BAYESIAN ANALYSIS OF MULTIVARIATE SAMPLE SELECTION MODELS USING GAUSSIAN COPULAS

Phillip Li and Mohammad Arshad Rahman

ABSTRACT

We consider the Bayes estimation of a multivariate sample selection model with p *pairs of selection and outcome variables. Each of the variables may be discrete or continuous with a parametric marginal distribution, and their dependence structure is modeled through a Gaussian copula function. Markov chain Monte Carlo methods are used to simulate from the posterior distribution of interest. The methods are illustrated in a simulation study and an application from transportation economics.*

Keywords: Bayesian estimation; Markov chain; Monte Carlo; Gaussian copulas; sample selection

JEL classifications: C10; C11; C13; C30; C31; C34; C35; R40; R41

Missing Data Methods: Cross-sectional Methods and Applications
Advances in Econometrics, Volume 27A, 269–288

ISSN: 0731-9053/doi:10.1108/S0731-9053(2011)000027A013

INTRODUCTION

This paper applies Bayesian methods to estimate a multivariate sample selection model that addresses the ubiquitous problem of sample selection. In general, sample selection occurs when a variable of interest is nonrandomly missing for a subset of the sample, resulting in a sample that is not representative of the desired population. A well-known application involves analyzing market wages that are only partially observed, depending on whether the individual is participating in the labor force or not. If inference is based only on the remaining observed sample, then specification errors may arise.

A widely used model to address sample selection involves modeling an observed binary selection variable, y_1, that determines whether a continuous outcome variable, y_2, is missing or observed (Heckman, 1976, 1979). Because the joint distribution for (y_1, y_2) is difficult to specify, the model is often re-parameterized in terms of y_1^* and y_2, where y_1^* is a continuous and latent representation of y_1, with the distributional assumption that $(y_1^*, y_2) \sim \mathcal{N}_2(\mu, \Sigma)$. The joint normality assumption is made to achieve tractable results and to obtain an explicit measure of dependence between the two variables through Σ.

Although many variations of this model have been developed and estimated for selection and outcome variables with different data types (e.g., count, ordered, censored, etc.) and distributional assumptions, they are often limited to the specific distributions assumed in the corresponding papers or to formulations with only two or three variables. For example, Terza (1998) studies a univariate count data regression subject to a binary selection variable, and Boyes, Hoffman, and Low (1989) analyze a binary regression with a separate binary selection variable. From a Bayesian perspective, Chib, Greenberg, and Jeliazkov (2009), Greenberg (2007), and van Hasselt (2009) provide analyses for a single Tobit or binary selection variable. For extensive surveys on other variations of sample selection models from a non-Bayesian perspective, refer to Vella (1998) and Greene (2008). However, certain theoretical and applied problems may require either different distributional assumptions or more dependent variables than these models and methods can accommodate, which limits the problems that can be studied.

To address these issues, we analyze a flexible multivariate sample selection model in which the desired marginal distributions are specified by the practitioner. The multivariate dependence is modeled through a copula function in conjunction with the specified marginal distributions. A copula, broadly speaking, is a function that links a multivariate distribution function to its univariate distribution functions with a particular

dependence structure (Sklar, 1959). In other words, there exists a copula function $\mathbb{C}$ such that $F(y_1, \ldots, y_n) = \mathbb{C}(F_1(y_1), \ldots, F_n(y_n))$, where $F(y_1, \ldots, y_n)$ is a multivariate distribution function with $F_1(y_1), \ldots, F_n(y_n)$ as the univariate distribution functions. This method is particularly useful when $F_1(y_1), \ldots, F_n(y_n)$ are known and $F(y_1, \ldots, y_n)$ is unknown, because the copula provides an alternative representation of the joint density.

This paper uses Gaussian copulas that are constructed using multivariate normal distribution functions and a theorem from Sklar (1959). While copulas have been used extensively in the statistics literature for several decades, their usage in econometrics has been relatively limited. Early work on copulas include Hoeffding (1940), Fréchet (1951), and Sklar (1959, 1973), with the latter proving an important theorem that states all continuous multivariate distribution functions have a unique copula representation; the reader is referred to Nelsen (1998) and Zimmer and Trivedi (2005) for comprehensive treatments on copula theory. For multivariate Gaussian copulas, Lee (2010) and Pitt, Chan, and Kohn (2006) respectively analyze multivariate count data models and general regression models using Bayesian simulation methods.

Recent work from econometrics pertaining to sample selection and copulas include Bhat and Eluru (2009), Genius and Strazzera (2008), Lee (1983), Prieger (2000), Smith (2003), and Zimmer and Trivedi (2006). Lee (1983) does not impose joint normality on the standard sample selection model but uses a bivariate Gaussian copula to link the two specified marginal distributions together. Similarly, Prieger (2000) and Bhat and Eluru (2009) develop a model based on a Farlie–Gumbel–Morgenstern copula, which only has moderate correlation coverage between the selection and outcome variables. The remaining authors analyze selection models using variations of Archimedean copulas, resulting in closed-form expressions that are relatively simple to estimate. The aforementioned papers on selection models mostly use maximum likelihood estimation and stay within a bivariate or trivariate structure.

