Journal of Econometrics 165 (2011) 221–232

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian inference in a sample selection model

Martijn van Hasselt *

Behavioral Health Economics Program, RTI International, 3040 Cornwallis Road,PO Box 12194, 27709 Research Triangle Park, NC, United States

ARTICLE INFO

Article history:
Received 11 November 2008
Received in revised form
29 June 2011
Accepted 1 August 2011
Available online 24 August 2011

JEL classification:
C11
C14
C15
C34

Keywords:
Sample selection
Gibbs sampling
Mixture distributions
Dirichlet process

ABSTRACT

This paper develops methods of Bayesian inference in a sample selection model. The main feature of this model is that the outcome variable is only partially observed. We first present a Gibbs sampling algorithm for a model in which the selection and outcome errors are normally distributed. The algorithm is then extended to analyze models that are characterized by nonnormality. Specifically, we use a Dirichlet process prior and model the distribution of the unobservables as a mixture of normal distributions with a random number of components. The posterior distribution in this model can simultaneously detect the presence of selection effects and departures from normality. Our methods are illustrated using some simulated data and an abstract from the RAND health insurance experiment.

1. Introduction

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

A selection model typically consists of two components. The first is an equation that determines the level of the outcome variable of interest. The second is an equation describing the selection mechanism: it determines whether we observe the outcome or not. The latter can sometimes be given a structural interpretation, in which the dependent variable in the selection equation represents an agent's latent utility. If this utility crosses a certain threshold level, the agent acts in such a way that his or her outcome is observed. If the threshold is not crossed, the agent acts differently and the outcome remains unobserved. Thus, a selection model can be viewed as a model for *potential* outcomes that are only *partially* realized and observed. This interpretation applies most directly to the context of modeling a wage offer distribution. Here the wage offer is a potential outcome that is realized only when an individual actually participates in the labor force.

The importance of selection issues in the analysis of labor markets was recognized early on by, among others, Gronau (1974) and Heckman (1974). In his seminal contribution Heckman (1979) treats sample selection as a potential specification error and proposes a two-step estimator that corrects for omitted variable bias. Both Heckman's two-step procedure and full-information maximum likelihood have since been widely used in applied work, and are readily available routines in many statistical software packages. An obvious problem, however, is that these estimation methods rely on strong parametric assumptions about the distribution of unobservables. When these assumptions are violated the estimators may become inconsistent. To overcome this problem a number of semiparametric methods have been proposed. Examples include Cosslett (1991), Ichimura and Lee (1991), Ahn and Powell (1993) and Lee (1994). An excellent survey of this literature is Vella (1998).

Despite the numerous contributions in classical econometrics, the Bayesian literature on selection models has remained relatively sparse. Bayarri and DeGroot (1987); Bayarri and Berger (1998) and Lee and Berger (2001) consider inference based on a univariate selected sample. More recently, Chib et al. (2009) develop a Bayesian method of inference in regression models that are subject to sample selection and endogeneity of some of the covariates. They consider models that are potentially nonlinear, but have normally distributed structural errors.

* Tel.: +1 919 541 6925; fax: +1 919 485 5555.
E-mail address: mvhasselt@rti.org.

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.08.003

Our paper adds to this literature by developing a Bayesian approach to inference in a type 2 Tobit model (e.g., Amemiya, 1985, Ch. 10). In this model the selection rule is binary: we only observe whether the latent selection variable crosses a threshold or not.[1] Although we do not explicitly treat alternative selection mechanisms, it is relatively easy to modify the methods presented here to cover such cases. We provide Gibbs sampling algorithms that produce an approximate sample from the posterior distribution of the model parameters. Our paper differs from Chib et al. (2009) in that we also consider a model with a flexible distribution for the unobserved heterogeneity (i.e. the residuals or 'errors' in the two model equations). The starting point for our analysis is a bivariate normal distribution. Gibbs sampling in this case is fairly straightforward. The basic model may, of course, be misspecified. We therefore extend the analysis to a semiparametric model, based on the Dirichlet process prior of Ferguson (1973, 1974). This prior implies that the unobserved heterogeneity follows a mixture distribution with a random number of components. It has become increasingly popular in Bayesian semiparametric analyses, and our contribution is to incorporate it into a sample selection framework.[2]

A Bayesian approach to inference has two attractive features. First, the output of the sampling algorithm not only provides the Bayesian analogue of confidence intervals for the model parameters, it also gives an immediate indication of the presence (or absence) of a selection effect and departures from normality. Second, if the econometrician has prior information, e.g. restrictions on the parameters, then this information can be easily incorporated through the prior distribution.

The remainder of this paper is organized as follows. Section 2 presents the selection model with bivariate normal errors and a Gibbs sampling algorithm. In Section 3 we develop the extension to a mixture model. We discuss identification issues, the Dirichlet process prior and present the corresponding algorithm to approximate the posterior. The use of the Dirichlet mixture model is illustrated in Section 4 with some simulated data, whereas Section 5 contains an application to estimating a model for health care expenditures, using an abstract of the RAND health insurance experiment. Section 6 concludes and details regarding the various Gibbs samplers are collected in Appendix. With regard to notation, $\mathcal{N}_k(\mu, \Sigma)$ denotes a k-dimensional normal distribution with mean μ and variance Σ. Unless there is ambiguity about the dimension, we will usually omit the subscript k. We use $\mathcal{TN}_{(a,b)}(\mu, \Sigma)$ to denote a $\mathcal{N}(\mu, \Sigma)$ distribution, truncated to the interval (a, b). The standard normal density and distribution functions are $\phi(\cdot)$ and $\Phi(\cdot)$, respectively. Finally, $\mathcal{G}(c_0, d_0)$ denotes the gamma distribution with parameters (c_0, d_0) and expected value c_0/d_0.

2. A sample selection model

2.1. Likelihood and prior

We use the following selection model for an individual member i of the population:

$$\begin{aligned} s_i^* &= x_{i1}'\beta_1 + u_{i1}, \\ s_i &= \mathbb{I}\{s_i^* > 0\}, \\ y_i &= \begin{cases} x_{i2}'\beta_2 + u_{i2} & \text{if } s_i = 1 \\ \text{missing} & \text{if } s_i = 0, \end{cases} \end{aligned} \tag{1}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. The row vectors x_{i1}' and x_{i2}' contain k_1 and k_2 variables, respectively. If x_i denotes the vector of distinct covariates in (x_{i1}', x_{i2}'), the econometrician observes an i.i.d. sample $\{x_i, y_i, s_i\}_{i=1}^n$ of size n from the population model.[3] Note that the outcome y_i is observed if and only if $s_i = 1$. We define $N_1 = \{i : s_i = 1\}$ and $N_0 = \{i : s_i = 0\}$ as the index sets of the observed and missing outcomes, respectively.

Letting $u_i = (u_{i1}, u_{i2})'$ be the vector of errors, a simple parametric model is obtained when we assume that $u_i|x_i \sim \mathcal{N}(0, \Sigma)$. This rules out the case where some of the covariates in x_i are endogenous. Provided valid instrumental variables are available, the selection model can be expanded with a set of reduced-form equations that relate instruments to endogenous variables. A parametric model then specifies a joint distribution (e.g. multivariate normal) for u_i and the reduced-form errors. This approach to modeling endogeneity is taken by Chib et al. (2009), and can be adapted for the models we discuss in this paper. To save space and keep the notation relatively simple, we do not present such an extension here.

Similar to Koop and Poirier (1997), we parameterize the covariance matrix of u_i as

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

where σ_{12} is the covariance and ξ^2 the conditional variance of u_{i2}, given u_{i1}. When the covariance is zero, u_{i1} is independent of y_i^* and we can conduct inference about β_2 based on the subsample indexed by N_1. This strategy would lead to selection bias when $\sigma_{12} \neq 0$. Setting the variance of u_{i1} equal to one is the typical identification constraint for a binary choice model. It should be noted that in a Bayesian treatment of this model it is not necessary to impose this constraint. We could proceed with an unrestricted covariance matrix and conduct inference in a way similar to McCulloch and Rossi (1994). The main difficulty, however, lies with selecting a prior for the unidentified parameters. This prior will induce a prior for the identified parameters, and needs to be carefully checked to ensure that it appropriately reflects a researcher's beliefs. The advantage of the current model formulation is that a prior is placed directly on the identified parameters; see Li (1998) and McCulloch et al. (2000), who proposed this strategy before.

In what follows let $\theta' = (\beta_1', \beta_2', \sigma_{12}, \xi^2)$ be the vector of model parameters. For the observed outcomes we know that $y_i|\theta \sim \mathcal{N}(x_{i2}'\beta_2, \xi^2 + \sigma_{12}^2)$.[4] It follows from the bivariate normality assumption that

$$\Pr\{s_i = 1|y_i, \theta\} = \Phi\left(x_{i1}'\beta_1\sqrt{1 + \sigma_{12}^2/\xi^2} + \frac{\sigma_{12}(y_i - x_{i2}'\beta_2)}{\xi\sqrt{\xi^2 + \sigma_{12}^2}}\right).$$

On the other hand, when the outcome is missing it does not contribute to the likelihood, and the probability that this occurs is $\Pr\{s_i = 0|\theta\} = 1 - \Phi(x_{i1}'\beta_1)$.

If y and s are the n-dimensional sample vectors of (y_i, s_i) values, the likelihood is given by

$$\begin{aligned} f(y, s|\theta) = &\prod_{i \in N_0}[1 - \Phi(x_{i1}'\beta_1)] \\ &\times \prod_{i \in N_1}(\xi^2 + \sigma_{12}^2)^{-1/2}\phi\left(\frac{y_i - x_{i2}'\beta_2}{\sqrt{\xi^2 + \sigma_{12}^2}}\right) \\ &\times \Phi\left(x_{i1}'\beta_1\sqrt{1 + \sigma_{12}^2/\xi^2} + \frac{\sigma_{12}(y_i - x_{i2}'\beta_2)}{\xi\sqrt{\xi^2 + \sigma_{12}^2}}\right). \end{aligned} \tag{3}$$

[1] In some cases the selection process contains more information. Lee (1994) uses a Tobit model for the selection equation. The outcome of interest is then observed based on a selection variable which itself is partially observed.

[2] A recent example is Conley et al. (2008) who use the Dirichlet process prior in an instrumental variable model.

[3] We allow for x_{i2} to be unobserved as well when $s_i = 0$. However, x_{i1} is observed for all sampling units.

[4] Throughout this paper we will omit conditioning on x_i for notational simplicity.

