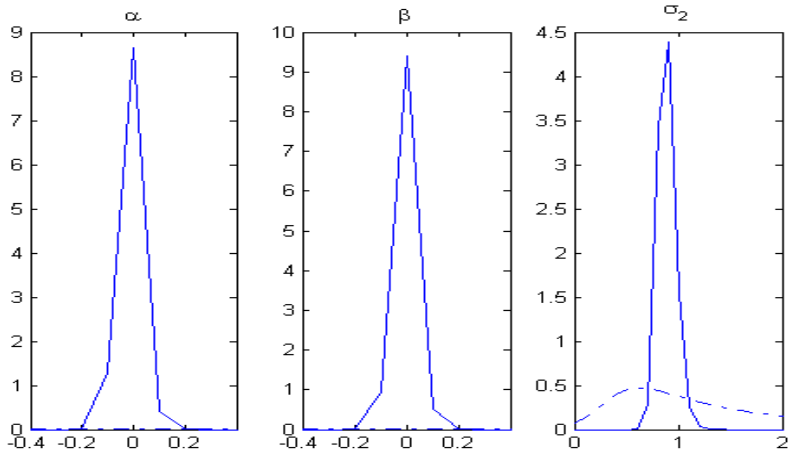

Figure 5.5: Prior (dashed) and posterior (solid) of α, β and σ_2 in algorithm 3.

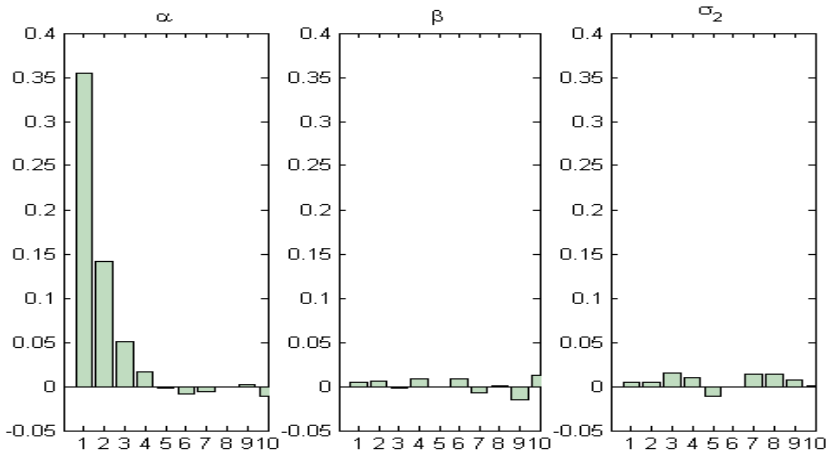

Figure 5.6: Autocorrelation function of α, β and σ_2 in algorithm 3.

Parameter	Mean	Median	St.Dev.	2.5%	97.5%
α	-0.0104	-0.0105	0.0442	-0.0971	0.0765
β	-0.0053	-0.0047	0.0431	-0.0918	0.0794
σ_2	0.881	0.8749	0.085	0.7357	1.0663

Table 5.3: Output summary for algorithm 3

By comparing tables 5.3 and 5.2 we see that the samples of $(\alpha, \beta, \sigma_2)$ are very similar. Ignoring the correlation does not effect the estimated posterior distribution of the remaining parameters.

5.4 Bayes Factors

We compute the Bayes factor of the 2PM versus the SSM using the two methods described in section 4. For the Savage density ratio method algorithm 2 with identified parameters is used to estimate the ratio of posterior to prior density of the covariance σ_{12}, evaluated at zero.[18] Estimating the marginal likelihoods directly is done with the aid of algorithms 2 and 3. Their ratio is the second estimate of the Bayes factor. Because it is unclear how variable these estimates are the samplers are run 20 times with differing starting values.

[18] We did not try this for algorithm 1 because the marginal prior and posterior of σ_{12} (or ρ) are hard to derive then Σ has an inverse-Wishart distribution.