This paper has two purposes. First, we analyze and estimate a multivariate sample selection model with p pairs of selection and outcome variables using Gaussian copulas, where each variable may be discrete or continuous with any parametric marginal distribution specified by the practitioner. We thereby move beyond the bivariate or trivariate structure of the preceding papers to accommodate a larger class of sample selection models. Second, we show how the Bayesian Markov chain Monte Carlo (MCMC) simulation methods from Lee (2010) and Pitt et al. (2006) can be applied to accommodate copula models with missing data. The proposed

estimation method has two main advantages. By using Bayesian simulation methods, it is not necessary to repeatedly compute the high-dimensional copula distribution functions that are needed with non-Bayesian methods. Even though there are methods to calculate these distribution functions (Bőrsch-Supan & Hajivassiliou, 1993; Geweke, 1991; Hajivassiliou & McFadden, 1998; Jeliazkov & Lee, 2010; Keane, 1994), the resulting likelihood is difficult to maximize, even for low-dimensional problems (Zimmer & Trivedi, 2005). Next, our proposed algorithm does not require simulation of the missing data and their associated quantities, which has been shown to improve the efficiency of the Markov chain (Chib et al., 2009; Li, 2011). Careful consideration is needed in this context since the amount and complexity of missing data grow simultaneously with the number of variables modeled (e.g., a model with p partially observed outcomes can have 2^p different combinations of missing data for each observation).

The methods are applied to study the effects of residential density on vehicle miles traveled and vehicle holdings for households in California. Residential density and household demographic variables are used to explain the number of miles a household drives with trucks and cars and the number of trucks and cars a household owns.

The rest of the paper is organized as follows. The second section provides a brief introduction to copulas, and the third section describes the proposed multivariate sample selection model. The fourth section presents the estimation algorithm while the fifth section illustrates the methods on simulated and actual data. The paper is concluded in the sixth section.

COPULAS

This section provides a brief introduction to copulas. Intuitively, a copula is a function that links a multivariate joint distribution to its univariate distribution functions. This approach allows joint modeling of outcomes for which the multivariate distributions are difficult to specify, which is often the case in econometric modeling (e.g., models for discrete choice, count data, and combinations of discrete and continuous data).

More formally, a copula $\mathbb{C}$ has the following definition from Zimmer and Trivedi (2005):

> **Definition.** An n-dimensional copula (or n-copula) is a function $\mathbb{C}$ from the unit n-cube $[0,1]^n$ to the unit interval $[0,1]$ which satisfies the following conditions:

1. $\mathbb{C}(1,\ldots,1,u_k,1,\ldots,1) = u_k$ for every $k \leq n$ and for all u_k in [0,1];
2. $\mathbb{C}(u_1,\ldots,u_n) = 0$ if $u_k = 0$ for any $k \leq n$;
3. $\mathbb{C}$ is n-increasing.

From this definition, a copula can be viewed as an n-dimensional distribution function for $U_1,\ldots,U_n$ defined over $[0,1]^n$, where U_i is uniformly distributed over [0,1] ($i = 1,\ldots,n$).

An important result is that multivariate distribution functions can be expressed in terms of a copula function and its univariate distribution functions. Let $Y_1,\ldots,Y_n$ be n continuous random variables with an n-dimensional distribution function $F(y_1,\ldots,y_n)$ and marginal distribution functions $F_1(y_1),\ldots,F_n(y_n)$. Then

$$F(y_1,\ldots,y_n) = \Pr(Y_1 < y_1,\ldots,Y_n < y_n) \tag{1}$$

$$= \Pr(U_1 < F_1(y_1),\ldots,U_n < F_n(y_n)) \tag{2}$$

$$= \mathbb{C}(u_1 = F_1(y_1),\ldots,u_n = F_n(y_n)) \tag{3}$$

since $F_i(Y_i) \sim U_i$ by the integral transformation result. The dependence between the marginal distributions is introduced through a dependence parameter specific to the chosen copula function. Note that the copula function in Eq. (3) is unique if $F_1(y_1),\ldots,F_n(y_n)$ are continuous distribution functions. The relationship in Eq. (3) still holds for discrete distributions, but the copula function is not unique.

Although many copulas exist, this paper uses a multivariate Gaussian copula of the form

$$\mathbb{C}(u_1,\ldots,u_n|\Omega) = \Phi_n(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_n)|\Omega) \tag{4}$$

where $\Phi_n(\cdot)$ is an n-dimensional distribution function for a multivariate normal vector z with mean zero and correlation matrix Ω, and $\Phi^{-1}(\cdot)$ is the inverse distribution function of a univariate standard normal random variable. Intuitively, the proposed approach is to transform the original variables with prespecified margins into uniform random variables and then into a new set of correlated random variables, z, that is distributed $\mathcal{N}(0,\Omega)$. The advantage of this approach is that dependence is easier to handle through the transformed data z than through the original or uniform random variables. From Song (2000), the density for this copula is proportional to

$$|\Omega|^{-(1/2)} \exp(0.5z'(I - \Omega^{-1})z) \tag{5}$$

where $z_i = \Phi^{-1}(u_i)$ $(i = 1, \ldots, n)$, and I is an identity matrix with the same dimensions as Ω.

The Gaussian copula has several desirable properties. It is one of the few multivariate copulas with $n(n-1)/2$ dependence parameters (the off-diagonals of Ω), one for each pair of variables. This feature is especially desirable in this context since the dependence between the selection and outcome variables is of interest. Furthermore, unlike some copulas, the dependence measures for this copula can be positive or negative. This property is also attractive as the sign of the dependence between the selection and outcome variables is not known a priori. Lastly, Gaussian copulas attain the Fréchet lower and upper bounds when the dependence parameters approach -1 and 1, respectively. This last property is an important factor when choosing a copula and implies that the Gaussian copula can cover the space between the Fréchet bounds.