It remains to specify a prior distribution $f(\theta)$. For computational convenience we take β_1, β_2 and (σ_{12}, ξ^2) to be independent in the prior and use $\beta_1 \sim \mathcal{N}(b_1, B_1)$, $\beta_2 \sim \mathcal{N}(b_2, B_2)$, $\xi^2 \sim \mathcal{IG}(c_0, d_0)$ and $\sigma_{12}|\xi^2 \sim \mathcal{N}(0, \tau\xi^2)$. Here $\mathcal{IG}$ denotes the inverse-gamma distribution.[5] The prior dependence between σ_{12} and ξ^2 through $\tau > 0$ allows various shapes of the prior on the correlation between u_{i1} and u_{i2}.[6]

2.2. Posterior

An application of Bayes' rule leads to the joint posterior distribution:

$$f(\theta|y, s) \propto f(y, s|\theta)f(\theta).$$

Given the likelihood function in (3), however, the posterior does not belong to any well-known parametric family. It is therefore relatively difficult, for example, to generate random draws from the posterior or calculate its marginals. Fortunately, a number of simulation algorithms can be used to approximate the posterior. Chen et al. (2000) provide an excellent overview and discussion of available methods. In this paper we approximate the posterior distribution of θ via Gibbs sampling, when the data is augmented with the vector of latent selection variables $s^* = (s_1^*, \ldots, s_n^*)$. Data augmentation, first proposed by Tanner and Wong (1987), has become a useful tool in Bayesian treatments of latent variable models. Examples include Albert and Chib (1993), McCulloch and Rossi (1994) and Munkin and Trivedi (2003). If θ is partitioned into M 'blocks' of parameters $\{\theta_m\}_{m=1}^M$, then the joint posterior $f(\theta, s^*|y, s)$ is approximated by generating random draws of $\{\theta_m\}_{m=1}^M$ and s^* from their respective conditional posteriors, and repeating this cycle many times. The advantage of data augmentation is that, given a natural conjugate prior, the conditional posteriors are standard parametric distributions. This, in turn, makes Gibbs sampling straightforward.[7]

Sampling of s^* is based on

$$f(s^*|y, s, \theta) = \prod_{i\in N_1} f(s_i^*|y_i, s_i, \theta) \prod_{i\in N_0} f(s_i^*|s_i, \theta).$$

Note that for $i \in N_0$ the outcome y_i is missing, so that only (s_i, θ) determines the conditional distribution of s_i^*. If $\theta_{-m} = \theta \setminus \{\theta_m\}$ is the collection of parameters, excluding θ_m, then sampling of θ_m for $m = 1, \ldots, M$ is based on

$$\begin{aligned} f(\theta_m|y, s, s^*, \theta_{-m}) &= f(\theta_m|y, s^*, \theta_{-m}) \\ &\propto \left[\prod_{i\in N_1} f(y_i, s_i^*|\theta)\right]\left[\prod_{i\in N_0} f(s_i^*|\theta)\right] f(\theta). \end{aligned}$$

The equality in the first line holds because s is a function of s^*. We can now formulate a Gibbs sampler for this model. Additional details and expressions for the parameters of the various posteriors are given in the Appendix.

Algorithm 1 (*Normal Selection Model*)**.** For given starting values of θ and s^*:

1. sample s_i^* for $i = 1, \ldots, n$ from

$$s_i^*|[y_i, \theta] \sim \mathcal{TN}_{(-\infty,0)}(x_{i1}'\beta_1, 1), \quad i \in N_0,$$

$$s_i^*|[y_i, \theta] \sim \mathcal{TN}_{(0,\infty)}\left(x_{i1}'\beta_1 + \frac{\sigma_{12}}{\xi^2 + \sigma_{12}^2} \times (y_i - x_{i2}'\beta_2), \frac{\xi^2}{\xi^2 + \sigma_{12}^2}\right), \quad i \in N_1,$$

2. sample β_1 from

$$\beta_1|[y, s^*, \beta_2, \sigma_{12}, \xi^2] \sim \mathcal{N}(\bar{b}_1, \bar{B}_1),$$

3. sample (β_2, σ_{12}) from

$$(\beta_2, \sigma_{12})|[y, s^*, \beta_1, \xi^2] \sim \mathcal{N}(\bar{g}, \bar{G}),$$

4. sample ξ^2 from

$$\xi^2|[y, s^*, \beta_1, \beta_2, \sigma_{12}] \sim \mathcal{IG}(\bar{c}, \bar{d}),$$

5. return to Step 1 and repeat.

Several remarks are in order. First, all conditional posteriors are standard distributions from which it is easy to generate a random draw. Running the algorithm for a large number of iterations yields a realization from a Markov chain. Under standard regularity conditions (Tierney, 1994) the stationary distribution of this chain is the joint posterior of θ and s^*. In practice one can monitor the simulated values of θ to assess convergence of the chain. For inference the first part of the simulated chain is often discarded as a so-called burn-in period, to ensure that the chosen starting values have a negligible impact. Also, when multiple chains are simulated from different starting values, certain diagnostic statistics can be calculated to monitor convergence (Gelman et al., 1995, Chapter 11).

Second, the algorithm given above only augments the data with the selection variable s_i^*. An alternative Gibbs sampler can be based on also simulating the missing outcomes. Instead of sampling s_i^* for $i \in N_0$, we could generate a random draw (s_i^*, y_i) from a bivariate normal distribution. This yields a balanced augmented sample. The subsequent steps for sampling θ in Algorithm 1 should then be modified by deriving conditional posteriors that also condition on the generated values $\{y_i : i \in N_0\}$. As noted by Chib et al. (2009), however, there are three advantages to not simulating unobserved outcomes. First, the computational and storage demands of Algorithm 1 are lower. Second, less data augmentation is likely to improve the mixing behavior of the Markov chain. Finally, Algorithm 1 is applicable when some or all of the covariates in x_{i2} are unobserved for $i \in N_0$. An algorithm based on full data augmentation would also require a model for x_{i2} and hence, a way to generate missing values.[8]

Third, it is often of interest to determine whether the covariance σ_{12} is zero or not. In the classical approach to inference this can be done via a t-test on the coefficient of the inverse Mills ratio (e.g. Wooldridge, 2002, Chapter 17). An alternative is to compute the Bayes factor. Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be the restricted $(\sigma_{12} = 0)$ and unrestricted models, respectively. If $\theta_j \in \Theta_j$ is the set of parameters in model $\mathcal{M}_j$ with prior distribution $f(\theta_j|\mathcal{M}_j)$, the Bayes factor is the ratio of marginal likelihoods:

$$B_{01} = \frac{f(y, s|\mathcal{M}_0)}{f(y, s|\mathcal{M}_1)},$$

$$f(y, s|\mathcal{M}_j) = \int_{\Theta_j} f(y, s|\theta_j, \mathcal{M}_j) f(\theta_j|\mathcal{M}_j) d\theta_j, \quad j = 0, 1.$$

Small values of B_{01} are evidence for the presence of a selection effect. Chib (1995) shows how to estimate the marginal likelihood from the output of the Gibbs sampler.[9] Note that Gibbs sampling in the restricted model is considerably easier than in Algorithm 1, because the likelihood in (3) is then separable in β_1 and (β_2, ξ^2). Provided that these parameters are independent in the prior, we can combine the Gibbs sampler of Albert and Chib (1993) for the Probit model with a standard Gibbs sampler for a normal linear model.

[5] We use the convention that $X \sim \mathcal{IG}(c_0, d_0)$ if and only if $X^{-1} \sim \mathcal{G}(c_0, d_0)$.

[6] In contrast, if σ_{12} has prior mean zero and is independent of ξ^2 in the prior $(\tau = 0)$, the induced prior on the correlation puts most of its probability near ± 1.

[7] This is certainly the case when u_i has a bivariate normal distribution. As we will show in Section 3, however, the Gibbs sampling algorithm can be modified to allow for non-normal likelihood functions.

[8] An earlier version of this paper used Gibbs sampling based on full data augmentation. I thank an anonymous referee for pointing out the disadvantage of this approach and referring me to Chib et al. (2009).

[9] For alternative methods and an extensive survey of Bayes factors, see Kass and Raftery (1995).

3. A Bayesian semiparametric model

3.1. Mixtures of normal distributions

The assumption of bivariate normality in (1) is convenient but often has no theoretical justification. Moreover, if the model is misspecified the posterior distribution may provide little information about the parameters of interest. Starting with Cosslett (1991), the classical literature on selection models has provided estimators that are consistent under much weaker, semiparametric restrictions on the distribution of u_i. In particular, the distribution of u_i does not have to belong to a parametric family. Suppose that u_i is independent of x_i, that u_{i2} has density f_{u_2}, and let F_{u_1} and $F_{u_1|u_2}$ be the marginal and conditional CDFs of u_{i1}, respectively. The likelihood for model (1) is then

$$f(y_i, s_i|x_i, \beta_1, \beta_2) = [F_{u_1}(-x_{i1}'\beta_1)]^{1-s_i}[f_{u_2}(y_i - x_{i2}'\beta_2) \times (1 - F_{u_1|u_2}(-x_{i1}'\beta_1|y_i - x_{i2}'\beta))]^{s_i}.$$

The method of inference in our paper – similar in spirit to the work of Gallant and Nychka (1987) and Van der Klaauw and Koning (2003) – is based on a likelihood function that is a mixture of normal distributions:

$$u_i|x_i \sim \sum_{j=1}^{k} \gamma_j \mathcal{N}_2(\mu(j), \Sigma(j)), \qquad \gamma_j \geq 0, \ \sum_{j=1}^{k} \gamma_j = 1. \tag{4}$$

Here $\mu(j)$ and $\Sigma(j)$ are the cluster-specific parameters and $\gamma = (\gamma_1, \ldots, \gamma_k)$ contains the mixing weights. Mixtures of normals form a very flexible family of distributions: even a small number k of components can generate distributions with skewness, excess kurtosis and multimodality. In some applications the mixture distribution has a clear structural interpretation. The clusters then correspond to heterogeneous groups within a larger population. Such an interpretation is absent here: the mixture distribution is used purely as a flexible modeling device.

