Information Criteria for Matrix Exponential Spatial Specifications

Osman Doğan* Ye Yang† Süleyman Taşpınar‡

August 23, 2023

Abstract

In this study, we suggest using information criteria for nested and non-nested model selection problems for the matrix exponential spatial specifications (MESS) under both homoskedasticity and heteroskedasticity. To this end, we consider the deviance information criterion, the Akaike information criterion and the Bayesian information criterion in a Bayesian setting. In the heteroskedastic case, we assume that the error terms have a scale mixture of normal distributions, where the scale mixture variables are latent variables that lead to different distributions. We demonstrate how the integrated likelihood function can be obtained analytically by integrating out the scale mixture variables from the complete-data likelihood function, and how this integrated likelihood function can be used to formulate the information criteria. We investigate the finite sample performance of these criteria in selecting the true model in a simulation study. The results show that these criteria perform satisfactorily and can be useful for selecting the correct model in specification search exercises. Finally, we apply the proposed information criteria to a spatially augmented growth model and a carbon emission model to show their usefulness for both nested and non-nested model selection problems.

JEL-Classification: C11, C21, C22.
Keywords: MESS, Information criteria, AIC, BIC, DIC, Bayesian inference, Integrated likelihood.

*Department of Economics, Istanbul Technical University, Istanbul, Türkiye, email: osmandogan@itu.edu.tr.

†School of Accounting, Capital University of Accounting and Business, Beijing, China, email: yeyang557@hotmail.com.

‡Department of Economics, Queens College, CUNY, New York, U.S., email: staspinar@qc.cuny.edu.

1 Introduction

The matrix exponential spatial specification (MESS) offers a convenient way of modeling weak cross-sectional dependence in a variable of interest through a matrix exponential term. As the name suggests, it implies an exponential rate of decay for the cross-sectional dependence among spatial units. An attractive property of the MESS-type models is that the matrix exponential terms are always invertible. Therefore, the reduced forms for these models always exist, and there are no restrictions on the parameter space of spatial parameters. Moreover, the likelihood based estimation of these models has the computational advantage over the estimation of popular spatial autoregressive (SAR) models since the likelihood functions of MESS-type models are free of any Jacobian terms. See Debarsy et al. (2015), Han and Lee (2013), LeSage and Pace (2007), and Yang et al. (2021, 2022) for further properties of the MESS-type models.

The spatial weights matrices in the MESS-type models specify spatial relationships among spatial units over the relevant space. The elements of these matrices can be determined in various ways depending on the nature of interaction among the spatial units. In the literature, the geographic information-based matrices such as such as contiguity-based (sharing a common border) or distance-based matrices, including nearest neighbor distance-based matrices, are widely used because these types of weights matrices are "exogenous". See Anselin (1988) and Getis and Aldstadt (2004) for some examples of spatial weights matrices created from location information. Alternatively, some notions of economic distance between spatial units can be employed to specify the elements of the weights matrices. These types of spatial weights matrices are usually time-varying, and can also be "endogenous". For some examples, among others, see Behrens et al. (2012), Brueckner (1998), Brueckner and Saavedra (2001), Case et al. (1993), Conley and Ligon (2002), Conley and Topa (2002), and Parent and LeSage (2008). In some studies, it is assumed that the elements of a weights matrix are generated through an auxiliary equation that depends on some underlying economic variables, e.g., see Qu and Lee (2015), Han and Lee (2016), and Qu et al. (2017). In some recent papers, regularized estimation strategies are suggested to estimate the elements of the spatial weights matrices (Ahrens and Bhattacharjee, 2015; Lam and Souza, 2020; Merk and Otto, 2022).

In this paper, we consider model selection problems for the MESS-type models and suggest various information criteria for these problems. When using a MESS-type model, applied researchers commonly encounter two specification problems: (i) how to choose a spatial weights matrix from a pool of candidates, which constitutes a non-nested model selection problem, and (ii) how to choose between nested or non-nested alternative model specifications. Often, it is the case that the model is specified in an ad-hoc manner and there is no guidance from an underlying structural model to address these issues. Our focus in this paper is to address these model specification problems for the MESS-type models with both homoskedastic and heteroskedastic error terms.

In the literature, Anselin (1984a,b, 1988) consider several econometric approaches, including the Cox test and the J-test, for the model selection exercises. The J-test is widely used for testing a null model against a non-nested alternative model (Davidson and MacKinnon, 1981; MacKinnon

et al., 1983). This test simply tests whether predictors from the non-nested alternative model can be statistically significant regressors in the null-model. Kelejian (2008) and Kelejian and Piras (2011) formally extend the J-test approach to a spatial setting and demonstrate how it can be used to test a null spatial model that has spatial lags in the outcome variable and the error term (for short SARAR(1,1)) against a set of alternative non-nested models. Burridge and Fingleton (2010) show that a bootstrap version of J-test performs better than the asymptotic version suggested in Kelejian (2008). Han and Lee (2013) use the J-test procedure and its bootstrap versions for the non-nested model selection problem between the SAR and MESS models in two-stage least squares and generalized method of moments frameworks. Jin and Lee (2013) consider the Cox-type and J-type tests as well as their bootstrap versions for testing the null SARAR(1,1) model against another SARAR(1,1) model with different spatial weights matrices. Liu and Lee (2019) develop a non-degenerate likelihood-ratio test for model selection between the SARAR(1,1) model and the MESS that has spatial dependence in the outcome variable and the error term (for short MESS(1,1)).

In this paper, instead of employing a testing approach, we explore various information criteria for model selection problems for the MESS-type models. More specifically, we consider the deviance information criterion (DIC) (Spiegelhalter et al., 2002), the Akaike information criterion (AIC) (Akaike, 1973), and the Bayesian information criterion (BIC) (Schwarz, 1978) for the model selection problems. From a decision-theoretic perspective, the AIC and various forms of DIC provide the asymptotically unbiased estimator of the expected Kullback-Leibler (KL) divergence between the true data generating process (DGP) and a suitable plug-in predictive distribution of hypothetically replicate data (Burnham and Anderson, 2002; Li et al., 2020). Thus, these measures select the candidate model that yields a better predictive performance, i.e., the smaller the value of these measures, the better the predictive performance of the candidate model. On the other hand, the BIC is based on a large sample approximation to the log-marginal likelihood and, therefore, selects the model that best explains the observed data.

In a Bayesian estimation framework, we consider these information criteria under both homoskedastic and heteroskedastic error terms. In the case of heteroskedasticity, we assume that the error terms follow a scale mixture of normal distributions, where the latent variables generate different distributions with distinct variance terms. Although this latent variable representation facilitates the estimation through the data augmentation techniques (Geweke, 1993), the readily available conditional likelihood function (i.e., the likelihood function obtained by conditioning on the latent variables) cannot be used to formulate the information criteria, as these functions undermine the theoretical requirements for the validity of criteria. See Li et al. (2020) for the theoretical evidence, and Chan and Grant (2016a,b) and Millar (2009) for the simulation evidence. In the latent variable models, these criteria should be formulated with the integrated likelihood function obtained by integrating out the latent variables from the complete-data likelihood function (i.e., the joint likelihood functions of data and latent variables). Notably, for the MESS-type models, we show how the integrated likelihood functions can be obtained analytically by integrating out the

scale mixture variables from the complete-data likelihood functions.

In an extensive simulation study, we investigate the performance of the suggested information criteria in selecting the true model. Our results show that these criteria can be useful for both nested and non-nested model selection problems for the MESS-type models. We use two empirical illustrations to demonstrate how to use these criteria. In the first illustration, we consider the MESS version of the spatially augmented growth model suggested by Ertur and Koch (2007) under both homoskedastic and heteroskedastic error terms. We use the information criteria to select a spatial weights matrix that leads to a better fit to the sample data. In the second application, we consider a model of carbon emissions to investigate the relationship between carbon emissions and economic activity in the United States. Following the related literature (Aldy, 2005; Auffhammer and Steinhauser, 2007; Burnett and Madariaga, 2017; Burnett et al., 2013; Spinoni et al., 2018), we consider the spatial extensions of the reduced-form energy demand, and use the information criteria to a select a suitable spatial specification. Our results show that a spatial Durbin version of the MESS model can provide a relatively better fit to our sample data.

The rest of the paper is organized as follows. In Section 2, we provide the details of the specifications under consideration and discuss the derivation of the likelihood functions. In Section 3, we suggest two Gibbs samplers for the estimation of our model under homoskedastic and heteroskedastic error terms. In Section 4, we show how the AIC, BIC and DICs can be formulated in the context of the MESS-type models. We also demonstrate the relationships among these information criteria. In Section 5, we investigate the performance of these information criteria in selecting the true model through an extensive simulation study. In Section 6, we show how to apply the information criteria to a spatially augmented growth model and a carbon emission model. In Section 7, we offer concluding remarks. Some additional simulation results are presented in a web appendix.

2 Model Specifications and the Likelihood Functions

The MESS-type models account for weak cross-sectional/spatial dependence using matrix exponential terms. The MESS(1,1) is given by

$$e^{\lambda W}Y = X\beta + U, \quad e^{\rho M}U = V, \tag{2.1}$$

where $Y = (y_1, \ldots, y_n)$ is the $n \times 1$ vector of an outcome variable, X is the $n \times k$ matrix of exogenous variables with the associated parameter vector β, $U = (u_1, \ldots, u_n)^{'}$ is the $n \times 1$ vector of regression error terms, and $V = (v_1, \ldots, v_n)^{'}$ is the $n \times 1$ vector of idiosyncratic error terms. The cross-sectional dependence in (2.1) is modeled through the matrix exponential terms $e^{\lambda W}$ and $e^{\rho M}$, where W and M are the $n \times n$ spatial weights matrices, and λ and ρ are the scalar spatial parameters. The $n \times n$ weights matrices W and M specify the relative weight of links between cross-sectional units and they have zero diagonal elements.

The matrix exponential terms $e^{\lambda W}$ and $e^{\rho M}$ in (2.1) are defined as $e^{\lambda W} = \sum_{i=0}^{\infty}(\lambda W)^i/i!$ and $e^{\rho M} = \sum_{i=0}^{\infty}(\rho M)^i/i!$, and are always invertible with the inverses $e^{-\lambda W}$ and $e^{-\rho M}$ (Chiu et al.,

1996). Thus, the reduced form of (2.1) always exists and is given by $Y = e^{-\lambda W}X\beta + e^{-\lambda W}e^{-\rho M}V$. This is in contrast to the spatial autoregressive type (SAR-type) models, because the reduced form for the SAR-type models exists under certain restrictions imposed on the parameter space of the spatial parameters.[1]

We consider both homoskedastic and heteroskedastic idiosyncratic error tems. In the homoskedastic case, we assume that $V \sim N(0, \sigma^2 I_n)$, where σ^2 is the unknown variance term and I_n is the $n \times n$ identity matrix. Then, (2.1) implies $Y \sim N\left(e^{-\lambda W}X\beta, \sigma^2 e^{-\lambda W}e^{-\rho M}e^{-\rho M'}e^{-\lambda W'}\right)$. Thus, the log-likelihood function of the model can be expressed as

$$\ln p(Y|\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(e^{\lambda W}Y - X\beta\right)' e^{\rho M'}e^{\rho M}(e^{\lambda W}Y - X\beta), \tag{2.2}$$

where $\theta = (\lambda, \rho, \beta', \sigma^2)'$. Note that the log-likelihood function does not involve any log-Jacobian terms since $\ln\left(|e^{\lambda W}|\right) = \ln\left(e^{\lambda \mathrm{tr}(W)}\right) = \ln(1) = 0$, and $\ln\left(|e^{\rho M}|\right) = \ln\left(e^{\rho \mathrm{tr}(M)}\right) = \ln(1) = 0$, where $|\cdot|$ and $\mathrm{tr}(\cdot)$ denote the determinant and the trace operators, respectively.

