Bayesian Inference in Spatial Sample Selection Models

Dogan, Osman and Taspinar, Suleyman

16 December 2016

Bayesian Inference in Spatial Sample Selection Models*

Osman Doğan† Süleyman Taşpınar‡

December16, 2016

Abstract

In this study, we consider Bayesian methods for the estimation of a sample selection model with spatially correlated disturbance terms. We design a set of Markov chain Monte Carlo (MCMC) algorithms based on the method of data augmentation. The natural parameterization for the covariance structure of our model involves an unidentified parameter that complicates posterior analysis. The unidentified parameter – the variance of the disturbance term in the selection equation – is handled in different ways in these algorithms to achieve identification for other parameters. The Bayesian estimator based on these algorithms can account for the selection bias and the full covariance structure implied by the spatial correlation. We illustrate the implementation of these algorithms through a simulation study.

JEL-Classification: C13, C21, C31
Keywords: Spatial dependence, Spatial sample selection model, Bayesian analysis, Data augmentation.

*This research was supported, in part, by a grant of computer time from the City University of New York High Performance Computing Center under NSF Grants CNS-0855217, CNS-0958379 and ACI-1126113.

†Project Department, Metro Istanbul, Esenler, Istanbul, Turkey, email: ODogan10@gmail.com.

‡Economics Program, Queens College, The City University of New York, New York, United States, email: STaspinar@qc.cuny.edu.

1 Introduction

A typical sample selection model consists of (i) a selection equation that models the selection mechanism through which we observe the level of outcome, and (ii) an outcome equation that describes the process that is generating the outcome. The model structure is characterized by the correlation between the disturbances of these equations, for which estimation requires special methods (Heckman, 1979, 1990; Lee, 1978, 1994; Newey, 2009; Olsen, 1980). Besides the cross equation correlation in the disturbance terms, spatial correlation may also be present in the disturbance terms of each equation. The spatial sample selection model considered in the present study accommodates both type of correlations.

Selection models, or more generally Type-I or Type-II type Tobit models, may arise frequently in urban economics, regional science, labor economics, agricultural economics, and social interaction models. It is natural to consider a notion of spatial correlation in the unobservables so long as data is organized by a notion of location in the relevant space. The presence of common shocks, factors, cluster effects provides a natural motivation. For example, McMillen, (1995) studies residential land values in urban areas through a sample selection model. He conjectures that unobserved variables that make a parcel more likely to receive residential zoning may increase the value of residential land. It is also plausible to allow for spatially correlated disturbance terms because nearby parcels are likely to be affected by the same neighborhood factors and spillovers. Büchel and Ham, (2003) studies overeducation– a job seeker's overqualification for a job she accepts due to her location constraints– through a Heckit model. The selection problem arises because overeducation observed only for the employed. The authors state that the risk of overeducation largely determined by spatial flexibility of a job seeker in combination with the spatial heterogeneity in suitable job opportunities, relative to the place of residence. They try to control for spatial correlations in the disturbance terms by clustering methods. Flores-Lagunes and Schnier, (2012) study (for details, see our empirical illustration) the spatial production of fisheries in the Eastern Bering Sea through a sample selection model. Since a negative shock that affects the fish population in a certain location would affect the production of all vessels in other locations by displacing fishing effort into more efficient surrounding locations, they allow disturbance terms to be spatially correlated.

Another motivation for considering spatial correlation is related to measurement error. The mismatch between the spatial unit of observations (e.g., census tract, county, state, peer groups, farm location, fishing zone etc.) and the unit of a study (e.g., student, household, housing market, labor market, farm, fishing vessel, etc.) can cause measurement errors in the variables of interest (Anselin, 2007). Since these measurement errors may vary systematically across space, the disturbance terms of a regression model over the same space are likely to be correlated. Ward et al., (2014) consider a sample selection model of cereal production where the selection equation specifies a farmer's endogenous decision about whether to plant cereals. They employ a first-order spatial autoregressive model for the disturbance terms, because data on yields or climate are aggregated for large administrative units, and correlation among unobservables may be driven by unobserved environmental, geographical and climatological clusters. Rabovič and Čížek, (2016) provides another example in the context of peer effects, where the outcome equation models a student's achievement on a test and the selection equation models the student's decision whether to take the exam. It is plausible that decision to take the exam and the score from the exam may depend on a student's ability that is likely to be similar to her peers' abilities. Therefore, social interaction literature often incorporate what is known as the "correlated effects" in the model (Lee et al., 2010).

The limited dependent variable models that accommodate spatial dependence have been studied in terms of both estimation and testing issues. The maximum likelihood estimator (MLE) of a probit model with a spatial autoregressive process requires evaluation of a multivariate normal cumulative

distribution function, which often leads to numerical estimation problems. To circumvent this shortcoming of the MLE, alternative methods have been suggested in the literature. For example, McMillen, (1992) uses an expectation maximization (EM) algorithm to circumvent the evaluation of the multivariate normal distribution function and suggests a tractable iterative estimation approach. Beron and Vijverberg, (2004) use the recursive importance sampling (RIS) simulator to approximate the log-likelihood function of the spatial probit model. Pinkse and Slade, (1998) formulate moment functions from the score vector of a partial MLE for the generalized method of moments (GMM) estimation of a spatial probit model. Instead of using entire joint distribution of observations implied by the spatial dependence, Wang et al., (2013) formulate a partial MLE based on the partial joint distribution of observations to reduce computational difficulties.

McMillen, (1995) extends the EM algorithm suggested in McMillen, (1992) to a sample selection model that has a first order spatial autoregressive process in the disturbance term. The estimation scheme requires inversion of an $n \times n$ matrix at each iteration which makes this approach impractical for large samples. Flores-Lagunes and Schnier, (2012) combine the GMM approach in Pinkse and Slade, (1998) and Kelejian and Prucha, (1998) and suggest a GMM method for a sample selection model that has a first order spatial autoregressive process in the disturbance terms. The moment functions for the estimation of the selection equation are the ones suggested by Pinkse and Slade, (1998) for the probit model. These moment functions are combined with the moment functions formulated for the outcome equation to form a joint GMME. The simulation studies reported in Flores-Lagunes and Schnier, (2012) show that the bias present in the selection equation parameters adversely affects the estimation of the parameter of the outcome equation. Rabovič and Čížek, (2016) extends the partial maximum likelihood (ML) method of Wang et al., (2013) to a sample selection model with a spatial lag of a latent dependent variable and a sample selection model with spatially correlated disturbance terms. They establish the large sample properties of the proposed partial MLE and provide a finite sample bias and precision comparison to the Heckit and the GMM approach of Flores-Lagunes and Schnier, (2012). Overall, the proposed partial MLE outperforms the Heckit and the GMM based estimator.

In this paper, we consider a sample selection model that has a first order spatial autoregressive process for the disturbance terms of the selection and outcome equations.[1] We consider the Bayesian MCMC estimation approach for this model with various alternative Gibbs samplers. In comparison with the GMM and partial ML approaches, the Bayesian approach with data augmentation formulates estimators that can exploit the full information on the spatial correlation structure. The data augmentation method treats the underlying latent variable as an additional parameter to be estimated and this treatment of latent variables facilitates the posterior simulation through an MCMC sampler (Albert and Chib, 1993; van Dyk and Meng, 2001). The natural parameterization for the covariance of our model involves an unidentified parameter which can complicate posterior analysis. The unidentified parameter, i.e., the variance of the disturbance terms in the selection equation, is handled in different ways in the suggested Gibbs samplers.

In the first algorithm, we specify prior distributions for the identified parameters and consider the method suggested in Nobile, (2000) that can be used to construct a Markov chain that fixes the unidentified parameter during the posterior analysis. In the second algorithm, we consider the re-parameterization approach suggested in Li, (1998), McCulloch et al., (2000) and van Hasselt, (2011) for the covariance matrix of the model. Given the bivariate normal distribution assumption for the disturbance terms, the covariance matrix is re-parameterized in terms of conditional variance and covariance of disturbance terms. In the third algorithm, we consider a different blocking scheme for

[1] To the best of our knowledge, McMillen, (1995), Flores-Lagunes and Schnier, (2012) and Rabovič and Čížek, (2016) are the only studies that consider estimation of sample selection models with spatial dependence.

the full set of conditional posterior distributions. The Gibbs sampler operates by splitting the full set of parameters into different blocks which can affect the mixing properties of the sampler. Hence, we consider an alternative sampler where the parameter vector of the regressors in the outcome equation and covariance term of the disturbance terms are sampled through a single block.

In the fourth and fifth algorithms, we consider samplers based on the parameter expansion method (or the marginal data augmentation method) for the estimation of our model (Liu and Wu, 1999; Meng and van Dyk, 1999). The unidentified parameter, i.e., the variance of the disturbance term in the selection equation, is introduced in the sampler through an appropriate prior distribution to improve the convergence properties of the sampler. The proper prior distributions that can be assigned to the unidentified parameter do not affect the posterior distribution of the identified model parameters in these samplers but can improve the characteristics of the resulting Markov chains. As a result, these algorithms accommodate the normalization constraint in the estimation while improving the convergence rates of the resulting Markov chains.

Through a simulation study, we illustrate the implementation of these algorithms in the context of our spatial sample selection model. Our results show that the Bayesian estimator in all algorithms reports estimates that are close to the true parameter value for the autoregressive parameter of the selection equation. For the autoregressive parameter in the outcome equation, the deviation of the posterior mean estimate from the true parameter value is negligible for the Bayesian estimator in Algorithms 1–4. As for the parameters of the exogenous variables in the selection and outcome equations, the Bayesian estimator in Algorithms 1 and 4 performs relatively better in terms of reported deviations between the point estimates and the true parameter values. Our results also indicate that all algorithms have similar mixing properties under our priors specifications. For an empirical illustration, we use the application in the area of natural resource economics considered in Flores-Lagunes and Schnier, (2012) to model the spatial production within a fishery with a spatial sample selection model. Our Bayesian estimator reports much precise estimates for this application as it accounts for the full covariance structure implied by the spatial correlation.

The remainder of this article is divided into five sections. In Section 2, we provide the model specification. In Section 3, we provide the posterior analysis including prior specifications. We present the details of our simulation design in Section 4. We evaluate the relative performance of algorithms in this section. In Section 5, we provide an empirical illustration and examine the relevance of Bayesian estimates in comparison with the estimates from a GMME suggested by Flores-Lagunes and Schnier, (2012). Section 6 concludes. Some technical results and figures are left to a web appendix.

2 Model Specification

We consider the following Type II Tobit model with a first order spatial autoregressive process:

$$Y_{1i}^{\star} = X_{1i}^{'}\beta + U_{1i}, \quad U_{1i} = \lambda \sum_{j=1}^{n} W_{ij}U_{1j} + \varepsilon_{1i}, \tag{2.1}$$

$$Y_{2i}^{\star} = X_{2i}^{'}\delta + U_{2i}, \quad U_{2i} = \rho \sum_{j=1}^{n} M_{ij}U_{2j} + \varepsilon_{2i}, \tag{2.2}$$

for $i = 1, \ldots, n$. Here, $Y_{1i}^{\star}$ and $Y_{2i}^{\star}$ are, respectively, the latent variables of selection and outcome equations, X_{1i} and X_{2i} are, respectively, $k_1 \times 1$ and $k_2 \times 1$ vectors of non-stochastic exogenous variables with associated vectors of coefficients β and δ. Let W and M be $n \times n$ spatial weight

matrices of known constants with zero diagonal elements. The (i,j)th element of these matrices are respectively denoted by W_{ij} and M_{ij} in (2.1) and (2.2). Hence, the regression disturbance term U_{1i} and U_{2i} are allowed to depend on the disturbance terms of other observations. The parameters λ and ρ are known as the spatial autoregressive parameters. The innovations terms ε_{1i} and ε_{2i} are assumed to be i.i.d $N(0,\Sigma)$ with

$$\Sigma = \begin{pmatrix} \sigma_1^2 = 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & \varrho\sigma_2 \\ \varrho\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{2.3}$$

where ϱ is the correlation coefficient between ε_{1i} and ε_{2i}. The observed variable Y_{1i} for the selection equation is related to the latent variable $Y_{1i}^{\star}$ by $Y_{1i} = \mathbb{I}(Y_{1i}^{\star} > 0)$ for $i = 1,\ldots,n$, where $\mathbb{I}(\cdot)$ is the indicator function. The observed variable Y_{2i} for the outcome equation is related to both the latent variables of the selection and outcome equations by

$$Y_{2i} = \begin{cases} Y_{2i}^{\star} = X_{2i}^{'}\delta + U_{2i} & \text{if} \quad Y_{1i} = 1 \\ \text{missing} & \text{if} \quad Y_{1i} = 0 \end{cases} \tag{2.4}$$

for $i = 1,\ldots,n$. For the non-spatial model where $\lambda = \rho = 0$, the OLS estimator of δ based on the subsample obtained when $Y_{1i} = 1$ is inconsistent when $\sigma_{12} \neq 0$. This result is the well-known selectivity bias problem (Heckman, 1979).[2]

There are two identification issues related to sample selection models: (i) the normalization imposed on the variance of ε_{1i}, and (ii) the exclusion restriction. Without the normalization $\sigma_1^2 = 1$, multiple values for the model parameters give rise to the same value for the likelihood function. Hence, the normalization can be considered as a way to achieve identification. The exclusion restriction, on the other hand, is relevant for the precision of estimator rather than for the identification of parameters. When there is no exclusion restriction, i.e., $X_{1i} = X_{2i}$, the parameters are still identified within the ML framework (Lee, 2003; Wooldridge, 2002).[3] However, if an excluded variable from the outcome equation is a relevant variable, i.e., when the exclusion restriction is false, then the suggested estimator will suffer from the omitted variable bias (Lee, 2003).[4]

For the Bayesian estimation of non-spatial sample selection model, the data is usually stacked for each pair of the latent observations such that the ith observation is a vector denoted by $Y_i^{\star} = (Y_{1i}^{\star}, Y_{2i}^{\star})^{'}$. Hence, the model can be written in a compact way by stacking over the pair of the

[2]In the literature, there are other variants of the non-spatial sample selection model. For example, van Hasselt, (2005) and Wooldridge, (2002) suggest a two-part model where the outcome equation is different than that of the Type II Tobit model. The difference arises because the two-part model assumes that the disturbance terms U_{1i} and U_{2i} are independent conditional on $Y_{1i}^{\star} > 0$. Another closely related variant is the Roy Model (or Type V Tobit model) which is a three equation model, where the first equation determines a selection outcome and the remaining two equations describe the main outcome for the cases of $Y_{1i} = 1$ and $Y_{1i} = 0$, respectively. Li, (1998) considers a variant in which the outcome equation is a Tobit equation. Hence, the observed variable for the outcome equation does not depend on the observed variable of the selection equation. In the variant considered by Chib et al., (2009), the selection equation is a Tobit type equation, and the observed variable of the outcome equation depends on the observed variable of the selection equation.

[3]There is also no identification problem for the Heckit of Heckman, (1979). However, this situation can introduce severe collinearity between X_{2i} and the inverse Mills ratio, which can lead to imprecise estimators. The ML estimator may also show poor performance when there is no exclusion restriction. For some simulation evidence, see Leung and Yu, (1996).