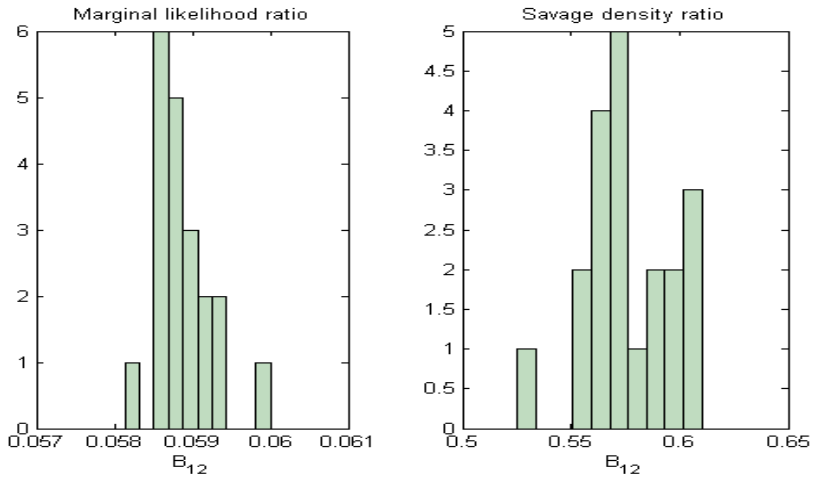

Figure 5.7: Bayes factor estimates for 2PM versus SSM.

The mean and standard deviation of the estimate in the left panel of figure 5.7 are 0.0589 and 0.0004. The corresponding numbers for the right panel are 0.5764 and 0.0216. Therefore it seems that Chib's (1995) method yields an estimate that is both more 'accurate' (i.e. it provides much stronger evidence against the hypothesis of zero correlation) and less variable.

6 Effects of Multicollinearity

A common way to determine whether a selection effect is present is to use two-step estimation of the parameters in (2.1) and a t-test on the coefficient of the inverse Mills ratio (e.g. Wooldridge 2002, pp.560-564). To be more precise, let x_i be the total set of covariates in the model. Previsouly section we had $x_i = (x_{i1}, x_{i2})$ but in general (with a slight abuse of notation) $x_i = (x_{i1} \cup x_{i2}) - (x_{i1} \cap x_{i2})$. The regression function of the logarithmic outcome, conditional on being positive is then

$$\begin{aligned} E\left[\ln y_i | x_i, I_i > 0\right] &= E\left[m_i | x_i, I_i > 0\right] \\ &= E\left[E\left[m_i | x_i, u_{i1}\right] \ | x_i, I_i > 0\right] \\ &= E\left[x_{i2}'\beta + E\left[u_{i2}|u_{i1}\right] \ | x_i, I_i > 0\right] \\ &= x_{i2}'\beta + \rho\sigma_2 E\left[u_{i1} | x_i, u_{i1} > -x_{i1}'\alpha\right] \\ &= x_{i2}'\beta + \rho\sigma_2 \lambda\left(x_{i1}'\alpha\right), \end{aligned}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio, the second equality follows from the law of iterated expectations and the fourth from bivariate normality. Note that to arrive at this equation bivariate normality is not strictly necessary: u_{i2} can be nonnormal but as long as the regression of u_{i2} on u_{i1} is linear with coefficient γ, then $\rho\sigma_2$ in the last line would simply be replaced by γ. Heckman's (1979) two-step method now involves estimating α by $\hat{\alpha}$ via a Probit model and then regressing the subset of $\ln y_i$ for which $y_i > 0$ on x_{i2} and $\lambda(x'_{i1}\hat{\alpha})$ to obtain estimates $\hat{\beta}$ and $\widehat{\rho\sigma_2}$. A t-test can then be used to test the hypothesis $H_0 : \rho = 0$.

As noted, among others, by Manning, Duan, and Rogers (1987) and Leung and Yu (1996) the effectiveness of two-step methods depend on exclusion restrictions between the selection and outcome equations and sufficient variation in the covariates of the selection part. If there are variables in x_{i1} that do not appear in x_{i2} and/or x_{i1} varies substantially, two-step estimators tend to work better. Conversely, when there are no exclusions restrictions and little variation in x_{i1} two-step estimators perform poorly due to multicollinearity between the regressors x_{i2} and the correction term for sample selection $\lambda(x'_{i1}\alpha)$. This problem is exacerbated when the fraction of zeros in the data increases. The goal in this section is to assess the performance of the various sampling algorithms in cases where multicollinearity renders the t-test based on Heckman's two-step estimator useless. More specifically we use Leung and Yu's (1996) designs [1] and [2].

In design [1] x_{i1} and x_{i2} are equal, containing a constant and a uniform random variable. That is, $x'_{i1} = x'_{i2} = x'_i = (1, x)$ where $x \sim U(0, 3)$. The sample size is $n = 1,000$ and the error distribution as in (5.1). Let $\alpha = (\alpha_1, \alpha_2)'$ and $\beta = (\beta_1, \beta_2)'$, where $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2 = 1$. The value of α_1 effects the probability p_0 of observing a zero outcome. We take $\alpha_1 = -0.58, -1.50, -2.42$ corresponding to $p_0 = 0.25, 0.50, 0.75$. In the simulated data set the fraction of zeros will differ slightly from p_0. In what follows we mainly present results concerning ρ.

