# Cross-Sectional Matrix Exponential Spatial Models: A Comprehensive Review and Some New Results

Ye Yang* Osman Dogan† Suleyman Taspinar‡ Fei Jin§

January 27, 2025

**Abstract**

In this paper, we provide a comprehensive review of the literature on estimation, inference, and model selection approaches for cross-sectional matrix exponential spatial models. We first discuss the properties of the matrix exponential specification in modeling cross-sectional dependence in comparison to the spatial autoregressive specification. We then provide a survey of the existing estimation and inference methods for cross-sectional matrix exponential spatial models. We carefully discuss summary measures for the marginal effects of regressors, detail the matrix-vector product method for efficient computation of matrix exponential terms, and then explore model selection approaches. Our aim is not only to summarize the main findings from the spatial econometric literature but also to make them more accessible to applied researchers. Additionally, we contribute to the literature by presenting several new results. We propose an M-estimation approach for models with heteroskedastic error terms and demonstrate that the resulting M-estimator is consistent and asymptotically normally distributed. Moreover, we provide additional results for model selection exercises. Finally, in a Monte Carlo study, we evaluate the finite sample properties of various estimators from the literature alongside the M-estimator.

JEL-Classification: C10, C13, C15, C21.

Keywords: Matrix exponential spatial specification, MESS, spatial autoregression, SAR, heteroskedasticity, Bayesian estimation, model selection, impact measures.

*Corresponding author, School of Accounting, Capital University of Economics and Business, 121 Zhangjialukou, Fengtai District, Beijing, 100070, China, Email: yang.ye@cueb.edu.cn.

†Department of Economics, Istanbul Technical University, Maslak, 34485 Istanbul, Turkey, Email: osmandogan@itu.edu.tr.

‡Department of Economics, Queens College, The City University of New York, 65-30 Kissena Blvd, Queens, NY, 11367, U.S.A., Email: staspinar@qc.cuny.edu.

§School of Economics, Fudan University, 220 Handan Road, Yangpu district, Shanghai, 200437, China, Email: jinfei@fudan.edu.cn.

# 1 Introduction and motivation

Spatial econometric models deal with estimation and inference problems that arise from (weak) cross-sectional dependence or correlation in data marked with location stamps. The spatial autoregressive (SAR) model has been a widely used approach for modeling spatial dependence since its inception in Whittle (1954) and Cliff and Ord (1969, 1973). The matrix exponential spatial specification (MESS) was introduced by LeSage and Pace (2007) as an alternative to the SAR specification, primarily due to its computationally appealing properties in likelihood-based estimation schemes. Despite various estimation and inference methods proposed in the econometrics literature for models using either specification, the empirical literature is predominantly populated with papers utilizing the SAR specification. For instance, a Google Scholar search for "spatial autoregressive model" yields 13,100 results, whereas a search for "matrix exponential spatial specification" returns only 212 results.

This highly skewed preference towards the SAR specification by applied researchers is unfortunate in the sense that the MESS attains some attractive properties. First, we must emphasize that the MESS and the SAR imply different rates of decay for cross-sectional dependence. While it is a geometric rate in the case of SAR, the MESS implies an exponential rate of decay for spatial correlation. Consequently, they also imply different reduced forms for the cross-sectional models. Second, contrary to the SAR specification, the MESS does not require any restrictions on the parameter space of the spatial autoregressive parameters as the reduced form of the MESS always exists. In particular, the MESS always yields a positive definite covariance matrix for the outcome variable. Third, in the likelihood-based estimation, the log-likelihood function of a SAR specification includes Jacobian determinant terms that can be difficult to compute when the number of cross-sections is large. The MESS log-likelihood function, on the other hand, does not involve such Jacobian terms.

In this paper, our aim is to provide a complete comprehensive review of the econometric literature on the estimation, inference, and model selection methods for the cross-sectional MESS-type models.[1] More specifically, we aim to present the existing results from the literature in a more accessible way so that they can be utilized easily in empirical applications by applied researchers. Furthermore, we extend the existing literature in some important respects. Firstly, we propose a new estimation and inference methodology for the cross-sectional MESS-type models with an unknown form of heteroskedasticity. Second, for the model selection problems involving cross-sectional MESS-type models, we consider a new method for computing the marginal likelihoods of the competing models in a Bayesian framework. In a Monte Carlo study, we also assess the finite sample properties of some existing estimators from the literature along with our proposed method. The simulation results show that the suggested M-estimator performs satisfactorily in finite samples.

[1] We focus on cross-sectional MESS models because there are a few papers on panel data MESS models in the literature. See, e.g., LeSage and Chih (2018), Zhang et al. (2019), Yang (2022) and Yang et al. (2024).

Various estimation methods for the MESS models have been considered in the literature (LeSage and Pace, 2007; Debarsy et al., 2015; Yang et al., 2021, 2024). LeSage and Pace (2007) consider both the maximum likelihood and Bayesian estimation approaches. Debarsy et al. (2015) formally investigate the large sample properties of the quasi-maximum likelihood estimator (QMLE) and the generalized method of moments estimator (GMME). Although both estimators have the standard large sample properties, the GMME can be more efficient than the QMLE when the innovations are non-normal or heteroskedastic. Debarsy et al. (2015) showed that, unlike the SAR-type models, in the presence of an unknown form of heteroskedasticity, the QMLE of the MESS model can remain consistent if the spatial weights matrices used in the model are commutative. In this paper, we extend on their results by introducing an M-estimation methodology that is robust to heteroskedasticity when the spatial weights matrices do not commute. We also formally establish the large sample properties of the resulting M-estimator.

We provide Bayesian estimation algorithms for the MESS models in the case of both homoskedastic and heteroskedastic error terms (LeSage and Pace, 2007; Yang et al., 2021; Doğan et al., 2023). In the case of heteroskedasticity, we assume that the error terms follow a scale mixture of normal distributions, where the latent scale variables generate distributions with different variances. The latent variable representation facilitates the estimation through data augmentation techniques. In both homoskedastic and heteroskedastic models, the conditional posterior distributions of parameters take known forms, except for those of the spatial parameters. The posterior draws for the spatial parameters can be generated either by using the random-walk or the independence-chain Metropolis-Hastings algorithms (Lesage, 1997; LeSage and Parent, 2007; LeSage and Pace, 2009; Yang et al., 2021). We also consider the estimation of the MESS with endogenous and Durbin's regressors. Jin and Lee (2018) show that the popular nonlinear two stage least squares (N2SLS) estimator, although consistent, may suffer from slow rates of convergence, and may attain nonstandard asymptotic distributions when the true value of a subset of model parameters is zero. We highlight how an adaptive group lasso estimator can provide a solution to these problems.

One prominent issue for the estimation of MESS-type model relates to the computation of the matrix exponential terms. Although there are various methods suggested in the literature, there is no single method that outperforms the rest in all cases (Moler and Van Loan, 2003). As such, we visit the computation of the matrix exponential terms and exhibit how the matrix-vector product approach originally suggested by LeSage and Pace (2007) can be utilized for quick computation of these terms. A further issue for the MESS models relates to the interpretation of the coefficient estimates for the explanatory variables. In spatial models, the interpretation of the coefficient estimates for the explanatory variables becomes more complicated due to the cross-sectional interactions. To this end, we review the existing results on the estimation and inference results for the impact measures for the

MESS-type models (LeSage and Pace, 2009; Jin and Lee, 2018; Arbia et al., 2020).

When modeling spatial dependence, researchers may encounter specification problems related to choosing a spatial weights matrix from a pool of candidates or choosing between nested or non-nested alternative model specifications. Often, modeling is done in an ad hoc manner and there is no guidance from an underlying structural model to address these issues. To this end, we provide a complete review of the literature on testing based, criterion based and marginal likelihood-based approaches for model selection problems involving MESS-type models (LeSage and Pace, 2009; Han and Lee, 2013b; Liu and Lee, 2019; Yang et al., 2022; Doğan et al., 2023). In this regard, we also visit the Bayesian approaches and consider the modified harmonic mean method of Gelfand and Dey (1994) for the computation of the marginal likelihoods of competing models.

The rest of this paper is organized as follows. Section 2 reviews several cross-sectional MESS-type models. This section also introduces the main properties of a matrix exponential term. Section 3 details the matrix-vector product approach for the efficient computation of matrix exponential terms. Sections 4–8 discuss various estimation and inference techniques for the MESS-type models. Section 9 presents the impact measures for the MESS-type models and illustrates the inference methods for the impact measures. Section 10 considers various kinds of model selection approaches involving the MESS-type models. Section 11 presents results of a Monte Carlo study, focusing on the M-estimation of the MESS-type models. Section 12 ends our review with some concluding remarks for future research. Some technical results are relegated to an appendix.