3.2. Identification

In the absence of any information about the clusters, it is immediate that the distribution in (4) can at most be identified up to permutations of the cluster indices. For example, consider a two-component mixture where clusters 1 and 2 are characterized by $\{\gamma, \mu, \Sigma\}$ and $\{1 - \gamma, \tilde{\mu}, \tilde{\Sigma}\}$, respectively. An alternative model with parameters $\{1 - \gamma, \tilde{\mu}, \tilde{\Sigma}\}$ for cluster 1 and $\{\gamma, \mu, \Sigma\}$ for cluster 2 has the same likelihood function. Hence, without further restrictions the cluster labels are not identified. Teicher (1963) and Yakowitz and Spragins (1968) show that except for the cluster labels the parameters of finite Gaussian mixtures are identified. The results in these papers, however, do not account for the presence of covariates. In particular, they do not imply that Gaussian mixtures of regression models are identified.[10] For linear models Hennig (2000, Theorem 2.2) shows that identification – or the lack thereof – is determined by the richness of the support of x_i relative to the number of mixture components. In the selection model considered here the assumed independence between u_i and x_i has identifying power, as illustrated by the following example adapted from Hennig (2000).

Example. Consider a two-component location mixture with $\gamma_1 = \gamma_2 = \frac{1}{2}$. With probability γ_j $(j = 1, 2)$ the model is given by

$$s_i^* = \alpha_j + \beta_j D_i + \varepsilon_{i1},$$
$$y_i = \begin{cases} \delta_j + \varepsilon_{i2} & \text{if } s_i = 1 \\ \text{missing} & \text{if } s_i = 0, \end{cases}$$

where $D_i \in \{0, 1\}$ is binary, $(\varepsilon_{i1}, \varepsilon_{i2}) \sim \mathcal{N}_2(0, \Sigma)$ and Σ contains unit variances and correlation ρ. The unknown mixture parameters are $\theta = \{\alpha_j, \beta_j, \delta_j\}_{j=1}^2$. The likelihood of (s_i, y_i) is then

$$f(s_i, y_i|D_i, \theta) = \left[1 - \frac{1}{2}\Phi(\alpha_1 + \beta_1 D_i) - \frac{1}{2}\Phi(\alpha_2 + \beta_2 D_i)\right]^{1-s_i} \times \left[\frac{1}{2}\phi(y_i - \delta_1)g_1(y_i, D_i) + \frac{1}{2}\phi(y_i - \delta_2)g_2(y_i, D_i)\right]^{s_i},$$

$$g_j(y_i, D_i) = \Phi\left(\frac{\alpha_j + \beta_j D_i + \rho(y_i - \delta_j)}{\sqrt{1 - \rho^2}}\right), \quad j = 1, 2.$$

Consider a competing set of mixture parameters $\tilde{\theta}$, where $\tilde{\alpha}_1 = \alpha_1, \tilde{\alpha}_2 = \alpha_2, \tilde{\delta}_1 = \delta_2$ and $\tilde{\delta}_2 = \delta_1$. For observations with $D_i = 0$ both θ and $\tilde{\theta}$ generate the same likelihood. If (β_1, β_2) and $(\tilde{\beta}_1, \tilde{\beta}_2)$ are such that

$$\alpha_1 + \beta_1 = \alpha_2 + \tilde{\beta}_2,$$
$$\alpha_2 + \beta_2 = \alpha_1 + \tilde{\beta}_1,$$

then the same is true when $D_i = 1$. Hence, θ is not identified. Note also that $\tilde{\theta}$ is not obtained from θ by a relabeling of clusters. If D_i had a multinomial distribution instead, then θ would be identified. The same is true when β_j does not vary across j: in that case we cannot find a $\tilde{\theta}$ – apart from relabeling clusters – that generates the same likelihood for almost all y_i. Put differently, we can define $u_i = (\alpha_j + \varepsilon_{i1}, \delta_j + \varepsilon_{i2})$ with probability γ_j. Then u_i has a mixture distribution and is independent of D_i. □

In this example the independence between u_i and the covariate leads to identification, but this may not be true in general. Parameters in the two parts of the model, namely the selection and outcome equations, face distinct identification issues. First, in a semiparametric binary choice model where u_{i1} is independent of x_{i1}, the vector β_1 is at most identified up to scale. Moreover, without further restrictions on the distribution of u_{i1}, an intercept in β_1 is generally not identified; see Manski (1988) for a detailed discussion. Second, classical analyses of semiparametric identification in selection models have focused on the regression function for observed outcomes:

$$\mathrm{E}(y_i|x_i, s_i = 1) = x_{i2}'\beta_2 + g(x_{i1}'\beta_1),$$

where $g(x_{i1}'\beta_1) = \mathrm{E}(u_{i2}|u_{i1} \geq -x_{i1}'\beta_1)$ is the selection correction. If $g(\cdot)$ is unrestricted, an intercept in β_2 is not identified. Moreover, if $g(\cdot)$ is (close to) linear, the vector β_2 is not identified without an exclusion restriction.[11]

Although strictly speaking a semiparametric identification analysis does not apply to a mixture model, we believe that such an analysis is relevant for our purpose as well. Many distributions for the structural errors can be closely approximated by a normal mixture, provided the number of clusters is sufficiently large. If one or more of the aforementioned identification conditions is violated, e.g. the absence of an exclusion restriction in the outcome equation, we expect that some parameters in a mixture model may be poorly identified.

3.3. Prior and likelihood

Given a prior distribution for the parameters in (1) and (4), and with a fixed and known value of k, a suitable MCMC algorithm can

[10] Again, identified up to permutations of the cluster indices.

[11] A number of different necessary and sufficient conditions for identification are given in Cosslett (1991), Lee (1994) and Newey (2009).

often be constructed to approximate the posterior. See Früwirth-Schnatter (2006) for a thorough discussion and examples. An obvious complication is that in many cases k is not known. To infer the most likely number of mixture components, the econometrician could estimate the mixture model for various values of k and select a model on the basis of Bayes factors. In practice, however, such calculations can be quite complex and time-consuming.

We propose to estimate the selection model using a Dirichlet process (DP) prior, originally developed by Ferguson (1973) and Antoniak (1974). This prior introduces an additional layer of uncertainty into a Bayesian model, relative to the fixed prior used in Section 2. Detailed expositions of the DP prior are given in Escobar and West (1998) and MacEachern (1998). In our context the DP prior gives rise to an error distribution which is a mixture of normals with a random number of components. The resulting posterior distribution has the advantage that we can directly infer the likely number of mixture components, thereby circumventing the need to estimate many different models.[12]

In what follows, let $\vartheta_i = \{\mu_i, \sigma_{12i}, \xi_i^2\}$ and $\mu_i = (\mu_{i1}, \mu_{i2})'$. Our selection model based on the DP prior is given by (1) and

$$\begin{aligned} &u_i|x_i,\ \vartheta_i \sim \mathcal{N}_2(\mu_i, \Sigma_i),\\ &\vartheta_i|G \sim G,\\ &G|\alpha,\ G_0 \sim \mathcal{DP}(\alpha, G_0), \end{aligned} \tag{5}$$

where $\mathcal{DP}(\alpha, G_0)$ denotes the Dirichlet process with precision $\alpha > 0$ and continuous base measure G_0. The variance Σ_i is parameterized in terms of σ_{12i} and ξ_i^2, as in Eq. (2). Here the u_i's are potentially drawn from distinct normal distributions, and $\{\vartheta_i\}_{i=1}^n$ are i.i.d. draws from G. In addition G itself is treated as unknown and given a DP prior. Thus, G can be viewed as a random probability measure.[13] Recall from our earlier discussion that an intercept in (β_1, β_2) and the mean of u_i are not separately identifiable. In Section 2 the intercept was included in (β_1, β_2) and the mean of u_i was normalized to zero. Here, for reasons that will be explained shortly, the constants are excluded from (β_1, β_2) and absorbed into the mean of u_i.

A sample $\{\vartheta_i\}_{i=1}^n$ from a Dirichlet process can be generated as follows. First, sample a value ϑ_1 from the base distribution. Then, for $i > 1$:

$$\vartheta_i|\vartheta_1, \ldots, \vartheta_{i-1} \begin{cases} = \vartheta_j & \text{with prob. } \dfrac{1}{\alpha + i - 1},\ j = 1, \ldots, i-1,\\ \sim G_0 & \text{with prob. } \dfrac{\alpha}{\alpha + i - 1}. \end{cases} \tag{6}$$

That is, ϑ_i either equals an existing value or is a new value drawn from G_0. The full-sample likelihood is therefore a random mixture: the number of mixture components Z_n is equal to the number of unique values in $\{\vartheta_i\}_{i=1}^n$. Note that if $\mu_i = 0$ for all i, the mixture becomes a unimodal distribution, symmetric around zero. In order to also allow for multiple modes and asymmetry, we therefore do not restrict μ_i to be zero.

The representation in (6) highlights the role of the precision parameter and base measure. As α increases, Z_n is likely to be large and $\{\vartheta_i\}_{i=1}^n$ will start to resemble an i.i.d. sample from G_0. On the other hand, as α decreases, Z_n tends to be small; in the limit as α approaches zero, all ϑ_i's will be equal, $Z_n = 1$ and the u_i's all follow the same normal distribution. Antoniak (1974) shows that

$$\Pr\{Z_n = k|\alpha\} = |s(n,k)|\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad k = 1, 2, \ldots, n,$$

where Γ is the gamma function and $s(n,k)$ is a Stirling number of the first kind. For any fixed value of α this distribution is easily calculated.

The choice of a prior for α is related to identification. If the prior places a large probability on large values of α, there is a risk of overfitting. In this case, for example, the absence of an exclusion restriction may lead to identification problems for β_2. This problem can be 'solved' by introducing identifying information via the prior. A prior that concentrates its mass around small values of α would lead with high probability to an error distribution that contains fewer mixture components. In the limit as $\alpha \to 0$ the model becomes identified (though, of course, it may then be misspecified).