[4]Note that the bivariate normality assumption facilitates the posterior analysis for our model. The performance of the Bayesian estimator under a distributional misspecification requires further investigation, which is beyond the scope of this paper. In this respect, a non-parametric approach that relaxes the bivariate normality assumption can be a direction for future studies.

latent observations. Due to the presence of spatial dependence in disturbance terms, we do not use the same format for our model. Instead, we stack each equation independently and get

$$Y_1^\star = X_1\beta + S^{-1}(\lambda)\varepsilon_1, \quad Y_2^\star = X_2\delta + R^{-1}(\rho)\varepsilon_2, \tag{2.5}$$

where $S(\lambda) = (I_n - \lambda W)$, $R(\rho) = (I_n - \rho M)$, $Y_j^\star = (Y_{j1}^\star, Y_{j2}^\star, \ldots, Y_{jn}^\star)^{'}$, $X_j = (X_{j1}^{'}, X_{j2}^{'}, \ldots, X_{jn}^{'})^{'}$ and $\varepsilon_j = (\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jn})^{'}$ for $j = 1, 2$. Let $\theta = (\beta^{'}, \delta^{'})^{'}$ be the augmented parameter vector. Furthermore, let $Y^\star = \big(Y_1^{\star'}, Y_2^{\star'}\big)^{'}$, $X = \begin{pmatrix} X_1 & 0_{n\times k_2} \\ 0_{n\times k_1} & X_2 \end{pmatrix}$ and $\varepsilon = \big(\varepsilon_1^{'} S^{-1'}(\lambda), \varepsilon_2^{'} R^{-1'}(\rho)\big)^{'}$. Then, our model can be more compactly written as

$$Y^\star = X\theta + \varepsilon, \quad \text{and} \quad \mathrm{E}(\varepsilon\varepsilon^{'}) = \Omega = \mathcal{K}(\lambda,\rho)\big(\Sigma \otimes I_n\big)\mathcal{K}^{'}(\lambda,\rho), \tag{2.6}$$

where $\mathcal{K}(\lambda,\rho) = \begin{pmatrix} S^{-1}(\lambda) & 0_{n\times n} \\ 0_{n\times n} & R^{-1}(\rho) \end{pmatrix}$.

3 Posterior Analysis

For the posterior analysis, the prior distributions need to be assigned to parameters of the model. We assume that the prior distribution functions of parameters of the model are independent. Let $p(\theta,\lambda,\rho,\Sigma) = p(\theta)p(\lambda)p(\rho)p(\Sigma)$ be the joint prior distribution function. We assume the following prior distribution functions:

$$p(\theta) = (2\pi)^{-k/2}|V_0|^{-1/2}\exp\Big\{-\frac{1}{2}(\theta-\mu_0)^{'}V_0^{-1}(\theta-\mu_0)\Big\}, \tag{3.1}$$

$$p(\Sigma) = \Big(2^{v_0}\pi^{1/2}\prod_{i=1}^{2}\Gamma\big(\frac{v_0+1-i}{2}\big)\Big)^{-1}|T_0|^{v_0/2}|\Sigma|^{-(v_0+3)/2}\exp\Big\{-\frac{1}{2}tr\big(T_0\Sigma^{-1}\big)\Big\}\mathbb{I}(\Sigma_{11}=1), \tag{3.2}$$

$$p(\lambda) = \begin{cases} \kappa_1/2 & \text{if } \lambda \in (-1/\kappa_1,\ 1/\kappa_1) \\ 0 & \text{otherwise} \end{cases}, \tag{3.3}$$

$$p(\rho) = \begin{cases} \kappa_2/2 & \text{if } \rho \in (-1/\kappa_2,\ 1/\kappa_2) \\ 0 & \text{otherwise} \end{cases}. \tag{3.4}$$

The prior in (3.2), denoted with $\mathrm{InvWish}(T_0, v_0)$, is the inverse Wishart distribution function which can be considered as the multivariate extension of the inverse-gamma distribution. As shown in Nobile, (2000), the normalization constraint in the selection equation is imposed on the prior through the indicator function $\mathbb{I}(\Sigma_{11} = 1)$, where $(1,1)$th element Σ_{11} is set to 1.[5] The uniform prior distributions for the autoregressive parameters in (3.3) and (3.4) indicate that all values in the corresponding intervals are equally probable. In these formulations, κ_1 and κ_2 are the spectral radius of W and M, respectively.[6] For the posterior analysis, the vectors of observed variables corresponding to the selection and the outcome equation are respectively denoted by Y_1 and Y_2.

[5] Following Nobile, (2000), we provide an algorithm for sampling from the inverse Wishart distribution subject to $\mathbb{I}(\Sigma_{11} = 1)$ in the web appendix.

[6] Note that the intervals $(-1/\kappa_1,\ 1/\kappa_1)$ and $(-1/\kappa_2,\ 1/\kappa_2)$ can be considered as the parameter spaces for λ and ρ, respectively (Kelejian and Prucha, 2010). An alternative to the uniform prior is the four parameter Beta prior introduced in LeSage and Pace, (2009) for autoregressive parameters.

Let $Y = (Y_1^{'}, Y_2^{'})$ be the combined vector on observed variables. The joint augmented posterior kernel including the latent variable $Y^{\star}$ can be expressed as

$$p(\theta, \lambda, \rho, \Sigma, Y^{\star}\big|Y) \propto p(\theta)p(\lambda)p(\rho)p(\Sigma)p(Y^{\star}|\theta, \lambda, \rho, \Sigma)p(Y|\theta, \lambda, \rho, \Sigma, Y^{\star}). \tag{3.5}$$

From our stacked model in (2.6), it can be easily determined that

$$p(Y^{\star}|\theta, \lambda, \rho, \Sigma) = (2\pi)^{-n}|S(\lambda)||R(\rho)||\Sigma|^{-n/2} \times \exp\left\{ -\frac{1}{2}(Y^{\star} - X\theta)^{'}\Omega^{-1}(Y^{\star} - X\theta)\right\}. \tag{3.6}$$

Given (3.6), we can infer the conditional distributions of blocks $Y_1^{\star}$ and $Y_2^{\star}$. These conditional distributions are normal distributions with the following means and covariances (Geweke, 2005, Theorem 5.3.1, p.171)

$$Y_1^{\star}|Y_2^{\star}, \theta, \lambda, \rho, \Sigma \sim N(\psi, \Psi), \quad Y_2^{\star}|Y_1^{\star}, \lambda, \rho, \Sigma \sim N(\varphi, \Upsilon), \tag{3.7}$$

where

$$\begin{aligned}
&\psi = X_1\beta + \big(\sigma_{12}/\sigma_2^2\big)S^{-1}(\lambda)R(\rho)\big(Y_2^{\star} - X_2\delta\big), \quad \Psi = \big(1 - \sigma_{12}^2/\sigma_2^2\big)S^{-1}(\lambda)S^{-1^{'}}(\lambda),\\
&\varphi = X_2\delta + \sigma_{12}R^{-1}(\rho)S(\lambda)\big(Y_1^{\star} - X_1\beta\big), \quad \Upsilon = \big(\sigma_2^2 - \sigma_{12}^2\big)R^{-1}(\rho)R^{-1^{'}}(\rho).
\end{aligned}$$

From the augmented joint posterior distribution in (3.5), the kernel of the conditional posterior of θ can be determined by collecting all terms that are not multiplicatively separable from θ. Thus

$$p\big(\theta|\rho, \lambda, \Sigma, Y^{\star}\big) \propto \exp\left\{ -\frac{1}{2}\big(\theta - \mu_0\big)^{'}V_0^{-1}\big(\theta - \mu_0\big)\right\}\exp\left\{ -\frac{1}{2}\big(Y^{\star} - X\theta\big)^{'}\Omega^{-1}\big(Y^{\star} - X\theta\big)\right\}. \tag{3.8}$$

This result in (3.8) implies that

$$\theta|\rho, \lambda, \Sigma, Y^{\star}, Y \sim N(\mu_1, V_1) \tag{3.9}$$

where $V_1 = \big(V_0^{-1} + X^{'}\Omega^{-1}X\big)^{-1}$ and $\mu_1 = V_1\big(V_0^{-1}\mu_0 + X^{'}\Omega^{-1}Y^{\star}\big)$.

The conditional posterior kernel of Σ is

$$\begin{aligned}
p(\Sigma|\rho, \lambda, Y^{\star}, Y) \propto |\Sigma|^{-(v_0+3+n)/2}\exp\left\{ -\frac{1}{2}\big[\mathrm{tr}\big(T_0\Sigma^{-1}\big) + (Y^{\star} - X\theta)^{'}\Omega^{-1}(Y^{\star} - X\theta)\big]\right\}\\
\times \mathbb{I}\big(\Sigma_{11} = 1\big).
\end{aligned} \tag{3.10}$$

The exponent term in $p\big(Y^{\star}|\theta, \lambda, \rho, \Sigma\big)$ can be written in a more compact way to facilitate the derivation of the conditional posterior of Σ. Using $\Omega^{-1} = \mathcal{K}^{'-1}(\lambda, \rho)\big(\Sigma^{-1} \otimes I_n\big)\mathcal{K}^{-1}(\lambda, \rho)$, $\varepsilon_1 = S(\lambda)Y_1^{\star} - S(\lambda)X_1\beta$, $\varepsilon_2 = R(\rho)Y_2^{\star} - R(\rho)X_1\delta$ and the matrix trace operator, we obtain

$$\big(Y^{\star} - X\theta\big)^{'}\Omega^{-1}(Y^{\star} - X\theta\big) = \mathrm{tr}\Big(\big(Y^{\star} - X\theta\big)^{'}\Omega^{-1}\big(Y^{\star} - X\theta\big)\Big) = \mathrm{tr}(A_1 \times \Sigma^{-1}), \tag{3.11}$$

where $A_1 = \begin{pmatrix} \varepsilon_1'\varepsilon_1 & \varepsilon_1'\varepsilon_2 \\ \varepsilon_2'\varepsilon_1 & \varepsilon_2'\varepsilon_2 \end{pmatrix}$. Substituting (3.11) into (3.10) yields

$$p(\Sigma|\rho,\lambda,Y^\star,Y) \propto |\Sigma|^{-(v_0+3+n)/2} \exp\left\{ -\frac{1}{2}\text{tr}\Big(\big(T_0+B\big)\Sigma^{-1} \Big) \right\} \times \mathbb{I}\big(\Sigma_{11}=1\big) \tag{3.12}$$

The above result shows that the conditional posterior density of Σ is given by

$$\Sigma|\theta,\rho,\lambda,Y^\star,Y \sim \text{InvWish}(v_1,T_1) \times \mathbb{I}(\Sigma_{11}=1), \tag{3.13}$$

where $v_1 = v_0 + n$ and $T_1 = T_0 + A_1$. The random draws from $\text{InvWish}(v_1,T_1)$ should also be conditional on the normalization constraint of $\Sigma_{11}=1$. An algorithm similar to the one suggested in Nobile, (2000) can be designed for imposing this constraint on the random draws.

Finally, the conditional posterior distributions for autoregressive parameters take the following forms:[7]

$$\begin{aligned} p(\lambda|\theta,\rho,\Sigma,Y^\star,Y) \propto |S(\lambda)| \times \exp\left\{ -\frac{1}{2}\left(\frac{\sigma_2^2}{\sigma_2^2-\sigma_{12}^2}\varepsilon_1'\varepsilon_1 - \frac{2\sigma_{12}}{\sigma_2^2-\sigma_{12}^2}\varepsilon_2'\varepsilon_1 \right) \right\} \\ \times \mathbb{I}\big(\lambda \in (-1/\kappa_1,\, 1/\kappa_1)\big), \end{aligned} \tag{3.14}$$

$$\begin{aligned} p(\rho|\theta,\lambda,\Sigma,Y^\star,Y) \propto |R(\rho)| \times \exp\left\{ -\frac{1}{2}\left(\frac{1}{\sigma_2^2-\sigma_{12}^2}\varepsilon_2'\varepsilon_2 - \frac{2\sigma_{12}}{\sigma_2^2-\sigma_{12}^2}\varepsilon_2'\varepsilon_1 \right) \right\} \\ \times \mathbb{I}\big(\rho \in (-1/\kappa_2,\, 1/\kappa_2)\big). \end{aligned} \tag{3.15}$$

Both conditional posterior densities in (3.14) and (3.15) are not in the form of known densities. Sampling for both λ and ρ can be accomplished through a Metropolis-Hasting algorithm. LeSage and Pace, (2009) suggest a random walk Metropolis-Hasting algorithm in which a normal distribution is used as the proposal distribution to generate random draws for these parameters.[8] According to this method, the candidate values $(\lambda^{new},\rho^{new})$ are generated by

$$\begin{pmatrix} \lambda^{new} \\ \rho^{new} \end{pmatrix} = \begin{pmatrix} \lambda^{old} \\ \rho^{old} \end{pmatrix} + \begin{pmatrix} z_\lambda & 0 \\ 0 & z_\rho \end{pmatrix} \times \begin{pmatrix} Z_\lambda \\ Z_\rho \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} Z_\lambda \\ Z_\rho \end{pmatrix} \sim N\left(0_{2\times 1},\, I_2\right). \tag{3.16}$$

The parameters z_λ and z_ρ in (3.16) are called the tuning parameters which ensure that the sampler moves over the entire conditional posterior distributions of the autoregressive parameters.[9] The candidate values generated through (3.16) are subject to the parameter space constraints. That is, the candidate values, for which $\lambda^{new} \notin (-1/\kappa_1, 1/\kappa_1)$ and $\rho^{new} \notin (-1/\kappa_2, 1/\kappa_2)$, are rejected as a way of imposing the parameter space constraints. Since the increment random variables Z_λ and Z_ρ are standard normal, the acceptance probability values for the candidate $(\lambda^{new},\rho^{new})$ take the following forms: $\left(\Pr\left(\lambda^{new}\right), \Pr\left(\rho^{new}\right)\right) = \left(\min\left\{1,\, \frac{p(\lambda^{new}|\theta,\rho,\Sigma,Y^*,Y)}{p\left(\lambda^{old}|\theta,\rho,\Sigma,Y^*,Y\right)}\right\}, \min\left\{1,\, \frac{p(\rho^{new}|\theta,\lambda,\Sigma,Y^*,Y)}{p\left(\rho^{old}|\theta,\lambda,\Sigma,Y^*,Y\right)}\right\} \right)$. The candidates λ^{new} and ρ^{new} are accepted, respectively, with probabilities $\Pr\left(\lambda^{new}\right)$ and $\Pr\left(\rho^{new}\right)$.

Now we turn the conditional posterior distribution of latent variables. From the augmented joint

[7] With the parameterization in (2.3), we have $(Y^\star - X\theta)'\Omega^{-1}(Y^\star - X\theta) = \frac{\sigma_2^2}{\sigma_2^2-\sigma_{12}^2}\varepsilon_1'\varepsilon_1 - \frac{2\sigma_{12}}{\sigma_2^2-\sigma_{12}^2}\varepsilon_2'\varepsilon_1 + \frac{1}{\sigma_2^2-\sigma_{12}^2}\varepsilon_2'\varepsilon_2$. We use this result in (3.14) and (3.15).

[8] For some Monte Carlo results on the performance of this algorithm, see Doğan and Taşpınar, (2014).