## 2 Model Specification

We consider the following first-order matrix exponential spatial model (for short MESS(1, 1))

$$e^{\lambda_0 \mathbf{W}} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{U}, \quad e^{\rho_0 \mathbf{M}} \mathbf{U} = \mathbf{V}, \tag{2.1}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)'$ is the $n \times 1$ vector of observations on a dependent variable, $\mathbf{X}$ is the $n \times k$ matrix of non-stochastic exogenous variables with the associated parameter vector $\boldsymbol{\beta}_0$, $\mathbf{U} = (u_1, \ldots, u_n)'$ is the $n \times 1$ vector of regression error terms, and $\mathbf{V} = (v_1, \ldots, v_n)'$ is the $n \times 1$ vector of idiosyncratic error terms. We follow the literature to assume that the elements of $\mathbf{X}$ are non-stochastic for simplicity (Kelejian and Prucha, 1998; Lee, 2004). Alternatively, the elements of $\mathbf{X}$ can be assumed to be stochastic with a finite moment of certain order. The matrix exponential term $e^{\lambda_0 \mathbf{W}}$ is defined by $e^{\lambda_0 \mathbf{W}} = \sum_{i=0}^{\infty} \frac{\lambda_0^i \mathbf{W}^i}{i!}$, where $\mathbf{W}$ is an $n \times n$ spatial weights matrix with zero diagonal elements and $\lambda_0$ is a scalar spatial parameter. The matrix exponential $e^{\rho_0 \mathbf{M}}$ is defined in a similar way, where $\mathbf{M}$ is another $n \times n$ spatial weights matrix and $\rho_0$ is a scalar spatial parameter.

The MESS(1,1) in (2.1) can be considered as the matrix exponential counterpart of the spatial

autoregressive model with spatial autoregressive disturbances (SARAR(1,1)),

$$(\mathbf{I}_n - \alpha_0\mathbf{W})\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{U}, \quad (\mathbf{I}_n - \tau_0\mathbf{M})\mathbf{U} = \mathbf{V}, \tag{2.2}$$

where $\alpha_0$ and $\tau_0$ are scalar spatial autoregressive parameters. The MESS(1,1) specification is obtained from (2.2) by replacing $(\mathbf{I}_n - \alpha_0\mathbf{W})$ and $(\mathbf{I}_n - \tau_0\mathbf{M})$ with $e^{\lambda_0\mathbf{W}}$ and $e^{\rho_0\mathbf{M}}$, respectively. The matrix exponential terms satisfy the following properties (LeSage and Pace, 2007):

1. $e^{c\mathbf{A}}$ is non-singular, where $\mathbf{A}$ is an $n \times n$ matrix and $c$ is a scalar constant,
2. $(e^{c\mathbf{A}})^{-1} = e^{-c\mathbf{A}}$,
3. $\left|e^{c\mathbf{A}}\right| = e^{c\,\mathrm{tr}(\mathbf{A})}$, where $|\cdot|$ is the determinant operator and $\mathrm{tr}(\cdot)$ is the trace operator,
4. $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$, where $\mathbf{A}$ and $\mathbf{B}$ are two $n \times n$ matrices satisfying the commutative property $\mathbf{AB} = \mathbf{BA}$.

Because of these properties, the spatial models formulated with the matrix exponential terms have some advantages over the spatial models formulated with the spatial autoregressive processes. The first and second properties ensure that the reduced form of matrix exponential models always exists and does not require any restrictions for the spatial parameters. In the context of (2.1), the reduced form can be expressed as

$$\mathbf{Y} = e^{-\lambda_0\mathbf{W}}\mathbf{X}\boldsymbol{\beta}_0 + e^{-\lambda_0\mathbf{W}}e^{-\rho_0\mathbf{M}}\mathbf{V}. \tag{2.3}$$

This reduced form suggests an exponential pattern of decay for the influence of high-order neighboring characteristics while the reduced form of a SAR process gives a geometric decay for the influence of high-order neighboring characteristics. We note that analogous to time series literature, fractionally differenced and fractionally integrated processes can also be considered for allowing possible slowly decaying rates for high-order neighborhood characteristics (LeSage and Pace, 2009; Otto and Sibbertsen, 2023).

The third property implies that $\left|e^{\lambda\mathbf{W}}\right| = e^{\lambda\mathrm{tr}(\mathbf{W})} = 1$ because $\mathbf{W}$ has zero diagonal elements. This property ensures that the log-likelihood functions of matrix exponential models are free of any Jacobian terms that need to be computed many times during estimation (see Section 4 for the details). On the other hand, the likelihood functions of spatial autoregressive models are not free of Jacobian terms. For example, the likelihood function of the SARAR(1,1) model involves $|\mathbf{I}_n - \tau\mathbf{M}|$ and $|\mathbf{I}_n - \alpha\mathbf{W}|$, which must be computed in each iteration during estimation.

The MESS(1,1) specification nests two alternative specifications, namely, the MESS(1,0) and MESS(0,1), which can be obtained by setting $\lambda_0 = 0$ and $\rho_0 = 0$, respectively. A spatial Durbin

extension can be obtained by including the spatial lags of the explanatory variables as regressors:

$$e^{\lambda_0 \mathbf{W}}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{W}\mathbf{X}\boldsymbol{\delta}_0 + \mathbf{U}, \quad e^{\rho_0 \mathbf{M}}\mathbf{U} = \mathbf{V}, \tag{2.4}$$

where $\mathbf{WX}$ denotes the spatial lag of $\mathbf{X}$ and $\boldsymbol{\delta}_0$ is the corresponding vector of coefficients.

In the MESS(1,1) model, spatial interactions in the outcome variable arise only trough $\mathbf{W}$, and in the disturbance terms only through $\mathbf{M}$. In some cases, spatial dependence may arise from different sources, requiring different matrix exponential terms formulated with different spatial weights matrices. Let $\{\mathbf{W}_i\}_{i=1}^p$ and $\{\mathbf{M}_j\}_{j=1}^q$ be two sequences of spatial weights matrices. Then, following LeSage and Pace (2009) and Debarsy et al. (2015), a high-order version including the matrix exponential terms formulated with $\{\mathbf{W}_i\}_{i=1}^p$ and $\{\mathbf{M}_j\}_{j=1}^q$ can be specified as

$$e^{\sum_{i=1}^p \lambda_{i0} \mathbf{W}_i}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{U}, \quad e^{\sum_{j=1}^q \rho_{j0} \mathbf{M}_j}\mathbf{U} = \mathbf{V}, \tag{2.5}$$

where $\{\lambda_{i0}\}_{i=1}^p$ and $\{\rho_{j0}\}_{j=1}^q$ are sequences of spatial parameters. This model can be called the MESS$(p, q)$ model. We can alternatively define the high-order version in the following way:

$$\left(\prod_{i=1}^{p} e^{\lambda_{i0}\mathbf{W}_i}\right)\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{U}, \quad \left(\prod_{j=1}^{q} e^{\rho_{j0}\mathbf{M}_j}\right)\mathbf{U} = \mathbf{V}. \tag{2.6}$$

This version may not coincide with (2.5) because the fourth property mentioned above states that $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$ holds when $\mathbf{A}$ and $\mathbf{B}$ are commutative, i.e., $\mathbf{AB} = \mathbf{BA}$.

Finally, we specify the distribution of the elements of $\mathbf{V}$. We can consider both homoskedastic and heteroskedastic error terms as specified in the following assumptions.

**Assumption 1.** *$v_i$'s are independent and identically distributed (i.i.d.) across $i$ with mean zero and variance $\sigma_0^2$, and $\mathrm{E}\,|v_i|^{4+\varrho} < \infty$ for some $\varrho > 0$.*

**Assumption 2.** *$v_i$'s are independently distributed over $i$ with $\mathrm{E}\,(v_i) = 0$ and $\mathrm{Var}\,(v_i) = \sigma_i^2$, and $\mathrm{E}\,|v_i|^{4+\varrho} < \infty$ for some $\varrho > 0$.*

Both assumptions require that the first $4 + \varrho$ moments of the error terms exist and are finite, which is required by the central limit theorem (CLT) considered by Kelejian and Prucha (2001, 2010) for the linear and quadratic forms of $\mathbf{V}$ (see Lemma 4 in the Appendix).

## 3 Computation of matrix exponential terms

A prominent issue in the estimation of MESS-type models is the computation of matrix exponential terms. To this end, there are several methods suggested in the literature such as the Taylor series

approximation, Padé approximation, ordinary differential equation methods, polynomial methods, matrix decomposition methods, splitting methods and Krylov space methods. Popular software such as Python, R, MATLAB and Mathematica provide functions that can be used to compute the matrix exponential of a given matrix. For example, MATLAB (function `expm`), Mathematica (function `MatrixExp`) and Python (function `scipy.linalg.expm`) utilize a scaling and squaring method combined with a Padé approximation for the computation of matrix exponential terms.