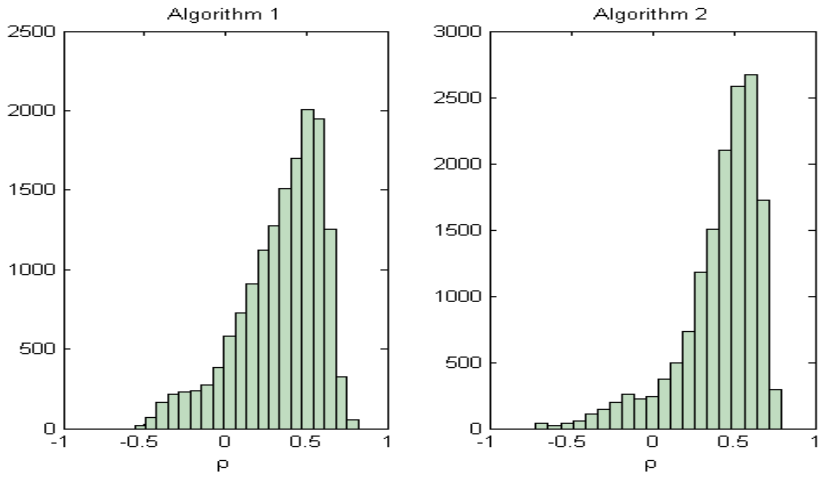

Figure 6.1: Histograms of ρ for $p_0 = 0.25$.

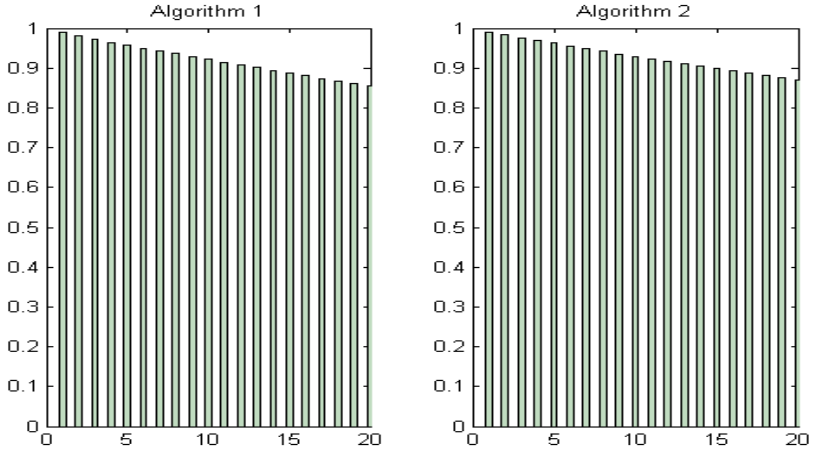

Figure 6.2: Autocorrelation function of ρ for $p_0 = 0.25$.

Algorithm	Mean	Median	St.Dev.	2.5%	97.5%
1	0.3397	0.3959	0.2576	-0.3164	0.6825
2	0.415	0.4821	0.2505	-0.2814	0.7069

Table 6.1: Output summary for ρ when $p_0 = 0.25$.

Figure 6.1 indicates that although the histograms of both samples have a large mode around 0.5 the

left tail is pretty thick. The autocorrelation function in figure 6.2 reveals that the autocorrelation in the Markov chain is large and only decreases very slowly. This may effect the rate of convergence of the chain to its stationary distribution.

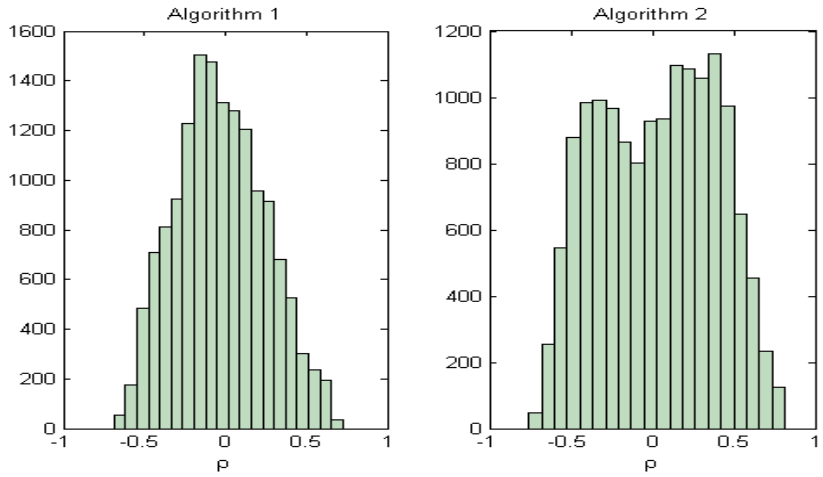

Figure 6.3: Histograms of ρ for $p_0 = 0.50$.