MODEL

Suppose we have $2p$ variables with N observations for each. Let the first p variables denote the selection variables that determine whether the remaining p variables of interest are observed. Following Pitt et al. (2006), for observational units $i = 1, \ldots, N$ and variables $j = 1, \ldots, 2p$, the proposed model is

$$y_{i,j} = F_{i,j}^{-1}(\Phi(z_{i,j})), \quad z_i = (z_{i,1}, \ldots, z_{i,2p})' \overset{\text{iid}}{\sim} \mathcal{N}_{2p}(0, \Omega) \tag{6}$$

In Eq. (6), $y_{i,j}$ is a discrete or continuous variable with distribution function $F_{i,j}(\cdot)$ that depends on the covariates $x_{i,j}$ and a vector of unknown parameters $\theta_{i,j}$. Also, denote $f_{i,j}(\cdot|\theta_{i,j})$ as the density function for $y_{i,j}$. As an example, if $y_{i,j} \sim \mathcal{N}(x_{i,j}'\beta, \sigma^2)$, then $\theta_{i,j} = (\beta, \sigma^2)$, and $F_{i,j}(\cdot)$ is a distribution function for a normal random variable with mean $x_{i,j}'\beta$ and variance σ^2. Furthermore, each $y_{i,j}$ is modeled with a corresponding Gaussian latent variable $z_{i,j}$, along with a column vector z_i that is distributed multivariate normal with mean zero and correlation matrix Ω. For simplicity, let $F_{i,j}(\cdot) = F_j(\cdot), f_{i,j}(\cdot|\theta_{i,j}) = f_j(\cdot|\theta_j)$, and $\theta_{i,j} = \theta_j$ for all i and j, which imply that the jth distribution function and vector of unknown parameters are the same across all observational units. For Eq. (6), it is important to note that the mapping $F_{i,j}^{-1}(\Phi(\cdot))$ is one-to-one when $y_{i,j}$ is continuous and many-to-one when $y_{i,j}$ is discrete. Also, for identification reasons, the set of covariates for

Table 1. Four Cases of Variable Observability When $p = 2$.

Variables	Case 1	Case 2	Case 3	Case 4
$y_{i,1}$	$\surd$	$\surd$	$\surd$	$\surd$
$y_{i,2}$	$\surd$	$\surd$	$\surd$	$\surd$
$y_{i,3}$	$\surd$	○	$\surd$	○
$y_{i,4}$	$\surd$	$\surd$	○	○

The symbols ○ and $\surd$ respectively denote whether the variable is missing or observed.

each selection variable should include at least one additional covariate than the corresponding variable of interest.

Sample selection is incorporated by assuming that the first p selection variables, $y_{i,1}, \ldots, y_{i,p}$, are always observed and determine whether the remaining variables of interest, $y_{i,p+1}, \ldots, y_{i,2p}$, are missing or observed. That is, $y_{i,1}$ determines whether $y_{i,p+1}$ is missing or observed, $y_{i,2}$ determines whether $y_{i,p+2}$ is missing or observed, etc. This paired sample selection structure implies that for any observational unit i, there are 2^p possible combinations of missing variables. Table 1 illustrates the combinations when $p = 2$. We observe either $(y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4})$, $(y_{i,1}, y_{i,2}, y_{i,4})$, $(y_{i,1}, y_{i,2}, y_{i,3})$, or $(y_{i,1}, y_{i,2})$. In the context of the transportation economics application, $y_{i,1}$ and $y_{i,2}$ are the number of trucks and cars the ith household owns, and $y_{i,3}$ and $y_{i,4}$ are the mileage driven with these vehicles. The mileage variables are missing when the number of vehicles is zero.

ESTIMATION

Posterior Density and Priors

The posterior density of interest is proportional to the data-augmented likelihood multiplied by the prior densities: $\pi(\theta, \Omega, z|y) \propto f(z, y|\theta, \Omega)\pi(\theta, \Omega)$. In this expression, θ contains θ_j for all variables, y contains all the observed $y_{i,j}$ variables, and z contains all the Gaussian latent variables from the copula function corresponding to y. The form of $f(z, y|\theta, \Omega)$ will be described in the next subsection. For Bayesians, this posterior density summarizes all the information available for the unknown parameters after seeing the data. It combines prior information on the parameters before seeing data with information from the observed data through the likelihood function.

We assume independent priors such that $\pi(\theta, \Omega) = \pi(\theta)\,\pi(\Omega)$ for convenience. The prior for Ω is $\mathcal{IW}(v, Q)$, an inverse-Wishart distribution with scalar hyperparameter v and $2p \times 2p$ hyperparameter Q. Because θ is application-specific, we will leave prior specifications to the practitioner. Note that conjugate priors do not aid tractability when using copulas in this context. Therefore, we suggest practitioners choose priors with simple functional forms that accurately reflect prior knowledge and proper priors if model comparisons with Bayes factors are desired.

Estimation Algorithm

The posterior distribution is approximated by MCMC methods, largely following Lee (2010) and Pitt et al. (2006). For the algorithm that follows, define y_j and z_j to be the elements of y and z corresponding to the jth variable, respectively. Also, let z_{-j} be $z \backslash z_j$ and θ_{-j} be $\theta \backslash \theta_j$. The algorithm to sample from $\pi(\theta, \Omega, z|y)$ is summarized as follows:

1. Sample Ω in one block from $f(\Omega|z)$.
2. Sample (θ_j, z_j) jointly for all discrete marginal distributions from $f(\theta_j, z_j|y, z_{-j}, \theta_{-j}, \Omega)$ as follows
   (a) Sample θ_j without z_j from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$.
   (b) Sample z_j conditioned on θ_j from $f(z_j|y, z_{-j}, \theta, \Omega)$.
3. Sample θ_j for all continuous marginal distributions from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$ and solve for z_j with y_j and θ_j through the one-to-one transformation in Eq. (6).

The Metropolis–Hastings algorithm is used to sample from the preceding distributions since random draws cannot be easily obtained from the posterior distribution using direct sampling. Broadly speaking, this algorithm generates samples from the posterior distribution by first proposing candidate values from a known proposal distribution and then accepting them with a certain probability. If a proposed value is rejected, then the previous value is used. This method constructs a Markov chain such that after a sufficient burn-in period, the draws can be shown to come from the posterior distribution of interest by Metropolis–Hastings convergence results (Chib & Greenberg, 1995; Tierney, 1994) as the number of iterations approaches infinity.