It remains to specify G_0, and the prior of $(\alpha, \beta_1, \beta_2)$. Excluding the intercept from β_1 and β_2, we use (as before) $\beta_1 \sim \mathcal{N}(b_1, B_1)$, $\beta_2 \sim \mathcal{N}(b_2, B_2)$. For the base measure G_0 we take μ_i to be independent of (σ_{12i}, ξ_i^2), and use $\mu_i \sim \mathcal{N}(0, M)$, $\sigma_{12i}|\xi_i^2 \sim \mathcal{N}(0, \tau\xi_i^2)$ and $\xi_i^2 \sim \mathcal{IG}(c_0, d_0)$. This is the same prior as in Section 2. The DP selection model therefore reduces to the normal model as α goes to zero. Finally, by placing a prior distribution on α it is possible to update beliefs about the distribution of Z_n. For computational convenience we follow Escobar and West (1995) and use $\alpha \sim \mathcal{G}(c_1, c_2)$, independent of (β_1, β_2).[14]

3.4. Posterior

Define $\theta = \{\alpha, \beta_1, \beta_2\}$, $\vartheta = \{\vartheta_i\}_{i=1}^n$, $\vartheta_{-i} = \{\vartheta_j\}_{j \neq i}$ and let k be the number of distinct values in ϑ. The set of distinct values in ϑ is indexed by j and denoted by $\vartheta^* = \{\vartheta_j^*\}_{j=1}^k$. The vector of cluster indicators is $\zeta = (\zeta_1, \ldots, \zeta_n)$, where $\zeta_i = j$ if and only if $\vartheta_i = \vartheta_j^*$. Finally, n_{ζ_i} is the number of observations belonging to cluster ζ_i. The strategy for approximating the posterior is the same as before: the data is augmented with s^*, and the model parameters are sampled consecutively from their conditional posterior distributions. This leads to the following MCMC algorithm. Additional details and formulas for the posterior parameters are given in the Appendix.

Algorithm 2 (*Semiparametric Selection Model*)**.** For given starting values of θ, ϑ and s^*:

1. sample s_i^* for $i = 1, \ldots, n$ from
$$s_i^*|[y_i, \theta, \vartheta] \sim \mathcal{TN}_{(-\infty,0)}(x_{i1}'\beta_1 + \mu_{i1}, 1), \quad i \in N_0,$$
$$s_i^*|[y_i, \theta, \vartheta] \sim \mathcal{TN}_{(0,\infty)}\left(x_{i1}'\beta_1 + \mu_{i1} + \frac{\sigma_{12i}}{\xi_i^2 + \sigma_{12i}^2} \times (y_i - x_{i2}'\beta_2 - \mu_{i2}), \frac{\xi_i^2}{\xi_i^2 + \sigma_{12i}^2}\right), \quad i \in N_1,$$
2. sample β_1 from
$$\beta_1|[y, s^*, \alpha, \beta_2, \vartheta] \sim \mathcal{N}(\bar{b}_1, \bar{B}_1),$$
3. sample β_2 from
$$\beta_2|[y, s^*, \alpha, \beta_1, \vartheta] \sim \mathcal{N}(\bar{b}_2, \bar{B}_2),$$
4. sample an auxiliary variable $\eta \sim Beta(\alpha + 1, n)$ and sample α from the mixture distribution
$$\alpha|k \sim p_\eta \mathcal{G}(c_1 + k, c_2 - \log\eta) + (1 - p_\eta)\mathcal{G}(c_1 + k - 1, c_2 - \log\eta),$$
where p_η is the mixing probability,

[12] Richardson and Green (1997) discuss an alternative to the Dirichlet process prior. In their reversible-jump MCMC algorithm mixture components can merge or be split into two. Implementing this method is beyond the scope of the current paper.

[13] 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.

[14] For alternatives, see Escobar (1994) and Conley et al. (2008).

5. for $i \in N_1$:
   (a) if $n_{\zeta_i} > 1$ sample an auxiliary value $\tilde{\vartheta} \sim G_0$. Sample ϑ_i according to
   $$\vartheta_i = \begin{cases} \vartheta_j & \text{with prob. } \dfrac{C}{\alpha+n-1} f(s_i^*, y_i|\theta, \vartheta_j), \quad j \neq i, \\ \tilde{\vartheta} & \text{with prob. } \dfrac{C}{\alpha+n-1} \alpha f(s_i^*, y_i|\theta, \tilde{\vartheta}), \end{cases}$$
   where C is a normalizing constant,
   (b) if $n_{\zeta_i} = 1$, sample ϑ_i according to
   $$\vartheta_i = \begin{cases} \vartheta_j & \text{with prob. } \dfrac{C}{\alpha+n-1} f(s_i^*, y_i|\theta, \vartheta_j), \ j \neq i, \\ \text{unchanged} & \text{with prob. } \dfrac{C}{\alpha+n-1} \alpha f(s_i^*, y_i|\theta, \vartheta_i). \end{cases}$$
   For $i \in N_0$: sample ϑ_i according to (a) and (b), with the bivariate likelihoods replaced by the univariate ones of s_i^*,
6. for $j = 1, \ldots, k$: either sample the entire vector ϑ_j^* from $f(\vartheta_j^*|y, s^*, \theta, \zeta)$ in one step, or blocks of ϑ_j^* from their conditional posterior,
7. return to Step 1 and repeat.

Regarding Algorithm 2 we remark the following. First, Steps 1–3 are largely similar to those in Algorithm 1, except that observation-specific parameters appear in Step 1, and the posterior means and variances of β_1 and β_2 are different (formulas are given in Appendix). In Step 4 we update α, and hence the probability distribution of the number of mixture components. Since G_0 is not a natural conjugate distribution, we cannot directly sample ϑ_i from its conditional posterior. Algorithm 8 of Neal (2000) is therefore used to update the collection $\{\vartheta_i\}_{i=1}^n$ in Step 5. In Step 6 the set of unique parameter values ϑ^* that determine the clusters are resampled, or 'remixed'. Strictly speaking this is not necessary, but a Markov chain without remixing may converge very slowly. The reason for this is that clusters of 'similar' observations can get stuck at a fixed value of ϑ_j^*, when ϑ is only updated one observation at a time. Remixing allows an entire cluster of observations to change parameters at once, which leads to a better exploration of the posterior of ϑ and ϑ^*.

The Dirichlet process selection model can accommodate departures from normality, since Algorithm 2 provides information about the number of components in the mixture distribution of u_i. At each iteration we can determine and store the number k of unique elements in ϑ. The sampled values of k provide an approximate sample from the posterior. The ratio of the posterior to prior probability of $k = 1$ then quantifies any evidence of non-normality.

The Dirichlet process selection model can also be used to approximate the posterior predictive distribution of u_i. This is the Bayesian analogue of a classical density estimate. At iteration $t \in \{1, \ldots, T\}$ of the Markov chain, and given the current state $\{\vartheta_{i,t}\}_{i=1}^n$, generate an out-of-sample value $\vartheta_{n+1,t}$ according to the procedure in (6):

$$\vartheta_{n+1,t} \begin{cases} = \vartheta_{i,t} & \text{with prob. } \dfrac{1}{\alpha+n}, \ i = 1, \ldots, n \\ \sim G_0 & \text{with prob. } \dfrac{\alpha}{\alpha+n}. \end{cases}$$

An estimate of the posterior predictive distribution is then

$$\hat{f}(u_1, u_2|y, s) = T^{-1} \sum_{t=1}^{T} f(u_1, u_2|\vartheta_{n+1,t}),$$

which can be calculated on a two-dimensional grid.

Finally, it is possible to make inference about the degree of dependence between u_{i1} and u_{i2}. Dependence implies that selection into the sample is related to the outcome, even after controlling for observables. In Algorithm 2 at iteration t we can add the step of generating a pseudo-sample $\{u_{i,t}\}_{i=1}^n$, where $u_{i,t} \sim \mathcal{N}(\mu_{i,t}, \Sigma_{i,t})$, and calculating a dependence measure.[15] The T realizations of this measure are then an approximate sample from its posterior distribution.

4. Simulation evidence

We now estimate a semiparametric selection model using Algorithm 2 and some simulated data. A sample of size $n = 1000$ is generated from (1) with

$$s_i^* = 2 - x_{i1,1} + x_{i1,2} + u_{i1},$$
$$y_i = 1 + 0.5x_{i2,1} - 0.5x_{i2,2} + u_{i2},$$

where $x_{i1,1} \sim \mathcal{N}(0, 3)$, $x_{i1,2} \sim \mathcal{U}(-3, 3)$, $x_{i2,1} \sim \mathcal{N}(0, 3)$ and $x_{i2,2} = x_{i1,2}$. Let $\hat{p}_0$ be the fraction of the sample in which y_i is missing. We consider two distributions for u_i. The first is a bivariate normal with mean zero, $\sigma_{12} = 0.5$ and $\xi^2 = 0.75$. Hence, the correlation is 0.5. The second distribution is a location-mixture of two normals:

$$u_i|x_i \sim \gamma \mathcal{N}(\mu_1, \Sigma) + (1-\gamma)\mathcal{N}(\mu_2, \Sigma), \qquad \Sigma = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$$

where $\gamma = 0.3$, $\mu_1 = (0, -2.1)'$ and $\mu_2 = (0, 0.9)'$. The correlation between u_{i1} and u_{i2} in this case is again 0.5. The base measure G_0 we use is defined by $\xi_i^2 \sim \mathcal{IG}(3, 6)$, $\sigma_{12i}|\xi_i^2 \sim \mathcal{N}(0, 0.5\xi_i^2)$, $\mu_i \sim \mathcal{N}(0, 10I_2)$. The priors for (β_1, β_2) – excluding the intercept – and α are $\beta_1 \sim \mathcal{N}(0, 10I_2)$, $\beta_2 \sim \mathcal{N}(0, 10I_2)$ and $\alpha \sim \mathcal{G}(2, 2)$. The implied prior of the number of mixture components is given in Figs. 1 and 2: it is fairly diffuse between 1 and 10 mixture components. For values larger than 10 the prior probability rapidly approaches zero. We are therefore confident that we are not overfitting, while at the same time allowing for significant departures from normality ($k = 1$). For each data set Algorithm 2 is run for 20,000 iterations. The first 2500 draws are discarded as a burn in period. In addition, the algorithm is started from three sets of initial values. The total approximate sample from the posterior thus contains 52,500 values.

Results for the normal sample ($\hat{p}_0 = 0.23$) are given in Table 1. The values of $\beta_{2,1}$, $\beta_{2,2}$ and ρ used to generate the data all lie well within the 95% highest posterior density (HPD) interval.[16] The Gelman–Rubin statistic suggests that the Markov chain is mixing well. In particular, the autocorrelation function of $\beta_{2,1}$, $\beta_{2,2}$ decreases rapidly, leading to a low value of the inefficiency factor.[17] The simulated draws of the correlation coefficient ρ display more autocorrelation, but its posterior offers strong evidence that u_{i1} and u_{i2} are dependent. The prior and posterior distributions of k, the number of mixture components, are plotted in Fig. 1. The posterior probability of a single normal component is more than 70%, and the ratio of the posterior to prior probability that k equals one is 15.7. The density contours of the posterior predictive distribution of u_i are given in the same figure. The plot suggests that the distribution of u_i is unimodal and elliptical. Note also that the contours are centered around the point (2, 1), which is the intercept in our simulation design.