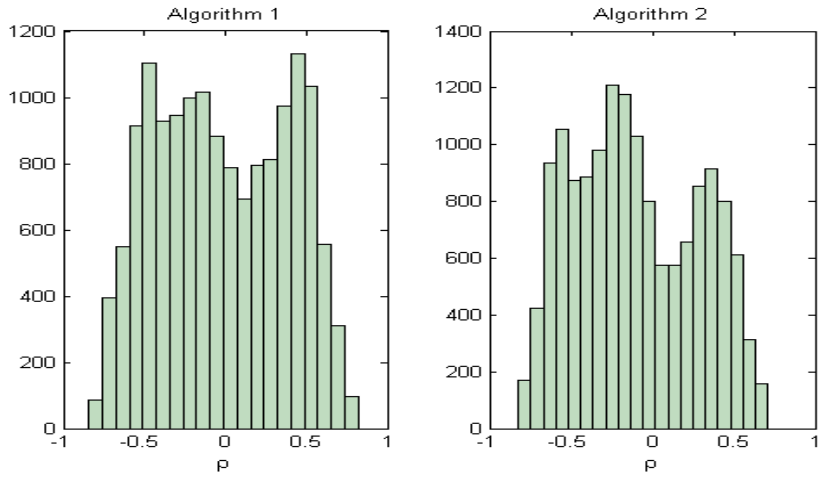

Figure 6.4: Histograms of ρ for $p_0 = 0.75$.

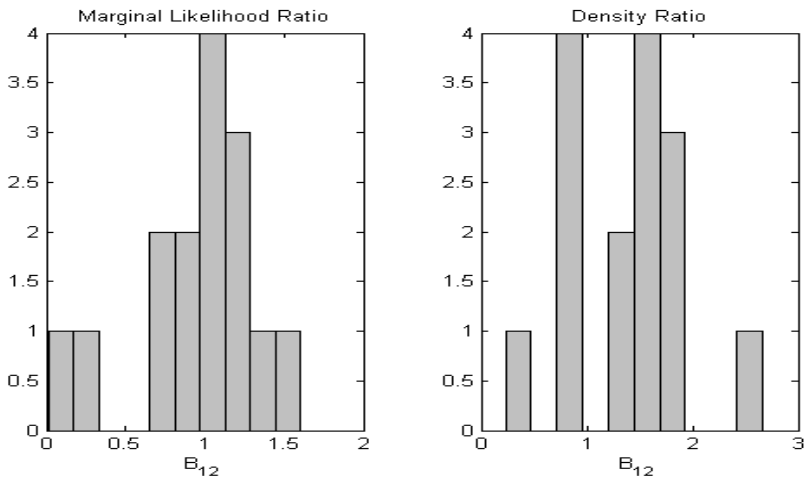

Figure 6.5: Bayes factor estimates in design [1] when $p_0 = 0.75$.

From figures 6.3 and 6.4 it becomes clear that as the probability of observing a zero increases the estimated posterior distributions of ρ become less and less reliable. This holds for both algorithms. The same autocorrelation patterns continue to emerge. Trying to distinguish between the 2PM and SSM also becomes a hopeless task: figure 6.5 shows the highly variable Bayes factor estimates from running 20 different Markov chains.[19] Interestingly the distribution of α, not shown here, is still centered around its true value. Inference about β becomes problematic as the estimated posterior becomes very diffuse.

Design [2] is similar except that now $x \sim U(0, 10)$. The values of the intercept are $\alpha_1 = -2.50, -5.00, -7.50$ which correspond to $p_0 = 0.25, 0.50, 0.75$, respectively.

[19] In 5 cases we ran into numerical problems, so the histograms are based on 15 estimates.