In the heteroskedastic case, following Geweke (1993), we assume a scale mixture of Gaussian distributions for the elements of V: $v_i|\eta_i \sim N(0, \eta_i\sigma^2)$, where η_i's are independent scale mixture variables. We assume that these scale mixture components are independently identically distributed with $\eta_i \sim \mathrm{IG}(\nu/2, \nu/2)$ for $i = 1, \ldots, n$, where IG denotes the inverse gamma distribution and ν is an unknown scalar parameter. Under this setting, the marginal distribution of v_i can be obtained in the following way:[2]

$$\begin{aligned} p(v_i) &= \int_0^\infty p(v_i, \eta_i)\mathrm{d}\eta_i = \int_0^\infty (2\pi\eta_i\sigma^2)^{-1/2} e^{-\frac{v_i^2}{2\eta_i\sigma^2}} \times \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\eta_i^{-(\nu/2+1)}e^{-\nu/2\eta_i}\mathrm{d}\eta_i \\ &= \frac{(\nu/2)^{\nu/2}}{\sqrt{2\pi\sigma^2}\Gamma(\nu/2)}\int_0^\infty \eta_i^{-\left(\frac{\nu+1}{2}+1\right)} e^{-\frac{\nu}{2\eta_i}\left(1+\frac{v_i^2}{\nu\sigma^2}\right)}\mathrm{d}\eta_i. \end{aligned} \tag{2.3}$$

The integrand in the last equation is the kernel of $\mathrm{IG}(\alpha_1, \alpha_2)$, where $\alpha_1 = (\nu+1)/2$ and $\alpha_2 = \nu\left(1 + v_i^2/\nu\sigma^2\right)/2$. Thus,

$$\begin{aligned} p(v_i) &= \frac{(\nu/2)^{\nu/2}}{\sqrt{2\pi\sigma^2}\Gamma(\nu/2)}\Gamma((\nu+1)/2)\left(\frac{\nu}{2}\right)^{-(\nu+1)/2}\left(1 + \frac{v_i^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} \\ &= \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi\sigma^2}\Gamma(\nu/2)}\left(1 + \frac{v_i^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}, \end{aligned} \tag{2.4}$$

which is the density of $t_\nu(0, \sigma^2)$, where $t_\nu(0, \sigma^2)$ is the t distribution with location 0, scale parameter σ^2, and ν degrees of freedom.

Let $\eta = (\eta_1, \ldots, \eta_n)'$ and $\theta = (\lambda, \rho, \beta', \sigma^2, \nu)'$. In the heteroskedastic case, the scale mixture of

[1] See Elhorst (2014), Kelejian and Prucha (2010), Lee (2004), and LeSage and Pace (2009) on the parameter space of spatial parameters. For example, the reduced form of the SAR model $Y = \lambda WY + X\beta + U$ exists under the assumption that $\varrho(\lambda W) < 1$, where $\varrho(\cdot)$ denotes the spectral radius, and is given as $Y = S^{-1}(\lambda)X\beta + S^{-1}(\lambda)U$, where $S(\lambda) = (I_n - \lambda W)$, and I_n is the $n \times n$ identity matrix.

[2] To denote the relevant density functions, we use $p(\cdot)$ and omit X in the conditional sets for the sake of simplicity.

Gaussian distributions representation allows us to consider three types of likelihood functions: (i) the conditional likelihood function denoted by $p(Y|\theta,\eta)$, (ii) the complete-data likelihood function denoted by $p(Y,\eta|\theta)$, and (iii) the integrated (or observed) likelihood function denoted by $p(Y|\theta) = \int p(Y,\eta|\theta)\mathrm{d}\eta$. The log-conditional likelihood function is readily available and given by

$$\begin{aligned}\ln p(Y|\theta,\eta) = &-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\ln\eta_i \\ &- \frac{1}{2\sigma^2}\left(e^{\lambda W}Y - X\beta\right)' e^{\rho M'} H^{-1}(\eta) e^{\rho M}\left(e^{\lambda W}Y - X\beta\right),\end{aligned} \tag{2.5}$$

where $H(\eta) = \mathrm{Diag}\,(\eta_1,\ldots,\eta_n)$ is the $n\times n$ diagonal matrix with the ith diagonal element η_i. As shown in Algorithm 2 of Section 3, this conditional likelihood function facilitates the MCMC estimation of our model through a data augmentation scheme.

The log-integrated likelihood function can be obtained by analytically integrating out the scale mixture variables η from the complete-data likelihood function, i.e., $p(Y|\theta) = \int p(Y,\eta|\theta)\mathrm{d}\eta = \int p(Y|\eta,\theta)p(\eta|\theta)\mathrm{d}\eta$. The following proposition gives the analytical expression for this function.

Proposition 1. *Let* $Y(\delta) = e^{\rho M}\left(e^{\lambda W}Y - X\beta\right)$, *where* $\delta = (\lambda,\rho,\beta')'$. *Then,*

$$\begin{aligned}\ln p(Y|\theta) = &-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \frac{n\nu}{2}\ln(\nu/2) \\ &+ n\ln\Gamma\left(\frac{\nu+1}{2}\right) - n\ln\Gamma(\nu/2) - \frac{\nu+1}{2}\sum_{i=1}^{n}\ln\left(\frac{\nu}{2} + \frac{y_i^2(\delta)}{2\sigma^2}\right),\end{aligned}$$

where $y_i(\delta)$ *is the* i*th element of* $Y(\delta)$.

Proof. See Appendix A. □

In Section 4, we will show that this integrated likelihood function is essential for the formulation of the information criteria for the heteroskedastic model.

3 A Bayesian Estimation Approach

In this section, we suggest Gibbs samplers for the Bayesian estimation of (2.1) under both homoskedasticity and heteroskedasticity. We assume the following prior distributions: $\lambda \sim N(\mu_\lambda, V_\lambda)$, $\rho \sim N(\mu_\rho, V_\rho)$, $\beta \sim N(\mu_\beta, V_\beta)$, $\sigma^2 \sim IG(a,b)$, and $\nu \sim \mathrm{Uniform}(2,\bar{\nu})$, where IG denotes the inverse-gamma distribution and Uniform$(a,\,b)$ denotes the uniform distribution over (a,b). As shown in (2.4), the marginal distribution of v_i is a t distribution with mean zero, scale parameter σ^2 and ν degrees of freedom, i.e., $v_i \sim t_\nu(0,\sigma^2)$. Thus, by choosing $\nu \sim \mathrm{Uniform}(2,\bar{\nu})$ for ν, we ensure that the variance of v_i exists. We set $\bar{\nu} = 50$ so that the t distribution can approximate the normal distribution well-enough. In our simulation, we consider the following values for the remaining hyperparameters: $\mu_\beta = 0$, $V_\beta = 10I_k$, $\mu_\lambda = \mu_\rho = 0$, $V_\lambda = V_\rho = 10$, $a = 0.01$ and $b = 0.01$.

Under these prior distributions, the posterior distribution of parameters in the homoskedastic case is given by

$$p(\theta|Y) \propto p(Y|\theta)p(\theta) = p(Y|\theta)p(\beta)p(\sigma^2)p(\lambda)p(\rho), \tag{3.1}$$

where $p(\theta)$ is the joint prior distribution of θ. To generate random draws from $p(\theta|Y)$, we suggest the Gibbs sampler presented in Algorithm 1.

Algorithm 1 (Estimation of (2.1) under homoskedasticity)**.**

1. *Sampling step for β:*

$$\beta|Y, \lambda, \rho, \sigma^2 \sim N(\widehat{\beta}, K_\beta), \tag{3.2}$$

where $K_\beta = (V_\beta^{-1} + \sigma^{-2}X'e^{\rho M'}e^{\rho M}X)^{-1}$ and $\widehat{\beta} = K_\beta(\sigma^{-2}X'e^{\rho M'}e^{\rho M}e^{\lambda W}Y + V_\beta^{-1}\mu_\beta)$.

2. *Sampling step for σ^2:*

$$\sigma^2|Y, \lambda, \rho, \beta \sim IG(\widehat{\sigma}^2, K_{\sigma^2}), \tag{3.3}$$

where $\widehat{\sigma}^2 = a + \frac{n}{2}$ and $K_{\sigma^2} = b + \frac{1}{2}(e^{\lambda W}Y - X\beta)'e^{\rho M'}e^{\rho M}(e^{\lambda W}Y - X\beta)$.

3. *Sampling step for λ:*

$$\begin{aligned} &p(\lambda|Y, \beta, \rho, \sigma^2) \\ &\propto \exp\left(-\frac{1}{2}\left(\sigma^{-2}(e^{\lambda W}y - X\beta)'e^{\rho M'}e^{\rho M}(e^{\lambda W}Y - X\beta) + V_\lambda^{-1}(\lambda^2 - 2\mu_\lambda\lambda)\right)\right), \end{aligned} \tag{3.4}$$

which is a non-standard distribution. We use the random-walk Metropolis-Hastings algorithm suggested in LeSage and Pace (2009). We generate a candidate value λ^{new} is according to

$$\lambda^{new} = \lambda^{old} + c_\lambda \times N(0, 1), \tag{3.5}$$

where c_λ is a tuning parameter.[3] *Then, we accept the candidate value λ^{new} with probability*

$$\mathbb{P}(\lambda^{new}, \lambda^{old}) = \min\left(1, \frac{p(\lambda^{new}|Y, \beta, \sigma^2, \rho)}{p(\lambda^{old}|Y, \beta, \sigma^2, \rho)}\right). \tag{3.6}$$

4. *Sampling step for ρ:*

$$\begin{aligned} &p(\rho|Y, \beta, \lambda, \sigma^2) \\ &\propto \exp\left(-\frac{1}{2}\left(\sigma^{-2}(e^{\lambda W}y - X\beta)'e^{\rho M'}e^{\rho M}(e^{\lambda W}Y - X\beta) + V_\rho^{-1}(\rho^2 - 2\mu_\rho\rho)\right)\right). \end{aligned} \tag{3.7}$$

[3]The tuning parameter is determined during the estimation such that the acceptance rate falls between 40% and 60%.

We use the random-walk Metropolis-Hastings algorithm described in Step 3 to generate random draws from $p(\rho|Y,\beta,\lambda,\sigma^2)$.

In Algorithm 1, the conditional posterior distributions of β and σ^2 are obtained respectively from $p(\beta|Y,\lambda,\rho,\sigma^2) \propto p(Y|\theta)p(\beta)$ and $p(\sigma^2|Y,\lambda,\rho,\beta) \propto p(Y|\theta)p(\sigma^2)$, where $p(Y|\theta)$ is the likelihood function, $p(\beta)$ and $p(\sigma^2)$ are the prior distributions. Then, using an analogous analysis to that of the standard Bayesian analysis for a linear regression model, we obtain $\beta|Y,\lambda,\rho,\sigma^2,\eta \sim N(\widehat{\beta}, K_\beta)$ and $\sigma^2|Y,\lambda,\rho,\beta,\eta \sim IG(\widehat{\sigma}^2, K_{\sigma^2})$. On the other hand, the conditional posterior distributions of spatial parameters take unknown forms as shown in Steps 3 and 4 of Algorithm 1. We use the random walk Metropolis-Hastings algorithm suggested LeSage and Pace (2009) to generate random draws for these parameters.

In the heteroskedastic case, the posterior distribution of parameters takes the following form:

$$p(\theta,\eta|Y) \propto p(Y|\theta,\eta)p(\theta,\eta) = p(Y|\theta,\eta)p(\beta)p(\sigma^2)p(\lambda)p(\rho)p(\eta|\nu)p(\nu), \tag{3.8}$$

where $p(Y|\theta,\eta)$ is the conditional likelihood function and $p(\theta,\eta)$ is the joint prior distribution of θ and η. Algorithm 2 presents a Gibbs sampler that can be used to generate random draws from $p(\theta,\eta|Y)$.

Algorithm 2 (Estimation of (2.1) under heteroskedasticity)**.**

1. *Sampling step for* β*:*

$$\beta|Y,\lambda,\rho,\sigma^2,\eta \sim N(\widehat{\beta}, K_\beta), \tag{3.9}$$

where $K_\beta = (V_\beta^{-1} + \sigma^{-2}X'e^{\rho M'}H^{-1}(\eta)e^{\rho M}X)^{-1}$ *and* $\widehat{\beta} = K_\beta(\sigma^{-2}X'e^{\rho M'}H^{-1}(\eta)e^{\rho M}e^{\lambda W}Y + V_\beta^{-1}\mu_\beta)$.

2. *Sampling step for* σ^2*:*

$$\sigma^2|Y,\lambda,\rho,\beta,\eta \sim IG(\widehat{\sigma}^2, K_{\sigma^2}), \tag{3.10}$$

where $\widehat{\sigma}^2 = a + \frac{n}{2}$ *and* $K_{\sigma^2} = b + \frac{1}{2}(e^{\lambda W}Y - X\beta)'e^{\rho M'}H^{-1}(\eta)e^{\rho M}(e^{\lambda W}Y - X\beta)$.

3. *Sampling step for* λ*:*

$$\begin{aligned} &p(\lambda|Y,\beta,\rho,\sigma^2,\eta) \\ &\propto \exp\left(-\frac{1}{2}\left(\sigma^{-2}(e^{\lambda W}y - X\beta)'e^{\rho M'}H^{-1}(\eta)e^{\rho M}(e^{\lambda W}Y - X\beta) + V_\lambda^{-1}(\lambda^2 - 2\mu_\lambda\lambda)\right)\right), \end{aligned} \tag{3.11}$$

which is a non-standard distribution. We use the random-walk Metropolis-Hastings algorithm described in Step 3 of Algorithm 1 to sample this parameter.

4. *Sampling step for ρ:*

$$p(\rho|Y,\beta,\lambda,\sigma^2,\eta) \propto \exp\left(-\frac{1}{2}\left(\sigma^{-2}(e^{\lambda W}y - X\beta)' e^{\rho M'} e^{\rho M}(e^{\lambda W}Y - X\beta) + V_\rho^{-1}(\rho^2 - 2\mu_\rho\rho)\right)\right). \tag{3.12}$$

We use the random-walk Metropolis-Hastings algorithm described in Step 3 of Algorithm 1 to generate random draws from $p(\rho|Y,\beta,\lambda,\sigma^2,\eta)$.