[9] We set $z_\lambda = z_\rho = z^\star = 0.5$ at the initial step. As suggested in LeSage and Pace, (2009), we adjust $z^\star$ so that the acceptance rates fall between 40% and 60% during the sampling process. A modified version of the tuned random walk procedure is to fix $z^\star$ after the burn-in period. During the burn-in period, the values of $z^\star$ can be collected and then the mean of these collected values can be used as the tuning parameter for the sampler.

posterior distribution function in (3.5), the conditional posterior of latent variable is proportional to $p\left(Y^{\star}|\theta,\lambda,\rho,\Sigma\right)p\left(Y|\theta,\lambda,\rho,\Sigma,Y^{\star}\right)=p\left(Y^{\star}|\theta,\lambda,\rho,\Sigma\right)p\left(Y|Y^{\star}\right)$, where we use the fact that the observable variable Y is determined with certainty given the information on latent variable $Y^{\star}$ regardless of θ, λ, ρ, and Σ. Depending on the sign of the latent variable of the selection equation, this conditional posterior will be in the from of a truncated multivariate normal distribution. As in the case of non-spatial sample selection model, there are two cases: (i) $Y_{1i}^{\star}\leq 0$, i.e., $Y_{1i}=0$ and (ii) $Y_{1i}^{\star}>0$, i.e, $Y_{1i}=1$. Let $\mathcal{N}_0=\{i:Y_{1i}=0\}$ be the index set of observations for which $Y_{1i}^{\star}\leq 0$. Similarly, let $\mathcal{N}_1=\{i:Y_{1i}=1\}$ be the index set of observations for which $Y_{1i}^{\star}>0$. For the outcome equation, $\mathcal{N}_0$ contains indices of missing outcomes whereas $\mathcal{N}_1$ contains indices of observed outcomes.

The conditional posterior of latent variable can be written as

$$p\big(Y^{\star}|\theta,\lambda,\rho,\Sigma,Y\big)\propto p\big(Y_1^{\star},Y_2^{\star}\big|\theta,\lambda,\rho,\Sigma\big)p\big(Y|Y^{\star}\big)=p\big(Y_2^{\star}\big|Y_1^{\star},\theta,\lambda,\rho,\Sigma\big)p\big(Y_1^{\star}|\theta,\lambda,\rho,\Sigma\big)p\big(Y|Y^{\star}\big) \tag{3.17}$$

where the proportionality in the first line follows from (3.5). The result in (3.17) also suggests a computational implementation based on the method of composition for sampling from $p\big(Y^{\star}\big|\theta,\lambda,\rho,\Sigma,Y\big)$. We can first draw $Y_1^{\star}$ from $p\big(Y_1^{\star}|\theta,\lambda,\rho,\Sigma\big)=N\big(X_1\beta,S^{-1}(\lambda)S^{-1'}(\lambda)\big)$ subject to $-\infty<Y_{1i}^{\star}\leq 0$ for $i\in\mathcal{N}_0$, and $0<Y_{1i}^{\star}<\infty$ for $i\in\mathcal{N}_1$. Then, this draw can be used to generate a draw of $Y_2^{\star}$ from the conditional distribution $p\big(Y_2^{\star}\big|Y_1^{\star},\theta,\lambda,\rho,\Sigma\big)=N\big(\varphi,\Upsilon\big)$ for the observations corresponding to indices in $\mathcal{N}_0$.[10].

In the first step of method of composition, we need to sample from the truncated multivariate normal distribution. The algorithm suggested in Geweke, (1991) can be used to generate random draws from the truncated multivariate normal distributions. Consider a random vector that has a multivariate normal distribution subject to a linear constraint. The conditional distribution of an individual element on all other elements of the vector is a univariate truncated normal distribution.[11] Geweke, (1991) uses this result and suggests a Gibbs sampler to generate random draws from the truncated multivariate normal distributions. In our simulation study, we use the same approach. The sampling steps for the latent variable can be summarized as follows.

Imputation Step:

1. Use the Gibbs sampler suggested in Geweke, (1991) to generate $Y_1^{\star}$ from $N\big(X_1\beta,S^{-1}(\lambda)S^{-1'}(\lambda)\big)$ subject to $-\infty<Y_{1i}^{\star}\leq 0$ for $i\in\mathcal{N}_0$, and $0<Y_{1i}^{\star}<\infty$ for $i\in\mathcal{N}_1$.

2. Use the vector $Y_1^{\star}$ obtained in Step 1 to generate $Y_2^{\star}$ from

$$MVN\left(X_2\delta+\sigma_{12}R^{-1}(\rho)S(\lambda)\left(Y_1^{\star}-X_1\beta\right),\ \left(\sigma_2^2-\sigma_{12}^2\right)R^{-1}(\rho)R^{-1'}(\rho)\right). \tag{3.18}$$

3. Use draws from Step 1 to update $Y_1^{\star}$. For the case of $Y_2^{\star}$, use draws from Step 2 to update $Y_2^{\star}$ only for $i\in\mathcal{N}_0$ since $Y_2^{\star}$ for $i\in\mathcal{N}_1$ is observed.

In the case of a non-spatial sample selection model, the draws of $Y_{1i}^{\star}$ for $i\in\mathcal{N}_1$ are generated conditional on the observed outcome values. As shown above, these observations are not sampled

[10]The explicit form of $N\left(\varphi,\Upsilon\right)$ is given in (3.7). This multi-step method for sampling from $p(Y^{\star}|\theta,\lambda,\rho,\Sigma,Y)$ is called the method of composition. For details, see Chib, (2001, p.3576)

[11]Chopin, (2011) describes an alternative numerical scheme which is computationally faster than some other alternative algorithms for sampling from a univariate truncated normal distribution.

conditional on the observed values of the outcome equation, instead they are sampled from the truncated marginal distribution of $Y_1^\star$. This difference arises because of the presence of spatial dependence in our model.[12]

For an alternative algorithm, we consider the re-parametrization approach used in Li, (1998), McCulloch et al., (2000) and van Hasselt, (2011) for Σ. Given our bivariate normality assumption for $\left(\varepsilon_{1i}, \varepsilon_{2i}\right)^{'}$, the conditional variance of ε_{2i} given ε_{1i} is $\text{Var}\left(\varepsilon_{2i}|\varepsilon_{1i}\right) = \sigma_2^2 - \sigma_{12}^2/\sigma_1^2$, where σ_1^2 is the variance of ε_{1i}. Imposing the normalization restriction of $\sigma_1^2 = 1$ yields $\text{Var}\left(\varepsilon_{2i}|\varepsilon_{1i}\right) = \sigma_2^2 - \sigma_{12}^2$. Let $\text{Var}\left(\varepsilon_{2i}|\varepsilon_{1i}\right) = \xi^2$, then $\sigma_2^2 = \sigma_{12}^2 + \xi^2$. In this approach, the population expectation of ε_{2i} given ε_{1i} is formulated with

$$\varepsilon_{2i} = \sigma_{12}\varepsilon_{1i} + \eta_i, \tag{3.19}$$

where η_i is i.i.d $N(0, \xi^2)$. In (3.19), the linearity is assumed in the population regression of ε_{2i} on ε_{1i}. This result is always implied by the bivariate normal assumption. This re-parameterization allows us to work with the model in terms of parameters σ_{12} and ξ^2 such that

$$\Sigma = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{12}^2 + \xi^2 \end{pmatrix}. \tag{3.20}$$

For the posterior analysis, we assume the following prior distributions: $\sigma_{12}|\xi^2 \sim N\left(0, \tau\xi^2\right)$, and $\xi^2 \sim IG\left(a_0, b_0\right)$, where $IG\left(a_0, b_0\right)$ is the inverse gamma density function with shape parameter a_0 and scale parameter b_0. The prior dependence between ξ and σ_{12} is suggested in van Hasselt, (2011), and the parameter τ is used to adjust the shape of prior implied for the correlation coefficient between ε_{1i} and ε_{2i}. In Section 4, we provide the details about the elicitation of this parameter.

From the joint augmented posterior distribution, the conditional posterior of ξ^2 is determined by collecting all terms that are not multiplicatively separable from ξ^2, which is given by

$$p(\xi^2|\theta, \lambda, \rho, \sigma_{12}, Y^\star, Y) \propto \left(\xi^2\right)^{-(\frac{2a_0+n+1}{2}+1)} \exp\left\{ -\frac{1}{\xi^2}\left(b_0 + \frac{\sigma_{12}^2}{2\tau} + \frac{\sigma_{12}^2}{2}\varepsilon_1^{'}\varepsilon_1 - \sigma_{12}\varepsilon_2^{'}\varepsilon_1 + \frac{1}{2}\varepsilon_2^{'}\varepsilon_2\right)\right\}, \tag{3.21}$$

where $\varepsilon_1 = S(\lambda)Y_1^\star - S(\lambda)X_1\beta$, and $\varepsilon_2 = R(\rho)Y_2^\star - R(\rho)X_2\delta$.[13] The result (3.21) implies that

$$\xi^2|\theta, \lambda, \rho, \sigma_{12}, Y^\star, Y \sim IG\left(a_1, b_1\right), \tag{3.22}$$

where $a_1 = \frac{2a_0+n+1}{2}$ and $b_1 = b_0 + \frac{\sigma_{12}^2}{2\tau} + \frac{\sigma_{12}^2}{2}\varepsilon_1^{'}\varepsilon_1 - \sigma_{12}\varepsilon_2^{'}\varepsilon_1 + \frac{1}{2}\varepsilon_2^{'}\varepsilon_2$. Similarly, the conditional posterior of σ_{12} is given as

$$p\left(\sigma_{12}|\theta, \lambda, \rho, \xi^2, Y^\star, Y\right) \propto \exp\left\{ -\frac{1}{2}\frac{\left(\sigma_{12} - \frac{\tau\varepsilon_2^{'}\varepsilon_1}{1+\tau\varepsilon_1^{'}\varepsilon_1}\right)^2}{\tau\xi^2(1+\tau\varepsilon_1^{'}\varepsilon_1)^{-1}}\right\}. \tag{3.23}$$

[12] We provide an alternative sampling scheme in which these observations can be drawn conditional on the observed outcome values. For details, see the web appendix.

[13] Using parameterization in (3.20), we have $(Y^\star - X\theta)^{'}\Omega^{-1}(Y^\star - X\theta) = \varepsilon_1^{'}\varepsilon_1 + \frac{\sigma_{12}^2}{\xi^2}\varepsilon_1^{'}\varepsilon_1 - \frac{2\sigma_{12}}{\xi^2}\varepsilon_2^{'}\varepsilon_1 + \frac{1}{\xi^2}\varepsilon_2^{'}\varepsilon_2$. We use this result in (3.21).

The above result implies that

$$\sigma_{12}|\theta,\lambda,\rho,\xi^2,Y^\star,Y \sim N\left(\frac{\tau\varepsilon_2^{'}\varepsilon_1}{1+\tau\varepsilon_1^{'}\varepsilon_1},\ \frac{\tau\xi^2}{(1+\tau\varepsilon_1^{'}\varepsilon_1)}\right). \tag{3.24}$$

With this new parameterization, the conditional posterior distributions for autoregressive parameters take the following forms:

$$p(\lambda|\theta,\rho,\sigma_{12},\xi^2,Y^\star,Y) \propto |S(\lambda)| \times \exp\left\{-\frac{1}{2}\left(\varepsilon_1^{'}\varepsilon_1 + \frac{\sigma_{12}^2}{\xi^2}\varepsilon_1^{'}\varepsilon_1 - \frac{2\sigma_{12}}{\xi^2}\varepsilon_2^{'}\varepsilon_1\right)\right\}, \tag{3.25}$$

$$p\left(\rho|\theta,\lambda,\sigma_{12},\xi^2,Y^\star,Y\right) \propto |R(\rho)| \times \exp\left\{-\frac{1}{2}\left(\frac{1}{\xi^2}\varepsilon_2^{'}\varepsilon_2 - \frac{2\sigma_{12}}{\xi^2}\varepsilon_2^{'}\varepsilon_1\right)\right\}. \tag{3.26}$$

Again, the random walk Metropolis-Hasting algorithm described in (3.16) can be used to generate random draws for these parameters.

The Gibbs samplers outlined so far are summarized in the following algorithms.

Algorithm 1:

1. Let $\left(\theta^0,\Sigma^0,\lambda^0,\rho^0\right)$ be the initial parameter values.
2. Update θ: Sample θ from (3.9).
3. Update Σ: Sample Σ from (3.13).
4. Update λ and ρ: Sample these parameters from (3.14) and (3.15) using a Metropolis-Hasting algorithm.
5. Update $Y_1^\star$ and $Y_2^\star$ using the imputation step.

Algorithm 2:

1. Let $(\theta^0,\sigma_{12}^0,\xi^{2,0},\lambda^0,\rho^0)$ be the initial parameter values.
2. Update θ: Sample θ from (3.9).
3. Update σ_{12}: Sample σ_{12} from (3.24).
4. Update ξ^2: Sample ξ^2 from (3.22).
5. Update λ and ρ: Sample these parameters from (3.25) and (3.26) using a Metropolis-Hasting algorithm.
6. Update $Y_1^\star$ and $Y_2^\star$ using the imputation step.

The Gibbs samplers outlined in Algorithms 1 and 2 operate by splitting the full set of parameters into different blocks. In general, the design of blocks is determined whether the conditional density of each block takes a known form. Gilks et al., (1995) and Chib, (2001) suggest that the set of parameters that are highly correlated should be treated as one block to improve mixing. For the non-spatial sample selection model, van Hasselt, (2005) shows that σ_{12} has a direct effect on δ, implying a high correlation between these parameters. Therefore, we consider an alternative

sampler in which the sampling for δ and σ_{12} is carried out jointly. Given the re-parameterization in (3.19), the model can be written as

$$Y_1^\star = X_1\beta + S^{-1}(\lambda)\varepsilon_1 \tag{3.27}$$

$$Y_2^\star = X_2\delta + R^{-1}(\rho)\left[\sigma_{12}\varepsilon_1 + \eta\right] = Z\gamma + R^{-1}(\rho)\eta \tag{3.28}$$

where $\eta \sim N\left(0, \xi^2 I_n\right)$, $Z = \left(X_2,\, R^{-1}(\rho)\varepsilon_1\right)$, and $\gamma = \left(\delta^{'}, \sigma_{12}\right)^{'}$. Following van Hasselt, (2011), we assume the following prior distributions: $\beta \sim N\left(h_0, H_0\right)$, $\xi^2 \sim IG\left(a_0, b_0\right)$, and $\gamma|\xi^2 \sim N\left(p_0, P_0\right)$ where $p_0 = \left(d_0^{'}, g_0\right)^{'}$ and $P_0 = \begin{pmatrix} D_0 & 0 \\ 0 & \tau\xi^2 \end{pmatrix}$. The prior distributions for the autoregressive parameters are the uniform distributions stated in (3.3) and (3.4). Starting from β, the posterior distribution is given by

$$p\big(\beta|\lambda, Y_1^\star, Y_1\big) \propto N\big(h_0, H_0\big) \times \exp\left\{ -\frac{1}{2}\big(Y_1^\star - X_1\beta\big)^{'} S^{'}(\lambda)S(\lambda)\big(Y_1^\star - X_1\beta\big)\right\}. \tag{3.29}$$

Then, it follows that

$$\beta|\lambda, Y_1^\star, Y_1 \sim N\big(h_1, H_1\big), \tag{3.30}$$

where $h_1 = H_1\big(X_1^{'}S^{'}(\lambda)S(\lambda)Y_1^\star + H_0^{-1}h_0\big)$, and $H_1 = \big(X_1^{'}S^{'}(\lambda)S(\lambda)X_1 + H_0^{-1}\big)^{-1}$. Similarly, the conditional posterior of γ is stated as

$$p\left(\gamma|\beta, \lambda, \rho, \sigma_{12}, \xi^2, Y, Y^\star\right) \propto N(p_0, P_0) \times \exp\left\{ -\frac{1}{2}\xi^{-2}\big(Y_2^\star - Z\gamma\big)^{'} R^{'}(\rho)R(\rho)\big(Y_2^\star - Z\gamma\big)\right\}. \tag{3.31}$$

With the same argument used to derive the result in (3.30), it can be shown that

$$\gamma|\beta, \lambda, \rho, \sigma_{12}, \xi^2, Y, Y^\star \sim N\big(p_1, P_1\big), \tag{3.32}$$

where $p_1 = P_1\big(\xi^{-2}Z^{'}R^{'}(\rho)R(\rho)Y_2^\star + P_0^{-1}p_0\big)$, and $P_1 = \big(\xi^{-2}Z^{'}R^{'}(\rho)R(\rho)Z + P_0^{-1}\big)^{-1}$. Using $\eta = R(\rho)Y_2^\star - R(\rho)Z\gamma$, the conditional posterior of ξ^2 is given by

$$p\left(\xi^2|\beta, \gamma, \lambda, \rho, \sigma_{12}, Y, Y^\star\right) \propto (\xi^2)^{-(\frac{2a_0+n-2}{2}+1)} \exp\left\{ -\frac{1}{\xi^2}\left(b_0 + \frac{1}{2\tau}(\sigma_{12} - g_0)^2 + \frac{1}{2}\eta^{'}\eta\right)\right\}. \tag{3.33}$$

The above result implies that

$$\xi^2|\beta, \gamma, \lambda, \rho, \sigma_{12}, Y, Y^* \sim IG(a_1, b_1), \tag{3.34}$$

where $a_1 = \frac{2a_0+n-2}{2}$ and $b_1 = b_0 + \frac{1}{2\tau}\big(\sigma_{12} - g_0\big)^2 + \frac{1}{2}\eta^{'}\eta$. For the spatial autoregressive parameters,

we have

$$p\left(\lambda|\beta, Y, Y^{\star}\right) \propto |S(\lambda)| \exp\left\{ \left(S(\lambda)Y_1^{\star} - S(\lambda)X_1\beta\right)' \left(S(\lambda)Y_1^{\star} - S(\lambda)X_1\beta\right) \right\} \tag{3.35}$$

$$p\left(\rho|\beta, \gamma, \lambda, \sigma_{12}, \xi^2, Y, Y^{\star}\right) \propto |R(\rho)| \exp\left\{ -\frac{1}{2}\xi^{-2} \left(R(\rho)Y_2^{\star} - R(\rho)Z\gamma\right)' \left(R(\rho)Y_2^{\star} - R(\rho)Z\gamma\right) \right\}. \tag{3.36}$$

Again, the results in (3.35) and (3.36) do not correspond to any known densities. The sampling for these parameters can be accomplished thorough the random walk Metropolis-Hasting algorithm described in (3.16).