Moler and Van Loan (1978, 2003) assess the effectiveness of nineteen methods according to the following attributes: (i) generality, (ii) reliability, (iii) stability, (iv) accuracy, (v) efficiency, (vi) storage requirements, (vii) ease of use, and (viii) simplicity. They conclude that though "none (of the methods in their paper) are completely satisfactory," a scaling and squaring method with either the rational Padé or Taylor approximants can be the most effective one to compute the matrix exponential terms.

As pointed out by Moler and Van Loan (1978, 2003), all methods suggested in the literature are dubious in the sense that a sole method may not be entirely reliable for all applications. For example, in the context of MESS-type models, the scaling and squaring method combined with the Padé approximation as implemented in MATLAB through `expm` function can be highly costly in terms of computation time (Yang et al., 2021). Our ensuing analysis on the estimation of the MESS(1,1) model indicates that we need to compute terms such as $e^{\lambda \mathbf{W}} e^{\rho \mathbf{M}} \mathbf{Y}$ and $e^{\rho \mathbf{M}} \mathbf{X}$. Because the matrix exponential terms show up as premultiplying a conformable vector, i.e., the matrix-vector product structure, instead of trying to approximate $e^{\lambda \mathbf{W}}$ and $e^{\rho \mathbf{M}}$, approximations to $e^{\lambda \mathbf{W}} e^{\rho \mathbf{M}} \mathbf{Y}$ and $e^{\rho \mathbf{M}} \mathbf{X}$ can be computed (LeSage and Pace, 2007). In fact, the matrix-vector product approximation can reduce the computation time significantly.

In the following, we show how to approximate the matrix-vector product terms $e^{\lambda \mathbf{W}} e^{\rho \mathbf{M}} \mathbf{Y}$ and $e^{\rho \mathbf{M}} \mathbf{X}$. Let $\text{Diag}(a_1, \ldots, a_n)$ be the $n \times n$ diagonal matrix with the $i$th diagonal element $a_i$. We first consider $e^{\lambda \mathbf{W}} e^{\rho \mathbf{M}} \mathbf{Y}$. We can truncate the matrix exponential terms at the $(q+1)$th order and express $e^{\lambda \mathbf{W}} e^{\rho \mathbf{M}} \mathbf{Y}$ as

$$
\begin{aligned}
e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}} \mathbf{Y} &\approx \sum_{i=0}^{q} \frac{\rho^i \mathbf{M}^i}{i!} \sum_{j=0}^{q} \frac{\lambda^j \mathbf{W}^j}{j!} \mathbf{Y} \\
&= \sum_{i=1}^{q} \sum_{j=0}^{i-1} \frac{\rho^i \lambda^j \mathbf{M}^i \mathbf{W}^j}{i! j!} \mathbf{Y} + \sum_{i=1}^{q} \sum_{j=0}^{i-1} \frac{\rho^j \lambda^i \mathbf{M}^j \mathbf{W}^i}{i! j!} \mathbf{Y} + \sum_{i=0}^{q} \frac{\rho^i \lambda^i \mathbf{M}^i \mathbf{W}^i}{(i!)^2} \mathbf{Y} \\
&= \mathbf{Y}_1 \mathbf{D}_1 \boldsymbol{\kappa}_1(\lambda, \rho) + \mathbf{Y}_2 \mathbf{D}_2 \boldsymbol{\kappa}_2(\lambda, \rho) + \mathbf{Y}_3 \mathbf{D}_3 \boldsymbol{\kappa}_3(\lambda, \rho), \qquad (3.1)
\end{aligned}
$$

where

$$\mathbf{Y}_1 = \left[\mathbf{MY}, \mathbf{M}^2\mathbf{Y}, \mathbf{M}^2\mathbf{WY}, \mathbf{M}^3\mathbf{Y}, \mathbf{M}^3\mathbf{WY}, \mathbf{M}^3\mathbf{W}^2\mathbf{Y}, \ldots, \mathbf{M}^q\mathbf{Y}, \mathbf{M}^q\mathbf{WY}, \ldots, \mathbf{M}^q\mathbf{W}^{q-1}\mathbf{Y}\right],$$
$$\mathbf{Y}_2 = \left[\mathbf{WY}, \mathbf{W}^2\mathbf{Y}, \mathbf{MW}^2\mathbf{Y}, \mathbf{W}^3\mathbf{Y}, \mathbf{MW}^3\mathbf{Y}, \mathbf{M}^2\mathbf{W}^3\mathbf{Y}, \ldots, \mathbf{W}^q\mathbf{Y}, \mathbf{MW}^q\mathbf{Y}, \ldots, \mathbf{M}^{q-1}\mathbf{W}^q\mathbf{Y}\right],$$
$$\mathbf{Y}_3 = \left[\mathbf{Y}, \mathbf{MWY}, \mathbf{M}^2\mathbf{W}^2\mathbf{Y}, \ldots, \mathbf{M}^q\mathbf{W}^q\mathbf{Y}\right],$$
$$\mathbf{D}_1 = \mathbf{D}_2 = \text{Diag}\left(\frac{1}{0!1!}, \frac{1}{0!2!}, \frac{1}{1!2!}, \ldots, \frac{1}{0!q!}, \ldots, \frac{1}{(q-1)!q!}\right),$$
$$\mathbf{D}_3 = \text{Diag}\left(\frac{1}{(0!)^2}, \frac{1}{(1!)^2}, \frac{1}{(2!)^2}, \ldots, \frac{1}{(q!)^2}\right),$$
$$\boldsymbol{\kappa}_1(\lambda, \rho) = \left[\rho, \rho^2, \rho^2\lambda, \rho^3, \rho^3\lambda, \rho^3\lambda^2, \ldots, \rho^q, \rho^q\lambda, \ldots, \rho^q\lambda^{q-1}\right]',$$
$$\boldsymbol{\kappa}_2(\lambda, \rho) = \left[\lambda, \lambda^2, \lambda^2\rho, \lambda^3, \lambda^3\rho, \lambda^3\rho^2, \ldots, \lambda^q, \lambda^q\rho, \ldots, \lambda^q\rho^{q-1}\right]',$$
$$\boldsymbol{\kappa}_3(\lambda, \rho) = \left[1, \rho\lambda, \rho^2\lambda^2, \rho^3\lambda^3, \ldots, \rho^q\lambda^q\right]'.$$

The result in (3.1) expresses $e^{\rho\mathbf{M}}e^{\lambda\mathbf{W}}\mathbf{Y}$ in terms of $\mathbf{Y}_j$ and $\mathbf{D}_j$ for $j \in \{1, 2, 3\}$. These terms can be computed once, and then supplied as inputs for the objective function in an optimization solver.

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k]$, where $\mathbf{X}_i$ is the $i$th column of $\mathbf{X}$. Then, we can express $e^{\rho\mathbf{M}}\mathbf{X}$ as

$$e^{\rho\mathbf{M}}\mathbf{X} = \left[e^{\rho\mathbf{M}}\mathbf{X}_1, e^{\rho\mathbf{M}}\mathbf{X}_2, \ldots, e^{\rho\mathbf{M}}\mathbf{X}_k\right] \approx \mathbb{X}\mathbf{D}_4\boldsymbol{\kappa}_4(\rho), \tag{3.2}$$

where

$$\mathbb{X} = [\mathbf{X}_1, \mathbf{MX}_1, \ldots, \mathbf{M}^q\mathbf{X}_1, \mathbf{X}_2, \mathbf{MX}_2, \ldots, \mathbf{M}^q\mathbf{X}_2, \ldots, \mathbf{X}_k, \mathbf{MX}_k, \ldots, \mathbf{M}^q\mathbf{X}_k],$$
$$\mathbf{D}_4 = \mathbf{I}_k \otimes \text{Diag}\left(\frac{1}{0!}, \frac{1}{1!}, \frac{1}{2!}, \ldots, \frac{1}{q!}\right),$$
$$\boldsymbol{\kappa}_4(\rho) = \mathbf{I}_k \otimes \left[1, \rho, \rho^2, \ldots, \rho^q\right]'.$$

The approximation in (3.2) indicates that the computation of $e^{\rho\mathbf{M}}\mathbf{X}$ also requires only the matrix-vector product operations. We can compute $\mathbb{X}$ and $\mathbf{D}_4$ only one time and then pass these terms as the inputs of the objective function in an optimization solver.

In an extensive Monte Carlo simulation study, Yang et al. (2021) compared the computation time required by the matrix-vector product method with the `expm` function of MATLAB. For the QMLE, they demonstrated that the matrix-vector product method reduced computation time by 98% to 99% compared to the `expm` function. In the case of GMME, the computation time decreased by 95% to 97%. In the context of the Bayesian estimator, the computation time was reduced by at least 99%.