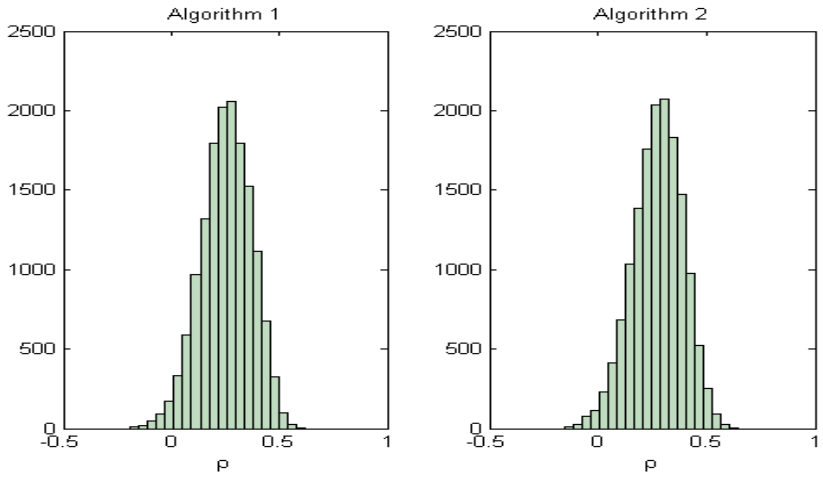

Figure 6.6: Histograms of ρ when $p_0 = 0.25$.

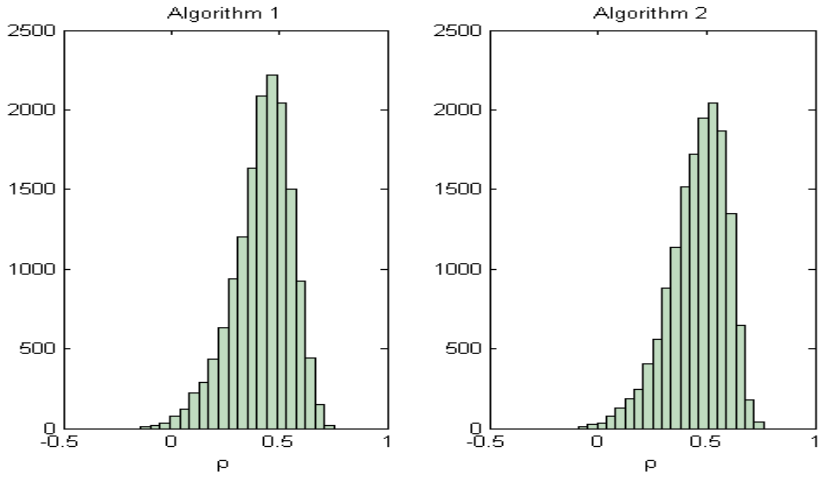

Figure 6.7: Histograms of ρ for $p_0 = 0.50$.

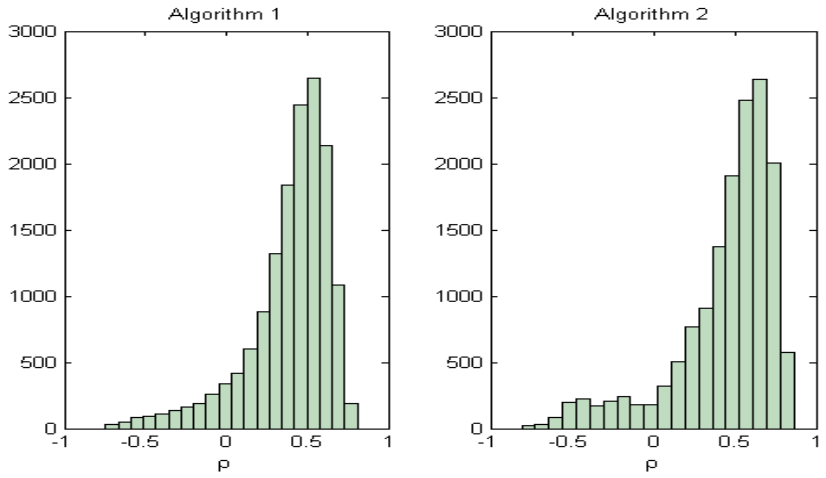

Figure 6.8: Histograms of ρ for $p_0 = 0.75$.

Bayes Factor	$p_0 = 0.25$	$p_0 = 0.50$	$p_0 = 0.75$
Density Ratio	0.7567	0.1260	0.5269
Marginal Likelihood Ratio	0.0300	0.0568	0.0438

Table 6.2: Bayes factors for design [2].

It appears that the posterior distribution of ρ obtained from either algorithm allows much better inference in design [2]. In the absence of exclusion restrictions, as is the case in both simulation designs, identification of the selection effect as measured by ρ comes from sufficient variation in the covariates. Leung and Yu (1996) find that in this case frequentist two-step methods perform well. The Bayes factor estimates in table 6.2 are also indicative of a selection effect. It remains to assess the variability of these estimates.