Finally, the conditional posteriors of latent variables are required to complete the Gibbs sampler. The model stated in (3.27) and (3.28) suggests that the conditional posterior of $Y_1^{\star}$ is truncated multivariate normal distribution, where the bounds of truncation are determined by cases (i) $Y_{1i} = 0$ and (ii) $Y_{1i} = 1$. Given the formulation in (3.27), we have

$$Y_1^{\star}|\beta, \lambda, Y_1 \sim N\left(X_1\beta,\, S^{-1}(\lambda)S^{'-1}(\lambda)\right), \tag{3.37}$$

subject to (i) $-\infty < Y_{1i}^{\star} \leq 0$ for $i \in \mathcal{N}_0$ and (ii) $0 < Y_{1i}^{\star} < \infty$ for $i \in \mathcal{N}_1$. Once $Y_1^{\star}$ is updated, $Y_2^{\star}$ can be updated by using ε_1 from the current iteration. For the outcome equation, we only need draws for the missing observations. We consider two approaches. In the first approach, we simply draw $Y_2^{\star}$ from $N\left(Z\gamma, \xi^2 R^{-1}(\rho)R^{-1'}(\rho)\right)$ and update $Y_{2i}^{\star}$ for $i \in \mathcal{N}_0$. For the second approach, we assume that $Y_2^{\star} = \left(Y_{2,1}^{\star'}, Y_{2,2}^{\star'}\right)'$, where the first $n_1 \times 1$ block of $Y_{2,1}^{\star}$ contains missing values and the second $n_2 \times 1$ block of $Y_{2,2}^{\star}$ contains the observed observations. As suggested in LeSage and Pace, (2009) for the case of spatial Tobit model, we can generate $Y_{2,1}^{\star}$ conditional on $Y_{2,2}^{\star}$. This conditional distribution can be determined from the following marginal distribution:

$$\begin{pmatrix} Y_{2,1}^{\star} \\ Y_{2,2}^{\star} \end{pmatrix} \Bigg| \beta, \gamma, \lambda, \sigma_{12}, \xi^2 \sim N\Bigg(Z\gamma, \xi^2 R^{-1}(\rho)R^{-1'}(\rho)\Bigg). \tag{3.38}$$

Using (3.38), we have the following partition:

$$\begin{pmatrix} Y_{2,1}^{\star} \\ Y_{2,2}^{\star} \end{pmatrix} \Bigg| \beta, \gamma, \lambda, \sigma_{12}, \xi^2 \sim N\left(\begin{pmatrix} Z_1\gamma_1 \\ Z_2\gamma_2 \end{pmatrix},\, \begin{pmatrix} R_{11}(\rho) & R_{12}(\rho) \\ R_{21}(\rho) & R_{22}(\rho) \end{pmatrix}\right)$$

where $Z_1\gamma_1$ and $Z_2\gamma_2$ are respectively the first $n_1 \times 1$ block and the second $n_2 \times 1$ block of $Z\gamma$. $R_{11}(\rho)$ is the first $n_1 \times n_1$ block of $\xi^2 R^{-1}(\rho)R^{-1'}(\rho)$, and the other elements are defined similarly. The above partition implies the following conditional posterior distribution:

$$\begin{aligned} &Y_{2,1}^{\star}\big|\beta, \gamma, \lambda, \sigma_{12}, \xi^2, Y_{2,2}^{\star}, Y \\ &\sim N\Bigg(Z_1\gamma_1 + R_{12}(\rho)R_{22}^{-1}(\rho)\left[Y_{2,2}^{\star} - Z_2\gamma_2\right],\, R_{11}(\rho) - R_{12}(\rho)R_{22}^{-1}(\rho)R_{21}(\rho)\Bigg). \end{aligned} \tag{3.39}$$

We summarize this alternative sampler in the following algorithm.

Algorithm 3:

1. Let $(\beta^0, \gamma^0, \xi^{2,0}, \lambda^0, \rho^0)$ be the initial parameter values.
2. Update β: Sample β from (3.30).

3. Update γ: Sample γ from (3.32).

4. Update ξ^2: Sample ξ^2 from (3.34).

5. Update λ and ρ: Sample these parameters from (3.35) and (3.36) using a Metropolis-Hasting algorithm.

6. Update $Y_1^\star$ and $Y_2^\star$: Sample latent variables using (3.37) and (3.39).

Algorithms 2 and 3 are based on the re-parameterization scheme considered for Σ. The re-parameterization approach can be seen as a special form of the more general method called "the parameter expansion method" or "the marginal data augmentation method" considered in Liu and Wu, (1999), Meng and van Dyk, (1999), and van Dyk and Meng, (2001). In this method, the model is expanded with an expansion parameter to improve the convergence rate of the resulting Markov chains. We consider this approach for our spatial sample selection model. Let α^2 be an expansion parameter that can be identified given the augmented data $(Y^\star, Y)$ with a prior of $p(\alpha^2|\Sigma)$. Then, we can introduce the expansion parameter into our data augmentation sampling scheme of (3.5) through

$$p(\theta, \lambda, \rho, \Sigma, Y^\star\big|Y) \propto p(\theta)p(\lambda)p(\rho)p(\Sigma) \int p(Y^\star, Y|\theta, \lambda, \rho, \Sigma, \alpha^2)p(\alpha^2|\Sigma)\mathrm{d}\alpha^2. \tag{3.40}$$

In (3.40), the marginalization over α^2 yields the same joint augmented posterior stated in (3.5) without any expansion parameter. The computational advantages of (3.40) are discussed in details in Liu and Wu, (1999) and Meng and van Dyk, (1999). To make the approach in (3.40) operational, Meng and van Dyk, (1999) suggest a general two-step marginalization strategy such that the convergence rate of the resulting Markov chains is better than any other scheme. In the context of our spatial model, this two-step strategy can also be seen as an alternative way to circumvent the computational problems that are due to the identification condition, $\Sigma_{11} = \sigma_1^2 = 1$. In the first step, the spatial sample selection model is transformed by the expansion parameter in such a way that the transformed model has an unconstrained covariance matrix so that the computational complications are circumvented for the posterior analysis. In the second step, an inverse Wishart distribution is assigned as a prior to the unconstrained covariance matrix. At this step, the priors for the expansion parameter and the constrained covariance matrix are also determined to complete a Gibbs sampler.

In the literature, this two-step strategy is applied to non-spatial multinomial probit models in Imai and van Dyk, (2005) and Burgette and Nordheim, (2012). These authors show that the algorithm based on this method is better in terms of convergence rate. Talhouk et al., (2012) suggest an efficient Bayesian MCMC algorithm based on the parameter expansion method for non-spatial multivariate probit models. Recently, Ding, (2014) applied a scheme based on the parameter expansion method to a non-spatial sample selection method and shows that an MCMC algorithm with this scheme can perform better. Hence, it is of interest to consider the same approach for our spatial model. In the following, we consider two new samplers in which the expansion parameter is handled differently.

In order to write the model in (2.6) in terms of unconstrained covariance matrix, define

$$E^\star = \begin{pmatrix} \sigma_1 I_n & 0_{n\times n} \\ 0_{n\times n} & I_n \end{pmatrix} \times \big(Y^\star - X\theta\big), \tag{3.41}$$

where the expansion parameter is $\alpha = \sigma_1$. Given our bivariate normal distribution assumption, we have $E^\star|\theta, \mathfrak{B} \sim N\left(0_{2n\times 2n},\, \mathcal{Q}\right)$, where $\mathcal{Q} = \mathcal{K}(\lambda, \rho)(\mathfrak{B} \otimes I_n)\mathcal{K}^{'}(\lambda, \rho)$, and $\mathfrak{B}$ is the unconstrained

covariance matrix of $(\varepsilon_{1i}, \varepsilon_{2i})'$, namely

$$\mathfrak{B} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & 1 \end{pmatrix} \times \Sigma \times \begin{pmatrix} \sigma_1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \tag{3.42}$$

where $\varrho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$. Now, the inverse Wishart distribution can be assigned as a prior to $\mathfrak{B}$, namely $\mathfrak{B} \sim \text{InvWish}(v_0, I_2)$. Given the transformation in (3.42), the priors for Σ and σ_1^2, i.e., the derived densities, can be deduced from $\text{InvWish}(v_0, I_2)$. These *derived* priors are given by

$$p(\Sigma) \propto \left(1-\varrho^2\right)^{-3/2} \sigma_2^{-(v_0+3)} \exp\left\{ -\frac{1}{2\sigma_2^2(1-\varrho^2)} \right\}, \tag{3.43}$$

$$\sigma_1^2|\Sigma \sim \left((1-\varrho^2)\chi^2_{v_0} \right)^{-1}. \tag{3.44}$$

Given our normal distribution assumption for $E^\star|\theta, \mathfrak{B} \sim N(0_{2n\times 2n},\, \mathcal{Q})$, the conditional posterior of $\mathfrak{B}$ is given by

$$p(\mathfrak{B}|\theta, \lambda, \rho, Y^\star, Y, E^\star) \propto |\mathfrak{B}|^{-(v_0+3)/2} \exp\left\{ -\frac{1}{2} tr\big(I_2\mathfrak{B}^{-1}\big) \right\} \times |\mathfrak{B}|^{-n/2} \exp\left\{ -\frac{1}{2} E^{\star\prime}\mathcal{Q}^{-1}E^\star \right\}. \tag{3.45}$$

Using $\varepsilon_1 = S(\lambda)Y_1^\star - S(\lambda)X_1\beta$ and $\varepsilon_2 = R(\rho)Y_2^\star - R(\rho)X_1\delta$, the quadratic term $E^{\star\prime}\mathcal{Q}^{-1}E^\star$ in (3.45) can be written as

$$E^{\star\prime}\mathcal{Q}^{-1}E^\star = tr\Big(E^{\star\prime}\mathcal{Q}^{-1}E^\star \Big) = \text{tr}\Big(A_2 \times \mathfrak{B}^{-1} \Big), \tag{3.46}$$

where $A_2 = \begin{pmatrix} \sigma_1^2\varepsilon_1'\varepsilon_1 & \sigma_1\varepsilon_1'\varepsilon_2 \\ \sigma_1\varepsilon_2'\varepsilon_1 & \varepsilon_2'\varepsilon_2 \end{pmatrix}$. Substituting (3.46) into (3.45) yields

$$\mathfrak{B}|\theta, \rho, \lambda, \Sigma, Y^\star, Y, E^\star \sim \text{InvWish}\big(v_1, T_1\big), \tag{3.47}$$

where $v_1 = n + v_0$ and $T_1 = I_2 + A_2$. To marginalize over σ_1^2, we draw $\mathfrak{B}$ from $\text{InvWish}(v_1, T_1)$ and set $\sigma_1^2 = \mathfrak{B}_{11}$, where $\mathfrak{B}_{11}$ is the $(1,1)$th element of $\mathfrak{B}$. Then, we set

$$\Sigma = \begin{pmatrix} 1/\sigma_1 & 0 \\ 0 & 1 \end{pmatrix} \times \mathfrak{B} \times \begin{pmatrix} 1/\sigma_1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.48}$$

This new approach can be summarized by the following algorithm.

Algorithm 4:

1. Let $(\theta^0, \Sigma^0, \lambda^0, \rho^0)$ be the initial parameter values.
2. Update θ: Sample θ from (3.9).
3. Update Σ: Use the following steps
   (a) Draw σ_1^2 from its prior $\left((1-\varrho^2)\chi^2_{v_0}\right)^{-1}$.
   (b) Apply the transformation in (3.41) to obtain $E^\star$.
   (c) Sample $\mathfrak{B}$ from (3.47).