# 4 Maximum likelihood estimation approach

The maximum likelihood (ML) estimation of the MESS model has been considered for both the homoskedastic case (LeSage and Pace, 2007, 2009; Debarsy et al., 2015) and the heteroskedastic case (Debarsy et al., 2015). In this section, we first introduce the estimation approach for the homoskedastic case and then for the heteroskedastic case.

## 4.1 Estimation under homoskedasticity

In this section, we will consider the quasi maximum likelihood estimation of the MESS(1,1) model with homoskedastic disturbances as maintained in Assumption 1. Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \sigma^2)'$, $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{\zeta}')'$ and $\boldsymbol{\zeta} = (\lambda, \rho)'$. Also let $\boldsymbol{\theta}_0 = (\boldsymbol{\gamma}_0', \sigma_0^2)'$ denote the true values of the parameters. Then, the quasi log-likelihood function for the MESS(1,1) is given by

$$\ln L(\boldsymbol{\theta}) = -\frac{n}{2}\ln(2\pi\sigma^2) + \ln\left|e^{\lambda\mathbf{W}}\right| + \ln\left|e^{\rho\mathbf{M}}\right| - \frac{1}{2\sigma^2}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' e^{\rho\mathbf{M}'} e^{\rho\mathbf{M}}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Since $\ln\left|e^{\lambda\mathbf{W}}\right| = \ln(e^{\lambda \mathrm{tr}(\mathbf{W})}) = \ln 1 = 0$ and $\ln\left|e^{\rho\mathbf{M}}\right| = \ln(e^{\rho \mathrm{tr}(\mathbf{M})}) = \ln 1 = 0$, the two Jacobian terms disappear in the quasi log-likelihood function. Thus, the quasi log-likelihood function simplifies to

$$\ln L(\boldsymbol{\theta}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' e^{\rho\mathbf{M}'} e^{\rho\mathbf{M}}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{4.1}$$

We can concentrate out $\sigma^2$ from the quasi log-likelihood function to obtain the concentrated quasi log-likelihood function only involving $\boldsymbol{\gamma}$. From the first order condition with respect to $\sigma^2$, the quasi maximum likelihood estimator of $\sigma^2$ is given by

$$\hat{\sigma}^2(\boldsymbol{\gamma}) = \frac{1}{n}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' e^{\rho\mathbf{M}'} e^{\rho\mathbf{M}}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{4.2}$$

Substituting (4.2) into (4.1), we obtain the concentrated quasi log-likelihood function as

$$\ln L(\boldsymbol{\gamma}) = -\frac{n}{2}\ln(2\pi + 1) - \frac{n}{2}\ln\hat{\sigma}^2(\boldsymbol{\gamma}). \tag{4.3}$$

Then, the QMLE $\hat{\boldsymbol{\gamma}}$ of $\gamma_0$ is defined as

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \ln L(\boldsymbol{\gamma}),$$

which is equivalent to

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} Q(\boldsymbol{\gamma}), \tag{4.4}$$

where $Q(\boldsymbol{\gamma}) = (e^{\lambda \mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' e^{\rho \mathbf{M}'} e^{\rho \mathbf{M}} (e^{\lambda \mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Substituting $\hat{\boldsymbol{\gamma}}$ into (4.2), we obtain the QMLE of $\sigma^2$ as $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\boldsymbol{\gamma}})$.

The large sample properties of the QMLE $\hat{\boldsymbol{\gamma}}$ can be established under some regularity conditions. For consistency, the necessary conditions are identifiable uniqueness of $\boldsymbol{\gamma}_0$ and the uniform stochastic convergence of the quasi maximum likelihood function to its population counterpart (White, 1994, Theorem 3.4). For asymptotic normality of $\hat{\boldsymbol{\gamma}}$, the CLT for linear and quadratic forms can be utilized (Kelejian and Prucha, 2001, 2010). The low level assumptions guaranteeing the large sample properties of $\hat{\boldsymbol{\gamma}}$ are (i) the existence of moments of error terms up to $4 + \varrho$ moment, (ii) a manageable degree of spatial correlation, (iii) a compact parameter space for $\boldsymbol{\zeta}$, (iv) the non-singularity of certain matrices in large samples, and (v) certain restrictions to guarantee identification of $\boldsymbol{\gamma}_0$ in large samples. A complete formal list of these low level assumptions is provided in Debarsy et al. (2015).

The score functions with respect to the elements of $\boldsymbol{\gamma}$ are given by

$$\frac{\partial Q(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \begin{cases} \boldsymbol{\beta}: & -2\mathbf{X}' e^{\rho_0 \mathbf{M}'} \mathbf{V}(\boldsymbol{\gamma}), \\ \lambda: & 2\mathbf{V}'(\boldsymbol{\gamma}) e^{\rho \mathbf{M}} \mathbf{W} e^{\lambda \mathbf{W}} \mathbf{Y}, \\ \rho: & 2\mathbf{V}'(\boldsymbol{\gamma}) \mathbf{M} \mathbf{V}(\boldsymbol{\gamma}), \end{cases} \tag{4.5}$$

where $\mathbf{V}(\boldsymbol{\gamma}) = e^{\rho \mathbf{M}}(e^{\lambda \mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Define $\mathbf{B} = \mathrm{Var}\left(\frac{1}{\sqrt{n}} \frac{\partial Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\right)$ and $\mathbf{A} = \mathrm{E}\left(-\frac{1}{n} \frac{\partial^2 Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}\right)$. To introduce the closed-forms of $\mathbf{A}$ and $\mathbf{B}$, let $\mu_3 = \mathrm{E}(v_i^3)$, $\mu_4 = \mathrm{E}(v_i^4)$, $\mathbb{W} = e^{\rho_0 \mathbf{M}} \mathbf{W} e^{-\rho_0 \mathbf{M}}$, $\mathbf{H}^s = \mathbf{H} + \mathbf{H}'$ for any square matrix $\mathbf{H}$ and $\mathrm{vec}_D(\mathbf{H})$ be a vector containing the diagonal elements of $\mathbf{H}$. Then, using (4.5) and Lemma 2 in Appendix A, we obtain

$$\mathbf{A} = -\frac{1}{n} \begin{pmatrix} 2\left(e^{\rho_0 \mathbf{M}} \mathbf{X}\right)' \left(e^{\rho_0 \mathbf{M}} \mathbf{X}\right) & * & * \\ -2\left(\mathbb{W} e^{\rho_0 \mathbf{M}} \mathbf{X} \boldsymbol{\beta}_0\right)' e^{\rho_0 \mathbf{M}} \mathbf{X} & \mathbf{A}_{\lambda\lambda} & * \\ \mathbf{0} & \sigma_0^2 \mathrm{tr}\left(\mathbb{W}^s \mathbf{M}^s\right) & \sigma_0^2 \mathrm{tr}\left(\mathbf{M}^s \mathbf{M}^s\right) \end{pmatrix},$$

$$\mathbf{B} = 2\sigma_0^2 \mathbf{A} + \frac{1}{n} \begin{pmatrix} \mathbf{0} & * & * \\ -2\mu_3 \left(\mathrm{vec}_D\left(\mathbb{W}^s\right)\right)' e^{\rho_0 \mathbf{M}} \mathbf{X} & \mathbf{0} & * \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{\rho\rho} \end{pmatrix},$$

where the elements are defined as $\mathbf{A}_{\lambda\lambda} = \sigma_0^2 \mathrm{tr}\left(\mathbb{W}^s \mathbb{W}^s\right) + 2\left(\mathbb{W} e^{\rho_0 \mathbf{M}} \mathbf{X} \boldsymbol{\beta}_0\right)' \left(\mathbb{W} e^{\rho_0 \mathbf{M}} \mathbf{X} \boldsymbol{\beta}_0\right)$ and $\mathbf{B}_{\rho\rho} = \left(\mu_4 - 3\sigma_0^4\right) \mathrm{vec}_D'\left(\mathbb{W}^s\right) \mathrm{vec}_D\left(\mathbb{W}^s\right) + 4\mu_3 \left(\mathbb{W} e^{\rho_0 \mathbf{M}} \mathbf{X} \boldsymbol{\beta}_0\right)' \mathrm{vec}_D\left(\mathbb{W}^s\right)$. The asymptotic distribution of $\hat{\boldsymbol{\gamma}}$ can be derived by applying the mean value theorem to $\frac{\partial Q(\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}}$ around $\boldsymbol{\gamma}_0$. By the mean value theorem, we can write $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = -\left(\frac{1}{n} \frac{\partial^2 Q(\bar{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}\right)^{-1} \frac{1}{\sqrt{n}} \frac{\partial Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}$, where $\bar{\boldsymbol{\gamma}}$ lies between $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}_0$ elementwise. Then, the desired result follows by showing that $\frac{1}{n} \frac{\partial^2 Q(\bar{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} - \frac{1}{n} \mathrm{E}\left(\frac{\partial^2 Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}\right) = o_p(1)$ and the asymptotic

normality of $\frac{1}{\sqrt{n}}\frac{\partial Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}$ by Lemma 4 in Appendix A. Thus, it follows that

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{d} N\Big(\mathbf{0}, \lim_{n\to\infty} \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\Big). \tag{4.6}$$

Note that there are two cases that yield $\mathbf{B} = 2\sigma_0^2\mathbf{A}$. The first case arises when $\mathbf{W}$ and $\mathbf{M}$ commute. Under the commutative property, we have $\mathbb{W} = \mathbf{W}$ and $\text{vec}_D(\mathbb{W}^s) = \mathbf{0}$, suggesting that $\mathbf{B} = 2\sigma_0^2\mathbf{A}$. The second case occurs when the disturbance terms are normally distributed. Under the normality, we have $\mu_4 = 3\sigma_0^4$ and $\mu_3 = 0$, yielding again $\mathbf{B} = 2\sigma_0^2\mathbf{A}$. In either case, the result in (4.6) reduces to $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{d} N(\mathbf{0}, \lim_{n\to\infty} 2\sigma_0^2\mathbf{A}^{-1})$.