5. *Sampling step for η:*

$$\eta_i|Y,\lambda,\rho,\beta,\sigma^2,\nu \sim IG\left(\frac{\nu+1}{2},\ \frac{\nu}{2} + \frac{y_i^2(\delta)}{2\sigma^2}\right) \quad for \quad i = 1,2,\ldots,n, \tag{3.13}$$

where $y_i(\delta)$ is the ith element of $Y(\delta)$.

6. *Sampling step for ν:*

$$p(\nu|\eta) \propto \frac{(\nu/2)^{n\nu/2}}{\Gamma^n(\nu/2)}\left(\prod_{i=1}^n \eta_i\right)^{-(\frac{\nu}{2}+1)} \exp\left(-\sum_{i=1}^n \frac{\nu}{2\eta_i}\right), \tag{3.14}$$

which is a non-standard density function. We use a Griddy-Gibbs sampler to sample this parameter.

In Algorithm 2, the conditional posterior distributions of β and σ^2 are obtained respectively from $p(\beta|Y,\lambda,\rho,\sigma^2,\eta) \propto p(Y|\theta,\eta)p(\beta)$ and $p(\sigma^2|Y,\lambda,\rho,\beta,\eta) \propto p(Y|\theta,\eta)p(\sigma^2)$, where $p(Y|\theta,\eta)$ is the conditional likelihood function. An analysis analogous to the one used in the case of Algorithm 1 yields the conditional posterior distributions of these parameters. As in the homoskedastic case, the conditional posterior distributions of spatial parameters take unknown forms as shown in Steps 3 and 4. We resort the random walk Metropolis-Hastings algorithm suggested LeSage and Pace (2009) to generate random draws for these parameters. In Algorithm 2, we have additional blocks for η and ν. Since the elements of η are i.i.d, we have

$$\begin{aligned} p(\eta|Y,\lambda,\rho,\beta,\sigma^2,\nu) &\propto p(Y|\theta,\eta)p(\eta|\nu) \\ &\propto \prod_{i=1}^n \eta_i^{-(\frac{\nu+1}{2}+1)} \exp\left(-\frac{1}{\eta_i}\left(\frac{\nu}{2} + \frac{y_i^2(\delta)}{2\sigma^2}\right)\right), \end{aligned}$$

which suggests that $\eta_i|Y,\lambda,\rho,\beta,\sigma^2,\nu \sim \text{IG}\left(\frac{\nu+1}{2},\ \frac{\nu}{2} + \frac{y_i^2(\delta)}{2\sigma^2}\right)$ for $i = 1,2,\ldots,n$. Finally, the conditional posterior distribution of ν is $p(\nu|Y,\lambda,\rho,\beta,\eta) = p(\nu|\eta) \propto p(\eta|\nu)p(\nu)$, which does not take a known form. Since this parameter has a support over $(2,\bar{v})$, we use the simulation method called the Gridy-Gibbs sampler (Ritter and Tanner, 1992) to generate draws from $p(\nu|\eta)$.

4 Information Criteria for Model Selection

The information criteria can be considered as measures of predictive accuracy, and are typically defined based on the deviance term $-2\ln p(Y|\theta)$ (Gelman et al., 2003). The popular information criterion AIC is defined by

$$\text{AIC} = -2\ln p(Y|\widehat{\theta}) + 2P, \tag{4.1}$$

where $\widehat{\theta}$ is the maximum likelihood estimator (MLE) of θ and P is the dimension of θ. Using a decision-theoretic perspective, we can show that the AIC chooses the model whose predictive distribution is close to the true DGP. Let $g(Y)$ be the DGP, and $Y_{\text{rep}} = (y_{1,\text{rep}}, \ldots, y_{n,\text{rep}})'$ be the $n \times 1$ vector of replicate data generated from $g(Y)$ independently from the observed data Y. Consider the Kullback-Leibler (KL) divergence between $g(Y_{\text{rep}})$ and the generic predictive distribution $p(Y_{\text{rep}}|Y)$:

$$\begin{aligned} KL\left(g(Y_{\text{rep}}), p(Y_{\text{rep}}|Y)\right) &= \mathbb{E}_{Y_{\text{rep}}}\left(\ln \frac{g(Y_{\text{rep}})}{p(Y_{\text{rep}}|Y)}\right) = \int \left(\ln \frac{g(Y_{\text{rep}})}{p(Y_{\text{rep}}|Y)}\right) g(Y_{\text{rep}}) dY_{\text{rep}} \\ &= \int \ln g(Y_{\text{rep}}) g(Y_{\text{rep}}) dY_{\text{rep}} - \int \ln p(Y_{\text{rep}}|Y) g(Y_{\text{rep}}) dY_{\text{rep}} \\ &= c - \int \ln p(Y_{\text{rep}}|Y) g(Y_{\text{rep}}) dY_{\text{rep}}, \end{aligned} \tag{4.2}$$

where the expectation $\mathbb{E}_{Y_{\text{rep}}}$ is with respect to $g(Y_{\text{rep}})$, and $c = \int \ln g(Y_{\text{rep}}) g(Y_{\text{rep}}) dY_{\text{rep}}$, which is constant across candidate models. In (4.2), if we replace the generic predictive distribution $p(Y_{\text{rep}}|Y)$ with the plug-in predictive distribution $p(Y_{\text{rep}}|\widehat{\theta})$, then it can be shown that (Burnham and Anderson, 2002; Li et al., 2020)

$$\begin{aligned} \mathbb{E}_Y\left(2 \times KL\left(g(Y_{\text{rep}}), p(Y_{\text{rep}}|\widehat{\theta})\right)\right) &= 2c + \mathbb{E}_Y\left(\int -2\ln p(Y_{\text{rep}}|\widehat{\theta}) g(Y_{\text{rep}}) dY_{\text{rep}}\right) \\ &= 2c + \mathbb{E}_Y\left(-2\ln p(Y_{\text{rep}}|\widehat{\theta}) + 2P\right) + o(1) \\ &= 2c + \mathbb{E}_Y\left(\text{AIC}\right) + o(1), \end{aligned} \tag{4.3}$$

where the expectation $\mathbb{E}_Y$ is with respect to $g(Y)$. Thus, the AIC measure is an asymptotically unbiased estimator of $\mathbb{E}_Y\left(2 \times KL\left(g(Y_{\text{rep}}), p(Y_{\text{rep}}|\widehat{\theta})\right)\right) - 2c$. This theoretical result indicates that a model with a smaller AIC value will perform better in terms of predictive performance.

Next, we consider the DIC measure suggested by Spiegelhalter et al. (2002). The Bayesian deviance term, as defined in Spiegelhalter et al. (2002), is given by $D(\theta) = -2\ln p(Y|\theta) + 2\ln f(Y)$, where $f(Y)$ is a standardizing term that depends solely on the data. Following Berg et al. (2004), we set $f(Y) = 1$ for the model comparison exercises. Then, the DIC suggested by Spiegelhalter et al. (2002) is defined by

$$\text{DIC} = \overline{D}(\theta) + P_D, \tag{4.4}$$

where $\overline{D}(\theta) = -2\mathbb{E}\left(\ln p(Y|\theta)|Y\right) = -2\int \ln p(Y|\theta)p(\theta|Y)\mathrm{d}\theta$ is the posterior mean deviance and serves as a Bayesian measure of model fit. The second term P_D is the difference between the posterior mean deviance and the deviance at the estimated parameters,

$$P_D = \overline{D}(\theta) - D(\bar{\theta}) = -2\mathbb{E}\left(\ln p(Y|\theta)|Y\right) + 2\ln p(Y|\bar{\theta}), \tag{4.5}$$

where $\bar{\theta}$ is the posterior mean.[4] P_D is used as a measure of the effective number of parameters in the model, i.e. it is a measure of model complexity. Thus, the DIC measure is given by

$$\text{DIC} = \overline{D}(\theta) + P_D = -4\mathbb{E}\left(\ln p(Y|\theta)|Y\right) + 2\ln p(Y|\bar{\theta}). \tag{4.6}$$

We can express the DIC measure in (4.4) in the following alternative way:

$$\text{DIC} = D(\bar{\theta}) + 2P_D, \tag{4.7}$$

which resembles to the AIC defined in (4.1). Under some regularity conditions, Li et al. (2020) show that if we replace the generic predictive distribution $p(Y_{\text{rep}}|Y)$ in (4.2) with the plug-in predictive distribution $p(Y_{\text{rep}}|\bar{\theta})$, the following result is obtained:

$$\begin{aligned}
\mathbb{E}_Y\left(2 \times KL\left(g(Y_{\text{rep}}), p(Y_{\text{rep}}|\bar{\theta})\right)\right) &= 2c + \mathbb{E}_Y\left(\int -2\ln p(Y_{\text{rep}}|\bar{\theta})g(Y_{\text{rep}})dY_{\text{rep}}\right) \\
&= 2c + \mathbb{E}_Y\left(-2\ln p(Y_{\text{rep}}|\bar{\theta}) + 2P_D\right) + o(1) \\
&= 2c + \mathbb{E}_Y\left(\text{DIC}\right) + o(1).
\end{aligned} \tag{4.8}$$

That is, the DIC is asymptotically an unbiased estimator of $\mathbb{E}_Y\left(2 \times KL\left(g(Y_{\text{rep}}), p(Y_{\text{rep}}|\bar{\theta})\right)\right) - 2c$. Thus, as in the case of AIC, a model with a smaller DIC value will perform better in terms of predictive performance. Therefore, the DIC is usually considered as a Bayesian counterpart of the AIC.

Using the form given in (4.7), Li et al. (2020) suggest the following version of the DIC,

$$\text{DIC}_L = D(\bar{\theta}) + 2P_L = D(\bar{\theta}) + 2\text{tr}\left(J(\bar{\theta})V(\bar{\theta})\right), \tag{4.9}$$

where $P_L = \text{tr}\left(J(\bar{\theta})V(\bar{\theta})\right)$, $J(\theta) = -\frac{\partial^2 \ln p(Y|\theta)}{\partial\theta\partial\theta'}$, and $V(\bar{\theta})$ is the posterior covariance of θ given by $V(\bar{\theta}) = \mathbb{E}\left((\theta-\bar{\theta})(\theta-\bar{\theta})'|Y\right)$. Under some regularity conditions, Li et al. (2020) show that $\text{DIC}_L = \text{AIC} + o_p(1)$ and $P_L = P + o_p(1)$.

Finally, we consider the Bayesian information criterion. This criterion is derived from a large sample approximation to the log-marginal likelihood function of the candidate model. The marginal likelihood function of model M_k can be expressed as $p(Y|M_k) = \int_{\Theta_k} p(Y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$, where

[4] Celeux et al. (2006) suggest $P_D = -2\mathbb{E}\left(\ln p(Y|\theta)|Y\right) + 2\ln p(Y|\tilde{\theta})$, where $\tilde{\theta}$ is the joint maximum a posterior (MAP) estimator . The MAP estimate can be approximated by the posterior draws of θ that yield the largest value for $p(Y|\theta)p(\theta)$, where $p(\theta)$ denotes the prior density of θ.

θ_k is the $P_k \times 1$ parameter vector in the model M_k. The Laplace approximation can be used to approximate $p(Y|M_k)$ in the following way (Schwarz, 1978):

$$\ln p(Y|M_k) = \ln p(Y|\widehat{\theta}_k, M_k) + \ln p(\widehat{\theta}_k|M_k) + \frac{P_k \pi}{2} - \frac{P_k \ln n}{2} - \frac{1}{2}|J_k(\widehat{\theta}_k)| + O_p(1/n), \quad (4.10)$$

where $\widehat{\theta}_k$ is the MLE of θ_k, and $J_k(\widehat{\theta}_k) = -\frac{1}{n}\frac{\partial^2 \ln p(Y|\widehat{\theta}_k, M_k)}{\partial \theta_k \partial \theta_k'}$. Under a non-informative prior distribution and ignoring all $O_p(1)$ terms in (4.10), Schwarz (1978) define the BIC for M_k as

$$\text{BIC}_k = -2 \ln p(Y|\widehat{\theta}_k, M_k) + P_k \ln n. \quad (4.11)$$

The Laplace approximation in (4.10) can also be used to show that the difference between the BIC's of two models is asymptotically equivalent to the log Bayes factor (Kass and Raftery, 1995). That is, for any $\epsilon > 0$, we have

$$\lim_{n \to \infty} P\left(\left|\frac{\text{BIC}_k - \text{BIC}_l}{\ln \text{BF}_{kl}} - 1\right| > \epsilon\right) = 0, \quad (4.12)$$

where $\text{BF}_{kl} = p(Y|M_k)/p(Y|M_l)$ is the Bayes factor of M_k against M_l. Thus, the BIC is also a consistent model selection criterion like the Bayes factor, i.e., both BIC and BF select the true model with probability approaching one when $n \to \infty$.