(d) Set $\sigma_1^2 = \mathfrak{B}_{11}$ and recover Σ by

$$\Sigma = \begin{pmatrix} 1/\sigma_1 & 0 \\ 0 & 1 \end{pmatrix} \times \mathfrak{B} \times \begin{pmatrix} 1/\sigma_1 & 0 \\ 0 & 1 \end{pmatrix}.$$

4. Update λ and ρ: Sample these parameters from (3.14) and (3.15) using a Metropolis-Hasting algorithm.

5. Update $Y_1^\star$ and $Y_2^\star$ using the imputation step.

In the above algorithm, although the sampled $\mathfrak{B}$ depends on the unidentifiable parameter σ_1^2, there is a complete marginalization over σ_1^2 in updating Σ. There is an alternative approach in Meng and van Dyk, (1999) and Imai and van Dyk, (2005), where the conditional posterior distribution of the expansion parameter, which is σ_1^2 in our case, is used in the sampler. Meng and van Dyk, (1999) and van Dyk and Meng, (2001) called the algorithm in which the distribution of the working parameters is used "the marginal data augmentation algorithm." Following Imai and van Dyk, (2005), we also consider a similar marginal data augmentation algorithm for our model. Transforming our model in (2.6) by multiplying with a positive scalar parameter α yields

$$\alpha Y^\star \equiv \tilde{Y}^\star = X\tilde{\theta} + \tilde{\varepsilon}, \tag{3.49}$$

where $\tilde{\theta} = \alpha\theta$ and $\tilde{\varepsilon} = \alpha\varepsilon$. Meng and van Dyk, (1999) called the parameter α a "working parameter" (it is called an "expansion parameter" in Liu and Wu, (1999)). A working parameter is an unidentifiable parameter but is identifiable in the expanded parameter space of a data augmentation algorithm. Denote the covariance of $\tilde{\varepsilon}$ with $\mathfrak{D}$. Given the transformation in (3.49), we have $\mathfrak{D} = \alpha^2\Omega = \mathcal{K}(\lambda,\rho)\big(\alpha^2\Sigma \otimes I_n\big)\mathcal{K}'(\lambda,\rho)$. Define $\tilde{\Sigma} = \alpha^2\Sigma$ as the new unconstrained covariance matrix. Now, the inverse Wishart distribution can be assigned as a prior to $\tilde{\Sigma}$, namely $\tilde{\Sigma} \sim \text{InvWish}(v_0, S_0)$. The transformation in (3.49) and the prior $\tilde{\Sigma} \sim \text{InvWish}(v_0, S_0)$ imply the following joint prior density:

$$p(\Sigma, \alpha^2) = \big(\alpha^2\big)^{-(\frac{2v_0}{2}+1)} |\Sigma|^{-(v_0+3)/2} \exp\Big\{ -\frac{1}{2\alpha^2}\text{tr}\big(S_0\Sigma^{-1}\big)\Big\}. \tag{3.50}$$

The joint prior in (3.50) implies the following derived prior densities (for details, see the web appendix)

$$p(\Sigma) \propto |\Sigma|^{-(v_0+3)/2} \times \big[\text{tr}\big(S_0\Sigma^{-1}\big)\big]^{-v_0}, \tag{3.51}$$

$$\alpha^2|\Sigma \sim \text{tr}\big(S_0\Sigma^{-1}\big)/\chi^2_{2v_0}. \tag{3.52}$$

For the posterior analysis, we consider the following prior specifications: $\tilde{\theta}|\alpha \sim N(0, \alpha^2 V_0)$, $\tilde{\Sigma} \sim \text{InvWish}(v_0, S_0)$, $\alpha^2|\Sigma \sim tr(S_0\Sigma^{-1})/\chi^2_{2v_0}$. With these priors, a new algorithm can be developed by considering the transformed model in (3.49). This algorithm, besides the imputation step and the conditional posterior distributions of autoregressive parameters, requires (i) the joint conditional posterior distribution of $\tilde{\theta}$ and α^2, and (ii) the conditional posterior distribution of $\tilde{\Sigma}$. To this end, $\tilde{\theta}$ and α^2 is sampled from the joint conditional posterior density, $p(\tilde{\theta}, \alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y) = p(\tilde{\theta}|\Sigma, \lambda, \rho, \alpha^2, \tilde{Y}^\star, Y) \times p(\alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y)$, which can be done by generating a draw of α^2 from $p(\alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y)$ and then using this draw to sample $\tilde{\theta}$ from $p(\tilde{\theta}|\Sigma, \lambda, \rho, \alpha^2, \tilde{Y}^\star, Y)$. Hence, the algorithm will be completed once $p(\tilde{\theta}|\Sigma, \lambda, \rho, \alpha^2, \tilde{Y}^\star, Y)$, $p(\alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y)$ and $p(\tilde{\Sigma}|\tilde{\theta}, \alpha^2, \lambda, \rho, \tilde{Y}^\star, Y)$ are determined. The conditional posterior densities of $\tilde{\theta}$ and $\tilde{\Sigma}$ can be easily determined with a

similar way used in the previous algorithms. The conditional posterior density, $p(\alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y)$ can be determined from

$$p(\alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y) = \frac{p(\tilde{\alpha}, \tilde{\theta}|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y)}{p(\tilde{\theta}|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y)}. \tag{3.53}$$

Using (3.53), it can be shown that

$$\begin{aligned} &\alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y \\ &\sim \frac{\big(\tilde{Y}^\star - X\theta_{gls}\big)^{'}\Omega^{-1}\big(\tilde{Y}^\star - X\theta_{gls}\big) + \theta'_{gls}\big(V_0 + \big[X'\Omega^{-1}X\big]^{-1}\big)^{-1}\theta_{gls} + \mathrm{tr}\big(S_0\Sigma^{-1}\big)}{\chi^2_{2(n+v_0)}}. \end{aligned} \tag{3.54}$$

where $\theta_{gls} = \big(X'\Omega^{-1}X\big)^{-1}X'\Omega^{-1}\tilde{Y}^\star$. The working parameter α is marginalized out in different ways to recover θ and Σ as shown in the following. The steps of this new algorithm at iteration t can be summarized in the following algorithm.[14]

Algorithm 5:

1. Let $(\theta^0, \Sigma^0, \lambda^0, \rho^0)$ be the initial parameter values.
2. Update $Y_1^\star$ and $Y_2^\star$ by using the imputation step. This step is the same as the last step of Algorithm 1. Let $Y^\star = (Y_1^\star, Y_2^\star)$ be the updated vector.
3. Draw α^2 from (3.52) and set $\tilde{Y}^\star = \alpha Y^\star$.
4. Update $\tilde{\theta}$ and α^2: Calculate $\theta_{gls} = \big(X'\Omega^{-1}X\big)^{-1}X'\Omega^{-1}\tilde{Y}^\star$ with $\Omega = \mathcal{K}(\lambda^{t-1}, \rho^{t-1})\big(\Sigma^{t-1} \otimes I_n\big)\mathcal{K}'(\lambda^{t-1}, \rho^{t-1})$, where λ^{t-1}, ρ^{t-1} and Σ^{t-1} are obtained from the previous iteration.
   (a) Draw α using

$$\begin{aligned} &\alpha^2|\Sigma, \lambda, \rho, \tilde{Y}^\star, Y \\ &\sim \frac{\big(\tilde{Y}^\star - X\theta_{gls}\big)^{'}\Omega^{-1}\big(\tilde{Y}^\star - X\theta_{gls}\big) + \theta'_{gls}\big(V_0 + \big[X'\Omega^{-1}X\big]^{-1}\big)^{-1}\theta_{gls} + \mathrm{tr}\big(S_0\Sigma^{-1}\big)}{\chi^2_{2(n+v_0)}}. \end{aligned} \tag{3.55}$$

   (b) Draw $\tilde{\theta}$ using

$$\tilde{\theta}|\Sigma, \alpha^2, \lambda, \rho, Y^\star, Y \sim N\Big(\mu_1,\, \alpha^2\big(V_0^{-1} + X'\Omega^{-1}X\big)^{-1}\Big). \tag{3.56}$$

   Set $\theta^t = \tilde{\theta}/\alpha$ as the sampled value at iteration t.
5. Draw $\tilde{\Sigma}$: Calculate $\tilde{\varepsilon}_1 = S(\lambda^{t-1})\tilde{Y}_1^\star - S(\lambda^{t-1})X_1\tilde{\beta}$, $\tilde{\varepsilon}_2 = R(\rho^{t-1})\tilde{Y}_2^\star - R(\rho^{t-1})X_1\tilde{\delta}$ and $\tilde{A}_1 = \begin{pmatrix} \tilde{\varepsilon}_1^{'}\tilde{\varepsilon}_1 & \tilde{\varepsilon}_1^{'}\tilde{\varepsilon}_2 \\ \tilde{\varepsilon}_2^{'}\tilde{\varepsilon}_1 & \tilde{\varepsilon}_2^{'}\tilde{\varepsilon}_2 \end{pmatrix}$. Draw $\tilde{\Sigma}$ using

$$\tilde{\Sigma}|\tilde{\theta}, \alpha^2, \lambda, \rho, \tilde{Y}^\star, Y \sim \mathrm{InvWish}(v_0 + n,\, S_0 + \tilde{A}_1), \tag{3.57}$$

   Finally set $\Sigma^t = \frac{1}{\tilde{\Sigma}_{11}} \times \tilde{\Sigma}$, and $Y^{\star,t} = \frac{1}{\sqrt{\tilde{\Sigma}_{11}}} \times \tilde{Y}^\star$ as the sampled values for iteration t.

[14] See the web appendix for details.

6. Update λ and ρ conditional on θ^t, Σ^t and $Y^{\star,t}$: Sample these parameters from (3.14) and (3.15) using a Metropolis-Hasting algorithm.

Note that the working parameter is completely marginalized after Step 5 in Algorithm 5. Meng and van Dyk, (1999) and Imai and van Dyk, (2005) consider another sampling scheme in which the working parameter is not sampled from its prior but instead is sampled from its conditional posterior distribution. For example, we could record draws of α^2 from (3.55) in the sampler for the next iterations. Imai and van Dyk, (2005) show that the scheme we outlined in Algorithm 5 outperforms some other schemes that can be considered for the sampler. Hence, we only consider the scheme in Algorithm 5.[15]

Algorithms 4 and 5 handle the working parameter in different ways. In terms of specification, the working parameter (or expansion parameter) is defined in a more general way in Algorithm 5. The difference in specifications imply different functional relationships between the constrained and the unconstrained covariances. Therefore, the *derived* conditional prior of working parameter stated in (3.44) for Algorithm 4 is different from the one stated in (3.52) for Algorithm 5. There are different ways in which the working parameter is marginalized out (or swept over) in these algorithms. It is obvious that the working parameter is more active in Algorithm 5 implying a higher variability in the augmented data. This additional variability in the augmented data may allow the Gibbs sampler to move around the parameter space more quickly (Imai and van Dyk, 2005).

Table 1: Prior hyper-parameters and Initial Values

	Hyper-parameters	Initial values
Algorithm 1:	$\theta_0 = 0_{4\times1}$, $V_0 = 10^3 \times I_4$ $v_0 = 3$, $T_0 = 8 \times I_2$ $\kappa_1 = \kappa_2 = 1$	$\theta^0 = (X^{'}X)^{-1}X^{'}Y$ $\Sigma^0 = [1, 0.25; 0.25, 1]$ $\lambda^0 = 0.25$, $\rho^0 = 0.25$
Algorithm 2:	$\theta_0 = 0_{4\times1}$, $V_0 = 10^3 \times I_4$ $\kappa_1 = \kappa_2 = 1$ $\tau = 0.7$, $a_0 = 1$, $b_0 = 1$	$\theta^0 = (X^{'}X)^{-1}X^{'}Y$ $\lambda^0 = 0.25$, $\rho^0 = 0.25$
Algorithm 3:	$h_0 = 0_{2\times1}$, $H_0 = 10^3 \times I_2$ $d_0 = 0_{2\times1}$,$g_0 = 0$, $D_0 = 10^3 \times I_2$ $\kappa_1 = \kappa_2 = 1$ $\tau = 0.7$, $a_0 = 1$, $b_0 = 1$	$\beta^0 = (X_1^{'}X_1)^{-1}X_1^{'}Y_1$,$\delta^0 = (X_2^{'}X_2)^{-1}X_2^{'}Y_2$ $\lambda^0 = 0.25$, $\rho^0 = 0.25$
Algorithm 4:	$\theta_0 = 0_{4\times1}$, $V_0 = 10^3 \times I_4$ $\kappa_1 = \kappa_2 = 1$, $v_0 = 3$	$\theta^0 = (X^{'}X)^{-1}X^{'}Y$ $\Sigma^0 = [1, 0.25; 0.25, 1]$ $\lambda^0 = 0.25$, $\rho^0 = 0.25$
Algorithm 5:	$\theta_0 = 0_{4\times1}$, $V_0 = 10^3 \times I_4$ $\kappa_1 = \kappa_2 = 1$ $v_0 = 8$, $S_0 = 8 \times I_2$	$\theta^0 = (X^{'}X)^{-1}X^{'}Y$ $\Sigma^0 = [1, 0.25; 0.25, 1]$ $\lambda^0 = 0.25$, $\rho^0 = 0.25$

[15]Our algorithm corresponds to Algorithm 1 of Imai and van Dyk, (2005) with their Scheme 1. Algorithm 1 with Scheme 1 in Imai and van Dyk, (2005) is same as with the PX-DA Algorithm with Scheme 1.1 in Liu and Wu, (1999). In the web appendix, we provide an alternative algorithm which corresponds to Algorithm 2 of Imai and van Dyk, (2005).

4 Simulation Study

To evaluate the finite sample properties of Gibbs samplers in Algorithms 1-5, we design a simulation study in this section. We consider the following data generating process (DGP):

$$\begin{pmatrix} Y_1^\star \\ Y_2^\star \end{pmatrix} = \begin{pmatrix} X_1 & 0_{n\times k_2} \\ 0_{n\times k_1} & X_2 \end{pmatrix} \begin{pmatrix} \beta \\ \delta \end{pmatrix} + \begin{pmatrix} S^{-1}(\lambda)\varepsilon_1 \\ R^{-1}(\rho)\varepsilon_2 \end{pmatrix}. \tag{4.1}$$

where $X_1 = (l_n', X_{1,1}')'$ with $\beta = (\beta_1, \beta_2)'$, and $X_2 = (l_n', X_{2,1}')'$ with $\delta = (\delta_1, \delta_2)'$. The exogenous variable $X_{1,1}$ consists of random draws from $U(0,1)$ whereas $X_{2,1}$ is generated from $N(0,1)$.[16] The innovations are generated from the bivariate normal distribution according to

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1.2 \end{pmatrix} \right) \tag{4.2}$$

for $i = 1, \ldots, n$. Hence, $\varrho = 0.25$. For the true parameter values, we set $(\delta_1, \delta_2)' = (0.4, 1.2)'$, and $\beta_2 = 2$. We use β_1 to control the amount of the sample selection in the model. We set $\beta_1 = -0.2$ to generate 25% censuring. We consider $(\lambda, \rho) = \{(0.1, 0.1), (0.4, 0.4)\}$ for the autoregressive parameters to allow for weak and moderate dependence in the error processes.

The row normalized W and M are based on the interaction scenario described in Liu and Lee, (2010). Both matrices are block diagonal matrices where each block represents the interaction structure of a group. Let the total sample involve R groups where the rth group has the groups size m_r. We consider an experiment where R is set to 30. We allow m_r to vary across R groups by randomly assigning a value from the set of integers $\{10, 11, 12, 13, 14, 15\}$ to each group size. Therefore, the total number of observations n varies between 300 and 450. The weight matrix W_r for the rth group is generated in two steps. First, an integer value τ_{ir} is uniformly drawn from the set of integer values $\{1, 2, 3, 4\}$. Then, if $\tau_{ir} + i \leq m_r$, the $(i+1)$th, $\ldots$, $(i+\tau_{ir})$th elements of the ith row of W_r are set to one and the rest of the elements in the ith row are set to zero. On the other hand, if $\tau_{ir} + i > m_r$, the first $(\tau_{ir} + i - m_r)$ entries of the ith row are set to one and the other elements of the ith row are set to zero. Then, $W = M = \text{Diag}(W_1, \ldots, W_R)$.

The hyper-parameter values and the initial parameter values we used in the simulation are stated in Table 1. These values are close to those used in Imai and van Dyk, (2005) in a simulation for a multinomial probit model. In Algorithms 2 and 3, we set $\tau = 0.7$, which generates a relatively diffuse prior for the correlation coefficient between ε_{1i} and ε_{2i}.[17] Finally, we run each MCMC algorithm for 60000 iterations. The first 10000 draws are discarded as a burn in period.

4.1 Simulation Results

The simulation results are presented in Tables 2-3, and Figures 1 – 4. We compare point estimates to gauge statistical inference implied by each algorithm in Tables 2 and 3. We provide (i) the mean of sampled draws, i.e., the estimated posterior means, (ii) the standard deviations of sampled draws (Std.dev.) and (iii) some other convergence diagnostics. The convergence diagnostics include, the Gelman-Rubin (GR) statistic, the first three lag-correlations in sampled draws (denoted by AC(1),

[16] Pace et al., (2012) show that the explanatory variables used in applied studies exhibit spatial dependence. As shown in Pace et al., (2012), the spatial dependence among regressors can also effect the performance of the likelihood-based estimators. It will be interesting to see the effect of spatial dependence among regressors on the Bayesian estimator. This issue is raised by a referee and can be explored in future studies.