Finally, for inference, the plug-in estimators of $\mathbf{A}$ and $\mathbf{B}$ can be utilized. To that end, $\sigma_0^2$ can be consistently estimated by evaluating (4.2) at $\hat{\boldsymbol{\gamma}}$, and $\mu_3$ and $\mu_4$ can be consistently estimated by their sample analogs using the residuals $\mathbf{V}(\hat{\boldsymbol{\gamma}})$. Thus, the standard error of $\hat{\boldsymbol{\gamma}}$ can be obtained as the square root of the diagonal elements of $\frac{1}{n}\mathbf{A}^{-1}(\hat{\boldsymbol{\gamma}})\mathbf{B}(\hat{\boldsymbol{\gamma}})\mathbf{A}^{-1}(\hat{\boldsymbol{\gamma}})$, where $\mathbf{A}(\hat{\boldsymbol{\gamma}})$ and $\mathbf{B}(\hat{\boldsymbol{\gamma}})$ are the plug-in estimators of $\mathbf{A}$ and $\mathbf{B}$, respectively.

### 4.2 Estimation under heteroskedasticity

In this subsection, we consider the quasi-maximum likelihood estimation of the MESS(1,1) under the assumption of heteroskedastic error terms. Let $\boldsymbol{\Sigma}$ be the variance covariance matrix of the error terms, i.e., $\boldsymbol{\Sigma} = \text{Diag}(\sigma_1^2, \ldots, \sigma_n^2)$, the diagonal matrix formed by $\sigma_i^2$'s. The score functions of the quasi likelihood function evaluated at $\boldsymbol{\gamma}_0$ are given by

$$\frac{\partial Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} = \begin{cases} \boldsymbol{\beta}: & -2\mathbf{X}' e^{\rho_0 \mathbf{M}'}\mathbf{V}, \\ \lambda: & 2\mathbf{V}' e^{\rho_0 \mathbf{M}}\mathbf{W}\left(\mathbf{X}\boldsymbol{\beta}_0 + e^{-\rho_0 \mathbf{M}}\mathbf{V}\right), \\ \rho: & 2\mathbf{V}'\mathbf{M}\mathbf{V}. \end{cases}$$

The expectation of the score functions with respect to $\boldsymbol{\beta}$ and $\rho$ at $\boldsymbol{\gamma}_0$ are zero by Lemma 2 in Appendix A. However, the expectation of the score function with respect to $\lambda$ at $\boldsymbol{\gamma}_0$ is $\text{tr}(\mathbb{W}\boldsymbol{\Sigma})$. By Lemma 1 in Appendix A, the order of this term is $O(n)$ under the assumption that $\mathbf{W}$ and $\mathbf{M}$ are bounded in matrix column sum and row sum norms. Hence, the QMLE $\hat{\boldsymbol{\gamma}}$ may not be consistent. However, when $\mathbf{W}$ and $\mathbf{M}$ commute, we have $\mathbb{W} = \mathbf{W}$, yielding $\text{tr}(\mathbf{W}\boldsymbol{\Sigma}) = 0$. Hence, when $\mathbf{W}$ and $\mathbf{M}$ commute, the QMLE of MESS(1,1) may remain consistent under the assumption of heteroskedastic disturbance terms.

The consistency and asymptotic normality of the QMLE $\hat{\boldsymbol{\gamma}}$ can be proved similarly to the homoskedastic case. Let $\mathbf{D} = \text{E}\left(-\frac{1}{n}\frac{\partial^2 Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'}\right)$ and $\mathbf{F} = \text{Var}\left(\frac{1}{\sqrt{n}}\frac{\partial Q(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\right)$. Using Lemma 2 in Appendix

A, we obtain

$$\mathbf{D} = -\frac{2}{n}\begin{pmatrix} \left(e^{\rho_0\mathbf{M}}\mathbf{X}\right)'\left(e^{\rho_0\mathbf{M}}\mathbf{X}\right) & * & * \\ -\left(\mathbf{W}e^{\rho_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0\right)' e^{\rho_0\mathbf{M}}\mathbf{X} & \mathbf{D}_{\lambda\lambda} & * \\ \mathbf{0} & \operatorname{tr}\left(\mathbf{M}^s\mathbf{W}\boldsymbol{\Sigma}\right) & \operatorname{tr}\left(\mathbf{M}^s\mathbf{M}\boldsymbol{\Sigma}\right) \end{pmatrix},$$

$$\mathbf{F} = \frac{2}{n}\begin{pmatrix} 2\left(e^{\rho_0\mathbf{M}}\mathbf{X}\right)'\boldsymbol{\Sigma}\left(e^{\rho_0\mathbf{M}}\mathbf{X}\right) & * & * \\ -2\left(\boldsymbol{\Sigma}\mathbf{W}e^{\rho_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0\right)' e^{\rho_0\mathbf{M}}\mathbf{X} & \mathbf{F}_{\lambda\lambda} & * \\ \mathbf{0} & \operatorname{tr}\left(\boldsymbol{\Sigma}\mathbf{M}^s\boldsymbol{\Sigma}\mathbf{W}^s\right) & \operatorname{tr}\left(\boldsymbol{\Sigma}\mathbf{M}^s\boldsymbol{\Sigma}\mathbf{M}^s\right) \end{pmatrix},$$

where

$$\mathbf{D}_{\lambda\lambda} = \operatorname{tr}\left(\mathbf{W}^s\mathbf{W}\boldsymbol{\Sigma}\right) + \left(\mathbf{W}e^{\rho_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0\right)'\left(\mathbf{W}e^{\rho_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0\right),$$
$$\mathbf{F}_{\lambda\lambda} = \operatorname{tr}\left(\boldsymbol{\Sigma}\mathbf{W}^s\boldsymbol{\Sigma}\mathbf{W}^s\right) + 2\left(\mathbf{W}e^{\rho_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0\right)'\boldsymbol{\Sigma}\left(\mathbf{W}e^{\rho_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0\right).$$

Then, it can be shown that

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{d} N\Big(\mathbf{0}, \lim_{n\to\infty} \mathbf{D}^{-1}\mathbf{F}\mathbf{D}^{-1}\Big). \tag{4.7}$$

For inference, the standard error of $\hat{\boldsymbol{\gamma}}$ can be obtained as the square root of the diagonal elements of $\frac{1}{n}\mathbf{D}^{-1}(\hat{\boldsymbol{\gamma}})\mathbf{F}(\hat{\boldsymbol{\gamma}})\mathbf{D}^{-1}(\hat{\boldsymbol{\gamma}})$, where $\mathbf{D}(\hat{\boldsymbol{\gamma}})$ and $\mathbf{F}(\hat{\boldsymbol{\gamma}})$ are the plug-in estimators of $\mathbf{D}$ and $\mathbf{F}$, respectively. Also, note that $\mathbf{D}$ and $\mathbf{F}$ involve the unknown diagonal matrix $\boldsymbol{\Sigma}$. As in White (1980), the terms involving $\boldsymbol{\Sigma}$ can be consistently estimated by replacing $\boldsymbol{\Sigma}$ with $\hat{\boldsymbol{\Sigma}} = \operatorname{Diag}\left(v_1^2(\hat{\boldsymbol{\gamma}}), \ldots, v_n^2(\hat{\boldsymbol{\gamma}})\right)$, where $v_i(\hat{\boldsymbol{\gamma}})$ is the $i$th element of $\mathbf{V}(\hat{\boldsymbol{\gamma}})$.