Remark 1. *The asymptotic results in this section require that (i) a Bernstein-von Mises type theorem holds for the posterior distribution, i.e., the posterior distribution $p(\theta|Y)$ converges to a normal distribution whose mean is the MLE $\widehat{\theta}$ and covariance is the inverse of the second derivative of the negative log-likelihood function, (ii) the MLE has the standard large sample properties, namely, consistency and asymptotic normality. Han et al. (2021) show that a Bernstein-von Mises type theorem as in Chernozhukov and Hong (2003) holds for the posterior distribution of a homoskedastic SAR model under certain conditions. Using a similar approach, we conjuncture that a Bernstein-von Mises type theorem will also hold for our MESS model under certain conditions. As for the second condition, Debarsy et al. (2015) and Liu and Lee (2019) show that the MLE of our MESS model has the standard large sample properties. However, it is known that these two conditions usually do not hold for the latent variables in the latent variable models (Gelman et al., 2003; Li et al., 2020). In particular, the latent variable representation we assumed for the heteroskedastic case undermines these conditions for the latent scale mixture variables. Therefore, the AIC and DIC measures based on the conditional likelihood functions will be biased estimators of the corresponding expected KL measures. These measures should be formulated based on the integrated likelihood function given in Proposition 1.*

Remark 2. *The computation of the AIC and BIC requires the MLE $\widehat{\theta}$ of θ, while that of the DIC's require the MCMC draws of θ from the posterior distribution $p(\theta|Y)$. Instead of using the MLE in the case of AIC and BIC, we can approximate $p(Y|\widehat{\theta})$ by the maximum of the likelihood function evaluated at the posterior draws, i.e., $p(Y|\widehat{\theta}) \approx \max\{p(Y|\theta^1), \ldots, p(Y|\theta^R)\}$, where $\{\theta^r\}_{r=1}^R$ is a*

sequence of posterior draws. The first term $\mathbb{E}\left(\ln p(Y|\theta)|Y\right)$ *in the DIC measures can be estimated by averaging the log-integrated likelihood function given in Proposition 1 over the posterior draws of* θ*, i.e.,* $\mathbb{E}\left(\ln p(Y|\theta)|Y\right) \approx \frac{1}{R}\sum_{r=1}^{R} \ln p(Y|\theta^r)$*. The second term* $\ln p(Y|\bar{\theta})$ *in DIC is simply obtained by evaluating the log-integrated likelihood function at the posterior mean* $\bar{\theta}$*. Finally, in the case of* DIC_L*, we need to compute* $\operatorname{tr}\left(J(\bar{\theta})V(\bar{\theta})\right)$*, where* $J(\bar{\theta})$ *is approximated by the numerical hessian and* $V(\bar{\theta})$ *is the covariance of the posterior draws.*[5] *As a result, in our simulation study, we use the MCMC algorithms given in Section 3 to compute all information criteria.*

5 Monte Carlo Simulations

5.1 Design

In this section, we use our suggested estimation algorithms given in Section 3 to investigate the finite sample performance of the information criteria stated in Section 4. To this end, we are interested in the performance of the AIC in (4.1), the DIC in (4.6), the $\mathrm{DIC_L}$ in (4.9) and the BIC in (4.11). We consider the following well-known DGPs:

$$
\begin{aligned}
M1: &\quad e^{\lambda W}Y = X\beta + V,\\
M2: &\quad Y = X\beta + U, \quad e^{\rho W}U = V,\\
M3: &\quad e^{\lambda W}Y = X\beta + U, \quad e^{\rho W}U = V,\\
M4: &\quad e^{\lambda W}Y = X\beta + WX_1\psi + V.
\end{aligned}
$$

We set $X = (l_n,\, X_1)$ and $(\beta_1, \beta_2, \psi)' = (-0.5, 1, 1)'$, where l_n is the $n \times 1$ vector of ones and $X_1 \sim N(0, I_n)$. For the spatial parameters, we consider $(\lambda, \rho)' = (-1.2, -0.4)'$. In the homoskedastic case, we set $V \sim N(0, \sigma^2 I_n)$, where $\sigma^2 = 1$. In the case of heteroskedasticity, we set $V \sim N\left(0, \mathrm{Diag}(\gamma_1, \ldots, \gamma_n)\right)$, where $\gamma_i = \exp(0.1 + 0.35X_{1i})$ and X_{1i} is the ith element of X_1. For the spatial weights matrix W, we consider the rook and queen contiguity cases. We will denote them by W_r and W_q, respectively. For the sample size, we use $n = \{100,\, 225\}$.

The length of the MCMC chain is set to 6000 draws, and the first 2000 draws are discarded as burn-ins. For readers interested in the computational complexity of the MESS-type model, we refer to Yang et al. (2021), who extensively study the computational time of the QML, GMM, and Bayesian estimation of the model. In this paper, the authors consider a fast estimation algorithm called the matrix-vector-product (mvp) method and find that its computational time is significantly lower than that of the default method utilized in a popular software (MATLAB). In various combinations of parameter values and sample sizes, they demonstrate that the computation time for the three estimators decreases by 95% to 99% compared to that of the default `expm` function in MATLAB.

In the first experiment, we consider the performance of information criteria in terms of selecting the correct spatial weights matrix under homoskedasticity. To that end, we use W_r to generate 300

[5] We use the `hessian` function provided in the Spatial Econometrics Toolbox to compute $J(\bar{\theta})$.

samples according to each DGP specified above. Then, we estimate each model with both W_r and W_q using the Algorithm 1 given in Section 3 for all samples. Using the estimation results, we then compute the corresponding information criteria for all samples.

In the second experiment, our objective is to evaluate the performance of the information criteria in distinguishing M3 from the other models under homoskedasticity. For this purpose, we employ W_r to generate 300 samples according to M3. Next, we use each sample to estimate each model using W_r and calculate the associated information criteria.

In the third experiment, our focus is on evaluating the performance of the information criteria in terms of selecting the true spatial weights matrix under heteroskedasticity. We use W_q to generate 300 samples according to each DGP and estimate each model with both W_r and W_q using Algorithm 2 in Section 3. Subsequently, we compute the information criteria for all samples.

In the final experiment, we generate 300 samples using W_q according to M3 under heteroskedasticity. Here, our interest lies in examining the performance of the information criteria in distinguishing M3 from the other DGPs under heteroskedasticity. For all samples, we estimate each model with W_q using Algorithm 2 given in Section 3. We then compute the corresponding information criteria for all samples.

5.2 Simulation Results

To present the simulation results in a meaningful concise manner, we resort to histogram plots as suggested by Chan and Grant (2016b). More specifically, for each experiment, for a given criterion, say the DIC, we subtract the DIC value of the true DGP from the DIC value of the false DGP. We then display these differences as histogram plots over the 300 samples. If the criterion under consideration perform satisfactorily, we expect majority of these differences to be positive over 300 samples, and we should observe most of the mass on the positive half line in the histogram plots. Each histogram provides the percentage of positive differences. For example, the notation "$> 0 : 95\%$" in a histogram indicates that 95 percent of differences are positive.[6]

We start with the results from the first experiment. Figure 1 presents the histogram plots for the four information criteria when $n = 100$ and the DGP is M1. Recall again that in the first experiment we are interested in the performance of the information criteria in terms of selecting the true spatial weights matrix W_r under homoskedasticity. We observe that all differences are positive for the AIC, DIC, DIC_L, and BIC. Therefore, all four information criteria choose the correct spatial weights matrix in all of the 300 samples. Figure 2 presents the histogram plots when $n = 225$ and the DGP is M1. We again observe that all four information criteria select the correct spatial weights matrix in all of the 300 samples. Figure 3 displays the histogram plots for the four information criteria when $n = 100$ and the DGP is M2. All four information criteria select the correct spatial weights matrix in about 85% of the 300 samples. However, the performances of all criteria increase when the sample size becomes $n = 225$ as shown in Figure 12 of the web appendix. Figure 4

[6]We will present the results for some selected cases in each experiment. The results for the rest of the cases provide similar results and are provided in the web appendix.

presents the histogram plots for the four information criteria when $n = 100$ and the DGP is M3. We observe that that the differences are all positive, which implies that all four information criteria choose the correct spatial weights matrix in all of the 300 samples. Finally, Figure 5 presents the histogram plots for the four information criteria when $n = 100$ and the DGP is M4. Once again, we observe that all four information criteria select the correct spatial weights matrix in all of the 300 samples.

Next, we present some results from the second experiment. Recall again that we are interested in the performance of the information criteria in terms of selecting the true DGP M3 against the other DGPs under homoskedasticity. Figure 6 presents the histogram plots for the four information criteria in selecting M3 against M1 when $n = 225$. We observe that all four information criteria choose M3 over M1 in most of the samples. Specifically, the percentage of positive differences is 94.3% for AIC, 93.7% for DIC, 93.7% for DIC_L, and 73.7% for BIC. Although BIC performs slightly worse than the other information criteria in this case, it is expected that this performance difference will diminish as the sample size increases. Figure 7 presents the histogram plots for the four information criteria in selecting M3 against M2 when $n = 225$. We observe that the differences are all positive, indicating that all four information criteria choose the correct DGP M3 against M2 in all of the 300 samples.

Moving on to the third experiment, our focus is on the performance of the information criteria in selecting the true spatial weights matrix under heteroskedasticity. Figure 8 presents the histogram plots for the four information criteria in terms of selecting the true spatial weights matrix W_q when the DGP is M1 and $n = 100$. We observe that all differences are positive for the AIC, DIC, DIC_L and BIC. Therefore, all four information criteria select the correct spatial weights matrix W_q in all of the 300 samples. Figure 9 presents the histogram plots for the four information criteria in terms of selecting the true spatial weights matrix W_q when the DGP is M3 and $n = 100$. Again, we observe that all four information criteria select the correct spatial weights matrix W_q in all samples.

Lastly, we present the results from the fourth experiment, focusing on the performance of the information criteria in selecting the true DGP M3 against the other DGPs under heteroskedasticity. Figure 10 presents the histogram plots for the four information criteria in selecting M3 against M1 when $n = 225$. We observe that all four information criteria choose M3 over M1 in the majority of samples. Specifically, the percentage of positive differences is 99.7% for AIC, DIC and DIC_L, and 95% for BIC. Figure 11 presents the histogram plots for the four information criteria in selecting M3 against M2 when $n = 225$. We observe that that the differences are all positive, indicating that all four information criteria choose the correct DGP M3 against M2 in all of the 300 samples.

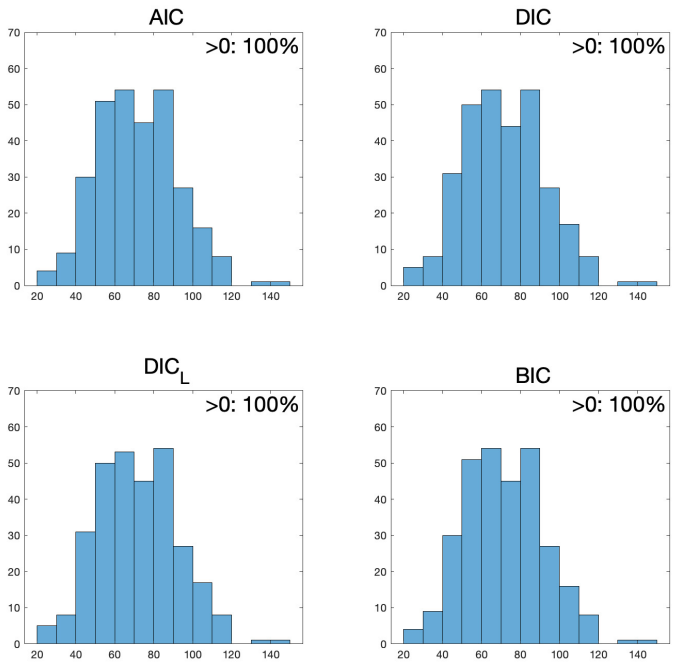

Figure 1: First experiment: Histogram plots for the information criteria of M1 with W_q minus the information criteria of M1 with W_r under homoskedasticity, $n = 100$.