A multivariate t proposal distribution with mean μ, scale matrix V, and degrees of freedom v is used in the subsequent sections. In order to obtain

parameter values such that this proposal density dominates the density of interest (also called the target density), μ and V are set to the maximum and inverse of the negative Hessian (evaluated at the maximum) of the density of interest, respectively. These quantities can be obtained by quasi-Newton methods. Lastly, the degrees of freedom parameter v is set to ensure heavy tails. Specific details are provided in the subsequent sampling sections.

Note that the missing variables of interest (e.g., the entries marked with a ○ in Table 1) and their corresponding Gaussian latent variables are not sampled. In many Bayesian MCMC algorithms for missing data problems, the missing data are often included in the sampling to facilitate the tractability of the sampling densities, however this strategy is not necessary and not always optimal. Chib et al. (2009) and Li (2011) have shown that the inclusion and conditioning of missing data in some applications can slow down the mixing of the Markov chain. In particular, for the semiparametric sample selection model of Chib et al. (2009), the inefficiency factors (a measure of how quickly the autocorrelations in a Markov chain chain taper off, where lower values indicate better performance) are at least four times greater when the missing data due to sample selection are included in the sampler. This issue is particularly problematic when the quantity of missing outcomes due to sample selection is large or when the model includes many parameters, both of which may be the case in this context. Therefore, the proposed algorithm does not sample the missing data and corresponding latent data.

Sampling Ω

Two different algorithms are presented to sample Ω. The first algorithm is based on the sampler from Chib and Greenberg (1998); it works well when the number of variables is small (less than four) and is easy to implement. For problems with more than four variables, we introduce an algorithm based on Chan and Jeliazkov (2009).

Since the observed variables can potentially change for every observational unit, additional notation will now be defined. Let s_i denote the indices of the observed variables for observation i, and let y_{s_i} and z_{s_i} respectively denote the columns of observed and latent variables corresponding to s_i such that $\text{Var}(z_{s_i}) = \Omega_{s_i}$. For example, if $y_{i,1}$, $y_{i,2}$, and $y_{i,4}$ are observed for $p = 2$, then

$$s_i = \{1,2,4\},\ y_{s_i} = (y_{i,1}, y_{i,2}, y_{i,4})',\ z_{s_i} = (z_{i,1}, z_{i,2}, z_{i,4})',\ \Omega_{s_i} = \begin{pmatrix} 1 & \omega_{12} & \omega_{14} \\ \omega_{21} & 1 & \omega_{24} \\ \omega_{41} & \omega_{42} & 1 \end{pmatrix}$$

The full conditional density $f(\Omega|z)$ is proportional to

$$f(z|\Omega)\pi(\Omega) \propto \pi(\Omega)\prod_{i=1}^{N}\left\{|\Omega_{s_i}|^{-(1/2)}\exp(-0.5\, z'_{s_i}\Omega_{s_i}^{-1}z_{s_i})\right\} \tag{7}$$

Since Ω is a $2p \times 2p$ correlation matrix, there are $2p(2p-1)/2$ unique off-diagonal terms, denoted by ω, that need to be sampled. To sample ω from Eq. (7), a Metropolis–Hastings step with a multivariate t proposal is used. The target density in Eq. (7) is first maximized with respect to ω using quasi-Newton methods; let $\hat{\omega}$ and $\hat{V}$ denote the maximizing vector and the inverse of the negative Hessian evaluated at the maximum. Next, propose ω' from a multivariate t distribution with mean vector $\hat{\omega}$, scale matrix $\hat{V}$, and degrees of freedom v. A proposed value for Ω' can now be constructed with ω'. If Ω' is not positive definite, then the previous value of Ω is used instead. Otherwise, the draw is accepted with probability

$$\alpha(\omega,\omega') = \min\left\{1, \frac{f(\Omega'|z)\, f_T(\omega|\hat{\omega},\hat{V},v)}{f(\Omega|z)\, f_T(\omega'|\hat{\omega},\hat{V},v)}\right\} \tag{8}$$

The second algorithm is based on the sampling strategy from Chan and Jeliazkov(2009). To introduce the technique, note that any positive definite covariance matrix Σ can be decomposed as $\Sigma = L'D^{-1}L$. The unit lower triangular matrix L contains ones on the diagonal and unrestricted elements on the lower off-diagonal, while the diagonal matrix D contains positive elements on the diagonal and zeros elsewhere. The insight of this algorithm is that we can sample the elements in L and D instead of the elements in Σ directly and reconstruct Σ through the decomposition.

Using similar notation to Chan and Jeliazkov (2009), denote λ_j $(j = 1,\ldots,2p)$ as the diagonal elements of D and $a_{j,k}$ $(1 \le k < j \le 2p)$ as the unrestricted elements on the lower off-diagonal of L. Similarly, denote $a^{j,k}$ as the (j, k)th element of L^{-1}. As an illustration

$$D = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_{2p} \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ a_{2,1} & 1 & 0 & \ldots & 0 \\ a_{3,1} & a_{3,2} & 1 & \ldots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{2p,1} & a_{2p,2} & \ldots & \ldots & 1 \end{pmatrix}$$