## 5 M-estimation approach

In this section, we study the consistent estimation of the MESS model under heteroskedasticity. Since the content of this section is new and has not been explored in previous literature, we present formal results along with the necessary assumptions. Under heteroskedasticity, when $\mathbf{W}$ and $\mathbf{M}$ do not commute, we can use the M-estimation method to formulate a consistent estimator of $\boldsymbol{\gamma}$ based on the adjusted score functions. We denote $\mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}) = e^{\rho\mathbf{M}}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ and $\mathbf{V} = \mathbf{V}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)$. Then, the

score functions based on (4.1) can be determined as[2]

$$S(\boldsymbol{\beta},\boldsymbol{\zeta}) = \begin{cases} \boldsymbol{\beta}: & \mathbf{X}'e^{\rho\mathbf{M}'}\mathbf{V}(\boldsymbol{\beta},\zeta), \\ \lambda: & -\mathbf{Y}'e^{\lambda\mathbf{W}'}\mathbf{W}'e^{\rho\mathbf{M}'}\mathbf{V}(\boldsymbol{\beta},\zeta), \\ \rho: & -\mathbf{V}'(\boldsymbol{\beta},\zeta)\mathbf{M}\mathbf{V}(\boldsymbol{\beta},\zeta). \end{cases} \tag{5.1}$$

The essential reason why the QMLE is not consistent is $\text{plim}_{n\to\infty}\frac{1}{n}S(\boldsymbol{\gamma}_0) \neq 0$. In the case of the score functions with respect to $\boldsymbol{\beta}$ and $\rho$, we have $\text{E}(\mathbf{X}'e^{\rho_0\mathbf{M}'}\mathbf{V}) = 0$, and $\text{E}(\mathbf{V}'\mathbf{M}\mathbf{V}) = \text{tr}(\boldsymbol{\Sigma}\mathbf{M}) = 0$ because $\boldsymbol{\Sigma}$ is a diagonal matrix and $\mathbf{M}$ has zero diagonal elements. In the case of the score function with respect to $\lambda$, we have

$$\begin{aligned} \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}\mathbf{W}'e^{\rho_0\mathbf{M}'}\mathbf{V}) &= \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}e^{-\rho_0\mathbf{M}'}\mathbf{W}'e^{\rho_0\mathbf{M}'}\mathbf{V}) = \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}\mathbb{W}'\mathbf{V}) \\ &= \text{E}(\mathbf{V}'\mathbb{W}\mathbf{V}) = \text{tr}(\boldsymbol{\Sigma}\mathbb{W}). \end{aligned} \tag{5.2}$$

If $\mathbf{W}$ and $\mathbf{M}$ commute, i.e., $\mathbf{W}\mathbf{M} = \mathbf{M}\mathbf{W}$, then we have $\mathbb{W} = \mathbf{W}$, which yields $\text{tr}(\boldsymbol{\Sigma}\mathbb{W}) = \text{tr}(\boldsymbol{\Sigma}\mathbf{W}) = 0$. Thus, when the commutative property holds, we have $\text{plim}_{n\to\infty}\frac{1}{n}S(\boldsymbol{\gamma}_0) = 0$, suggesting that the QMLE can be consistent under heteroskedasticity. However, if $\mathbf{W}\mathbf{M} \neq \mathbf{M}\mathbf{W}$, then we have $\text{tr}(\boldsymbol{\Sigma}\mathbb{W}) = O(n)$ and $\text{plim}_{n\to\infty}\frac{1}{n}S(\boldsymbol{\gamma}_0) \neq 0$ in general, indicating that the QMLE may not be consistent under heteroskedasticity. We will adjust the score function with respect to $\lambda$ so that $\text{plim}_{n\to\infty}\frac{1}{n}S(\boldsymbol{\gamma}_0) = 0$ holds in all cases.

To adjust the score function with respect to $\lambda$, we use the trace property $\text{tr}(\mathbf{D}\mathbf{A}) = \text{tr}(\mathbf{D}\,\text{Diag}(\mathbf{A}))$, where $\mathbf{D}$ is an $n \times n$ diagonal matrix and $\mathbf{A}$ is a conformable matrix. Using this property, we can express $\text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}\mathbf{W}'e^{\rho_0\mathbf{M}'}\mathbf{V})$ as

$$\begin{aligned} \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}\mathbf{W}'e^{\rho_0\mathbf{M}'}\mathbf{V}) &= \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}\mathbb{W}'\mathbf{V}) = \text{tr}(\boldsymbol{\Sigma}\mathbb{W}) = \text{tr}(\boldsymbol{\Sigma}\,\text{Diag}(\mathbb{W})) \\ &= \text{E}(\mathbf{V}'\,\text{Diag}(\mathbb{W})\mathbf{V}) = \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}\,\text{Diag}(\mathbb{W})\mathbf{V}). \end{aligned} \tag{5.3}$$

Then, subtracting the last term from the second term in (5.3), we obtain

$$\begin{aligned} &\text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}\mathbb{W}'\mathbf{V}) - \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}\,\text{Diag}(\mathbb{W})\mathbf{V}) = 0 \\ &\implies \text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}\mathbb{W}_D\mathbf{V}) = 0, \end{aligned} \tag{5.4}$$

where $\mathbb{W}_D = \mathbb{W} - \text{Diag}(\mathbb{W})$. Thus, we suggest using the sample counter part of $\text{E}(\mathbf{Y}'e^{\lambda_0\mathbf{W}'}e^{\rho_0\mathbf{M}'}\mathbb{W}_D\mathbf{V})$ as the adjusted score function with respect to $\lambda$. Then, the adjusted score functions take the following

[2]Under heteroskedasticity, there is no score function with respect to $\sigma^2$. Our aim is to construct adjusted score functions such that $\text{E}\left(S(\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\right) = 0$.

form:

$$S^*(\boldsymbol{\gamma}) = \begin{cases} \boldsymbol{\beta}: & \mathbf{X}' e^{\rho \mathbf{M}'} \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}), \\ \lambda: & -\mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbb{W}_D(\rho) \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}), \\ \rho: & -\mathbf{V}'(\boldsymbol{\beta}, \zeta) \mathbf{M} \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}), \end{cases} \tag{5.5}$$

where $\mathbb{W}_D(\rho) = \mathbb{W}(\rho) - \text{Diag}(\mathbb{W}(\rho))$ and $\mathbb{W}(\rho) = e^{\rho \mathbf{M}} \mathbf{W} e^{-\rho \mathbf{M}}$. Note that $\mathrm{E}\left(S^*(\boldsymbol{\gamma}_0)\right) = 0$ holds by construction. We first derive the estimator of $\boldsymbol{\beta}_0$ for a given $\boldsymbol{\zeta}$ value, which is given by

$$\hat{\boldsymbol{\beta}}_M(\boldsymbol{\zeta}) = (\mathbf{X}' e^{\rho \mathbf{M}'} e^{\rho \mathbf{M}} \mathbf{X})^{-1} \mathbf{X}' e^{\rho \mathbf{M}'} e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}} \mathbf{Y}. \tag{5.6}$$

Then, substituting $\hat{\boldsymbol{\beta}}_M(\boldsymbol{\zeta})$ into the $\lambda$ and $\rho$ elements of (5.5), we obtain the concentrated adjusted score functions as

$$S^{c*}(\boldsymbol{\zeta}) = \begin{cases} \lambda: & -\mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbb{W}_D(\rho) \hat{\mathbf{V}}(\boldsymbol{\zeta}), \\ \rho: & -\hat{\mathbf{V}}'(\boldsymbol{\zeta}) \mathbf{M} \hat{\mathbf{V}}(\boldsymbol{\zeta}), \end{cases} \tag{5.7}$$

where $\hat{\mathbf{V}}(\boldsymbol{\zeta}) = \mathbf{V}(\hat{\boldsymbol{\beta}}_M(\boldsymbol{\zeta}), \boldsymbol{\zeta})$. Then, the M-estimator (ME) of $\boldsymbol{\zeta}_0$ is defined by

$$\hat{\boldsymbol{\zeta}}_M = \text{argsolve}\{S^{c*}(\boldsymbol{\zeta}) = 0\}. \tag{5.8}$$