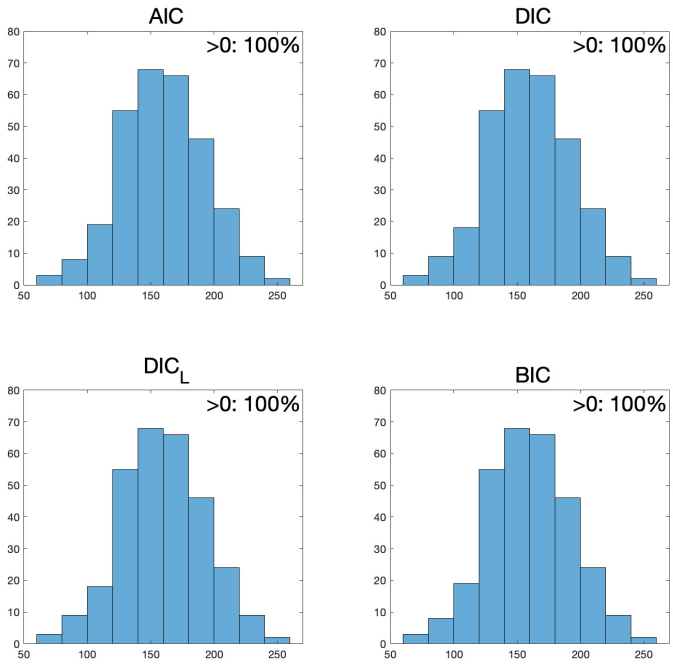

Figure 2: First experiment: Histogram plots for the information criteria of M1 with W_q minus the information criteria of M1 with W_r under homoskedasticity, $n = 225$.

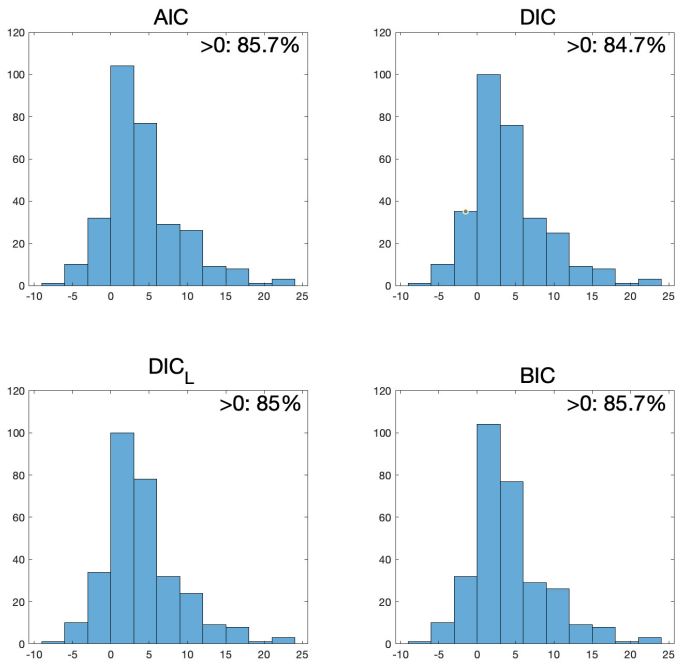

Figure 3: First experiment: Histogram plots for the information criteria of M2 with W_q minus the information criteria of M2 with W_r under homoskedasticity, $n = 100$.

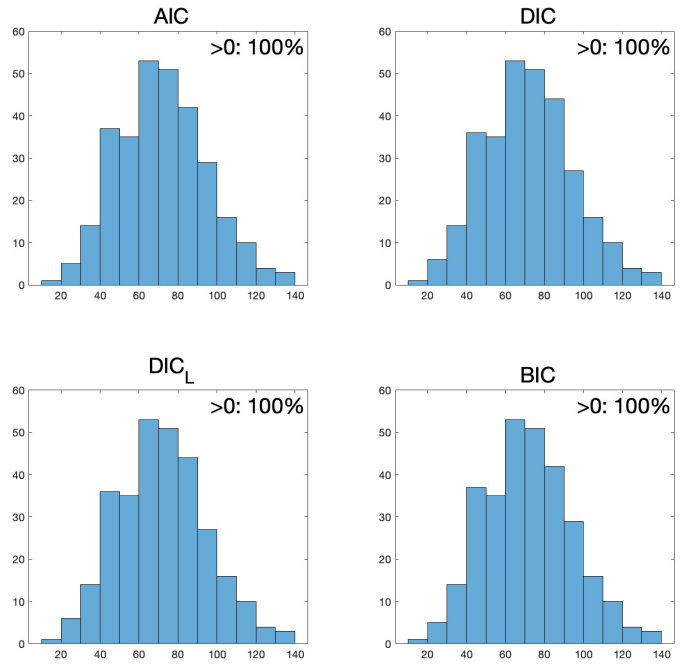

Figure 4: First experiment: Histogram plots for the information criteria of M3 with W_q minus the information criteria of M3 with W_r under homoskedasticity, $n = 100$.

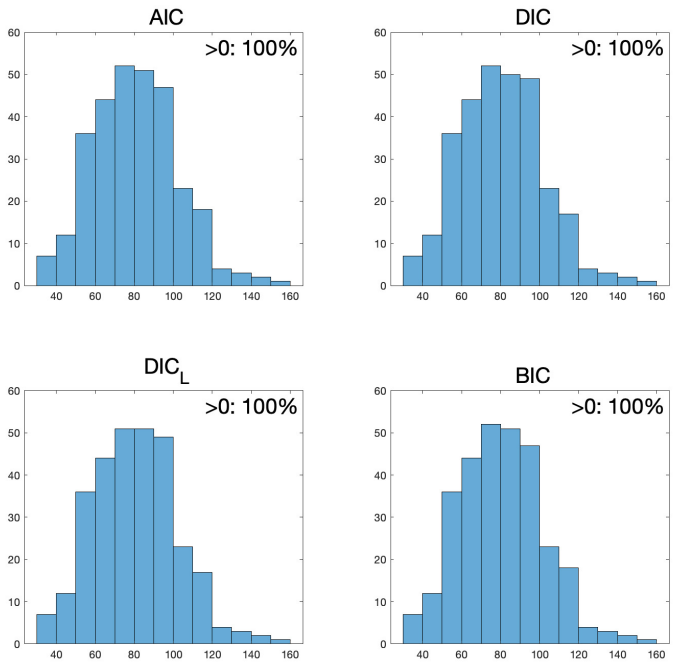

Figure 5: First experiment: Histogram plots for the information criteria of M4 with W_q minus the information criteria of M4 with W_r under homoskedasticity, $n = 100$.

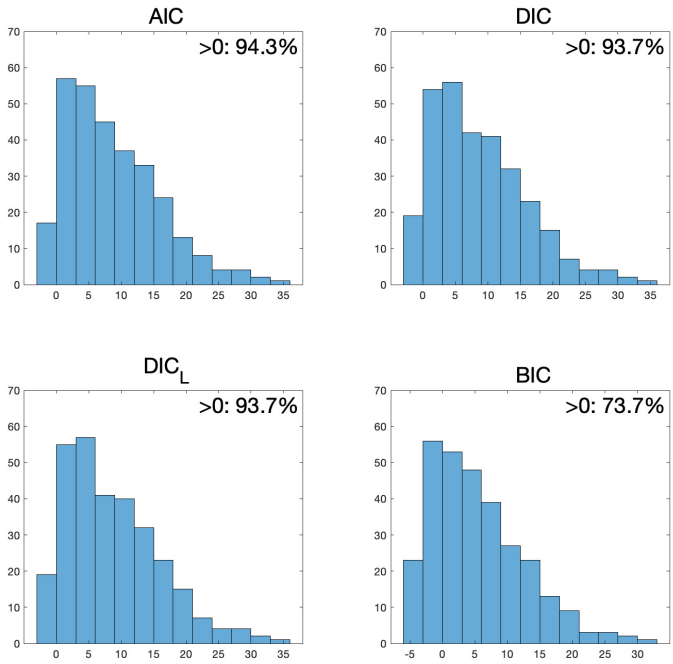

Figure 6: Second experiment: Histogram plots for the information criteria of M1 with W_r minus the information criteria of M3 with W_r under homoskedasticity, $n = 225$.

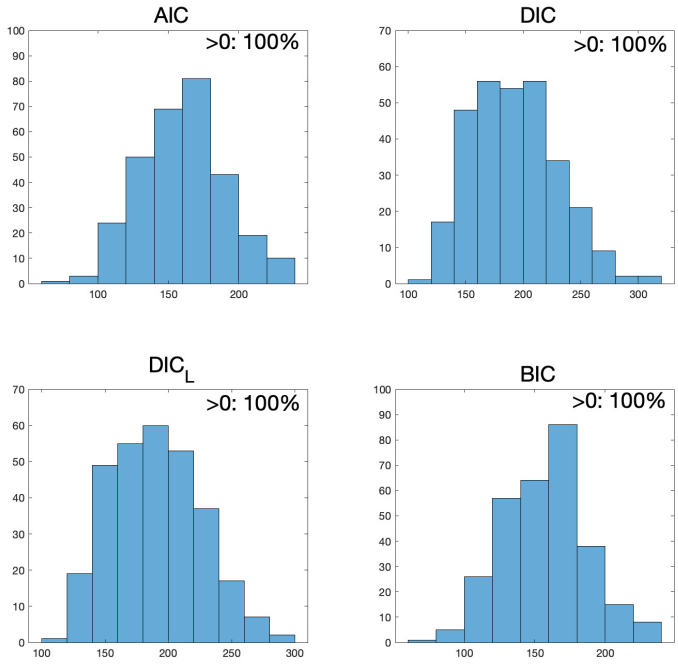

Figure 7: Second experiment: Histogram plots for the information criteria of M2 with W_r minus the information criteria of M3 with W_r under homoskedasticity, $n = 225$.

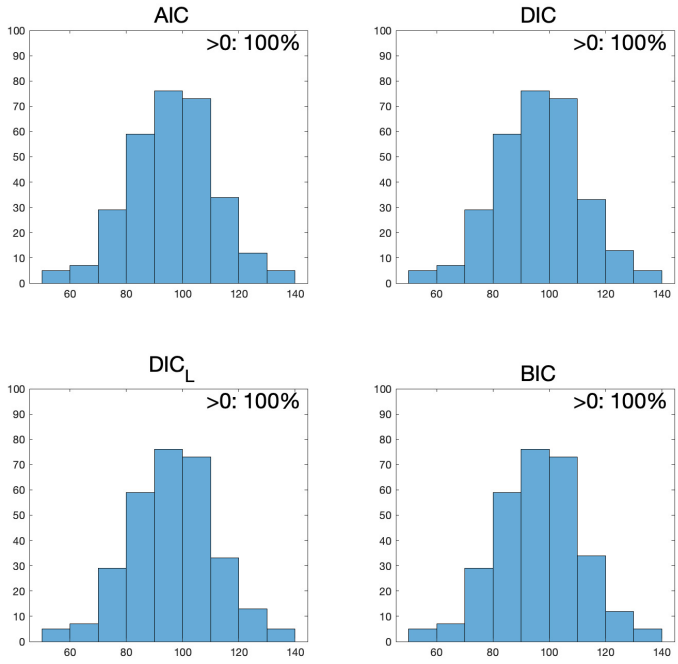

Figure 8: Third experiment: Histogram plots for the information criteria of M1 with W_r minus the information criteria of M1 with W_q under heteroskedasticity, $n = 100$.

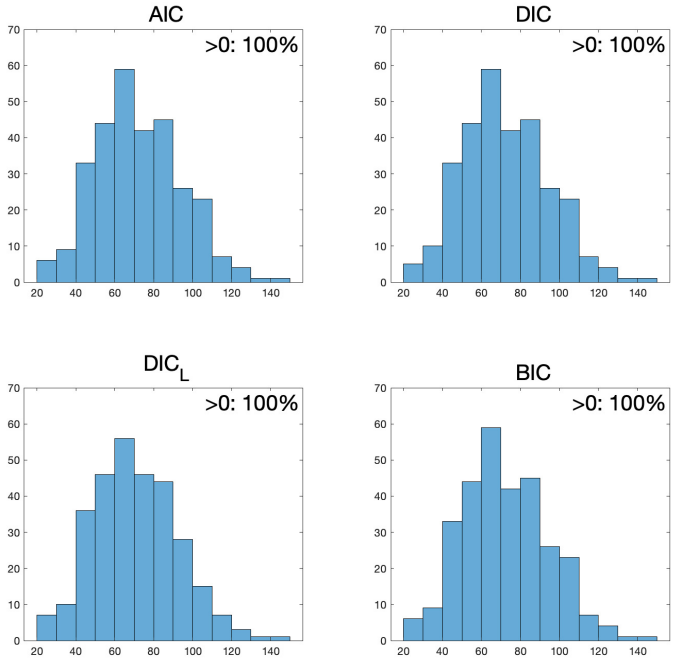

Figure 9: Third experiment: Histogram plots for the information criteria of M3 with W_r minus the information criteria of M3 with W_q under heteroskedasticity, $n = 100$.