Let $\sigma_{j,k}$ denote the element of Σ corresponding to the jth row and kth column. After imposing $\sigma_{1,1} = \sigma_{2,2} = \cdots = \sigma_{2p,2p} = 1$ in Σ to obtain correlation form and expanding $L'D^{-1}L$, the free elements of Σ must satisfy the constraints

$$\sigma_{j,k} = a^{j,k}\lambda_k + \sum_{h=1}^{k-1} a^{j,h}a^{k,h}\lambda_h, \quad 1 \leq k < j \leq 2p \tag{9}$$

and λ_j must satisfy

$$\lambda_1 = 1 \tag{10}$$

$$\lambda_j = 1 - \sum_{k=1}^{j-1}(a^{j,k})^2\lambda_k, \quad j = 2, \ldots, 2p \tag{11}$$

As noted in the referenced paper, when Σ is in correlation form, $\{\lambda_j\}$ is only a function of $\{a_{j,k}\}$. This implies that the off-diagonal elements in $\{\sigma_{j,k}\}$ are also functions of $\{a_{j,k}\}$ only. Consequently, we only need to sample $\{a_{j,k}\}$ when Σ is expressed as $L'D^{-1}L$.

The second algorithm will also utilize a Metropolis–Hastings step since Eq. (7) is not a recognizable distribution with respect to $\{a_{j,k}\}$. First, decompose Ω as $L'D^{-1}L$ and express Eq. (7) in terms of $\{a_{j,k}\}$ analogously to Eq. (9) through Eq. (11). Let $\hat{a}$ and $\hat{A}$ respectively denote the values that maximize Eq. (7) with respect to $a = \{a_{j,k} : 1 \leq k < j \leq 2p\}$ and the inverse of the negative Hessian evaluated at the maximum. Next, propose a' from a multivariate t distribution with mean $\hat{a}$, scale matrix $\hat{A}$, and degrees of freedom v, which can be used to construct the proposed value Ω'. The proposed value is accepted with probability

$$\alpha(a, a') = \min\left\{1, \frac{f(\Omega'|z)\, f_T(a|\hat{a}, \hat{A}, v)}{f(\Omega|z)\, f_T(a'|\hat{a}, \hat{A}, v)}\right\} \tag{12}$$

This approach differs slightly from the one in Chan and Jeliazkov (2009) since Eq. (7) is of a different form due to the missing data.

Both of these algorithms allow Ω to be sampled in one block with the positive definite constraint intact. The first algorithm is relatively easy to implement, however it is generally inefficient when the dimension of Ω is large. The positive definite constraint will become increasingly difficult to satisfy when p increases, resulting in proposed values that are frequently rejected and slower mixing of the Markov chain. The second algorithm is more involved but has been shown to be efficient (Chan & Jeliazkov, 2009), therefore it is recommended for models with more than four variables.

Sampling (θ_j, z_j) *for Discrete Marginals*

For discrete marginals, the pair (θ_j, z_j) is sampled jointly. Using the method of composition, θ_j is first sampled from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$, then z_j is sampled from $f(z_j|y, z_{-j}, \theta, \Omega)$ with θ_j conditioned on. The first density is

$$f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega) \propto \pi(\theta_j|\theta_{-j}) \prod_{i=1}^{N} \{f(y_{i,j}|z_{-j}, \theta_j, \Omega)\}^{\mathbb{I}(j \in s_i)} \tag{13}$$

where

$$f(y_{i,j}|z_{-j}, \theta_j, \Omega) = \int f(y_{i,j}|z_{i,j}, \theta_j) f(z_{i,j}|z_{-j}, \Omega)\, \mathrm{d}z_{i,j} \tag{14}$$

and $\mathbb{I}(A)$ is an indicator function that takes the value 1 when A is true and 0 otherwise. Upon defining $\mu_{i,j|-j}$ and $\sigma^2_{i,j|-j}$ as the conditional mean and variance of the normal density $f(z_{i,j}|z_{-j}, \Omega)$, it can be shown that

$$f(y_{i,j}|z_{-j}, \theta_j, \Omega) = \Phi\left(\frac{T^U - \mu_{i,j|-j}}{\sigma_{i,j|-j}}\right) - \Phi\left(\frac{T^L - \mu_{i,j|-j}}{\sigma_{i,j|-j}}\right) \tag{15}$$

with

$$T^L = \Phi^{-1}(F_j(y_{i,j} - 1)) \tag{16}$$

$$T^U = \Phi^{-1}(F_j(y_{i,j})) \tag{17}$$

The sampling of θ_j can proceed with the target density in Eq. (13) and a Metropolis–Hastings step just like the preceding sections.

Now, the density for $z_{i,j}$ is

$$f(z_{i,j}|y, z_{-j}, \theta, \Omega) \propto f(y_{i,j}|z_{i,j}, \theta_j)\, f(z_{i,j}|z_{-j}, \Omega) \tag{18}$$

where

$$f(y_{i,j}|z_{i,j}, \theta_j) = \mathbb{I}(T^L < z_{i,j} \leq T^U)$$

Thus, for any $j \in S_i$ that corresponds to a discrete $y_{i,j}$,

$$z_{i,j}|y, z_{-j}, \theta, \Omega \sim \mathcal{TN}_{(T^L, T^U)}(\mu_{i,j|-j}, \sigma^2_{i,j|-j}) \tag{19}$$

where $\mathcal{TN}_{(a,b)}(\mu, \sigma^2)$ denotes a univariate normal distribution with mean μ and variance σ^2 truncated to the region (a, b). Note that the conditional

moments depend on which latent variables from z_{-j} are available for observation i, indicated by s_i, and need to be adjusted accordingly.