7 An Alternative Sampling Algorithm

The evidence in the previous section suggests that the absence of an exclusion restriction between the selection and outcome equations leads to difficulties in estimating the correlation, especially when the regressors display little variation. The generated samples of α and σ_2, not considered in the previous section, were all centered around the true parameter values. Even when algorithms 1 and 2 are run for one million iterations and every 100^{th} value is retained to reduce autocorrelation,

the histograms of β, σ_{12} and ρ show substantial dispersion and/or bimodality. However, when the correlation (or covariance) is fixed at its true value, the samples of β are centered around the true value and look roughly normal. Having to sample σ_{12} has a direct effect on β. In the following we therefore focus on these problematic parameters. Recall that in algorithm 2 the full conditional of σ_{12} is given by

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

By checking the components of the mean and variance of this distribution throughout the Markov chain, we find that $u_1'u_2$ is not very stable. At the same time $u_1'u_1$ is always very large, so that the posterior draws of σ_{12} are close to $u_1'u_2$.

An alternative sampler is obtained by sampling (β, σ_{12}) jointly, rather than sequentially. Because of the bivariate normality of (u_{i1}, u_{i2}) we can write

$$\begin{aligned} I &= X_1\alpha + u_1, \\ m &= X_2\beta + u_1\sigma_{12} + \varepsilon, \\ \varepsilon &\sim N\left(0, \xi^2\right). \end{aligned}$$

Given that $\pi(\alpha) = N(\alpha_0, A_0)$ we sample α from (3.14) and I from (3.15) as in the 2PM. Using the equation for m and natural conjugate priors for (β, σ_{12}) and ξ it is straightforward to sample from the posterior. Let $Z = [X_2 : u_1]$ and $\gamma = \left(\beta', \sigma_{12}\right)'$ so that $m = Z\gamma + \varepsilon$. If the priors are

$$\begin{aligned} \pi(\xi) &= \Gamma^{-1}(c_0, d_0), \\ \pi(\gamma|\xi) &= N(\gamma_0, G_0), \\ \gamma_0 &= \left(\beta_0', g\right)', \quad G_0 = \begin{bmatrix} B_0 & 0 \\ 0 & \tau\xi^2 \end{bmatrix}, \end{aligned}$$

then the posteriors are

$$\begin{aligned}
\xi|Z,\gamma,m &\sim \Gamma^{-1}\left(\tilde{c}_0,\tilde{d}_0\right), && (7.1)\\
\tilde{c}_0 &= c_0+\frac{n+1}{2},\\
\tilde{d}_0 &= d_0+\frac{1}{2\tau}\left(\sigma_{12}-g\right)^2+\frac{1}{2}\varepsilon'\varepsilon,\\
\gamma|Z,\xi,m &\sim N\left(\bar{\gamma},\bar{G}\right), && (7.2)\\
\bar{\gamma} &= \left(G_0^{-1}+\xi^{-2}Z'Z\right)^{-1}\left(G_0^{-1}\gamma_0+\xi^{-2}Z'Z\hat{\gamma}\right),\\
\hat{\gamma} &= \left(Z'Z\right)^{-1}Z'm,\\
\bar{G} &= \left(G_0^{-1}+\xi^{-2}Z'Z\right)^{-1}.
\end{aligned}$$

Note that the zeros of m are missing values. Given the current iteration of the sampler, $u_1 = I - X_1\alpha$ is given. A missing value of m can then be generated according to

$$\left(m_i|\gamma,\xi,s_i=0\right)\sim N\left(x_{i2}'\beta+u_{i1}\sigma_{12},\xi^2\right). \tag{7.3}$$

The algorithm, labeled 'Two-Step SSM', can now be summarized as

Algorithm 4 (Two-Step SSM) *For given starting values of* $(\alpha,\beta,\sigma_{12},\xi)$*:*

1. *Sample* α *from (3.14) and* I *from (3.15);*
2. *Generate missing values of* m_i *from (7.3);*
3. *Sample* γ *from (7.2);*
4. *Sample* ξ *from (7.1);*
5. *Return to* 1 *and repeat* T *times.*

The data is generated according to design [1] with $p_0 = 0.25, 0.50, 0.75$. This corresponds to $\beta_1 = -0.58, -1.50, -2.42$. The value of ρ is 0.5 in all cases. The sampler is run for $100,000$ iterations with a burn-in period of $10,000$.