Results for the mixture data ($\hat{p}_0 = 24.1\%$) are given in Table 2. The posterior standard deviations of $\beta_{2,1}$ and $\beta_{2,2}$ are slightly

[15] In the bivariate normal model dependence is entirely captured by the correlation coefficient. Of course, in a mixture model zero correlation does not imply independence. It may then be useful to report alternative measures as well, such as Kendall's tau or Spearman's rho.

[16] This interval is calculated as the shortest interval that contains 95% of the sampled values. See Chen et al. (2000, Chapter 7).

[17] See Chib et al. (2009).

Table 1
Posterior summary, normal DGP.

Parameter	Mean	Std. dev.	95% HPD	GR[a]	AC(1)[b]	AC(2)	AC(3)	IF[c]
$\beta_{2,1}(0.5)$	0.5184	0.0197	[0.4789, 0.5560]	1.0000	0.2342	0.0517	0.0202	1.4130
$\beta_{2,2}(-0.5)$	−0.4945	0.0240	[−0.5415, −0.4474]	1.0000	0.4234	0.2269	0.1714	4.2766
$\rho(0.5)$	0.5016	0.0994	[0.3053, 0.6865]	1.0001	0.8651	0.8082	0.7551	17.5626

[a] Gelman–Rubin statistic.
[b] Autocorrelation.
[c] Inefficiency factor.

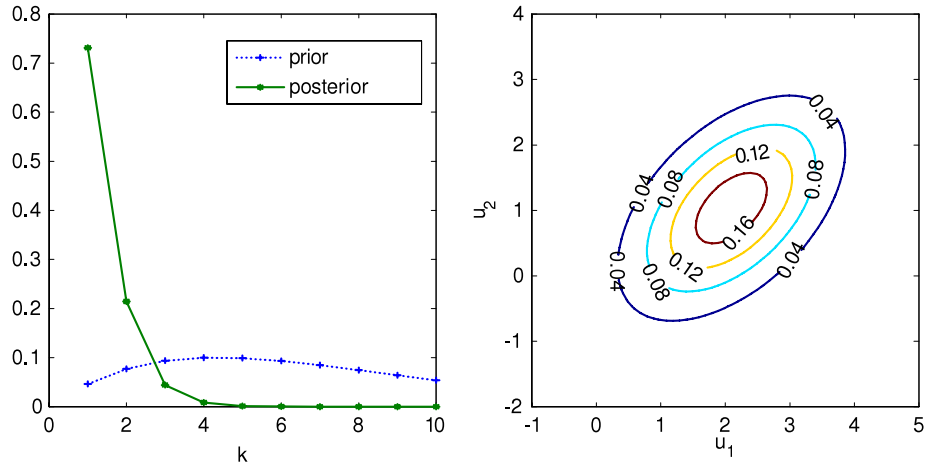

Fig. 1. Posterior of k and predictive distribution of u_i; normal DGP.

Table 2
Posterior summary, mixture DGP.

Parameter	Mean	Std. dev.	95% HPD	GR[a]	AC(1)[b]	AC(2)	AC(3)	IF[c]
$\beta_{2,1}(0.5)$	0.4778	0.0279	[0.4225, 0.5315]	1.0001	0.5923	0.3627	0.2336	4.2189
$\beta_{2,2}(-0.5)$	−0.5345	0.0281	[−0.5895, −0.4796]	1.0000	0.5928	0.3651	0.2381	4.3225
$\rho(0.5)$	0.4101	0.0959	[0.2234, 0.5964]	1.0018	0.8456	0.7905	0.7400	23.2745

[a] Gelman–Rubin statistic.
[b] Autocorrelation.
[c] Inefficiency factor.

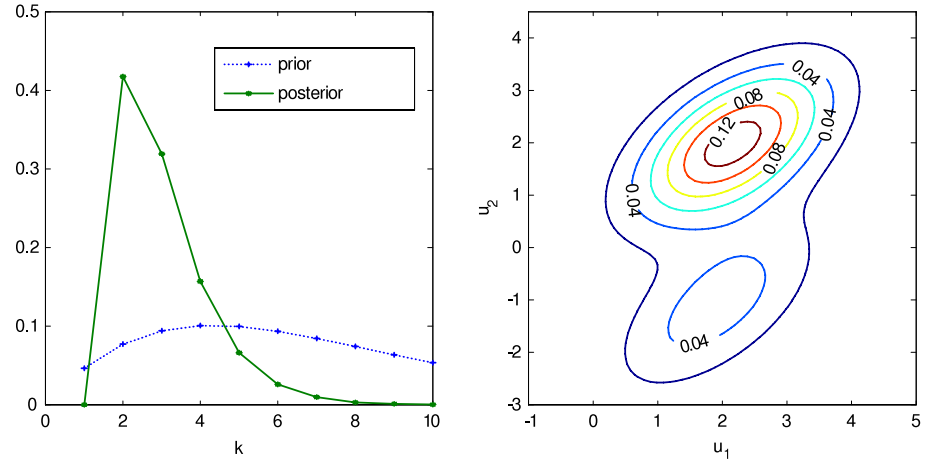

Fig. 2. Posterior of k and predictive distribution of u_i; mixture DGP.

larger compared to the normal case, but again the true parameter values lie well within the 95% HPD. From Fig. 2 it is clear that the Dirichlet model overwhelmingly fits two (42%) or three (32%) mixture components to the data, whereas the posterior probability of normality ($k = 1$) is effectively zero. The contour plot also reveals bimodality in the posterior predictive distribution of u_i.

5. Empirical application

The RAND Health Insurance Experiment (RHIE), conducted in the period 1974–1982, was a large scale experimental study of health care costs and utilization. Individuals were randomly assigned to different health insurance plans, with the goal of determining the causal effect of insurance characteristics on

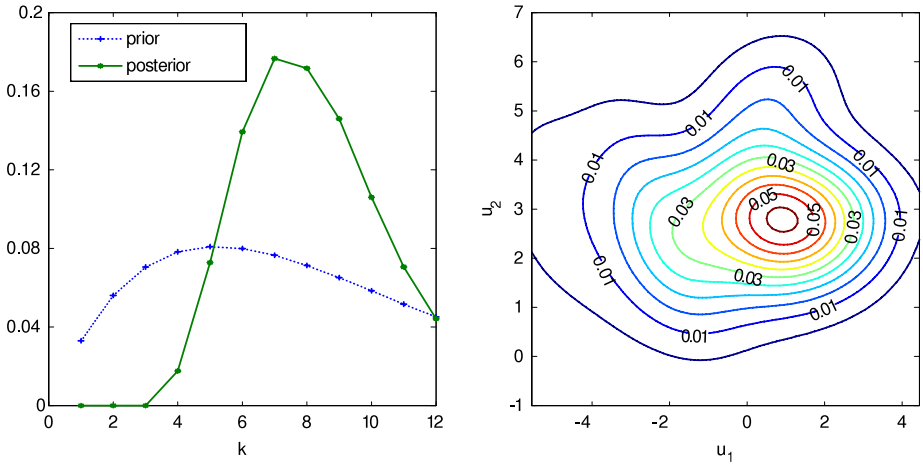

Fig. 3. Posterior of k and predictive distribution of u_i, RHIE data.

Table 3
RHIE data, normal model.

Variable	Mean	Std dev	95% HPD	GR	AC(1)	AC(2)	AC(3)	IF
LC	−0.0733	0.0335	[−0.1378, −0.0075]	1.0000	0.4684	0.2080	0.0913	2.0923
IDP	−0.1468	0.0656	[−0.2761, −0.0197]	1.0001	0.4563	0.2033	0.0844	1.9299
LPI	0.0144	0.0104	[−0.0059, 0.0347]	1.0003	0.4704	0.2141	0.1020	1.9706
FMDE	−0.0240	0.0193	[−0.0624, 0.0136]	1.0000	0.4709	0.2094	0.0939	1.9627
PHYSLIM	0.3519	0.0755	[0.2014, 0.4990]	1.0000	0.4962	0.2505	0.1204	2.2112
NDISEASE	0.0283	0.0038	[0.0207, 0.0355]	1.0000	0.4947	0.2545	0.1279	2.2239
HLTHG	0.1559	0.0519	[0.0540, 0.2572]	1.0001	0.4657	0.2004	0.0824	1.9401
HLTHF	0.4418	0.0958	[0.2562, 0.6322]	1.0001	0.4797	0.2144	0.0861	1.9771
HLTHP	0.9887	0.1880	[0.6261, 1.3651]	1.0000	0.5039	0.2538	0.1218	2.2310
ρ	0.7120	0.0390	[0.6342, 0.7840]	1.0006	0.9592	0.9230	0.8885	34.5570

patients' use of health care services. The data we use is the same as in Deb and Trivedi (2002), except that we restrict our analysis to the second year of data. This yields a cross section of 5574 individuals. Medical expenditures are zero for 23.3% of the sample, and a selection model is estimated for the logarithm of medical expenditures. The list of covariates we use, together with the variable definitions, are given in Deb and Trivedi (2002, Table 1). Since it is not immediately clear that there is a variable that affects selection, but not the level of expenditures, we do not impose an exclusion restriction. Hence, the vectors x_{i1} and x_{i2} in (1) are the same.

We estimate the normal and Dirichlet mixture models and report results for selected coefficients of the outcome equation below. In particular, we focus on the effects of insurance variables (LC, IDP, LPI, FMDE) and health status variables (PHYSLIM, NDISEASE, HLTHG, HLTHF, HLTP) on the logarithm of medical expenditures, and on the correlation between u_{i1} and u_{i2}. The priors used are the same as in the previous section.[18]

Fig. 3 shows that there is clear evidence of nonnormality: the posterior probability of normal errors is effectively zero, whereas around 35% of the time the error distribution is a mixture of 7 or 8 components.[19] Comparing Tables 3 and 4 we see that the posterior standard deviations of the coefficients – with the exception of NDISEASE – are smaller when the assumption of normal errors is relaxed. This results in narrower 95% HPD intervals. Conley et al. (2008) report a similar finding in an instrumental variables model. For some coefficients, most notably that of HLTHP, the posterior mean is substantially different as well. Finally, we note that the normal model suggests the presence of a selection effect: the 95% HPD of ρ is [0.6342, 0.7840]. In the model with the Dirichlet process prior on the other hand, there is no strong evidence of a nonzero correlation between u_{i1} and u_{i2}. The large posterior standard deviation and inefficiency factor of ρ suggest that the posterior uncertainty about ρ is similar to the prior uncertainty.