Sampling θ_j *for Continuous Marginals*
For continuous marginal distributions, θ_j is sampled from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$. The sampling density is proportional to

$$\pi(\theta_j|\theta_{-j}) \prod_{i=1}^{N} \{f(y_{i,j}|\theta_j)\}^{\mathbb{I}(j \in S_i)} \exp(0.5\, z'_{S_i}(I_{S_i} - \Omega_{S_i}^{-1}) z_{S_i}) \tag{20}$$

Note that the elements in z_{S_i} corresponding to the jth variable are also functions of θ_j, so the last term cannot be dropped. This is from the relationship $z_{i,j} = \Phi^{-1}(F_j(y_{i,j}))$, where $F_j(y_{i,j})$ depends on θ_j. A Metropolis–Hastings step is needed to obtain a draw from Eq. (20). Once θ_j is drawn, the elements in z_j can be recovered through the aforementioned relationship, therefore z_j does not need to be sampled for continuous marginals.

APPLICATIONS

Simulated Data

This section illustrates the estimation methods with simulated data. The purpose is to study the performance of the algorithm on a model that will be used in the next subsection and to demonstrate that the algorithm can correctly recover the parameters of interest. Specifically, the model from the third section is estimated with two Poisson selection variables ($y_{i,1}$ and $y_{i,2}$) and two normally distributed outcome variables ($y_{i,3}$ and $y_{i,4}$). To set the context, sample selection is incorporated as follows: $y_{i,3}$ is observed if and only if $y_{i,1} > 0$, and $y_{i,4}$ is observed if and only if $y_{i,2} > 0$.

For $i = 1, \ldots, 1000$, we have the following

$$y_{i,1} \sim Po(\lambda_{i,1}), \quad \log(\lambda_{i,1}) = x'_{i,1}\beta_1 \tag{21}$$

$$y_{i,2} \sim Po(\lambda_{i,2}), \quad \log(\lambda_{i,2}) = x'_{i,2}\beta_2$$

$$y_{i,3} \sim \mathcal{N}(x'_{i,3}\beta_3, \sigma_3^2)$$

$$y_{i,4} \sim \mathcal{N}(x'_{i,4}\beta_4, \sigma_4^2)$$

where $x'_{i,j}$ $(j = 1, \ldots, 4)$ are randomly drawn exogenous covariate vectors from standard normal distributions. The true generating values for the parameters of interest $\theta_1 = \beta_1$, $\theta_2 = \beta_2$, $\theta_3 = (\beta_3, \sigma_3^2)$, $\theta_4 = (\beta_4, \sigma_4^2)$, and Ω are presented in Table 2. The percentage of missing data for each outcome variable is 20%, similar to the real data. Proper priors are used with hyperparameters that reflect non-informativeness. For θ_1 and θ_2, multivariate normal priors are used with mean vector zero and a variance–covariance matrix of an identity matrix multiplied by 100; similar priors are used for β_3 and β_4. Lastly, inverse gamma priors are used for σ_3^2 and σ_4^2.

The algorithm is iterated 5,000 times with 500 iterations discarded for burn-in. Table 2 reports the posterior means and standard deviations along with their true generated values, and Fig. 1 illustrates the lagged

Table 2. Posterior Means and Standard Deviations for θ_j $(j = 1, \ldots, 4)$ and $\text{Vech}(\Omega) = (\omega_{2,1}, \omega_{3,1}, \omega_{3,2}, \omega_{4,1}, \omega_{4,2}, \omega_{4,3})$.

Parameter	Generated Value	$\mathbb{E}(\cdot \mid y)$	$\text{SD}(\cdot \mid y)$
$\beta_{1,1}$	0.30	0.31	0.06
$\beta_{1,2}$	0.30	0.28	0.05
$\beta_{1,3}$	0.30	0.30	0.04
$\beta_{1,4}$	0.30	0.31	0.05
$\beta_{1,5}$	0.30	0.30	0.04
$\beta_{2,1}$	0.20	0.20	0.06
$\beta_{2,2}$	0.20	0.17	0.05
$\beta_{2,3}$	0.20	0.23	0.05
$\beta_{2,4}$	0.20	0.19	0.05
$\beta_{2,5}$	0.20	0.21	0.05
$\beta_{3,1}$	0.50	0.81	0.23
$\beta_{3,2}$	0.50	0.54	0.07
$\beta_{3,3}$	0.50	0.40	0.06
$\beta_{3,4}$	0.50	0.44	0.07
σ_3^2	3.00	2.79	0.16
$\beta_{4,1}$	0.30	0.30	0.16
$\beta_{4,2}$	0.30	0.35	0.05
$\beta_{4,3}$	0.30	0.38	0.05
σ_4^2	2.00	1.92	0.10
$\omega_{2,1}$	0.28	0.25	0.04
$\omega_{3,1}$	0.28	0.27	0.03
$\omega_{3,2}$	0.28	0.27	0.04
$\omega_{4,1}$	0.28	0.31	0.04
$\omega_{4,2}$	0.28	0.28	0.04
$\omega_{4,3}$	0.28	0.27	0.04