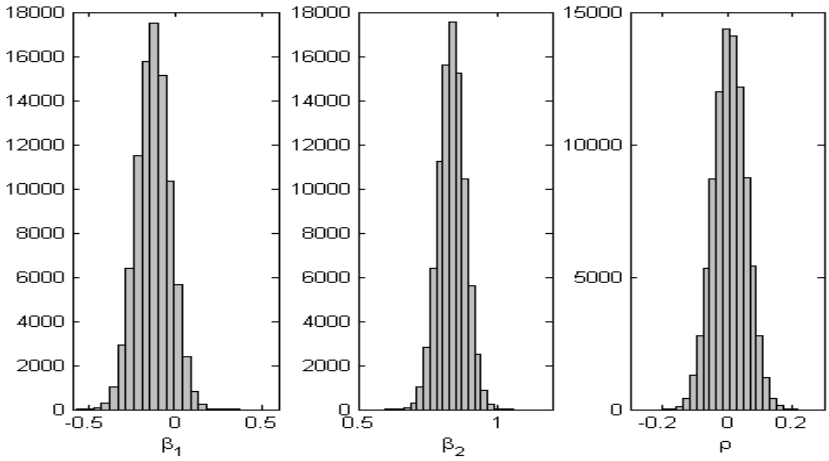

Figure 7.1: Histograms of β_1, β_2, ρ for $p_0 = 0.25$.

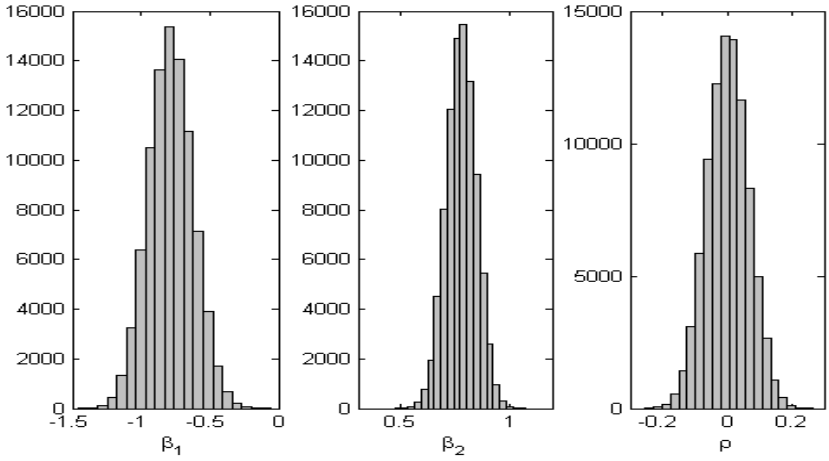

Figure 7.2: Histograms of β_1, β_2, ρ for $p_0 = 0.50$.

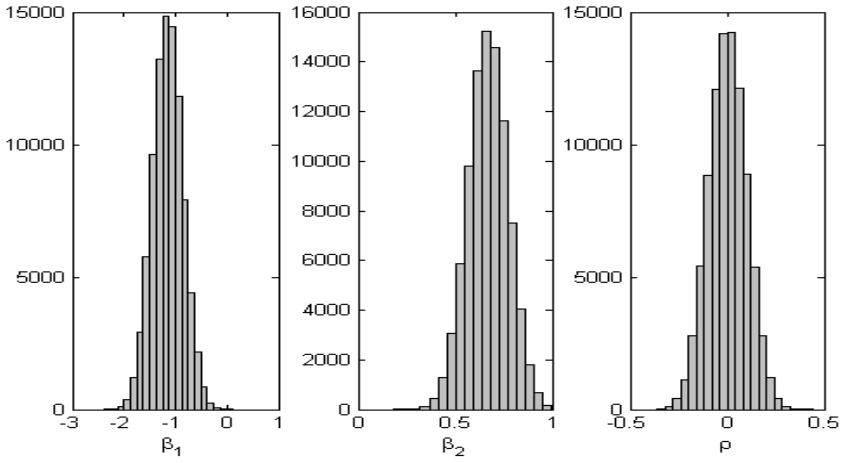

Figure 7.3: Histograms of β_1, β_2, ρ for $p_0 = 0.75$.

Although the autocorrelation functions of β_1, β_2 and ρ now go to zero much faster, the posteriors put no mass around the true parameter values. Both elements of β are biased towards zero, whereas the positive correlation is not detected. As the probability of observing a zero value of m increases, so does the variance of β_1 and ρ. A variation of this algorithm in which only the observed values of m are used, effectively reducing the sample size, yields similar results. Since the prior distributions for β and ρ are virtually flat (take $B_0 = diag\{100, 100\}$ and $\tau = 0.5$) problems arise because the likelihood has multiple local maxima.