6. Conclusion

In this paper we have developed Gibbs sampling algorithms that enable a Bayesian analysis of a model with sample selectivity. Such a model essentially consists of a latent structure, which is only partially observed. This paper has treated a model with exogenous covariates and a binary selection rule. The methods developed here, however, can be readily adapted to accommodate endogeneity and more complicated selection rules.

If the distribution of the unobserved heterogeneity is assumed to be normal, Gibbs sampling is straightforward. Without this assumption, a more flexible semiparametric Bayesian model can be based on the Dirichlet process prior. This prior implies that the error distribution is a mixture with an unknown number of components. In particular, we use mixtures of normal distributions, as a natural extension of the simple normal model. This paper develops a Gibbs sampling algorithm for the Dirichlet model that does not require the use of natural conjugate distributions, or augmenting the data with the missing outcomes.

The use of the Dirichlet process prior in a Bayesian model of sample selection is appealing for two main reasons. First, the unobserved heterogeneity follows a mixture distribution with a random number of components. From Bayes' rule we can make inference about the likely number of mixture components. The posterior distribution of this number can then be used to detect departures from the parametric normal model. Second, the posterior provides information about the potential dependence, after controlling for observables, between the selection mechanism and outcome process.

[18] The prior for the normal model is the base measure G_0 of Section 4.

[19] We experimented with different values of c_1 and c_2, the parameters of the gamma prior of α. The results were very similar and are therefore not reported here.

Table 4
RHIE data, Dirichlet mixture model.

Variable	Mean	Std dev	95% HPD	GR	AC(1)	AC(2)	AC(3)	IF
LC	−0.049	0.0320	[−0.1108, 0.0149]	1.0031	0.547	0.3855	0.3135	23.871
IDP	−0.0941	0.0601	[−0.2106, 0.0251]	1.0028	0.4872	0.3128	0.2367	7.1989
LPI	0.0052	0.0096	[−0.0136, 0.0242]	1.005	0.5038	0.3374	0.2571	8.0127
FMDE	−0.0196	0.0171	[−0.0530, 0.0138]	1.0001	0.4754	0.2946	0.2164	4.8717
PHYSLIM	0.2519	0.0683	[0.1181, 0.3849]	1.0038	0.4973	0.3244	0.2507	11.9877
NDISEASE	0.0217	0.0038	[0.0145, 0.0294]	1.0142	0.5698	0.4182	0.3545	73.814
HLTHG	0.1223	0.0451	[0.0327, 0.2094]	1.0003	0.4546	0.2789	0.195	4.3471
HLTHF	0.402	0.0857	[0.2287, 0.5656]	1.0002	0.4879	0.3136	0.2408	7.4734
HLTHP	0.5985	0.1725	[0.2615, 0.9374]	1.0008	0.5243	0.3596	0.2772	19.0882
ρ	0.0493	0.2182	[−0.3509, 0.4761]	1.0544	0.8641	0.8288	0.8042	313.1993

We have illustrated the use of the Dirichlet process prior with some simulated data. In these cases the posterior distribution assigns a high probability to the number of mixture components in the true data generating process. We have also estimated two models for individual medical expenditures, using a subset of the RHIE data. In the bivariate normal model there is evidence for the presence of a selection effect. The correlation coefficient in the distribution of the unobserved heterogeneity is positive with large posterior probability. The model based on the Dirichlet process prior, however, finds substantial evidence for nonnormality. Relaxing the assumption of normality results in smaller posterior standard deviations and narrower 95% HPD intervals for most parameters. Moreover, the posterior distribution of the correlation coefficient now has its probability mass centered around zero. This does not imply that there is no selection effect. Rather, in highly nonnormal distributions there may be forms of dependence (e.g., tail dependence) that are not easily detected by a simple correlation.

Acknowledgments

I am heavily indebted to two anonymous referees, whose comments and suggestions lead to a substantially different and improved paper. I have benefited from helpful discussions with William McCausland, Chris Bollinger and Youngki Shin. Frank Kleibergen and Tony Lancaster provided support and encouragement during the earlier stages of this work. I am, of course, solely responsible for any remaining errors.

Appendix. Gibbs sampling in the normal model

We provide some additional details about Algorithm 1. With regard to the selection variable:

$$s_i^*|[y_i,\theta] \sim \begin{cases} \mathcal{N}(x_{i1}'\beta_1,\,1), & i\in N_0 \\ \mathcal{N}\left(x_{i1}'\beta_1+\dfrac{\sigma_{12}}{\xi^2+\sigma_{12}^2}(y_i-x_{i2}'\beta_2),\,\dfrac{\xi^2}{\xi^2+\sigma_{12}^2}\right), & i\in N_1. \end{cases} \tag{7}$$

Conditioning on s_i yields the distributions in Step 1 of the algorithm. Next, consider β_1. Since β_1 is a priori independent of $(\beta_2,\sigma_{12},\xi^2)$, its posterior satisfies

$$f(\beta_1|y,s^*,\beta_2,\sigma_{12},\xi^2) \propto f(s^*|y,\theta)f(\beta_1) \propto \left[\prod_{i\in N_0} f(s_i^*|\theta)\right]\left[\prod_{i\in N_1} f(s_i^*|y_i,\theta)\right]f(\beta_1).$$

The two components of the likelihood are given in (7). With a normal prior distribution for β_1, the posterior is a $\mathcal{N}(\bar b_1,\bar B_1)$ distribution with

$$\bar B_1 = \left(B_1^{-1}+\sum_{i\in N_0}x_{i1}x_{i1}'+\left[\frac{\xi^2}{\xi^2+\sigma_{12}^2}\right]^{-1}\sum_{i\in N_1}x_{i1}x_{i1}'\right)^{-1},$$

$$\bar b_1 = \bar B_1\left(B_1^{-1}b_1+\sum_{i\in N_0}x_{i1}x_{i1}'\hat\beta_1(0)+\left[\frac{\xi^2}{\xi^2+\sigma_{12}^2}\right]^{-1}\sum_{i\in N_1}x_{i1}x_{i1}'\hat\beta_1(1)\right),$$

and $\hat\beta_1(0)$, $\hat\beta_1(1)$ are two least squares estimators:

$$\hat\beta_1(0)=\left[\sum_{i\in N_0}x_{i1}x_{i1}'\right]^{-1}\sum_{i\in N_0}x_{i1}s_i^*,$$

$$\hat\beta_1(1)=\left[\sum_{i\in N_1}x_{i1}x_{i1}'\right]^{-1}\sum_{i\in N_1}x_{i1}\left(s_i^*-\frac{\sigma_{12}}{\xi^2+\sigma_{12}^2}(y_i-x_{i2}'\beta_2)\right).$$

For the conditional posteriors of (β_2,σ_{12}) and ξ^2, note that only the observations with $i\in N_1$ are informative about these parameters. From Bayes' rule and the prior independence between β_1 and the remaining parameters, we get

$$f(\beta_2,\sigma_{12},\xi^2|y,s^*,\beta_1) = f\left(\beta_2,\sigma_{12},\xi^2|\{y_i,s_i^*\}_{i\in N_1},\beta_1\right) \propto \left[\prod_{i\in N_1}f(y_i|s_i^*,\theta)\right]f(\beta_2,\sigma_{12},\xi^2).$$

The likelihood in the previous display follows from the fact that for $i\in N_1$:

$$y_i = x_{i2}'\beta_2+\sigma_{12}(s_i^*-x_{i1}'\beta_1)+\eta_i,$$

is a normal linear regression model with $\eta_i|x_i\sim\mathcal{N}(0,\xi^2)$. Given the natural conjugate priors for (β_2,σ_{12}) and ξ^2, their conditional posteriors are completely standard. Let W be the $n_1\times(k_2+1)$ matrix with rows $\{x_{i2}',s_i^*-x_{i1}'\beta_1\}_{i\in N_1}$ and

$$g_0=\begin{pmatrix}b_2\\0\end{pmatrix},\qquad G_0=\begin{bmatrix}B_2&0\\0&\tau\xi^2\end{bmatrix}.$$

The conditional posterior of (β_2,σ_{12}) is then given in Step 3 of the algorithm, where

$$\bar G=(G_0^{-1}+\xi^{-2}W'W)^{-1},\qquad \bar g=\bar G(G_0^{-1}g_0+\xi^{-2}W'W\hat g),$$

and $\hat g = (W'W)^{-1}W'y^*$ is the least squares estimator. Finally, given the prior $\xi^2\sim\mathcal{IG}(c_0,d_0)$, the posterior of ξ^2 is an inverse-gamma distribution with $\bar c=c_0+(1+\sum_{i=1}^n s_i)/2$ and

$$\bar d = d_0+\frac{\sigma_{12}^2}{2\tau}+\frac{1}{2}\sum_{i\in N_1}(y_i-x_{i2}'\beta_2-\sigma_{12}(s_i^*-x_{i1}'\beta_1))^2.$$

Gibbs sampling in the Dirichlet mixture model

Updating (α,β_1,β_2)

Here we provide additional details and the posterior parameters for Algorithm 2. For conciseness we will write u_{i1} instead of $(s_i^*-x_{i1}'\beta_1)$, and u_{i2} instead of $(y_i-x_{i2}'\beta_2)$ for $i\in N_1$. The distributions in Step 1 follow directly from the fact that $u_i|\vartheta_i\sim\mathcal{N}_2(\mu_i,\Sigma_i)$. The conditional posterior of β_1 can be found through

$$f(\beta_1|y, s^*, \alpha, \beta_2, \vartheta) \propto f(\beta_1) \times \prod_{i\in N_0} f(s_i^*|\theta, \vartheta) \times \prod_{i\in N_1} f(s_i^*|y_i, \theta, \vartheta).$$