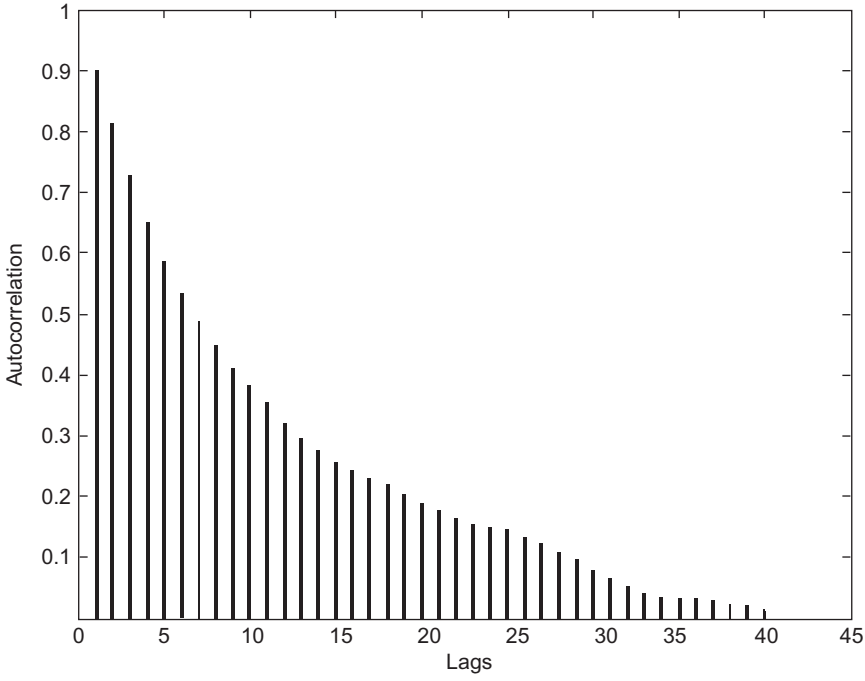

Fig. 1. Autocorrelation Plot for $\beta_{1,3}$.

autocorrelations for a randomly chosen parameter $\beta_{1,3}$ up to order 40. In general, the results from Table 2 suggest that all the parameters have been estimated well since the posterior means are reasonably close to their generated values with tight standard deviations. Furthermore, the autocorrelation plot, a way of assessing how well the Markov chain mixes, suggests that our algorithm performs well. The autocorrelations for $\beta_{1,3}$ decrease and taper off around lag 40, as do most of the autocorrelations for the remaining parameters. However, we suggest iterating the algorithm at least 15,000 times to obtain more precise results.

Vehicle Usage in California

The model from Eq. (21) is applied to analyze the effects of residential density on household vehicle usage in California. A sample selection

framework is utilized since vehicle usage is nonrandomly missing from the sample data with a probability that depends on whether the household owns a vehicle or not. Households may be selecting themselves into being vehicle owners for unobserved reasons that also affect how much they drive, creating differences in the observed and unobserved samples. Therefore, sample selection must be accounted for.

Some studies suggest that certain changes in urban spatial structure (e.g., residential density) may be effective in reducing fuel consumption of automobiles or in influencing travel behavior (Brownstone & Fang, 2009; Brownstone & Golob, 2009; Cervero & Kockelman, 1997; Dunphy & Fisher, 1996; Fang, 2008). For example, it may be more costly to maneuver around a location with higher residential density due to increased congestion and time spent in searching for parking spaces, resulting in households driving less and switching to more fuel-efficient vehicles. Consequently, understanding this potential relationship can provide alternative policies to control fuel consumption and congestion.

The dataset is from the 2001 National Household Travel Survey. It contains the daily and long-distance travel information between April 2001 and May 2002 for approximately 66,000 households across the nation, along with variables such as residential density, household size, residential location type, income, education, and other household characteristics. The dataset used contains 1,000 randomly sampled households that reside in California. The primary variables of interest are the number of trucks and cars owned by the household ($y_{i,1}$ and $y_{i,2}$) and the corresponding annual mileage driven with these vehicles ($y_{i,3}$ and $y_{i,4}$), where 20–30% of the mileage variables are missing. A truck is defined as a van, sports utility vehicle, or pickup truck, and a car is an automobile, car, or station wagon. These two categories have distinct differences in miles per gallon (MPG) requirements by the Corporate Average Fuel Economy (CAFE) standards. Covariates of interest include residential density (housing units per square mile at the census block level), household size, and dummy variables representing whether the household resides in an urban location, is low income, has a young child, and owns their home. Descriptive statistics are summarized in Table 3.

The results are presented in Tables 4 and 5. From Table 4, the estimated correlation between the truck equations is 0.37, suggesting that sample selection is not ignorable for these vehicles. This relationship is due to positive associations in unobserved factors that affect both truck ownership and utilization (e.g., a predisposition to travel more in spacious vehicles like trucks). On the other hand, the estimated correlation for the car equations is negligible and suggests that selection may not be an issue in this case.

Table 3. Descriptive Statistics Based on 1,000 Observations.

Variables	Description	Mean	SD
	Dependent variables		
Tnum	Number of trucks owned by the household	0.72	0.79
Cnum	Number of cars owned by the household	1.10	0.82
Tmile	Mileage per year driven with trucks (10,000 miles)	0.71	1.10
Cmile	Mileage per year driven with cars (10,000 miles)	0.89	1.00
	Exogenous covariates		
Density	Houses per square mile	2564.99	1886.09
Hhsize	Number of individuals in a household	2.69	1.44
Urb	Household is in an urban area	0.93	0.25
Lowinc	Household income is between 20 K and 30 K	0.11	0.31
Child	Youngest child is under 6 years old	0.17	0.37
Home	Household owns the home	0.26	0.44

Table 4. Posterior Means for Ω.

1.00	−0.41	0.37	−0.02
−0.41	1.00	−0.17	−0.02
0.37	−0.17	1.00	0.01
−0.02	−0.02	0.01	1.00

Table 5. Posterior Means and Standard Deviations of β_1, β_2, β_3, and β_4.

Covariates	Tnum		Cnum		Tmile		Cmile	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
log(Density)	−0.06	0.03	0.21	0.09	−0.01	0.26	0.05	0.17
Hhsize	0.09	0.05	0.12	0.61	0.15	0.23	0.09	0.16
Urb	−0.52	0.55	0.43	0.49	−0.13	1.13	−0.15	0.90
Lowinc	−0.23	0.58	−0.35	0.41	0.20	0.93	−0.07	0.74
Child	0.10	0.37	−0.24	0.32	0.10	0.78	−0.09	0.56
Home	0.26	0.27	0.34	0.20	–	–	–	–

The estimates in Table 5 suggest that the effect of residential density on vehicle usage is uncertain. The posterior standard deviations are large relative to the means, and the 95% probability intervals for these parameters contain 0 (not shown in the table). This result is consistent with the findings of Li (2011) in which a multivariate sample selection model is also used to

analyze a similar application. However, this conclusion differs from the ones presented in Fang (2008) and Brownstone and Golob (2009), where these authors generally find evidence for negative associations between truck usage and residential density. This difference arises due to the usage of a sample selection model and different distributional assumptions.