[17] The implied prior density of ϱ for different values of τ is illustrated in the web appendix. A roughly uniform prior for ϱ can be induced when $\tau \in [0.6, 0.7]$.

Table 2: Posterior summary: $\lambda_0 = 0.1$ and $\rho_0 = 0.1$

Parameter	Mean	Std.dev.	GR	AC(1)	AC(2)	AC(3)	IF
Algorithm 1							
λ	0.0836	0.1586	1.0006	0.9320	0.8684	0.8095	26.4242
ρ	0.1407	0.0761	1.0000	0.6840	0.4657	0.3181	5.2606
σ_{12}	-0.0006	0.0923	1.0000	0.4730	0.2808	0.1680	3.2418
σ_2^2	1.2653	0.1084	1.0000	0.2538	0.0710	0.0149	1.6795
β_1	-0.1066	0.1429	1.0000	0.4552	0.2283	0.1279	2.9592
β_2	1.8529	0.2798	1.0002	0.6312	0.4180	0.2864	4.7646
δ_1	0.5266	0.0841	1.0000	0.3530	0.2005	0.1231	2.5820
δ_2	1.2549	0.0685	1.0000	0.2609	0.0652	0.0239	1.6998
Algorithm 2							
λ	0.0840	0.1603	1.0003	0.9364	0.8770	0.8218	27.0106
ρ	0.1459	0.0659	1.0000	0.6888	0.4707	0.3192	5.0085
σ_{12}	0.0030	0.0822	1.0000	0.4763	0.2836	0.1720	3.2778
ξ^2	0.8752	0.0696	1.0000	-0.1739	0.0305	-0.0045	1.0000
β_1	-0.1070	0.1426	1.0000	0.4463	0.2192	0.1243	2.8113
β_2	1.8563	0.2800	1.0000	0.6251	0.4071	0.2769	4.8148
δ_1	0.5257	0.0704	1.0000	0.3557	0.2075	0.1192	2.5831
δ_2	1.2556	0.0571	1.0000	0.2646	0.0719	0.0257	1.7243
Algorithm 3							
λ	0.0856	0.1565	1.0002	0.9547	0.9125	0.8724	70.3474
ρ	0.1460	0.0647	1.0001	0.7127	0.5080	0.3624	5.8613
σ_{12}	0.0038	0.0816	1.0000	0.4660	0.2713	0.1636	3.2068
ξ^2	0.8689	0.0691	1.0000	-0.1770	0.0413	-0.0055	1.0000
β_1	-0.1071	0.1438	1.0001	0.4447	0.2186	0.1179	2.8074
β_2	1.8569	0.2806	1.0001	0.6222	0.4069	0.2728	4.6321
δ_1	0.5248	0.0703	1.0000	0.3576	0.2019	0.1153	2.5566
δ_2	1.2560	0.0571	1.0001	0.2628	0.0688	0.0221	1.7075
Algorithm 4							
λ	0.0819	0.1566	1.0008	0.9617	0.9251	0.8900	51.7106
ρ	0.1409	0.0760	1.0001	0.6945	0.4835	0.3391	5.4550
σ_{12}	0.0012	0.0931	1.0000	0.4767	0.2775	0.1648	3.1767
σ_2^2	1.2414	0.1061	1.0000	0.2480	0.0651	0.0190	1.6643
β_1	-0.1067	0.1422	1.0000	0.4426	0.2140	0.1125	2.7783
β_2	1.8545	0.2802	1.0000	0.6259	0.4100	0.2770	4.7014
δ_1	0.5262	0.0833	1.0000	0.3520	0.2032	0.1211	2.5676
δ_2	1.2548	0.0681	1.0000	0.2657	0.0750	0.0209	1.7232
Algorithm 5							
λ	0.2804	0.1427	1.0002	0.9029	0.8132	0.7333	17.8891
ρ	0.1427	0.0825	1.0003	0.7197	0.5149	0.3655	5.7368
σ_{12}	0.0091	0.1176	1.0000	0.5323	0.2566	0.0676	2.6462
σ_2^2	1.5528	1.1383	1.0000	0.7511	0.3920	0.1017	2.5470
β_1	-0.0835	0.1554	1.0000	0.4409	0.2074	0.0900	2.5324
β_2	1.7992	0.3668	1.0000	0.6350	0.2729	0.0058	2.8274
δ_1	0.5354	0.1672	1.0000	0.6900	0.4335	0.1940	3.6725
δ_2	1.2826	0.3755	1.0001	0.8096	0.4779	0.1509	2.6257

AC(2), and AC(3)), and the inefficiency factors (IF). The inefficiency factor (or autocorrelation time) is defined as the ratio of the squared numerical standard errors to the variance of the posterior mean based on the hypothetical independent draws (Chib, 2001, p. 3580).[18] An efficient sampler would generate a sequence of sampled draws with a small IF close to 1.

We also provide several graphical summaries of draws in Figures 1 – 4 to compare algorithms in terms of convergence and mixing properties of the corresponding Markov chains. For the sake of brevity, we only provide the figures for Algorithms 1 and 4 for the case of $\{\lambda, \rho\} = \{0.4, 0.4\}$.[19] These figures include the convergence plots (or time series plots), the autocorrelation plots, the lag-one scatter plots, and the Gelman and Rubin, (1992)'s $\sqrt{R}$ statistic. The convergence plots in these figures for a parameter simply show all draws generated by the samplers against iterations. The autocorrelation plots provide a simple visual inspection of the lag-correlation in sampled draws

[18] It is calculated as $\text{IF}(\overline{\theta}) = 1 + 2\sum_{k=1}^{M-1}\left(1 - \frac{k}{M}\right)\rho_k$, where M is the total number of draws, and ρ_k is the lag k sample autocorrelation.

[19] Figures for other algorithms and the case of $\{\lambda, \rho\} = \{0.1, 0.1\}$ are similar and provided in the web appendix.

Table 3: Posterior summary: $\lambda_0 = 0.4$ and $\rho_0 = 0.4$

Parameter	Mean	Std.dev.	GR	AC(1)	AC(2)	AC(3)	IF
Algorithm 1							
λ	0.4025	0.0993	1.0002	0.8975	0.8068	0.7256	17.8230
ρ	0.3147	0.0716	1.0001	0.6898	0.4727	0.3230	5.1795
σ_{12}	-0.0086	0.0998	1.0000	0.5205	0.3271	0.2084	3.7044
σ_2^2	1.2991	0.1194	1.0000	0.3122	0.1072	0.0306	1.9002
β_1	-0.2511	0.1604	1.0000	0.4170	0.2116	0.1251	2.8534
β_2	1.9199	0.2788	1.0001	0.6477	0.4286	0.2879	4.7719
δ_1	0.6006	0.1089	1.0000	0.3974	0.2420	0.1551	2.9608
δ_2	1.2609	0.0675	1.0000	0.2442	0.0618	0.0208	1.6535
Algorithm 2							
λ	0.3997	0.0991	1.0000	0.8886	0.7895	0.7030	16.7897
ρ	0.3271	0.0616	1.0000	0.6811	0.4569	0.3049	4.8734
σ_{12}	-0.0078	0.0879	1.0001	0.5261	0.3330	0.2169	3.7923
ξ^2	0.8738	0.0727	1.0000	-0.2088	0.0494	-0.0079	1.0000
β_1	-0.2500	0.1596	1.0000	0.4064	0.2024	0.1212	2.6977
β_2	1.9184	0.2793	1.0001	0.6426	0.4190	0.2817	4.8156
δ_1	0.6020	0.0910	1.0001	0.4019	0.2476	0.1534	2.9784
δ_2	1.2621	0.0557	1.0000	0.2464	0.0603	0.0188	1.6511
Algorithm 3							
λ	0.3976	0.0980	1.0002	0.9368	0.8782	0.8246	44.8648
ρ	0.3280	0.0631	1.0000	0.6722	0.4522	0.3052	4.9678
σ_{12}	-0.0063	0.0876	1.0001	0.5262	0.3336	0.2127	3.7699
ξ^2	0.8680	0.0724	1.0000	-0.2140	0.0509	-0.0073	1.0000
β_1	-0.2556	0.1591	1.0001	0.4044	0.2020	0.1176	2.6693
β_2	1.9301	0.2789	1.0002	0.6400	0.4184	0.2815	4.7345
δ_1	0.6008	0.0913	1.0000	0.4013	0.2469	0.1556	3.0066
δ_2	1.2625	0.0562	1.0001	0.2435	0.0640	0.0232	1.6613
Algorithm 4							
λ	0.3986	0.0998	1.0000	0.8787	0.7731	0.6815	14.7438
ρ	0.3159	0.0705	1.0000	0.7051	0.4998	0.3543	5.5756
σ_{12}	-0.0082	0.1003	1.0000	0.5302	0.3286	0.2100	3.7459
σ_2^2	1.2730	0.1179	1.0000	0.3270	0.1193	0.0457	1.9840
β_1	-0.2496	0.1604	1.0000	0.3985	0.1892	0.0996	2.5818
β_2	1.9155	0.2797	1.0000	0.6383	0.4147	0.2711	4.6257
δ_1	0.6015	0.1077	1.0000	0.3977	0.2442	0.1566	2.9564
δ_2	1.2608	0.0673	1.0000	0.2470	0.0605	0.0109	1.6368
Algorithm 5							
λ	0.4355	0.0968	1.0001	0.8525	0.7247	0.6160	11.5732
ρ	0.3024	0.0801	1.0000	0.6910	0.4745	0.3215	5.1716
σ_{12}	-0.0013	0.1300	1.0000	0.5915	0.3172	0.1106	3.0104
σ_2^2	1.5484	1.1086	1.0000	0.7708	0.4316	0.1425	2.6178
β_1	-0.2440	0.1703	1.0000	0.4299	0.2000	0.0833	2.4710
β_2	1.8995	0.3649	1.0000	0.6445	0.2926	0.0306	2.9356
δ_1	0.6024	0.1895	1.0000	0.6838	0.4593	0.2437	3.8644
δ_2	1.2664	0.3744	1.0000	0.8170	0.5018	0.1827	2.5420

and can be used to assess the mixing properties of the sampler. Low autocorrelation in sampled draws means that the sampler is expected to converge to the posterior distribution more quickly. Hence, slowly decaying autocorrelations in these figures indicate inefficiencies for a sampler. The $\sqrt{R}$ statistic is defined as the ratio of the between-chain variance and the within-chain variance; and values close to one indicate acceptable mixing.

We provide the salient features of our results in the following list.

1. The results in Tables 2 and 3 indicate that the estimated posterior means of autoregressive parameters are close to true parameter values, except in the case of Algorithm 5 for λ. These results become visually available through the convergence plots provided in Figures 1 – 4. The sampled draws generated for ρ are slightly off-centered from the true parameter value.

2. The Bayesian point estimates for σ_{12} are far away from the true value in all algorithms. As can be seen from the convergence plots provided in Figures 1 – 4, this parameter is underestimated in all algorithms. The samplers in Algorithms 2 and 3 produce similar estimates, in particular,