Given the normal prior for β_1, the posterior is $\mathcal{N}(\bar{b}_1, \bar{B}_1)$, where

$$\bar{B}_1 = \left(B_1^{-1} + \sum_{i\in N_0} x_{i1}x_{i1}' + \sum_{i\in N_1} c_i^{-2}x_{i1}x_{i1}'\right)^{-1},$$

$$\bar{b}_1 = \bar{B}_1\left(B_1^{-1}b_1 + \sum_{i\in N_0} x_{i1}x_{i1}'\hat{\beta}_1(0) + \sum_{i\in N_1} c_i^{-2}x_{i1}x_{i1}'\hat{\beta}_1(1)\right),$$

$$\hat{\beta}_1(0) = \left(\sum_{i\in N_0} x_{i1}x_{i1}'\right)^{-1}\sum_{i\in N_0} x_{i1}(s_i^* - \mu_{i1}),$$

$$\hat{\beta}_1(1) = \left(\sum_{i\in N_1} c_i^{-2}x_{i1}x_{i1}'\right)^{-1}\sum_{i\in N_1} c_i^{-2}x_{i1} \times \left(s_i^* - \mu_{i1} - \frac{\sigma_{12i}}{\xi_i^2 + \sigma_{12i}^2}(u_{i2} - \mu_{i2})\right),$$

and $c_i^2 = \xi_i^2/(\xi_i^2 + \sigma_{12i}^2)$ is the conditional variance of s_i^* given y_i, when $i \in N_1$. The conditional posterior of β_2 follows from

$$f(\beta_2|y, s^*, \alpha, \beta_1, \vartheta) \propto f(\beta_2)\prod_{i\in N_1} f(y_i|s_i^*, \theta, \vartheta),$$

$$y_i|s_i^*, \theta, \vartheta \sim \mathcal{N}\left(x_{i2}'\beta_2 + \mu_{i2} + \sigma_{12i}(u_{i1} - \mu_{i1}), \xi_i^2\right).$$

Given the normal prior, some tedious algebra now yields the normal distribution in Step 3, with

$$\bar{B}_2 = \left(B_2^{-1} + \sum_{i\in N_1} \xi_i^{-2}x_{i2}x_{i2}'\right)^{-1},$$

$$\bar{b}_2 = \bar{B}_2\left(B_2^{-1}b_2 + \sum_{i\in N_1} \xi_i^{-2}x_{i2}x_{i2}'\hat{\beta}_2\right),$$

$$\hat{\beta}_2 = \left(\sum_{i\in N_1} \xi_i^{-2}x_{i2}x_{i2}'\right)^{-1}\sum_{i\in N_1} \xi_i^{-2}x_{i2}(y_i - \mu_{i2} - \sigma_{12i}(u_{i1} - \mu_{i1})).$$

Next, consider Step 4 in the algorithm. Conditioning on (β_1, β_2) and ϑ, the distributions of s_i^* (when $i \in N_0$) and (s_i^*, y_i) (when $i \in N_1$) do not depend on α. Also, from (6) it is clear that α determines the number of unique values k in ϑ, but not the actual values. It then follows that

$$\begin{aligned} f(\alpha|y, s^*, \beta_1, \beta_2, \vartheta) &= f(\alpha|k) \\ &\propto f(k|\alpha)f(\alpha) \\ &\propto \frac{\alpha^k\Gamma(\alpha)}{\Gamma(\alpha+n)}f(\alpha). \end{aligned}$$

From the identity $\int_0^1 x^{a-1}(1-x)^{b-1}dx = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, the posterior can be written as

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

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

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

From this it is clear that $\eta|\alpha, k \sim Beta(\alpha+1, n)$. With a $\mathcal{G}(c_1, c_2)$ prior for α, we find

$$f(\alpha|\eta, k) \propto \alpha^{c_1+k-1}e^{-\alpha(c_2-\log\eta)} + n\alpha^{c_1+k-2}e^{-\alpha(c_2-\log\eta)}.$$

This is a mixture of the $\mathcal{G}(c_1+k, c_2-\log\eta)$ and $\mathcal{G}(c_1+k-1, c_2-\log\eta)$ distributions. The mixing proportion p_η can be solved from the relation

$$\frac{p_\eta}{1-p_\eta} = \frac{c_1+k-1}{n(c_2-\log\eta)}.$$

Updating ϑ and remixing

From (6) we write the prior of ϑ_i given ϑ_{-i} as

$$f(\vartheta_i|\vartheta_{-i}) = \frac{\alpha}{\alpha+n-1}dG_0(\vartheta_i) + \sum_{j\neq i}\frac{1}{\alpha+n-1}\delta_{\vartheta_j}(\vartheta_i),$$

where $\delta_{\vartheta_j}(\cdot)$ is the measure with unit mass at ϑ_j. Then for $i \in N_0$:

$$\begin{aligned} f(\vartheta_i|y, s^*, \theta, \vartheta_{-i}) &\propto \frac{\alpha}{\alpha+n-1}dG_0(\vartheta_i)f(s_i^*|\theta, \vartheta_i) \\ &\quad + \sum_{j\neq i}\frac{1}{\alpha+n-1}\delta_{\vartheta_j}(\vartheta_i)f(s_i^*|\theta, \vartheta_i) \\ &\propto \alpha f(s_i^*|\theta, \vartheta_i)dG_0(\vartheta_i) + \sum_{j\neq i}\delta_{\vartheta_j}(\vartheta_i)f(s_i^*|\theta, \vartheta_j) \\ &\propto \alpha f(s_i^*|\theta)f(\vartheta_i|s_i^*, \theta) + \sum_{j\neq i}\delta_{\vartheta_j}(\vartheta_i)f(s_i^*|\theta, \vartheta_j). \end{aligned} \tag{8}$$

The conditional posterior of ϑ_i therefore also takes the Pólya urn form. With probability proportional to the likelihood $f(s_i^*|\theta, \vartheta_j)$, set ϑ_i equal to an existing value ϑ_j in the sample. With probability proportional to the marginal likelihood $f(s_i^*|\theta)$, sample a value ϑ_i from $f(\vartheta_i|s_i^*, \theta)$, where

$$f(s_i^*|\theta) = \int f(s_i^*|\theta, \vartheta_i)dG_0(\vartheta_i)d\vartheta_i,$$

$$f(\vartheta_i|s_i^*, \theta) \propto f(s_i^*|\theta, \vartheta_i)dG_0(\vartheta_i).$$

Similarly, for $i \in N_1$:

$$\begin{aligned} f(\vartheta_i|y, s^*, \theta, \vartheta_{-i}) &\propto \alpha f(s_i^*, y_i|\theta)f(\vartheta_i|y_i, s_i^*, \theta) \\ &\quad + \sum_{j\neq i}\delta_{\vartheta_j}(\vartheta_i)f(s_i^*, y_i|\theta, \vartheta_j). \end{aligned} \tag{9}$$

With a natural conjugate measure G_0, the marginal likelihood can be explicitly calculated, and $f(\vartheta_i|s_i^*, \vartheta_j)$ takes a standard form. As a result one can directly sample from (8) or (9); see for example Conley et al. (2008). Without a natural conjugate measure, however, there is no closed-form solution for $f(s_i^*, y_i|\theta)$ and direct sampling from $f(\vartheta_i|y_i, s_i^*, \theta)$ is not possible. This is the case for our choice of $G_0(\vartheta_i)$. We therefore use Algorithm 8 of Neal (2000). As can be seen from step 5, this algorithm only requires likelihood evaluations and the ability to sample from G_0. It has the obvious advantage that one is not restricted to use natural conjugate priors.

To describe the remixing in Step 6, recall that $\zeta_i = j$ if and only if $\vartheta_i = \vartheta_j^*$. The model can therefore be parameterized by ϑ or $\{\vartheta^*, \zeta\}$. The latter is used for remixing. The distinct values in ϑ^* are a priori independent and drawn from G_0 (see Antoniak (1974)). Hence, the ϑ_j^*'s are (conditionally) independent in the posterior. Define the sets $N_{0j} \equiv \{i : \zeta_i = j, s_i = 0\}$ and $N_{1j} = \{i : \zeta_i = j, s_i = 1\}$. We next consider three cases.

First, suppose j is such that $N_{1j} = \varnothing$. Let $\vartheta_j^* = \{\mu_{j1}^*, \mu_{j2}^*, \sigma_{12j}^*, \xi_j^{*2}\}$. Then

$$\begin{aligned} f(\vartheta_j^*|y, s^*, \theta, \zeta) &\propto \left[\prod_{i:\zeta_i=j} f(s_i^*|\theta, \vartheta_j^*)\right]dG_0(\vartheta_j^*) \\ &\propto \exp\left\{-\frac{1}{2}\sum_{i:\zeta_i=j}(u_{i1} - \mu_{j1}^*)^2\right\}dG_0 \\ &\quad \times (\mu_{j1}^*, \mu_{j2}^*)dG_0(\sigma_{12j}^*, \xi_j^{*2}). \end{aligned}$$

It is then clear that

$$\begin{aligned} f(\vartheta_j^*|y, s^*, \theta, \zeta) &= f(\mu_{j1}^*|y, s^*, \theta, \zeta) \times dG_0(\mu_{j2}^*|\mu_{j1}^*) \\ &\quad \times dG_0(\sigma_{j12}^*, \xi_j^{*2}). \end{aligned}$$

A new draw ϑ_j^* from the posterior can be obtained by sampling from the following distributions:

$$\xi_j^{*2} \sim \mathcal{IG}(c_0, d_0),$$
$$\sigma_{12j}^*|\xi_j^{*2} \sim \mathcal{N}(0, \tau\xi_j^{*2}),$$
$$\mu_{j1}^*|y, s^*, \theta, \zeta \sim \mathcal{N}\left(\left[m_{11}^{-1} + n_j\right]^{-1} \sum_{i:\zeta_i=j} u_{i1}, [m_{11}^{-1} + n_j]^{-1}\right),$$
$$\mu_{j2}^*|\mu_{j1}^* \sim \mathcal{N}\left(\frac{m_{12}}{m_{11}}\mu_{j1}^*, m_{22} - \frac{m_{12}^2}{m_{11}}\right),$$

where $n_j = \sum_{i=1}^n \mathbb{I}\{\zeta_i = j\}$ and m_{11}, m_{12}, m_{22} are the elements of the prior covariance matrix M.