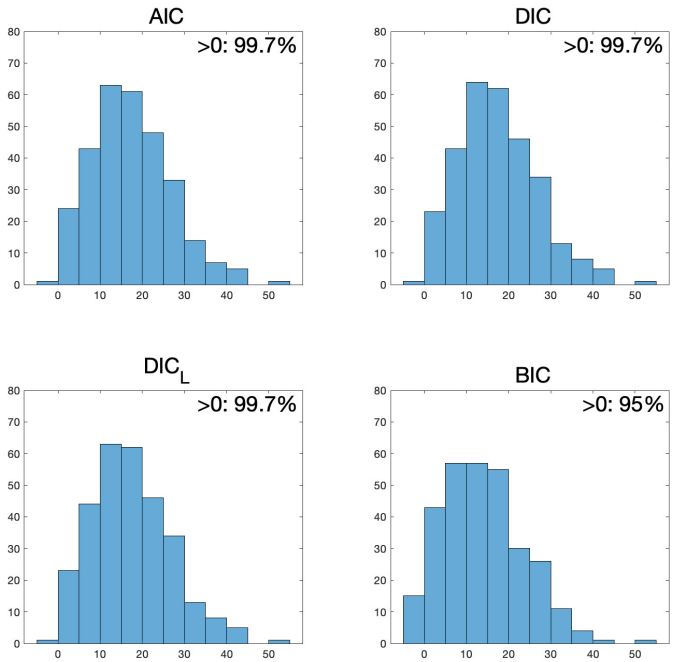

Figure 10: Fourth experiment: Histogram plots for the information criteria of M1 with W_q minus the information criteria of M3 with W_q under heteroskedasticity, $n = 225$.

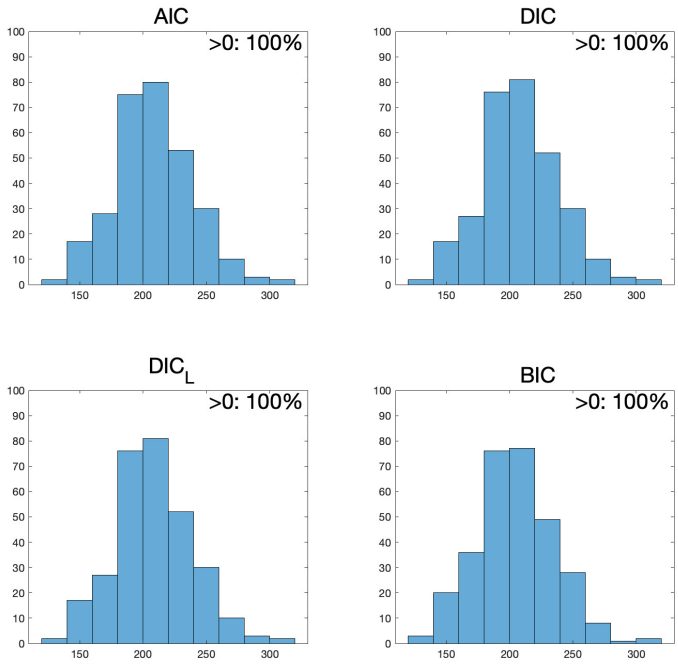

Figure 11: Fourth experiment: Histogram plots for the information criteria of M2 with W_q minus the information criteria of M3 with W_q under heteroskedasticity, $n = 225$.

6 Empirical Applications

6.1 A Spatial Growth Model

In this section, we show how the information criteria can be useful in model selection exercises. To this end, we consider the MESS-type counterpart of the spatial Durbin model considered in Ertur and Koch (2007) under both homoskedasticity and heteroskedasticity. Ertur and Koch (2007) (EK) incorporate technological interdependence into a neo-classical Solow growth model to explore the impact of technology spillover effects on economic growth. Their structural model yields a spatial Durbin model for the empirical analysis, and their findings indicate evidence for statistically significant spatial externalities. The empirical model suggested by EK takes the following form:

$$Y = \lambda WY + X\beta + \epsilon, \tag{6.1}$$

where Y is the logarithm of the $n \times 1$ vector of the output per-worker, X is the $n \times 5$ matrix containing the following variables: (i) the log of fraction of savings $\ln(s)$, (ii) the exogenous growth rate of labor variable $\ln(p + 0.05l_n)$ with l_n being a $n \times 1$ vector of ones, (iii) the spatial lag terms $W \ln(s)$ and $W \ln(p + 0.05l_n)$ and (iv) an intercept term, and ϵ is the $n \times 1$ vector of error terms.[7] We assume that the sum of the annual rate of depreciation of physical capital and the balanced growth rate of capital-output ratio is set to 0.05, which is a common assumption in the economic growth literature (Islam, 1995; Mankiw et al., 1992). EK consider two spatial weights matrices (i) $W_1 = (w_{1ij})$ and (ii) $W_2 = (w_{2ij})$, whose elements are specified as

$$w_{1ij} = \begin{cases} 0 & \text{if } i = j, \\ d_{ij}^{-2} & \text{if } i \neq j, \end{cases} \qquad w_{2ij} = \begin{cases} 0 & \text{if } i = j, \\ e^{-2\,d_{ij}} & \text{if } i \neq j, \end{cases} \tag{6.2}$$

where d_{ij} is the great-circle distance between country capitals. Both weights matrices are row normalized. In our context, we consider the MESS version of (6.1), which can be written as

$$e^{\lambda W} Y = X\beta + \nu, \tag{6.3}$$

where ν is the $n \times 1$ vector of error terms.

[7] In (6.1), we take the element-wise logarithm of vectors.

Table 1: List of countries

Angola	Argentina	Australia	Austria	Burundi
Belgium	Benin	Burkina Faso	Bangladesh	Bolivia
Brazil	Botswana	Central African Republic CAF	Canada	Congo, Republic of
Switzerland	Chile	Cote d'Ivoire	Cameroon	Colombia
Costa Rica	Denmark	Dominican Republic	Ecuador	Egypt
Spain	Ethiopia	Finland	France	United Kingdom
Ghana	Greece	Guatemala	Hong Kong	Honduras
Indonesia	India	Ireland	Israel	Italy
Jamaica	Jordan	Japan	Kenya	Korea, Republic of
Sri Lanka	Morocco	Madagascar	Mexico	Mali
Mozambique	Mauritania	Mauritius	Malawi	Malaysia
Niger	Nigeria	Nicaragua	Netherlands	Norway
Nepal	New Zealand	Pakistan	Panama	Peru
Philippines	Papua New Guinea	Portugal	Paraguay	Rwanda
Senegal	Singapore	Sierra Leone	El Salvador	Sweden
Syria	Chad	Togo	Thailand	Trinidad & Tobago
Tunisia	Turkey	Tanzania	Uganda	Uruguay
USA	Venezuela	South Africa	Congo, Dem. Rep.	Zambia
Zimbabwe				

Table 2: Descriptive Statistics

Variable	Obs	Mean	SD	Min	Max
log of output per worker (Y)	91	9.193	1.225	6.484	10.934
fraction of savings (s)	91	0.154	0.083	0.019	0.411
growth rate of labor (p)	91	0.022	0.009	0.003	0.043

Our sample data consist of the same data set used by EK. The data set is a cross-sectional data set on 91 countries for the year 1995. The list of countries is presented in Table 1, and the sample statistics for our variables are provided in Table 2. Using our estimation algorithms in Section 3, we estimate (6.3) with both W_1 and W_2 under both homoskedasticity and heteroskedasticity. The estimation results are reported in Table 3. In this table, we provide the mean and the standard deviation of the posterior draws. All four information criteria are reported in the bottom panel. In columns (1) and (2) of Table 3, we also reproduce EK's results for easy reference and compute the AIC and BIC. In columns (3) and (4), we present the estimation results for (6.3) under homoskedasticity. We observe that the estimates for $\ln s$ and $\ln(\mathrm{n} + 0.05l_n)$ are close to those from EK. However, while EK report statistically significant negative estimates for $W \ln s$, in columns (3) and (4), although the estimates are close, they are no longer statistically significant. The coefficient for $W \ln(\mathrm{n} + 0.05l_n)$ is estimated imprecisely similar to EK.

For the spatial parameter λ, we observe in columns (3) and (4) that estimates are negative around -0.85 and statistically significant. Note that these estimates are not directly comparable to the estimates of λ in columns (1) and (2). However, Debarsy et al. (2015) propose a relation between the spatial parameters of the SAR and MESS models, when the spatial weights matrix is row normalized: $\lambda_{\mathrm{SAR}} = 1 - e^{\lambda_{\mathrm{MESS}}}$. Here, we observe that $\lambda_{\mathrm{SAR}} = 0.740$ and $\lambda_{\mathrm{MESS}} = -0.857$

for W_1, and $\lambda_{\text{SAR}} = 0.658$ and $\lambda_{\text{MESS}} = -0.822$ for W_2. These coefficients have opposite signs, and approximately satisfy the relation $\lambda_{\text{SAR}} = 1 - e^{\lambda_{\text{MESS}}}$. For the information criteria, we can see that AIC and BIC have smaller values for W_1, which implies W_1 is preferred over W_2 for the SAR model. For the MESS model, AIC, DIC, DIC_{L} and BIC have smaller values for W_1, which also implies that W_1 is preferred over W_2.

Columns (5) and (6) present the estimation results under heteroskedasticity. The findings are in general very similar to those from the homoskedastic specifications in columns (3) and (4). One important difference occurs in the estimate of λ in column (6), which is smaller in magnitude. The estimates of ν indicate no significant deviations from the normality of the error terms. For the information criteria, we again observe that DIC, DIC_{L}, AIC and BIC have smaller values for W_1, which implies that W_1 is preferred over W_2. Across columns (3) through (6), the lowest values for all information criteria are observed in column (5).

Table 3: Estimation results and information criteria for the spatial growth model

	SAR		MESS			
	Homoskedasticity		Homoskedasticity		Heteroskedasticity	
	(1) W_1	(2) W_2	(3) W_1	(4) W_2	(5) W_1	(6) W_2
Constant	0.988	0.530	1.288	0.806	1.139	1.807
	(0.602)	(0.778)	(1.781)	(1.799)	(1.788)	(1.800)
$\ln(s)$	0.825^{***}	0.792^{***}	0.949^{***}	0.893^{***}	0.957^{***}	0.873^{***}
	(0.000)	(0.000)	(0.116)	(0.121)	(0.116)	(0.117)
$\ln(p+0.05l_n)$	-1.498^{***}	-1.451^{***}	-1.662^{***}	-1.614^{***}	-1.673^{***}	-1.556^{**}
	(0.008)	(0.009)	(0.628)	(0.619)	(0.629)	(0.609)
$W\ln(s)$	-0.322^{***}	-0.372^{***}	-0.292	-0.332^{*}	-0.338	-0.062
	(0.079)	(0.024)	(0.223)	(0.192)	(0.223)	(0.198)
$W\ln(p+0.05l_n)$	0.571	0.137	0.149	-0.050	0.189	-0.345
	(0.501)	(0.863)	(0.842)	(0.788)	(0.844)	(0.787)
λ	0.740^{***}	0.658^{***}	-0.857^{***}	-0.822^{***}	-0.894^{***}	-0.581^{***}
	(0.000)	(0.000)	(0.115)	(0.102)	(0.113)	(0.105)
σ^2			0.334^{***}	0.349^{***}	0.333^{***}	0.355^{***}
			(0.052)	(0.054)	(0.052)	(0.056)
ν					24.724^{**}	27.615^{**}
					(12.462)	(12.311)
AIC	161.06	173.49	164.87	169.07	**156.93**	163.45
DIC			163.12	167.23	**155.26**	161.71
DIC_L			167.15	171.07	**164.11**	166.62
BIC	**156.08**	168.51	182.44	186.65	177.02	183.54

Significance levels: *: 10%, **: 5%, and ***: 1%.

6.2 A Carbon Emission Model

In this section, we consider a model of carbon emissions to investigate the relationship between carbon emissions and economic activity in the United States. In the literature, the state-level emissions are measured from the state-level energy use, and therefore the estimation equations considered in the literature are usually derived from a reduced-form energy demand model that links energy consumption to energy prices and aggregate economic activity (Aldy, 2005; Auffhammer and Steinhauser, 2007; Burnett and Madariaga, 2017; Burnett et al., 2013; Spinoni et al., 2018). In this empirical application, we consider the spatial extensions of the reduced-form energy demand, and use the information criteria to a select a suitable spatial specification.

We use state level data from the United States on emissions, economic activity, and some observed characteristics of the states. Our final data set consists of observations on 48 contiguous U.S. States from 1997 to 2019, excluding Alaska, Hawaii, and the District of Columbia. Therefore, in our sample, we have $n = 48$ and $T = 23$, which yields 1104 observations in total. The data on CO2 emissions are obtained from the U.S. Department of Energy, and they are calculated by multiplying a state's coal, natural gas and petroleum consumption by their respective thermal conversion factors. Consequently, our CO2 emissions variable does not represent actual emissions, but are estimates of actual emissions. It is measured in units of metric tonnes per person and is available from 1970 onwards at the annual frequency.