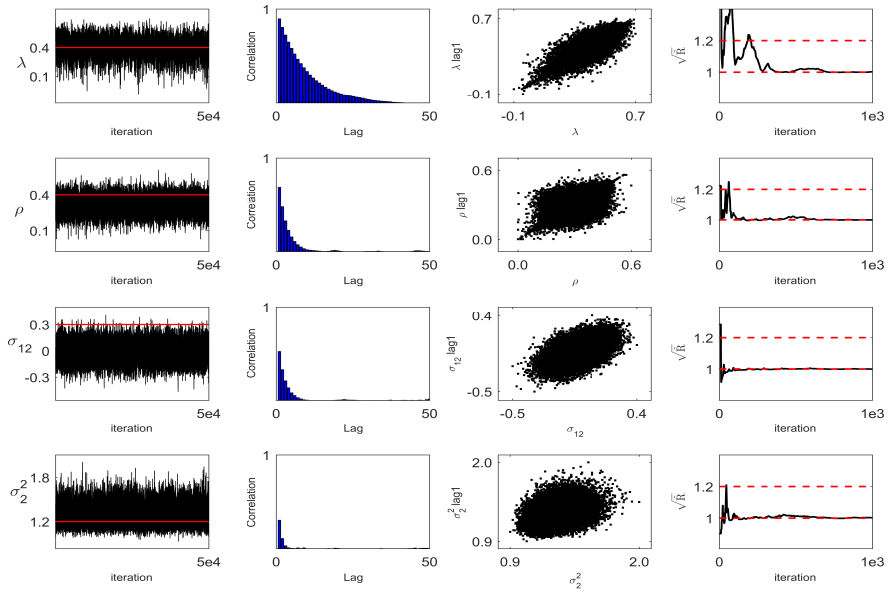


Figure 1: Algorithm 1

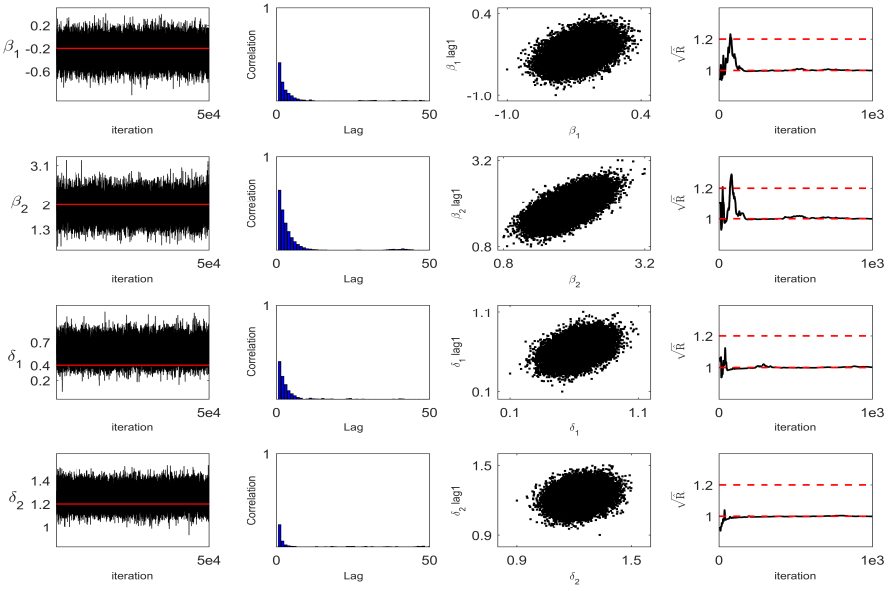


Figure 2: Algorithm 1

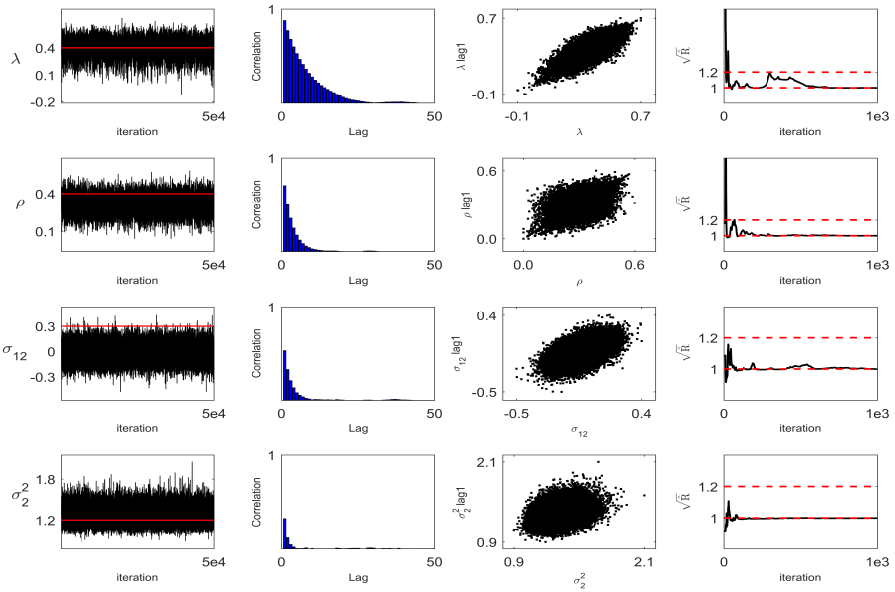


Figure 3: Algorithm 4

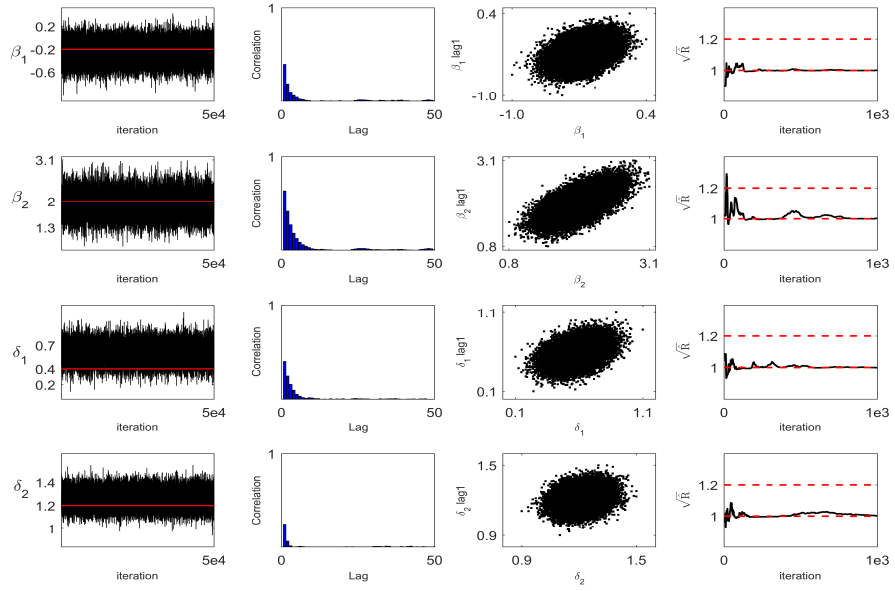


Figure 4: Algorithm 4

the estimates for ξ^2 deviate substantially from the true value. Algorithms 1 and 4 report estimates of σ_2^2 that are close to the true value, while Algorithm 5 does not.

3. Although all algorithms report estimates for β_1 and δ_2 that are close to the true values, there is some deviations in the case of β_2 and δ_1. Overall, the deviation between estimated posterior means and the true values is relatively smaller in Table 2, and the Bayesian estimator based on Algorithms 1 and 4 performs relatively better in both tables.

4. To give an overall picture on the deviation between the estimated posterior means and the true parameter values, we calculate average deviations over all parameters from Tables 2 and 3. In Table 2, the average absolute deviation is (i) 0.1056 from Algorithm 1, (ii) 0.1264 from Algorithm 2, (iii) 0.1268 from Algorithm 3, (iv) 0.1023 from Algorithm 4, and (v) 0.1752 from Algorithm 5. In Table 3, we have the average absolute bias of (i) 0.1110 from Algorithm 1, (ii) 0.1266 from Algorithm 2, (iii) 0.1264 from Algorithm 3, (iv) 0.1078 from Algorithm 4, and (v) 0.1495 from Algorithm 5. Hence, Algorithms 1 and 4 performs relatively better in both tables.

5. The standard deviations reported in both tables are close to each other for the first four algorithms. The standard deviations reported for Algorithm 5 are slightly larger, especially in the case of σ_2^2.

6. The reported Gelman-Rubin (GR) statistic, the IF statistics, the autocorrelation plots, the lag-one scatter plots and the $\sqrt{R}$ statistics can be used to assess the mixing properties of samplers. The Gelman-Rubin statistics in Tables 2 and 3 suggest that the Markov chains are mixing well. The $\sqrt{R}$ statistics in Figures 1 – 4 stabilize near one quickly. The inefficiency factor (IF) statistics are very similar across algorithms except for the case of λ. In particular, the autocorrelation plots and lag-one scatter plots for λ in Figures 1 – 4 indicate that the autocorrelation among draws decays slowly, leading to large values of IF in all algorithms. Overall our results indicate that there is no substantial differences among algorithms in terms of mixing properties.[20]

5 Empirical Illustration

For the empirical illustration, we use the application in the area of natural resource economics considered in Flores-Lagunes and Schnier, (2012). In this application, the authors model the spatial production within a fishery with a spatial sample selection model. The data set is collected from the Pacific cod fishery, located in the Eastern Bering Sea of Alaska. Among the groundfish fisheries of Alaska, the Pacific cod fishery is the second largest one with landings valued at more than 185 million dollars in 2012. For production purposes, the fishery is divided into 90 spatially different locations. The catch per unit effort (CPUE) which is measured as the metric tons of fish caught during the year 1997 in a fishing fleet is used to analyze the productivity and efficiency

[20]Note that the use of the deviations between the posterior mean estimates and the true parameter values does not necessarily measure the performance of our suggested algorithms. Instead, here, they serve as indicators for our algorithms under the given prior specifications within the context of our spatial sample selection model. For more details on the principle of unbiasedness in the Bayesian framework, see Gelman et al., (2003, pg. 248). In addition, our comparison of these algorithms in terms of mixing properties should be considered under the given prior definitions. It is also possible that the priors specified in these algorithms may not lead to the same marginal posteriors for the common parameters even though they are specified to be reasonably diffuse. In that case, the performance of these algorithms require further investigation, which is an issue beyond the scope of this study. We thank one referee for raising these issues, which is a limitation of our study.

within the fishery. A fishing fleet consists of vessels grouped according to the size of the vessel, gear utilized, and type of vessel (catcher-processor vs. catcher-vessel). Due to the confidentiality reasons, the CPUE of a fishing fleet that has less than four vessels in a location is not reported in this data set. In other words, the CPUE is observed only for those locations where four or more vessels with similar characteristics fish within that region. Since the CPUE at a certain region is likely to be an increasing function of unobserved variables that cause four or more vessels to fish at that region, a valid inference on the entire population of fishing regions should account for this selection problem. The data set contains 320 observations with a sample selection rate of 35%.[21] Moreover, because a negative shock that affects the fish population in a certain location would affect the production of all vessels in other locations by displacing fishing effort into more efficient surrounding locations, the disturbance terms are likely to be spatially correlated. Therefore, a valid model of fishing productions must account for the selection problem and the spatially correlated disturbances simultaneously.

For the outcome equation, the dependent variable is the logarithm of CPUE and the explanatory variable X_2 contains (i) the log-transformed bathymetric measurements corresponding to the maximum and minimum depth within the locations, (ii) the stock assessment data of locations received from annual biomass trawl survey, and (iii) the indicator variables for the vessel types: catcher-vessel (CV), hook-and-line gear (HAL), non-pelagic trawl gear (NPT), and vessel at least 125 feet long (Large). For the selection equation, X_1 contains X_2 and 1-year lagged stock assessment data received from the annual biomass trawl survey. The lower fish stock in a location in the previous year affects the number of vessels that will fish in that location in the upcoming year. Therefore, the time lag of the total biomass of a location will be a relevant variable for the selection equation. On the other hand, Flores-Lagunes and Schnier, (2012) assume that the time lag of the total biomass of a location may not affect the amount of hauls that will be conducted in the next year in the same location and hence, it is excluded from the outcome equation. Note that this exclusion restriction may not hold, that is, the time lag of the total biomass of a location may also be a relevant variable for the outcome equation. For example, it is possible that the lower biomass in a location can be improved by some favorable environmental factors in the upcoming years. If fishers are aware of this fact, then they will conduct a large amount of hauls in the same location in the upcoming years, hence the time lag of the total biomass of a location will be a relevant variable for the outcome equation.[22]

The specification for the weight matrices is distance based with a band. Let N_i be the set of observations in location i, where $i = 1, \ldots, 90$. Also, let d_{ij} denote the Euclidean distance between locations i and j. Then, the (i, j)th element of W and M is equal to $1/d_{ij}^2$ if $j \notin N_i$ and zero otherwise. To control the number of neighbors in a location, a band of 7 is used. For example, an observation in location i can have at most 6 neighbors in location j. Finally, both weight matrices are row normalized.

Estimation results from Algorithm 3 is presented in Table 4.[23] The tables include (i) the mean of sampled draws, (ii) the median of sampled draws, (iii) the standard deviation of sampled draws (sdv.), (iv) the 95% highest posterior density (HPD) intervals, (v) the numerical standard errors

[21] This data set is available in the Journal of Applied Econometrics Data Archive at `http://onlinelibrary.wiley.com/doi/10.1002/jae.1189/abstract`.

[22] Recall that it is not necessary to have an exclusion restriction for identification in our methodology. To investigate the effect of the exclusion restriction assumed by Flores-Lagunes and Schnier, (2012) on the parameter estimates, we also estimate the model without the exclusion restriction. The results in Table 4 in Section H of the web appendix indicate that there are not any significant changes in the results in terms of sign, magnitude and statistical significance. Therefore, we did not pursue this issue further.

[23] The results for the other algorithms are similar and left to the web appendix.

Table 4: Posterior Summary

	mean	median	sdv.	95% HPD	nse	M*	CD	IF	AC(1)
Algorithm 3									
Selection equation									
constant	-1.3116	-1.2797	0.9668	[-1.3354,-1.2260]	0.0079	14988.4887	-1.6385	3.3359	0.4103
Max. depth	0.3479	0.3468	0.1142	[0.3395,0.3538]	0.0010	13260.1456	0.9336	3.7707	0.5576
Min. depth	-0.1012	-0.1007	0.0720	[-0.1053,-0.0965]	0.0006	13843.2979	0.9260	3.6119	0.5784
Biomass	0.0671	0.0668	0.0819	[0.0618,0.0721]	0.0007	13700.5567	0.5354	3.6495	0.5753
Dum CV	-0.9748	-0.9743	0.2005	[-0.9866,-0.9619]	0.0016	16680.7041	-1.1661	2.9975	0.4734
Dum HAL	0.9582	0.9577	0.2322	[0.9434,0.9723]	0.0019	15350.6942	0.1981	3.2572	0.5017
Dum NPT	0.2703	0.2688	0.2831	[0.2518,0.2861]	0.0021	18393.1317	0.7652	2.7184	0.4636
Dum Large	-0.1454	-0.1446	0.1837	[-0.1561,-0.1331]	0.0014	16673.2626	-0.1435	2.9988	0.5171
Lag biomass	-0.0401	-0.0399	0.0799	[-0.0450,-0.0348]	0.0006	15366.8398	0.5221	3.2538	0.5478
λ	0.7724	0.7871	0.1111	[0.7803,0.7937]	0.0015	5706.4515	1.5449	8.7620	0.7867
Outcome equation									
constant	7.4684	7.4748	0.6771	[7.4358,7.5170]	0.0056	14697.7320	0.4096	3.4019	0.4417
Max. depth	0.0720	0.0716	0.0976	[0.0655,0.0776]	0.0007	18266.9821	-0.6309	2.7372	0.3936
Min. depth	0.0439	0.0442	0.0666	[0.0400,0.0484]	0.0005	21144.6629	0.9436	2.3647	0.3821
Biomass	0.1881	0.1882	0.0698	[0.1837,0.1926]	0.0005	23300.8059	-0.8908	2.1458	0.3551
Dum CV	1.2611	1.2639	0.2957	[1.2454,1.2820]	0.0034	7750.0968	0.3440	6.4515	0.7041
Dum HAL	0.0726	0.0744	0.2836	[0.0568,0.0913]	0.0028	10138.8031	-0.3507	4.9315	0.6256
Dum NPT	-0.5976	-0.5968	0.3114	[-0.6161,-0.5770]	0.0025	15099.9288	-0.2719	3.3113	0.5084
Dum Large	0.5983	0.5998	0.1652	[0.5897,0.6097]	0.0010	25722.3042	-0.4191	1.9438	0.3174
ρ	0.3502	0.3648	0.2333	[0.3505,0.3806]	0.0025	8450.8401	1.1637	5.9166	0.7201
σ_{12}	0.0183	0.0171	0.1314	[0.0091,0.0251]	0.0016	6877.1023	-0.0775	7.2705	0.6945
ξ^2	1.1313	1.1261	0.1327	[1.1183,1.1339]	0.0006	50000.0000	-1.0730	1.0000	-0.4378

(nse), (vi) the i.i.d equivalent number of iterations ($M^\star$), (vii) the Geweke, (1992)'s CD score, (viii) the inefficiency factor (IF), (ix) the first lag-correlations in sampled draws (AC(1)).

The numerical standard errors (*nse*) capture simulation noise surrounding posterior mean of each parameter and can be made arbitrarily small by choosing a sufficiently large number of iterations. Let $\{\theta_1, \theta_2, \ldots, \theta_M\}$ be a sequence of draws generated for parameter θ. Consider the mean $\overline{\theta} = \frac{1}{M}\sum_{i=1}^{M}\theta_i$. Then, the *nse* of $\overline{\theta}$ is calculated as (Koop et al., 2007, p. 145): $\text{nse}(\overline{\theta}) = \sqrt{S^2/M\left(1 + 2\sum_{k=1}^{M-1}(1 - k/M)\,\rho_k\right)}$, where S^2 is the sample variance of the sequence of draws and ρ_k is the lag k sample autocorrelation. As can be seen in Table 4, the numerical standard errors are very close to zero.

The i.i.d equivalent number of iterations ($M^\star$) is another diagnostic tool to assess the efficiency of the sampler and it is calculated from the IF. It is simply given by $M^\star = M/IF$. Hence, a very small $M^\star$ (a very large IF) is an indicative of an inefficient sampler. The results in Table 4 report large IF values and hence smaller $M^\star$ values for the autoregressive parameters as they have large AC(1) values.

Finally, the Geweke, (1992)'s CD score is a test statistic to determine if the chain of a parameter converges to the target posterior distribution. Let $M_1 = 0.1M$ and $M_2 = 0.6M$. Let $\overline{\theta}_1$ be the mean of the segment $\{\theta_1, \theta_2, \ldots, \theta_{M_1}\}$, and $\overline{\theta}_2$ be the mean of the last segment $\{\theta_{M_2+1}, \theta_2, \ldots, \theta_M\}$. Then, the CD score is given $\text{CD} = \left(\overline{\theta}_1 - \overline{\theta}_2\right)/\sqrt{\text{nse}^2(\overline{\theta}_1) - \text{nse}^2(\overline{\theta}_2)}$. Under the null hypothesis that the whole sequence of $\{\theta_1, \theta_2, \ldots, \theta_M\}$ contains random draws from the target posterior distribution, the CD score converges in distribution to $N(0, 1)$. Hence, a CD test statistic that is larger than 1.96 in absolute value indicates that the sequence of draws may not have converged to the target posterior distribution. The CD scores in Table 4 indicate that the sequence of draws converged to the target posterior distribution for all parameters.