On the other hand, there is evidence that households residing in denser neighborhoods tend to own fewer trucks and more cars. This can be attributable to the increased costs of operating vehicles with lower fuel efficiency in these areas, resulting in preferences for cars with better fuel economy. Also, larger households tend to have more trucks, presumably because these vehicles can fit more passengers.

CONCLUDING REMARKS

This paper analyzes a multivariate sample selection model with p pairs of selection and outcome variables. A unique feature of this model is that the variables can be discrete or continuous with any parametric distribution, resulting in a large class of multivariate selection models that can be accommodated. For example, the model may involve any combination of variables that are continuous, binary, ordered, or censored. Although the joint distribution can be difficult to specify, a multivariate Gaussian copula function is used to link the marginal distributions together and handle the multivariate dependence. The proposed estimation approach relies on the MCMC-based techniques from Lee (2010) and Pitt et al. (2006) and adapts the preceding methods to a missing data setting. An important aspect of this algorithm is that it does not require simulation of the missing outcomes, which has been shown in some cases to improve the mixing of the Markov chain. The methods are applied to both simulated and real data, and the results show that the algorithm works well and can reveal new conclusions in the data.

A copy of the Matlab code to estimate the model in the real data section is available upon request.

ACKNOWLEDGMENTS

We would like to thank Ivan Jeliazkov, David Brownstone, Dale Poirier, the participants at the Advances in Econometrics conference, the seminar participants at the econometrics seminar at UCI, the editor, and two anonymous referees for their helpful comments.

REFERENCES

Bhat, C., & Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, *43*(7), 749–765.

Boyes, W. J., Hoffman, D. L., & Low, S. A. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, *40*(1), 3–14.

Brownstone, D., & Fang, A. (2009). *A vehicle ownership and utilization choice model with endogenous residential density*. Working Paper. University of California, Irvine, CA.

Brownstone, D., & Golob, T. F. (2009). The impact of residential density on vehicle usage and energy consumption. *Journal of Urban Economics*, *65*(1), 91–98.

Bőrsch-Supan, A., & Hajivassiliou, V. A. (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics*, *58*(3), 347–368.

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219.

Chan, J. C.-C., & Jeliazkov, I. (2009). MCMC estimation of restricted covariance matrices. *Journal of Computational and Graphical Statistics*, *18*(2), 457–480.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician*, *49*(4), 327–335.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.

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*(2), 321–348.

Dunphy, R., & Fisher, K. (1996). Transportation, congestion, and density: New insights. *Transportation Research Record: Journal of the Transportation Research Board*, *1552*, 89–96.

Fang, H. A. (2008). A discrete-continuous model of households' vehicle choice and usage, with an application to the effects of residential density. *Transportation Research Part B: Methodological*, *42*(9), 736–758.

Fréchet, M. (1951). Sur les tableaux de correlation dont les marges son donnees. *Annals of the University of Lyon, Section A* (14), 53–77.

Genius, M., & Strazzera, E. (2008). Applying the copula approach to sample selection modelling. *Applied Economics*, *40*(11), 1443–1455.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. *Computing Science and Statistics*, 571–578.

Greenberg, E. (2007). *Introduction to Bayesian econometrics*. Cambridge University Press.

Greene, W. H. (2008). *Econometric analysis*. Prentice Hall.

Hajivassiliou, V. A., & McFadden, D. L. (1998). The method of simulated scores for the estimation of LDV models. *Econometrica*, *66*(4), 863–896.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement*, *5*(4). NBER Chapters. National Bureau of Economic Research, Inc., pp. 120–137.

Hoeffding, W. (1940). Masstabinvariente korrelationstheorie, schriften des Mathematischen Instituts und des Instituts fur Angewandt Mathematik der Universitat Berlin.

Jeliazkov, I., & Lee, E. (2010). MCMC perspectives on simulated likelihood estimation. *Advances in Econometrics: Maximum Simulated Likelihood*, 3–39.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica*, *62*(1), 95–116.

Lee, E. H. (2010). *Copula analysis of correlated counts*. Working Paper. University of California, Irvine, CA.

Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica*, *51*(2), 507–512.

Li, P. (2011). Estimation of sample selection models with two selection mechanisms. *Computational Statistics and Data Analysis*, *55*, 1099–1108.

Nelsen, R. B. (1998). *An introduction to copulas (lecture notes in statistics)* (1st ed). Springer.

Pitt, M., Chan, D., & Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, *93*(3), 537–554.

Prieger, J. (2000). *A generalized parametric selection model for non-normal data*. Working Paper no. 00-9. University of California at Davis, Department of Economics.

Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universit de Paris*, *8*, 229–231.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica*, *9*, 449–460.

Smith, M. D. (2003). Modelling sample selection using Archimedean copulas. *Econometrics Journal*, *6*(1), 99–123.

Song, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, *27*(2), 305–320.

Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics*, *84*(1), 129–154.

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

van Hasselt, M. (2009). *Bayesian inference in a sample selection model*. Working Paper. The University of Western Ontario, Ontario, London.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources*, *33*(1), 127–169.

Zimmer, D. M., & Trivedi, P. K. (2005). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics, 1*(1), 57.

Zimmer, D. M., & Trivedi, P. K. (2006). Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business & Economic Statistics*, *24*, 63–76.