The economic activity in a given state and a given year is proxied by the real GDP data from the Bureau of Economic Analysis (BEA). Although GDP data are available from 1963 to current by state, we do not consider the years prior to 1997 due to the fact that there is a discontinuity because of a switch from SIC industry definitions to NAICS industry definitions. The BEA strongly cautions against appending GDP data for before-the-break and after-the-break periods. The real GDP values are in millions of chained 2012 dollars. The energy price data are obtained from the U.S. Energy Information Administration. These are state-level annual average prices of petroleum products, and average coal, natural gas, and electricity prices over all sectors. They are measured dollars per British thermal unit (Btu), and are available from 1970 onward at the annual frequency. Annual state population figures are available from the U.S. Census Bureau, and they represent resident population (including armed forces) in thousands.

An important covariate for explaining CO2 emissions is residential energy consumption. It has been documented in the literature that cooling degree days (cdd) and heating degree days (hdd) are highly correlated with the residential energy consumption (Burnett et al., 2013; Quayle and Diaz, 1980; Spinoni et al., 2018). The cdd and hdd are measures to reflect demand for energy to heat or cool houses and businesses. A mean daily temperature 65° Fahrenheit is the base for both heating and cooling degree day calculations. The cdd and hdd are summations of deviations from the base, and can be considered as a measure of accumulated cold and accumulated heat, respectively. The data on cdd and hdd are obtained from the Climate Analysis Center within the National Oceanic and Atmospheric Administration. Table 4 presents the summary statistics for our dataset.

Our main estimation specifications are the following four MESS specifications, which are the

Table 4: Descriptive statistics

Statistic	Obs.	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Population (thousands)	1,104	6,253.747	6,738.524	489	1,857	7,220.2	39,438
GDP (millions of chained 2012 $)	1,104	314,977.2	381,015.5	20,072	85,916.5	395,561.2	2,800,505
CO2 (metric tonnes per person)	1,104	114.897	110.114	5.453	50.156	142.990	684.690
Coal price ($ per Btu)	1,104	2.025	0.912	0.010	1.380	2.580	6.110
Electricity price ($ per Btu)	1,104	25.709	8.397	11.430	19.298	29.275	54.680
Petroleum price ($ per Btu)	1,104	17.306	6.700	4.770	10.505	21.745	30.070
Natural gas price ($ per Btu)	1,104	7.085	2.232	2.420	5.390	8.567	14.000
cdd (deviations in Fahrenheit)	1,104	1,177.951	803.638	118	566.5	1,604	4,125
hdd (deviations in Fahrenheit)	1,104	1,959.505	791.857	34	1,384	2,555.5	4,087

same as the specifications considered in Section 5:

$$
\begin{aligned}
M1: &\quad (I_T \otimes e^{\lambda W})Y = X\beta + V, \\
M2: &\quad Y = X\beta + U, \quad (I_T \otimes e^{\rho W})U = V, \\
M3: &\quad (I_T \otimes e^{\lambda W})Y = X\beta + U, \quad (I_T \otimes e^{\rho W})U = V, \\
M4: &\quad (I_T \otimes e^{\lambda W})Y = X\beta + (I_T \otimes W)X_1\psi + V.
\end{aligned}
$$

where Y is the logarithm of CO2 emissions, I_T is the $T \times T$ identity matrix, $\otimes$ denotes the Kronecker product, and X contains the following explanatory variables: (i) the logarithm of coal price, (ii) the logarithm electricity price, (iii) the logarithm of natural gas price, (iv) the logarithm of petroleum price, (v) the logarithm of cdd, (vi) the logarithm of hdd, (vii) the logarithm of GDP per capita, and (viii) the square of the logarithm of GDP per capita. The spatial weights matrix W is specified based on the contiguity scheme such that its (i, j)th element is set to 1 if states i and j share a common border, otherwise to 0. The weights matrix created in this way is then row normalized.

We estimate each of the four specifications under homoskedastic and heteroskedastic assumptions. The estimated posterior means and standard deviations for each parameter are given in Table 5. There are two important observations.[8] First, all information criteria prefer the heteroskedastic models over the homoskedastic models. For example, comparing column (5) with (1), the AIC, DIC, DIC_L and BIC for the heteroskedastic models are 2299.75, 2305.48, 2002.73 and 2359.83 respectively, which are correspondingly smaller than those for the homoskedastic model, which are 2448.89, 2462.85, 2497.09 and 2503.97 respectively. This result is confirmed by the estimates of ν, which are around 2 for all heteroskedastic models. The second observation is that, among both homoskedastic and heteroskedastic models, all information criteria suggest the spatial Durbin version of the MESS model $M4$, except the DIC_L in the heteroskedastic case. This

[8]Though our data are spatiotemporal, we did not consider the time lag, the spatiotemporal lag, the unobserved spatial fixed effects, and the time fixed effects in M1-M4, because our focus is on cross-sectional MESS models. We report the Durbin-Watson (DW) and the Moran's I statistics in the table as well. These statistics are computed by using the residuals obtained from the non-spatial linear models. Both statistics suggest the extension of the non-spatial models to the spatial models. Although a spatiotemporal model may be more suitable for this application, we did not explore such a model in this paper as it is beyond the scope of our paper.

observation is not surprising because most of the spatial Durbin terms are statistically significant as can be seen from the fourth and eight columns. In column (8), the AIC, DIC and BIC for M4 are 2040.07, 2038.87 and 2140.21 respectively, which are smaller than those of M1, M2 and M3. The only exception is $\mathrm{DIC_L}$, which is slightly bigger than that of M1. Overall, the spatial Durbin version (M4) seems to provide a better fit for our sample data.

Table 5: Estimation results for the carbon emission model

	Homoskedasticity				Heteroskedasticity			
	(1) M1	(2) M2	(3) M3	(4) M4	(5) M1	(6) M2	(7) M3	(8) M4
Constant	-6.927^{***}	-5.823^{**}	2.968	-3.769	-6.118^{**}	-1.966	0.985	-4.265
	(2.644)	(2.624)	(2.697)	(2.913)	(2.578)	(2.506)	(2.431)	(2.863)
$\log(coal\_p)$	0.231^{***}	0.204^{***}	0.108^{***}	0.160^{***}	0.314^{***}	0.234^{***}	0.184^{***}	0.210^{***}
	(.027)	(.024)	(.016)	(.024)	(.018)	(.016)	(.016)	(.021)
$\log(elec\_p)$	-0.776^{***}	-0.402^{***}	0.266^{**}	0.819^{***}	-0.755^{***}	0.037	0.251^{**}	0.539^{***}
	(.096)	(.154)	(.116)	(.153)	(.072)	(.110)	(.102)	(.168)
$\log(natug\_p)$	0.054	0.179	0.321^{***}	0.385^{***}	0.218^{***}	0.571^{***}	0.483^{***}	0.572^{***}
	(.080)	(.120)	(.108)	(.146)	(.056)	(.087)	(.084)	(.116)
$\log(petro\_p)$	-0.105	-0.386^{***}	-1.413^{***}	-2.949^{***}	-0.262^{***}	-0.795^{***}	-0.961^{***}	-2.151^{***}
	(.067)	(.123)	(.226)	(.354)	(.052)	(.100)	(.124)	(.312)
$\log(cdd)$	0.047	-0.100	-0.272^{***}	-0.346^{***}	0.010	-0.104^{**}	-0.077^{*}	-0.067
	(.056)	(.070)	(.056)	(.066)	(.041)	(.046)	(.044)	(.058)
$\log(hdd)$	-0.585^{***}	-0.719^{***}	-0.485^{***}	-0.398^{***}	-0.331^{***}	-0.362^{***}	-0.311^{***}	-0.297^{***}
	(.067)	(.075)	(.064)	(.078)	(.042)	(.044)	(.042)	(.055)
$\log(gdp\_pc)$	6.980^{***}	7.664^{***}	8.555^{***}	4.596^{**}	5.161^{***}	3.874^{***}	3.889^{***}	0.708
	(1.401)	(1.405)	(1.339)	(1.913)	(1.352)	(1.313)	(1.236)	(1.830)
$(\log(gdp\_pc))^2$	-0.705^{***}	-0.768^{***}	-0.951^{***}	-0.445^{*}	-0.436^{**}	-0.270	-0.310^{*}	0.127
	(.181)	(.182)	(.174)	(.246)	(.177)	(.171)	(.161)	(.238)
$W\log(\text{coal\_}p)$				-0.058				0.161^{***}
				(.056)				(.060)
$W\log(elec\_p)$				-1.040^{***}				-0.796^{***}
				(.188)				(.192)
$W\log(natug\_p)$				-0.257				-0.338^{***}
				(.160)				(.124)
$W\log(petro\_p)$				3.010^{***}				1.937^{***}
				(.362)				(.324)
$W\log(cdd)$				1.004^{***}				0.589^{***}
				(.106)				(.103)
$W\log(hdd)$				0.689^{***}				0.804^{***}
				(.138)				(.111)
$W\log(gdp\_pc)$				-3.446				-0.264
				(2.061)				(1.990)
$W(\log(gdp\_pc))^2$				0.151				-0.247
				(.266)				(.257)
λ	-0.411^{***}		1.166^{***}	-0.457^{***}	-0.663^{***}		0.501^{***}	-0.670^{***}
	(.041)		(.097)	(.039)	(.040)		(.078)	(.050)
ρ		-0.637^{***}	-1.950^{***}			-0.953^{***}	-1.413^{***}	
		(.067)	(.100)			(.039)	(.081)	
σ^2	0.540^{***}	0.536^{***}	0.509^{***}	0.390^{***}	0.167^{***}	0.143^{***}	0.138^{***}	0.139^{***}
	(.023)	(.023)	(.021)	(.017)	(.012)	(.010)	(.010)	(.010)
ν					2.005^{***}	2.005^{***}	2.005^{***}	2.005^{***}
					(.003)	(.003)	(.003)	(.003)
Moran's I	7.78	7.78	7.78	9.52	7.78	7.78	7.78	9.52
DW	1.74	1.74	1.74	1.81	1.74	1.74	1.74	1.81
AIC	2448.89	2443.39	2400.10	2109.78	2299.75	2214.79	2187.12	**2040.07**
DIC	2462.85	2452.87	2396.81	2108.64	2305.48	2212.58	2181.24	**2038.87**
DIC_L	2497.09	2456.67	2396.49	2110.82	**2002.73**	2209.99	2180.39	2013.00
BIC	2503.97	2498.46	2460.18	2204.91	2359.83	2274.87	2252.21	**2140.21**

Significance levels: *: 10%, **: 5%, and ***: 1%.

7 Conclusion

In this paper, we focused on the problems related to model specification in the MESS-type models. Specifically, we addressed the challenges of selecting a spatial weights matrix from a set of candidates and choosing between nested (or non-nested) alternative specifications. To resolve these specification problems in a Bayesian setting, we proposed using the DIC, AIC and BIC measures. Our approach has the advantage of being based on the integrated likelihood function, which is obtained analytically by integrating out the latent variables from the complete data likelihood function, in the heteroskedastic case. Our simulation results demonstrate that all information criteria perform well in finite samples. In two empirical applications, we demonstrated how to apply the proposed information criteria to a spatial growth model and a carbon emission model. In future studies, our approach can be extended to more general MESS-type models, such as a panel data MESS model with both entity and time fixed effects, or a panel data MESS model with interactive fixed effects.

Acknowledgement: We thank the editor, the associate editor, and an anonymous referee for many useful comments on the earlier versions. Ye Yang gratefully acknowledges the financial support from the research fund for new professors (XRZ2023042) and the Special Research Fund of Beijing for Capital University of Economics and Business (ZD202104).