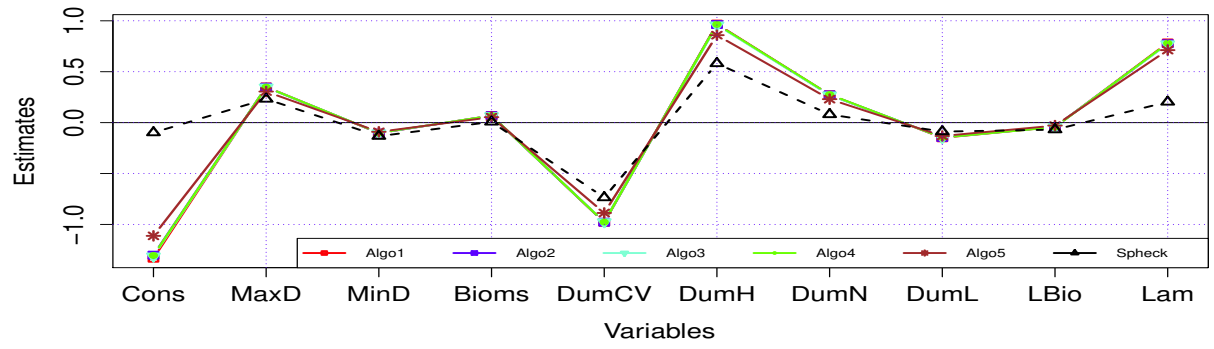
For an easy comparison of the point estimates from all algorithms, we provide the estimates of the posterior means in Figure 5. This figure also includes the estimates reported in Flores-Lagunes and Schnier, (2012) based on a GMME (denoted by Spheck). Figure 5(a) indicates that all algorithms report very similar estimates for the selection equation, and they are also very similar to those obtained from the Spheck estimator of Flores-Lagunes and Schnier, (2012) except for the spatial autoregressive parameter, λ. Our simulation results in Tables 2 – 3 indicate that the Bayesian estimator in all algorithms reports estimates of λ that are close to the true value, except in the case of Algorithm 5. Therefore, the estimates reported by our Bayesian estimator can be close to the true parameter value in this application. The Spheck estimator only provides statistically significant estimates for Max. depth, Min. depth, Dum CV and Dum HAL, while all estimates provided by our Bayesian estimator are significant as indicated by the 95% HPD intervals in Table 4. Our Bayesian estimator provides relatively more precise estimates, since it accounts for the full covariance structure implied by the spatial correlation.

The estimates for the outcome equation are displayed in Figure 5(b). Although estimates obtained from Algorithms 1–4 are in agreement, those from Algorithm 5 are slightly different in the case of Dum CV, Dum HAL, and the spatial autoregressive parameter ρ. The estimates reported by the Spheck estimator do not agree with our estimates in terms of magnitude for the case of Min. depth, Dum CV, Dum NPT, and especially for ρ, but they are in agreement in terms of signs except for the case of Min. depth. To see the effect of the selection problem on the estimate of autoregressive parameter, we consider the estimates from the following spatial error model that does not account for the selection problem: $Y_{2i} = X_{2i}^{'}\delta + \rho \sum_{j=1}^{n} M_{ij}U_{2j} + \varepsilon_{2i}$. Flores-Lagunes and Schnier, (2012) estimate this model by the GMME of Kelejian and Prucha, (1998), which is denoted by KP-SAE in their Table VIII. The Spheck estimator yields an estimate for ρ (close to 0.92) that is not so different in magnitude than the estimate of the KP-SAE estimator (close to 0.91) that only controls for the spatial correlations. Indeed, these estimates are close to the boundary of the parameter space for the spatial autoregressive parameter. As seen from Figure 5(b), our Bayesian estimator, on the other hand, yields estimates for the spatial autoregressive parameter that are much smaller in magnitude. Our simulation results in Tables 2 – 3 indicate that the Bayesian estimator corresponding to Algorithms 1–4 reports estimates of ρ that are close to the true parameter value, and therefore the true value of ρ for this application is more likely to be around 0.4.

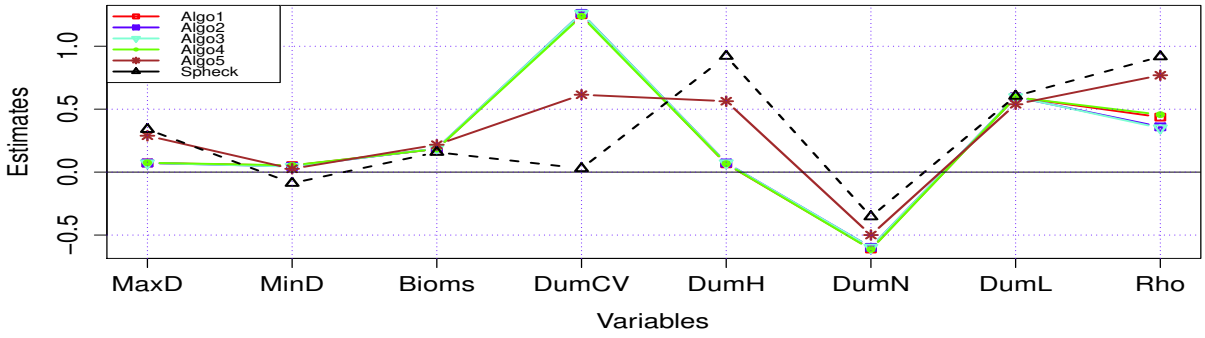
For the effect of spatial dependence on the parameter estimates, Flores-Lagunes and Schnier, (2012) show that the Spheck and the Heckit estimators largely agree in the magnitude of the estimates in most coefficients, although not in their statistical significance. The Spheck estimator provides insignificant estimates for Min.depth, Dum CV, Dum HAL, and Dum NPT, while all estimates are insignificant in the case of Heckit. Note that the Heckit estimator is inconsistent in the presence of spatial dependence. On the other hand, the 95% HPD intervals reported in Table 4 indicate that all estimates are significant. We think that the differences in the set of inference provided by the Spheck and our Bayesian estimators are due to the fact that our Bayesian estimator accounts for the full spatial correlation structure, whereas the Spheck estimator partially accounts for the spatial correlation.

6 Conclusion

In this study, we considered various Gibbs samplers for a sample selection model that accommodates spatial correlations in the disturbance terms of selection and outcome equations. To the best of our knowledge, this study is the first extensive study to illustrate the implementation of these Gibbs



(a) Selection equation



(b) Outcome equation

Figure 5: Estimates of Selection and Outcome Equations

samplers with the given prior specifications for a spatial sample selection model. These samplers are designed to account for both the sample selection bias and the spatial correlation structure implied by the model specification.

The natural parameterization of our model involved an unidentified parameter, i.e., σ_1^2. The unidentified parameter was handled in different ways in these algorithms to circumvent the computational problems. In the first algorithm, the identification constraint of $\sigma_1^2 = 1$ was directly imposed on the posterior distribution of covariance matrix of the model. In the second and third algorithms, the covariance matrix was re-parameterized in such a way that the resulting posterior distributions are not subject to the identification constraint. In the fourth and fifth algorithms, the marginal data augmentation (or the parameter expansion) method was used to handle the unidentified parameter in the posterior analysis.

Our simulation results demonstrated that for the autoregressive parameter of selection equation the Bayesian estimator reports point estimates that are close to the true parameter value in all algorithms. The results for the spatial autoregressive parameter of the outcome equation showed that the Bayesian estimates are very close to the true parameter values in Algorithms 1–4. As for the parameter of exogenous variables in the selection and outcome equations, the Bayesian estimator in Algorithms 1 and 4 performs relatively better in terms of deviations between point

estimates and the true parameter values. Finally, our results indicated that all algorithms have similar mixing properties.

References

Albert, James H. and Siddhartha Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data". English. In: *Journal of the American Statistical Association* 88.422.

Anselin, Luc (2007). "Spatial Econometrics". In: *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*. Ed. by Kerry Patterson and Terence C. Mills. Palgrave Macmillan.

Beron, Kurt J. and Wim P.M. Vijverberg (2004). "Probit in a Spatial Context: A Monte Carlo Analysis". In: *Advances in Spatial Econometrics*. Ed. by Luc Anselin, RaymondJ.G.M. Florax, and SergioJ. Rey. Advances in Spatial Science. Springer Berlin Heidelberg, pp. 169–195.

Büchel, Felix and Maarten van Ham (2003). "Overeducation, regional labor markets, and spatial flexibility". In: *Journal of Urban Economics* 53.3, pp. 482 –493.

Burgette, Lane F. and Erik V. Nordheim (2012). "The Trace Restriction: An Alternative Identification Strategy for the Bayesian Multinomial Probit Model". In: *Journal of Business & Economic Statistics* 30.3, pp. 404–410.

Chib, Siddhartha (2001). "Chapter 57 Markov Chain Monte Carlo Methods: Computation and Inference". In: *Handbook of Econometrics*. Ed. by J.J. Heckman and E. Leamer. Vol. 5. Handbook of Econometrics. Elsevier, pp. 3569 –3649.

Chib, Siddhartha, Edward Greenberg, and Ivan Jeliazkov (2009). "Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection". In: *Journal of Computational and Graphical Statistics* 18.2, pp. 321–348.

Chopin, Nicolas (2011). "Fast simulation of truncated Gaussian distributions". In: *Statistics and Computing* 21.2, pp. 275–288.

Ding, Peng (2014). "Bayesian robust inference of sample selection using selection-t models". In: *Journal of Multivariate Analysis* 124.0, pp. 451 –464.

Doğan, Osman and Süleyman Taşpınar (2014). "Spatial autoregressive models with unknown heteroskedasticity: A comparison of Bayesian and robust GMM approach". In: *Regional Science and Urban Economics* 45.0, pp. 1 –21.

Flores-Lagunes, Alfonso and Kurt Erik Schnier (2012). "Estimation of sample selection models with spatial dependence". In: *Journal of Applied Econometrics* 27.2, pp. 173–204.

Gelman, A. and D.B. Rubin (1992). "Inference from iterative simulation using multiple sequences". In: *Statistical Science* 7, pp. 457–511.

Gelman, A. et al. (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Geweke, John (1991). "Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities". In: *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Ed. by E. M. Keramidas. Interface Foundation of North America, Inc., pp. 571–578.

— (1992). "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments". In: *Bayesian Statistics 4*. Ed. by A. P. Dawid J. M. Bernardo J. O. Berger and A. F. M. Smith. Oxford University Press, pp. 169–193.

— (2005). *Contemporary Bayesian Econometrics and Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Gilks, Walter R., Sylvia Richardson, and David Spiegelhalter (1995). "Introducing Markov Chain Monte Carlo". In: *Markov Chain Monte Carlo in Practice*. Ed. by W.R. Gilks, S. Richardson, and D. Spiegelhalter. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, pp. 1–16.

Heckman, James J. (1979). "Sample Selection Bias as a Specification Error". In: *Econometrica* 47.1, pp. 153–161.

— (1990). "Varieties of Selection Bias". In: *The American Economic Review* 80.2, pp. 313–318.

Imai, Kosuke and David A. van Dyk (2005). "A Bayesian analysis of the multinomial probit model using marginal data augmentation". In: *Journal of Econometrics* 124.2, pp. 311 –334.

Kelejian, Harry H. and Ingmar R. Prucha (1998). "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances". In: *Journal of Real Estate Finance and Economics* 17.1, pp. 1899–1926.

— (2010). "Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances". In: *Journal of Econometrics* 157, pp. 53–67.

Koop, Gary, Dale J. Poirier, and Justin L. Tobias (2007). *Bayesian Econometric Methods*. New York, USA: Cambridge University Press.

Lee, Lung-fei (1978). "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables". In: *International Economic Review* 19.2, pp. 415–33.

— (1994). "Semiparametric two-stage estimation of sample selection models subject to Tobit-type selection rules". In: *Journal of Econometrics* 61.2.

Lee, Lung-fei, Xiaodong Liu, and Xu Lin (2010). "Specification and estimation of social interaction models with network structures". In: *The Econometrics Journal* 13, pp. 145–176.

Lee, Myoung-Jae (2003). "Exclusion Bias in Sample-Selection Model Estimators". In: *Japanese Economic Review* 54.2, pp. 229–236.

LeSage, James and Robert K. Pace (2009). *Introduction to Spatial Econometrics (Statistics: A Series of Textbooks and Monographs*. London: Chapman and Hall/CRC.

Leung, Siu Fai and Shihti Yu (1996). "On the choice between sample selection and two-part models". In: *Journal of Econometrics* 72.1âĂŞ2, pp. 197 –229.

Li, Kai (1998). "Bayesian inference in a simultaneous equation model with limited dependent variables". In: *Journal of Econometrics* 85.2, pp. 387 –400.

Liu, Jun S. and Ying Nian Wu (1999). "Parameter Expansion for Data Augmentation". In: *Journal of the American Statistical Association* 94.448, pp. 1264–1274.

Liu, Xiaodong and Lung-fei Lee (2010). "GMM estimation of social interaction models with centrality". In: *Journal of Econometrics* 159.1, pp. 99 –115.

McCulloch, Robert E., Nicholas G. Polson, and Peter E. Rossi (2000). "A Bayesian analysis of the multinomial probit model with fully identified parameters". In: *Journal of Econometrics* 99.1, pp. 173 –193.

McMillen, Daniel P. (1992). "Probit with Spatial Autocorrelation". In: *Journal of Regional Science* 32.3, pp. 335–348.

— (1995). "Selection bias in spatial econometrics models". In: *Journal of Regional Science* 35.3.

Meng, X-L and David A. van Dyk (1999). "Seeking efficient data augmentation schemes via conditional and marginal augmentation". In: *Biometrika* 86.2, pp. 301–320.

Newey, Whitney K. (2009). "Two-step series estimation of sample selection models". In: *The Econometrics Journal* 12.

Nobile, Agostino (2000). "Comment: Bayesian multinomial probit models with a normalization constraint". In: *Journal of Econometrics* 99.2, pp. 335 –345.

Olsen, Randall J. (1980). "A Least Squares Correction for Selectivity Bias". In: *Econometrica* 48.7, pp. 1815–1820.

Pace, Robert K., James P. LeSage, and Shuang Zhu (2012). "Spatial Dependence in Regressors and its Effect on Performance of Likelihood-Based and Instrumental Variable Estimators". In: ed. by Daniel Millimet Dek Terrell. 30th Anniversary Edition (Advances in Econometrics, Volume 30). Emerald Group Publishing Limited, pp. 257–295.

Pinkse, Joris and Margaret E. Slade (1998). "Contracting in space: An application of spatial statistics to discrete-choice models". In: *Journal of Econometrics* 85.1, pp. 125 –154.

Rabovič, Renata and Pavel Čížek (2016). *Estimation of Spatial Sample Selection Models: Partial Maximum Likelihood Approach*. CentER Discussion Paper Series.

Talhouk, Aline, Arnaud Doucet, and Kevin Murphy (2012). "Efficient Bayesian Inference for Multivariate Probit Models With Sparse Inverse Correlation Matrices". In: *Journal of Computational and Graphical Statistics* 21.3, pp. 739–757.

van Dyk, David A. and Xiao-Li Meng (2001). "The Art of Data Augmentation". In: *Journal of Computational and Graphical Statistics* 10.1, pp. 1–50.

van Hasselt, Martijn (2005). *Bayesian Sampling Algorithms for the Sample Selection and Two-Part Models*. Computing in Economics and Finance 2005 241. Society for Computational Economics.

— (2011). "Bayesian inference in a sample selection model". In: *Journal of Econometrics* 165.2, pp. 221 –232.

Wang, Honglin, Emma M. Iglesias, and Jeffrey M. Wooldridge (2013). "Partial maximum likelihood estimation of spatial probit models". In: *Journal of Econometrics* 172.1, pp. 77 –89.

Ward, Patrick S., Raymond J. G. M. Florax, and Alfonso Flores-Lagunes (2014). "Climate change and agricultural productivity in Sub-Saharan Africa: a spatial sample selection model". In: *European Review of Agricultural Economics* 41.2, pp. 199–226.

Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.