Substituting $\hat{\boldsymbol{\zeta}}_M$ into (5.6), we get the M-estimator for $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}_M = \hat{\boldsymbol{\beta}}_M(\hat{\boldsymbol{\zeta}}_M)$. To prove the consistency of $\hat{\boldsymbol{\gamma}}_M = (\hat{\boldsymbol{\beta}}_M', \hat{\boldsymbol{\zeta}}_M')'$, we only need to prove the consistency of $\hat{\boldsymbol{\zeta}}_M$ since $\hat{\boldsymbol{\beta}}_M = \hat{\boldsymbol{\beta}}_M(\hat{\boldsymbol{\zeta}}_M)$. To that end, we let $\bar{S}^*(\boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathrm{E}(S^*(\boldsymbol{\beta}, \boldsymbol{\zeta}))$ be the population counterpart of the adjusted score functions in (5.5). Given $\boldsymbol{\zeta}$, we can write $\bar{\boldsymbol{\beta}}_M(\boldsymbol{\zeta}) = (\mathbf{X}' e^{\rho \mathbf{M}'} e^{\rho \mathbf{M}} \mathbf{X})^{-1} \mathbf{X}' e^{\rho \mathbf{M}'} e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}} \mathrm{E}(\mathbf{Y})$, which can be substituted into the $\lambda$ and $\rho$ elements of $\bar{S}^*(\boldsymbol{\beta}, \boldsymbol{\zeta})$ to obtain

$$\bar{S}^{c*}(\boldsymbol{\zeta}) = \begin{cases} \lambda: & -\mathrm{E}\left(\mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbb{W}_D(\rho) \bar{\mathbf{V}}(\boldsymbol{\zeta})\right), \\ \rho: & -\mathrm{E}\left(\bar{\mathbf{V}}'(\boldsymbol{\zeta}) \mathbf{M} \bar{\mathbf{V}}(\boldsymbol{\zeta})\right), \end{cases} \tag{5.9}$$

where $\bar{\mathbf{V}}(\boldsymbol{\zeta}) = \mathbf{V}(\bar{\boldsymbol{\beta}}_M(\boldsymbol{\zeta}), \boldsymbol{\zeta})$. The uniform convergence $\sup_{\boldsymbol{\zeta} \in \boldsymbol{\Delta}} \frac{1}{n} \left\| S^{*c}(\boldsymbol{\zeta}) - \bar{S}^{*c}(\boldsymbol{\zeta}) \right\| \xrightarrow{p} 0$ and Assumption 6 in Appendix B ensure the consistency of $\hat{\boldsymbol{\zeta}}_M$.

**Theorem 5.1.** *Under Assumptions 2–6 stated in Appendix B, we have* $\hat{\boldsymbol{\gamma}}_M \xrightarrow{p} \boldsymbol{\gamma}_0$.

*Proof.* See Section C.1 in the Appendix. □

To derive the asymptotic distribution of $\hat{\boldsymbol{\gamma}}_M$, we apply the mean value theorem to $S^*(\hat{\boldsymbol{\gamma}}_M) = 0$ at $\boldsymbol{\gamma}_0$, to obtain $\sqrt{n}(\hat{\boldsymbol{\gamma}}_M - \boldsymbol{\gamma}_0) = -\left(\frac{1}{n} \frac{\partial S^*(\overline{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}'}\right)^{-1} \frac{1}{\sqrt{n}} S^*(\boldsymbol{\gamma}_0)$, where $\overline{\boldsymbol{\gamma}}$ lies between $\boldsymbol{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}_M$ elementwise. By substituting the reduced form $\mathbf{Y} = e^{-\lambda_0 \mathbf{W}} \left(\mathbf{X} \boldsymbol{\beta}_0 + e^{-\rho_0 \mathbf{M}} \mathbf{V}\right)$ into $S^*(\boldsymbol{\gamma}_0)$, we obtain a linear-

quadratic form in $\mathbf{V}$:

$$S^*(\boldsymbol{\gamma}_0) = \begin{cases} \boldsymbol{\beta}: & \mathbf{X}' e^{\rho_0 \mathbf{M}'} \mathbf{V}, \\ \lambda: & -\boldsymbol{\beta}_0' \mathbf{X}' e^{\rho_0 \mathbf{M}'} \mathbb{W}_D \mathbf{V} - \mathbf{V}' \mathbb{W}_D \mathbf{V}, \\ \rho: & -\mathbf{V}' \mathbf{M} \mathbf{V}, \end{cases} \tag{5.10}$$

where $\mathbb{W}_D = \mathbb{W}_D(\rho_0)$. Thus, the CLT for the linear-quadratic forms of $\mathbf{V}$ in Lemma 4 of the Appendix can be used to establish the asymptotic normality of $\frac{1}{\sqrt{n}} S^*(\boldsymbol{\gamma}_0)$. Also, our assumptions ensure that $\frac{1}{n} \frac{\partial S^*(\overline{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}'} - \frac{1}{n} \mathrm{E}\left(\frac{\partial S^*(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}\right) = o_p(1)$. Using these results, we determine the asymptotic distribution of $\hat{\boldsymbol{\gamma}}_M$ in Theorem 5.2.

**Theorem 5.2.** *Under Assumptions 2–6 stated in Appendix B, we have*

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}_M - \boldsymbol{\gamma}_0) \xrightarrow{d} N\left(0, \lim_{n\to\infty} \boldsymbol{\Psi}^{-1}(\boldsymbol{\gamma}_0) \boldsymbol{\Omega}(\boldsymbol{\gamma}_0) \boldsymbol{\Psi}^{-1'}(\boldsymbol{\gamma}_0)\right), \tag{5.11}$$

*where* $\boldsymbol{\Psi}(\boldsymbol{\gamma}_0) = -\frac{1}{n} \mathrm{E}\left(\frac{\partial S^*(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}\right)$ *and* $\boldsymbol{\Omega}(\boldsymbol{\gamma}_0) = \mathrm{Var}\left(\frac{1}{\sqrt{n}} S^*(\boldsymbol{\gamma}_0)\right)$ *are assumed to exist and* $\boldsymbol{\Psi}(\boldsymbol{\gamma}_0)$ *is assumed to be positive definite for sufficiently large* $n$.

*Proof.* See Section C.2 in the Appendix. □

To conduct inference, we need consistent estimators of $\boldsymbol{\Psi}(\boldsymbol{\gamma}_0)$ and $\boldsymbol{\Omega}(\boldsymbol{\gamma}_0)$. For $\boldsymbol{\Psi}(\boldsymbol{\gamma}_0)$, we can use its observed counterpart given by $\boldsymbol{\Psi}(\hat{\boldsymbol{\gamma}}_M) = -\frac{1}{n} \frac{\partial S^*(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_M}$. The elements of $\boldsymbol{\Psi}(\boldsymbol{\gamma})$ are given by

$$\begin{aligned}
\boldsymbol{\Psi}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\gamma}) &= \frac{1}{n} \mathbf{X}' e^{\rho \mathbf{M}'} e^{\rho \mathbf{M}} \mathbf{X}, \quad \boldsymbol{\Psi}_{\boldsymbol{\beta}\lambda}(\boldsymbol{\gamma}) = -\frac{1}{n} \mathbf{X}' e^{\rho \mathbf{M}'} \mathbf{Y}(\boldsymbol{\zeta}), \\
\boldsymbol{\Psi}_{\boldsymbol{\beta}\rho}(\boldsymbol{\gamma}) &= -\frac{1}{n} \mathbf{X}' e^{\rho \mathbf{M}'} \mathbf{M}^s \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}), \; \boldsymbol{\Psi}_{\lambda\boldsymbol{\beta}}(\boldsymbol{\gamma}) = -\frac{1}{n} \mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbb{W}_D(\rho) e^{\rho \mathbf{M}} \mathbf{X}, \\
\boldsymbol{\Psi}_{\lambda\lambda}(\boldsymbol{\gamma}) &= \frac{1}{n} \mathbf{Y}'(\boldsymbol{\zeta}) \mathbb{W}_D(\rho) \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}) + \frac{1}{n} \mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbb{W}_D(\rho) \mathbf{Y}(\boldsymbol{\zeta}), \\
\boldsymbol{\Psi}_{\lambda\rho}(\boldsymbol{\gamma}) &= \frac{1}{n} \mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbf{M}' \mathbb{W}_D(\rho) \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}) + \frac{1}{n} \mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \dot{\mathbb{W}}_D(\rho) \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}) \\
&\quad + \frac{1}{n} \mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbb{W}_D(\rho) \mathbf{M} \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}), \\
\boldsymbol{\Psi}_{\rho\boldsymbol{\beta}}(\boldsymbol{\gamma}) &= \boldsymbol{\Psi}_{\boldsymbol{\beta}\rho}(\boldsymbol{\gamma}), \quad \boldsymbol{\Psi}_{\rho\lambda}(\boldsymbol{\gamma}) = \frac{1}{n} \mathbf{Y}'(\boldsymbol{\zeta}) \mathbf{M}^s \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}), \\
\boldsymbol{\Psi}_{\rho\rho}(\boldsymbol{\gamma}) &= \frac{1}{n} \mathbf{V}'(\boldsymbol{\beta}, \boldsymbol{\zeta}) \mathbf{M}^s \mathbf{M} \mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\zeta}),
\end{aligned}$$

where $\dot{\mathbb{W}}_D(\rho) = \frac{\partial \mathbb{W}_D(\rho)}{\partial \rho} = \mathbf{M}\mathbb{W}_D(\rho) - \mathbb{W}_D(\rho)\mathbf{M} - \mathrm{Diag}\left(\mathbf{M}\mathbb{W}_D(\rho) - \mathbb{W}_D(\rho)\mathbf{M}\right)$ and $\mathbf{Y}(\boldsymbol{\zeta}) = e^{\rho \mathbf{M}} \mathbf{W} e^{\lambda \mathbf{W}} \mathbf{Y}$. In the proof of Theorem 5.2, we show that $\boldsymbol{\Psi}(\hat{\boldsymbol{\gamma}}_M)$ is a consistent estimator of $\boldsymbol{\Psi}(\boldsymbol{\gamma}_0)$. Using Lemma