Appendix

A Proof of Proposition 1

Let $Y(\delta) = e^{\rho M}\left(e^{\lambda W}Y - X\beta\right)$, and $y_i(\delta)$ be the ith element of $Y(\delta)$. We use some properties of the inverse-gamma distribution to determine $p(Y|\theta) = \int p(Y,\eta|\theta)\mathrm{d}\eta$.

$$
\begin{aligned}
p(Y|\theta) &= \int p(Y,\eta|\theta)\mathrm{d}\eta = \int p(Y|\eta,\theta)p(\eta|\theta)\mathrm{d}\eta \\
&= \int (2\pi)^{-n/2}(\sigma^2)^{-n/2}|H(\eta)|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}Y^{'}(\delta)H^{-1}(\eta)Y(\delta)\right) \\
&\times \prod_{i=1}^{n}\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\eta_i^{-(\frac{\nu}{2}+1)}\exp\left(-\frac{\nu}{2\eta_i}\right)\mathrm{d}\eta \\
&= \int (2\pi)^{-n/2}(\sigma^2)^{-n/2}\left(\prod_{i=1}^{n}\eta_i^{-1/2}\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{y_i^2(\delta)}{\eta_i}\right) \\
&\times \frac{(\nu/2)^{n\nu/2}}{\Gamma^n(\nu/2)}\left(\prod_{i=1}^{n}\eta_i^{-(\frac{\nu}{2}+1)}\right)\exp\left(\sum_{i=1}^{n}-\frac{\nu}{2\eta_i}\right)\mathrm{d}\eta \\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\times\frac{(\nu/2)^{n\nu/2}}{\Gamma^n(\nu/2)} \\
&\times \int\left(\prod_{i=1}^{n}\eta_i^{-(\frac{\nu+1}{2}+1)}\right)\exp\left(\sum_{i=1}^{n}-\frac{1}{\eta_i}\left(\frac{\nu}{2}+\frac{y_i^2(\delta)}{2\sigma^2}\right)\right)\mathrm{d}\eta \\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\times\frac{(\nu/2)^{n\nu/2}}{\Gamma^n(\nu/2)} \\
&\times \prod_{i=1}^{n}\int\left(\eta_i^{-(\frac{\nu+1}{2}+1)}\right)\exp\left(-\frac{1}{\eta_i}\left(\frac{\nu}{2}+\frac{y_i^2(\delta)}{2\sigma^2}\right)\right)\mathrm{d}\eta_i. \qquad \text{(A.1)}
\end{aligned}
$$

Note that if $z \sim \mathrm{IG}(\alpha,\beta)$, then $\int\frac{\beta^\alpha}{\Gamma(\alpha)}z^{-(\alpha+1)}\exp(-\frac{\beta}{z})\mathrm{d}z = 1$. Thus, we have $\int z^{-(\alpha+1)}\exp(-\frac{\beta}{z})dz = \beta^{-\alpha}\Gamma(\alpha)$. Then, setting $\alpha = \frac{\nu+1}{2}$ and $\beta_i = \left(\frac{\nu}{2}+\frac{y_i^2(\delta)}{2\sigma^2}\right)$ in (A.1), we obtain

$$
\prod_{i=1}^{n}\int\left(\eta_i^{-(\frac{\nu+1}{2}+1)}\right)\exp\left(-\frac{1}{\eta_i}\left(\frac{\nu}{2}+\frac{y_i^2(\delta)}{2\sigma^2}\right)\right)\mathrm{d}\eta = \Gamma^n(\frac{\nu+1}{2})\prod_{i=1}^{n}\left(\frac{\nu}{2}+\frac{y_i^2(\delta)}{2\sigma^2}\right)^{-(\frac{\nu+1}{2})}. \qquad \text{(A.2)}
$$

Using (A.2) in (A.1), we obtain

$$
p(Y|\theta) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\frac{(\nu/2)^{n\nu/2}\Gamma^n(\frac{\nu+1}{2})}{\Gamma^n(\nu/2)}\times\prod_{i=1}^{n}\left(\frac{\nu}{2}+\frac{y_i^2(\delta)}{2\sigma^2}\right)^{-(\frac{\nu+1}{2})}.
$$

Thus, the log-integrated likelihood function is

$$\begin{aligned}\ln p(Y|\theta) = &-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \frac{n\nu}{2}\ln(\nu/2) \\ &+ n\ln\Gamma(\frac{\nu+1}{2}) - n\ln\Gamma(\nu/2) - \frac{\nu+1}{2}\sum_{i=1}^{n}\ln\left(\frac{\nu}{2} + \frac{y_i^2(\delta)}{2\sigma^2}\right).\end{aligned}$$

References

Ahrens, Achim and Arnab Bhattacharjee (2015). "Two-Step Lasso Estimation of the Spatial Weights Matrix". In: *Econometrics* 3.1, pp. 128–155.

Akaike, Hirotogu (1973). "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. Ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. New York: Springer, pp. 199–213.

Aldy, Joseph E. (2005). "An Environmental Kuznets Curve Analysis of U.S. State-Level Carbon Dioxide Emissions". In: *The Journal of Environment & Development* 14.1, pp. 48–72.

Anselin, Luc (1984a). "Specification Tests and Model Selection for Aggregate Spatial Interaction: An Empirical Comparison". In: *Journal of Regional Science* 24.1, pp. 1–15.

— (1984b). "Specification tests on the structure of interaction in spatial econometric models". In: *Papers of the Regional Science Association* 54, pp. 165–182.

— (1988). *Spatial econometrics: Methods and Models*. New York: Springer.

Auffhammer, Maximilian and Ralf Steinhauser (2007). "The future trajectory of US CO2 emissions: the role of state vs aggregate information". In: *Journal of Regional Science* 47.1, pp. 47–61.

Behrens, Kristian, Cem Ertur, and Wilfried Koch (2012). "Dual Gravity: Using Spatial Econometrics to Control for Multilateral Resistance". In: *Journal of Applied Econometrics* 27.5, pp. 773–794.

Berg, Andreas, Renate Meyer, and Jun Yu (2004). "Deviance Information Criterion for Comparing Stochastic Volatility Models". In: *Journal of Business & Economic Statistics* 22.1, pp. 107–120.

Brueckner, Jan K. (1998). "Testing for Strategic Interaction Among Local Governments: The Case of Growth Controls". In: *Journal of Urban Economics* 44.3, pp. 438 –467.

Brueckner, Jan K. and Luz A. Saavedra (2001). "Do Local Governments Engage in Strategic Property-Tax Competition?" In: *National Tax Journal* 54.2, pp. 203 –230.

Burnett, J. Wesley and Jessica Madariaga (2017). "The convergence of U.S. state-level energy intensity". In: *Energy Economics* 62, pp. 357–370.

Burnett, J. Wesley, John C. Bergstrom, and Jeffrey H. Dorfman (2013). "A spatial panel data approach to estimating U.S. state-level energy emissions". In: *Energy Economics* 40, pp. 396–404.

Burnham, K.P. and D.R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Second Edition. New York: Springer.

Burridge, Peter and Bernard Fingleton (2010). "Bootstrap Inference in Spatial Econometrics: the J-test". In: *Spatial Economic Analysis* 5.1, pp. 93–119.

Case, Anne C., Harvey S. Rosen, and James R. Hines (1993). "Budget spillovers and fiscal policy interdependence". In: *Journal of Public Economics* 52.3, pp. 285 –307.

Celeux, G. et al. (Dec. 2006). "Deviance information criteria for missing data models". In: *Bayesian Analysis* 1.4, pp. 651–673.

Chan, Joshua C.C. and Angelia L. Grant (2016a). "Fast computation of the deviance information criterion for latent variable models". In: *Computational Statistics & Data Analysis* 100, pp. 847–859.

— (2016b). "On the Observed-Data Deviance Information Criterion for Volatility Modeling". In: *Journal of Financial Econometrics* 14.4, pp. 772–802.

Chernozhukov, Victor and Han Hong (2003). "An MCMC approach to classical estimation". In: *Journal of Econometrics* 115.2, pp. 293–346.

Chiu, Tom Y. M., Tom Leonard, and Kam-Wah Tsui (1996). "The Matrix-Logarithmic Covariance Model". In: *Journal of the American Statistical Association* 91.433, pp. 198–210.

Conley, Timothy G. and Ethan Ligon (2002). "Economic Distance and Cross-Country Spillovers". In: *Journal of Economic Growth* 7.2, pp. 157–187.

Conley, Timothy G. and Giorgio Topa (2002). "Socio-economic distance and spatial patterns in unemployment". In: *Journal of Applied Econometrics* 17.4, pp. 303–327.

Davidson, Russell and James G. MacKinnon (1981). "Several Tests for Model Specification in the Presence of Alternative Hypotheses". In: *Econometrica* 49.3, pp. 781–793.

Debarsy, Nicolas, Fei Jin, and Lung fei Lee (2015). "Large sample properties of the matrix exponential spatial specification with an application to FDI". In: *Journal of Econometrics* 188.1, pp. 1–21.

Elhorst, J. Paul (2014). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer Briefs in Regional Science. New York: Springer Berlin Heidelberg.

Ertur, Cem and Wilfried Koch (2007). "Growth, technological interdependence and spatial externalities: theory and evidence". In: *Journal of Applied Econometrics* 22.6, pp. 1033–1062.

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

Getis, Arthur and Jared Aldstadt (2004). "Constructing the Spatial Weights Matrix Using a Local Statistic". In: *Geographical Analysis* 36.2, pp. 90–104.

Geweke, J. (1993). "Bayesian treatment of the independent student-t linear model". In: *Journal of Applied Econometrics* 8.S1, S19–S40.

Han, Xiaoyi and Lung fei Lee (2013). "Model selection using J-test for the spatial autoregressive model vs. the matrix exponential spatial model". In: *Regional Science and Urban Economics* 43.2, pp. 250–271.

Han, Xiaoyi and Lung-Fei Lee (2016). "Bayesian Analysis of Spatial Panel Autoregressive Models With Time-Varying Endogenous Spatial Weight Matrices, Common Factors, and Random Coefficients". In: *Journal of Business & Economic Statistics* 34.4, pp. 642–660.

Han, Xiaoyi, Lung-Fei Lee, and Xingbai Xu (2021). "Large sample properties of Bayesian estimation of spatial econometric models". In: *Econometric Theory* 37.4, 708–746.

Islam, Nazrul (1995). "Growth Empirics: A Panel Data Approach". In: *The Quarterly Journal of Economics* 110.4, pp. 1127–1170. (Visited on 06/07/2023).

Jin, Fei and Lung fei Lee (2013). "Cox-type tests for competing spatial autoregressive models with spatial autoregressive disturbances". In: *Regional Science and Urban Economics* 43.4, pp. 590 –616.

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795.

Kelejian, Harry H. (2008). "A spatial J-test for model specification against a single or a set of non-nested alternatives". In: *Letters in Spatial and Resource Sciences* 1.1, pp. 3–11.

Kelejian, Harry H. and Gianfranco Piras (2011). "An extension of Kelejian's J-test for non-nested spatial models". In: *Regional Science and Urban Economics* 41.3, pp. 281–292.

Kelejian, Harry H. and Ingmar R. Prucha (2010). "Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances". In: *Journal of Econometrics* 157, pp. 53–67.

Lam, Clifford and Pedro C.L. Souza (2020). "Estimation and Selection of Spatial Weight Matrix in a Spatial Lag Model". In: *Journal of Business & Economic Statistics* 38.3, pp. 693–710.

Lee, Lung-fei (2004). "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models". In: *Econometrica* 72.6, pp. 1899–1925.

LeSage, James P. and R. Kelley Pace (2007). "A matrix exponential spatial specification". In: *Journal of Econometrics* 140, pp. 190–214.

— (2009). *Introduction to Spatial Econometrics*. London: Chapman and Hall/CRC.

Li, Yong, Jun Yu, and Tao Zeng (2020). "Deviance information criterion for latent variable models and misspecified models". In: *Journal of Econometrics* 216.2, pp. 450–493.

Liu, Tuo and Lung fei Lee (2019). "A likelihood ratio test for spatial model selection". In: *Journal of Econometrics* 213.2, pp. 434–458.

MacKinnon, James G., Halbert White, and Russell Davidson (1983). "Tests for model specification in the presence of alternative hypotheses: Some further results". In: *Journal of Econometrics* 21.1, pp. 53–70.

Mankiw, N. Gregory, David Romer, and David N. Weil (1992). "A Contribution to the Empirics of Economic Growth". In: *The Quarterly Journal of Economics* 107.2, pp. 407–437.

Merk, Miryam S. and Philipp Otto (2022). "Estimation of the spatial weighting matrix for regular lattice data-An adaptive lasso approach with cross-sectional resampling". In: *Environmetrics* 33.1.

Millar, Russell B. (2009). "Comparison of Hierarchical Bayesian Models for Overdispersed Count Data using DIC and Bayes' Factors". In: *Biometrics* 65.3, pp. 962–969.

Parent, Olivier and James P. LeSage (2008). "Using the variance structure of the conditional autoregressive spatial specification to model knowledge spillovers". In: *Journal of Applied Econometrics* 23.2, pp. 235–256.

Qu, Xi and Lung-fei Lee (2015). "Estimating a spatial autoregressive model with an endogenous spatial weight matrix". In: *Journal of Econometrics* 184.2, pp. 209 –232.

Qu, Xi, Lung fei Lee, and Jihai Yu (2017). "QML estimation of spatial dynamic panel data models with endogenous time varying spatial weights matrices". In: *Journal of Econometrics* 197.2, pp. 173 –201.

Quayle, Robert G. and Henry F. Diaz (1980). "Heating Degree Day Data Applied to Residential Heating Energy Consumption". In: *Journal of Applied Meteorology (1962-1982)* 19.3, pp. 241–246.

Ritter, Christian and Martin A. Tanner (1992). "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler". In: *Journal of the American Statistical Association* 87.419, pp. 861–868.

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461 –464.

Spiegelhalter, David J. et al. (2002). "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639.

Spinoni, Jonathan et al. (2018). "Changes of heating and cooling degree-days in Europe from 1981 to 2100". In: *International Journal of Climatology* 38.S1, e191–e208.

Yang, Ye, Osman Doğan, and Süleyman Taşpınar (2021). "Fast Estimation of Matrix Exponential Spatial Models". In: *Journal of Spatial Econometrics* 2.9.

— (2022). "Model selection and model averaging for matrix exponential spatial models". In: *Econometric Reviews* 41.8, pp. 827–858.