8 Conclusions and Directions for Future Research

This paper has developed sampling algorithms for the parameters of the sample selection and two-part models. Since the 2PM does not contain a correlation parameter the SSM has to be used to determine whether a selection effect is present. Since only the sign of I_i is observed and m_i is missing when $I_i \leq 0$, the Gibbs samplers are based on the idea of data augmentation. From the sampler output it is straightforward to approximate the posterior distribution of the correlation coefficient. Our first simulation experiment indicates that if there are exclusion restrictions and the covariates display substantial variation, the draws generated by the Markov chain are centered around the true parameter values. Because classical two-step methods and tests often work with generated regressors the covariance matrix of the estimates needs to be adjusted which can be

complicated in practice. Gibbs sampling is then a comparatively easy way to conduct inference.

Leung and Yu (1996) show that in simulation designs [1] and [2] two-step estimators of the SSM break down. Unfortunately, so do our first two sampling algorithms. We develop a third Gibbs sampler for the SSM that samples β and σ_{12} jointly, rather than sequentially. Although the autocorrelation in the resulting Markov chain is substantially reduced, the realizations of β are biased towards zero. At the same time the sampler does not pick up the positive correlation in our simulation design. Since the prior distribution are all relatively flat the likelihood presents a problem, in particular for β and ρ. The sample selection likelihood in formulas (2.4) does not have a unique maximum. Lee and Chesher (1986) analyze another way in which the likelihood can be problematic: when the true correlation is zero and the covariates satisfy certain conditions the score is identically zero. As a consequence the information matrix is singular and the conventional score test breaks down. Lee and Chesher (1986) also find that some parameters have an asymptotic nonnormal distribution and that convergence to that distribution can be at a rate much lower than $n^{1/2}$. However, ρ is still identified because it determines whether the distribution of positive outcomes is left-skewed, right-skewed or symmetric. We suspect that local maxima are responsible for very large autocorrelation in the Gibbs sampler and bimodality in the posterior of ρ.[20] Olsen (1982) noted that for a given value of ρ the likelihood does have a unique maximum and proposes a grid search method to find the global maximum.

As a direction for future research we intend to develop a more general Metropolis-Hastings type algorithm that would ideally jump away from local maxima in the likelihood. This should improve the mixing properties of the Markov chain and more fully explore the various posterior distributions. An improved algorithm should also decrease the amount of autocorrelation in the Markov chain and allow less variable Bayes factor estimates. In this context the hit-and-run algorithm of Chen and Schmeiser (1993) and the mode-jumping Metropolis algorithm of Tjelmeland and Hegstad (2001) and Tjelmeland and Eidsvik (2004) may prove useful. Another extension of this work is to formulate an algorithm that uses a mixture of normal distributions in the likelihood. Such a specification is much more flexible than the current one and would undoubtedly fit the observed distribution of the positive outcomes much better. Before even constructing such an algorithm it is

[20] In some simulations, not reported in the paper, the Gibbs sampler got 'stuck' around a value far away from the truth.

necessary to reformulate the sample selection model and determine how it really embodies a sample selection effect. Of course the two-part model may be extended in several directions, for example by adding nonlinear terms or using more general error distributions. It also remains to be seen to what extent a more general version of the 2PM is observationally equivalent to the SSM.

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.

Chen, M.-H., and B. Schmeiser (1993): "Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers," *Journal of Computational and Graphical Statistics*, 2(3), 251–272.

Chib, S. (1995): "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90(432), 1313–1321.

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.

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.

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.

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

Lee, L. F. (2003): "Self-Selection," in *A Companion to Theoretical Econometrics*, ed. by B. H. Baltagi, chap. 18. Blackwell Publishing.

Lee, L. F., and A. Chesher (1986): "Specification Testing when Score Test Statistics are Identically Zero," *Journal of Econometrics*, 31, 121–149.

Leung, S. F., and S. Yu (1996): "On the Choice Between Sample Selection and Two-Part Models," *Journal of Econometrics*, 72, 197–229.

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.

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.

Olsen, R. J. (1982): "Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator," *International Economic Review*, 23(1), 223–240.

Tjelmeland, H., and J. Eidsvik (2004): "On the Use of Local Optimizations within Metropolis-Hastings Updates," *Journal of the Royal Statistical Society, Ser. B*, 66, 411–427.

Tjelmeland, H., and B. K. Hegstad (2001): "Mode Jumping Proposals in MCMC," *Scandinavian Journal of Statistics*, 28, 205–223.

VELLA, F. (1998): "Estimating Models with Sample Selection Bias: a Survey," *Journal of Human Resources*, 33, 127–169.

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

ZELLNER, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, Inc., New York.