Second, suppose j is such that $N_{0j} = \varnothing$. Then exact sampling of ϑ_j^* from the posterior is not feasible, but the conditional posteriors (for a Gibbs iteration) are easily found. The likelihood is determined by

$$u_i|\zeta_i = j \sim \mathcal{N}(\mu_j^*, \Sigma_j^{*-1}), \qquad \mu_j^* = (\mu_{j1}^*, \mu_{j2}^*),$$
$$\Sigma_j^* = \begin{bmatrix} 1 & \sigma_{12j}^* \\ \sigma_{12j}^* & \xi_j^{*2} + \sigma_{12j}^{*2} \end{bmatrix},$$

so that

$$\mu_j^*|y, s^*, \theta, \zeta, \sigma_{12j}^*, \xi_j^{*2} \sim \mathcal{N}_2\left([M^{-1} + n_j\Sigma_j^{*-1}]^{-1}\Sigma_j^{*-1} \times \sum_{i:\zeta_i=j} u_i, \left[M^{-1} + n_j\Sigma_j^{*-1}\right]^{-1}\right). \tag{10}$$

Also,

$$f(\sigma_{12j}^*, \xi_j^{*2}|y, s^*, \theta, \zeta, \mu_j^*) \propto (\xi_j^{*2})^{-1/2} \exp\left\{-\frac{\sigma_{12j}^{*2}}{2\tau\xi_j^{*2}}\right\} (\xi_j^{*2})^{-(c_0+1)} e^{-d_0/\xi_j^{*2}} \times (\xi_j^{*2})^{-n_j/2} \exp\left\{-\frac{1}{2\xi_j^{*2}} \sum_{i:\zeta_i=j} (u_{i2} - \mu_{j2}^* - \sigma_{12j}^*(u_{i1} - \mu_{j1}^*))^2\right\},$$

so that

$$\xi_j^{*2}|y, s^*, \theta, \zeta, \mu_j^*, \sigma_{12j}^* \sim \mathcal{IG}(c_0 + (n_j + 1)/2, \bar{d}), \tag{11}$$
$$\bar{d} = d_0 + \frac{\sigma_{12j}^{*2}}{2\tau} + \frac{1}{2}\sum_{i:\zeta_i=j} (u_{i2} - \mu_{j2}^* - \sigma_{12j}^*(u_{i1} - \mu_{j1}^*))^2,$$

and

$$\sigma_{12j}^*|y, s^*, \theta, \zeta, \mu_j^*, \xi_j^{*2} \sim \mathcal{N}\left(\bar{\sigma}_{12}, \left(\tau^{-1} + \sum_{i:\zeta_i=j} (u_{i1} - \mu_{j1}^*)^2\right)^{-1} \xi_j^{*2}\right), \tag{12}$$
$$\bar{\sigma}_{12} = \left(\tau^{-1} + \sum_{i:\zeta_i=j} (u_{i1} - \mu_{j1}^*)^2\right)^{-1} \sum_{i:\zeta_i=j} (u_{i1} - \mu_{j1}^*)(u_{i2} - \mu_{j2}^*).$$

The remixing step now involves sampling μ_j^* from (10), ξ_j^{*2} from (11) and σ_{12j}^* from (12).

Finally, suppose j is such that $N_{0j} \neq \varnothing$ and $N_{1j} \neq \varnothing$. Then

$$f(\vartheta_j^*|y, s^*, \theta, \zeta) \propto \left[\prod_{i:\zeta_i=j} \exp\left\{-\frac{1}{2}(u_{i1} - \mu_{j1}^*)^2\right\}\right] dG_0(\mu_{j1}^*) \times (\xi_j^{*2})^{-n_{j1}/2} \exp\left\{-\frac{1}{2\xi_j^{*2}} \times \sum_{i:\zeta_i=j,s_i=1} (u_{i2} - \mu_{j2}^* - \sigma_{12j}^*(u_{i1} - \mu_{j1}^*))^2\right\} \times dG_0(\mu_{j2}^*|\mu_{j1}^*) dG_0(\sigma_{12j}^*, \xi_j^{*2}), \tag{13}$$

where $n_{j1} = \sum_{i=1}^n s_i\mathbb{I}\{\zeta_i = j\}$. The posterior can then be factored as

$$f(\vartheta_j^*|y, s^*, \theta, \zeta) = f(\mu_{j1}^*|y, s^*, \theta, \zeta) f(\mu_{j2}^*, \sigma_{12j}^*, \xi_j^{*2}|y, s^*, \theta, \zeta, \mu_{j1}^*),$$

and

$$\mu_{j1}^*|y, s^*, \theta, \zeta \sim \mathcal{N}\left([m_{11}^{-1} + n_j]^{-1} \sum_{i:\zeta_i=j} u_{i1}, [m_{11}^{-1} + n_j]^{-1}\right). \tag{14}$$

The remaining parameters cannot be sampled in a single step. The conditional posteriors follow easily from (13):

$$\mu_{j2}^*|y, s^*, \theta, \zeta, \mu_{j1}^*, \sigma_{12j}^*, \xi_j^{*2} \sim \mathcal{N}\left(\bar{\mu}_2, \bar{M}_2\right), \tag{15}$$
$$\bar{M}_2 = \left(\frac{m_{11}}{m_{11}m_{22} - m_{12}^2} + \frac{n_{j1}}{\xi_j^{*2}}\right)^{-1},$$
$$\bar{\mu}_2 = \bar{M}_2\left(\frac{m_{12}\mu_{j1}^*}{m_{11}m_{22} - m_{12}^2} + \frac{1}{\xi_j^{*2}} \times \sum_{i:\zeta_i=j,s_i=1} (u_{i2} - \sigma_{12j}^*(u_{i1} - \mu_{j1}^*))\right),$$
$$\sigma_{12j}^*|y, s^*, \theta, \zeta, \mu_j^*, \xi_j^{*2} \sim \mathcal{N}\left(\bar{\sigma}_{12}, \left(\tau^{-1} + \sum_{i:\zeta_i=j,s_i=1} (u_{i1} - \mu_{j1}^*)^2\right)^{-1} \xi_j^{*2}\right), \tag{16}$$
$$\bar{\sigma}_{12} = \left(\tau^{-1} + \sum_{i:\zeta_i=j,s_i=1} (u_{i1} - \mu_{j1}^*)^2\right)^{-1} \times \sum_{i:\zeta_i=j,s_i=1} (u_{i1} - \mu_{j1}^*)(u_{i2} - \mu_{j2}^*),$$
$$\xi_j^{*2}|y, s^*, \theta, \zeta, \mu_j^*, \sigma_{12j}^* \sim \mathcal{IG}(c_0 + (n_{j1} + 1)/2, \bar{d}), \tag{17}$$
$$\bar{d} = d_0 + \frac{\sigma_{12j}^{*2}}{2\tau} + \frac{1}{2}\sum_{i:\zeta_i=j,s_i=1} (u_{i2} - \mu_{j2}^* - \sigma_{12j}^*(u_{i1} - \mu_{j1}^*))^2.$$

The remixing in Step 6 of Algorithm 2 now involves sampling from (14)–(17).

References

Ahn, H., Powell, J.L., 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics 58, 3–29.

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press.

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Annals of Statistics 2, 1152–1174.

Bayarri, M., Berger, J., 1998. Robust Bayesian analysis of selection models. Annals of Statistics 26, 645–659.

Bayarri, M., DeGroot, M., 1987. Bayesian analysis of selection models. Journal of the Royal Statistical Society, Series D (The Statistician) 36, 137–146.

Chen, M.-H., Shao, Q.-M., Ibrahim, J.G., 2000. Monte Carlo Methods in Bayesian Computation. Springer-Verlag, New York.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

Chib, S., Greenberg, E., Jeliazkov, I., 2009. Estimation of semiparametric models in the presence of endogeneity and sample selection. Journal of Computational and Graphical Statistics 18, 321–348.

Conley, T.G., Hansen, C.B., McCulloch, R.E., Rossi, P.E., 2008. A semi-parametric Bayesian approach to the instrumental variable problem. Journal of Econometrics 144, 276–305.

Cosslett, S.R., 1991. Semiparametric estimation of a regression model with sample selectivity. In: W.A. Barnett, J. Powell, and G. Tauchen, (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics. Cambridge, UK, pp. 175–197.
Deb, P., Trivedi, P.K., 2002. The structure of demand for health care: latent class versus two-part models. Journal of Health Economics 21, 601–625.
Escobar, M.D., 1994. Estimating normal means with a Dirichlet process prior. Journal of the American Statistical Association 89, 268–277.
Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.
Escobar, M.D., West, M., 1998. Computing nonparametric hierarchical models. In: Dey, D., Müller, P., Sinha, D. (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics. Springer, New York, pp. 1–22.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1, 209–230.
Ferguson, T.S., 1974. Prior distributions on spaces of probability measures. Annals of Statistics 2, 615–629.
Früwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer.
Gallant, A., Nychka, D., 1987. Semi-nonparametric maximum likelihood estimation. Econometrica 55, 363–390.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall.
Gronau, R., 1974. Wage comparisons—a selectivity bias. The Journal of Political Economy 82, 1119–1143.
Heckman, J.J., 1974. Shadow prices, market wages, and labor supply. Econometrica 42, 679–694.
Heckman, J.J., 1979. Sample selection as a specification error. Econometrica 47, 153–162.
Hennig, C., 2000. Identifiability of models for clusterwise linear regression. Journal of Classification 17, 273–296.
Ichimura, H., Lee, L.-F., 1991. Semiparametric least squares estimation of multiple index models: single equation estimation. In: Barnett, W.A., Powell, J., and Tauchen, G. (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics, Cambridge, UK, pp. 3–49.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.
Koop, G., Poirier, D.J., 1997. Learning about the across-regime correlation in switching regression models. Journal of Econometrics 78, 217–227.
Lee, L.F., 1994. Semiparametric two-stage estimation of sample selection models subject to Tobi t-type selection rules. Journal of Econometrics 61, 305–344.
Lee, J., Berger, J., 2001. Semiparametric Bayesian analysis of selection models. Journal of the American Statistical Association 96, 1269–1276.
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: Dey, D., Müller, P., Sinha, D. (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics. Springer, New York, pp. 23–44.
Manski, C.F., 1988. Identification of binary response models. Journal of the American Statistical Association 83, 729–738.
McCulloch, R.E., Polson, N.G., Rossi, P.E., 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. Journal of Econometrics 99, 173–193.
McCulloch, R.E., Rossi, P.E., 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics 64, 207–240.
Munkin, M.K., Trivedi, P.K., 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.
Neal, R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9, 249–265.
Newey, W., 2009. Two-step series estimation of sample selection models. The Econometrics Journal 12, 217–229.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society. Series B 59, 731–792.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–550.
Teicher, H., 1963. Identifiability of finite mixtures. The Annals of Mathematical Statistics 34, 1265–1269.
Tierney, L., 1994. Markov chains for exploring posterior distributions. The Annals of Statistics 22, 1701–1728.
Van der Klaauw, B., Koning, R., 2003. Testing the normality assumption in the sample selection model with an application to travel demand. Journal of Business and Economic Statistics 21, 31–42.
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.
Yakowitz, S.J., Spragins, J.D., 1968. On the identifiability of finite mixtures. The Annals of Mathematical Statistics 39, 209–214.