2 in the Appendix, we determined the closed form of $\mathbf{\Omega}(\boldsymbol{\gamma}_0)$ as

$$\mathbf{\Omega}(\boldsymbol{\gamma}_0) = \begin{pmatrix} \frac{1}{n}\mathbf{X}' e^{\rho_0 \mathbf{M}'} \mathbf{\Sigma} e^{\rho_0 \mathbf{M}} \mathbf{X} & -\frac{1}{n}\mathbf{X}' e^{\rho_0 \mathbf{M}'} \mathbf{\Sigma} \mathbb{W}_D' e^{\rho_0 \mathbf{M}} \mathbf{X}\boldsymbol{\beta}_0 & \mathbf{0}_{k\times 1} \\ * & \mathbf{\Omega}_{22} & \frac{1}{n}\mathrm{tr}(\mathbf{\Sigma}\mathbb{W}_D\mathbf{\Sigma}\mathbf{M}^s) \\ * & * & \frac{1}{n}\mathrm{tr}(\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}\mathbf{M}^s) \end{pmatrix},$$

where $\mathbf{\Omega}_{22} = \frac{1}{n}\boldsymbol{\beta}_0' \mathbf{X}' e^{\rho_0 \mathbf{M}'} \mathbb{W}_D \mathbf{\Sigma} \mathbb{W}_D' e^{\rho_0 \mathbf{M}} \mathbf{X}\boldsymbol{\beta}_0 + \frac{1}{n}\mathrm{tr}(\mathbf{\Sigma}\mathbb{W}_D\mathbf{\Sigma}\mathbb{W}_D^s)$. Let $\mathbf{\Omega}(\hat{\boldsymbol{\gamma}}_M)$ be the plug-in estimator of $\mathbf{\Omega}(\boldsymbol{\gamma}_0)$, where we replace $\mathbf{\Sigma}$ with $\hat{\mathbf{\Sigma}} = \mathrm{Diag}\left(v_1^2(\hat{\boldsymbol{\gamma}}), \ldots, v_n^2(\hat{\boldsymbol{\gamma}})\right)$ and $v_i(\hat{\boldsymbol{\gamma}}_M)$ is the $i$th element of $\mathbf{V}(\hat{\boldsymbol{\gamma}}_M)$.

**Theorem 5.3.** *Under Assumptions 2–6 stated in Appendix B, we have* $\mathbf{\Omega}(\hat{\boldsymbol{\gamma}}_M) = \mathbf{\Omega}(\boldsymbol{\gamma}_0) + o_p(1)$.

*Proof.* See Section C.3 in the Appendix. □

Thus, the standard error of $\hat{\boldsymbol{\gamma}}_M$ can be obtained as the square root of the diagonal elements of $\frac{1}{n}\mathbf{\Psi}^{-1}(\hat{\boldsymbol{\gamma}}_M)\mathbf{\Omega}(\hat{\boldsymbol{\gamma}}_M)\mathbf{\Psi}^{-1'}(\hat{\boldsymbol{\gamma}}_M)$.

# 6 GMM estimation approach

In this section, we consider the GMM estimation of the MESS model, which can be more efficient than the QML estimation for either the homoskedastic case or the heteroskedastic case (Debarsy et al., 2015).

## 6.1 Estimation under homoskedasticity

In this subsection, we consider the GMM estimation of the MESS(1,1) model under Assumption 1. Recall again from the definition of MESS(1,1) that $\mathbf{V}(\boldsymbol{\gamma}) = e^{\rho \mathbf{M}}(e^{\lambda \mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, where $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{\zeta}')'$ and $\boldsymbol{\zeta} = (\lambda, \rho)'$. We consider the following vector of moment functions consisting of $k_p$ quadratic moments and $k_f$ linear moments:

$$g(\boldsymbol{\gamma}) = \frac{1}{n}\left(\mathbf{V}'(\boldsymbol{\gamma})\mathbf{P}_1\mathbf{V}(\boldsymbol{\gamma}), \ldots, \mathbf{V}'(\boldsymbol{\gamma})\mathbf{P}_{k_p}\mathbf{V}(\boldsymbol{\gamma}), \mathbf{V}'(\boldsymbol{\gamma})\mathbf{F}\right)',$$

where $\mathbf{P}_m$'s are $n \times n$ matrices of constants with $\mathrm{tr}(\mathbf{P}_m) = 0$ for $m = 1, \ldots, k_p$, and $\mathbf{F}$ is the $n \times k_f$ matrix of instrumental variables (IV). Given an arbitrary symmetric weighting matrix $\mathbf{\Phi}$, the GMM objective function is given by $g'(\boldsymbol{\gamma})\mathbf{\Phi}g(\boldsymbol{\gamma})$. Then, an initial GMME (IGMME) can be obtained by

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} g'(\boldsymbol{\gamma})\mathbf{\Phi}g(\boldsymbol{\gamma}). \tag{6.1}$$

Define $\mathbf{G} = \mathrm{E}\left(\frac{\partial g(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}\right)$ and $\mathbf{H} = n\mathrm{E}\left(g\left(\boldsymbol{\gamma}_0\right) g'\left(\boldsymbol{\gamma}_0\right)\right)$. Let $\mathrm{vec}(\mathbf{A})$ denote the column vector formed by stacking the columns of matrix $\mathbf{A}$ and recall that $\mathrm{vec}_D(\mathbf{A})$ denotes the column vector formed by the diagonal elements of the matrix $\mathbf{A}$. Then, by Lemma 2 in Appendix A, we obtain

$$\mathbf{H} = \frac{1}{n}\begin{pmatrix} \frac{\sigma_0^4}{2}\boldsymbol{\omega}'\boldsymbol{\omega} + \frac{1}{4}\left(\mu_4 - 3\sigma_0^4\right)\boldsymbol{\omega}_d'\boldsymbol{\omega}_d & \frac{1}{2}\mu_3\boldsymbol{\omega}_d'\mathbf{F} \\ \frac{1}{2}\mu_3\mathbf{F}'\omega_d & \sigma_0^2\mathbf{F}'\mathbf{F} \end{pmatrix},$$

and

$$\mathbf{G} = \frac{1}{n}\begin{pmatrix} \mathbf{0} & \frac{\sigma_0^2}{2}\boldsymbol{\omega}'\,\mathrm{vec}\left(\mathbb{W}^s\right) & \frac{\sigma_0^2}{2}\boldsymbol{\omega}'\,\mathrm{vec}\left(\mathbf{M}^s\right) \\ -\mathbf{F}'e^{\tau_0\mathbf{M}}\mathbf{X} & \mathbf{F}'\mathbb{W}e^{\tau_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0 & \mathbf{0} \end{pmatrix},$$

where $\boldsymbol{\omega} = (\mathrm{vec}(\mathbf{P}_1^s), \ldots, \mathrm{vec}(\mathbf{P}_{k_p}^s))$ and $\boldsymbol{\omega}_d = (\mathrm{vec}_D(\mathbf{P}_1^s), \ldots, \mathrm{vec}_D(\mathbf{P}_{k_p}^s))$.

The large sample properties of the IGMME $\hat{\boldsymbol{\gamma}}$ can be established under some regularity conditions. For consistency, the necessary conditions are identification of $\boldsymbol{\gamma}_0$ from the population moments and the uniform stochastic convergence of the generalized method of moments objective function to its population counterpart. For the asymptotic normality of $\hat{\boldsymbol{\gamma}}$, the central limit theorem for linear and quadratic forms can be utilized.[3] The asymptotic distribution of $\hat{\boldsymbol{\gamma}}$ can be derived by applying the mean value theorem to $\frac{\partial g'(\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}}\boldsymbol{\Phi}g(\hat{\boldsymbol{\gamma}}) = 0$ at $\boldsymbol{\gamma}_0$ to get $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = -\left(\frac{\partial g'(\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}}\boldsymbol{\Phi}\frac{\partial g(\bar{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}'}\right)^{-1}\frac{\partial g'(\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}}\boldsymbol{\Phi}\sqrt{n}g(\boldsymbol{\gamma}_0)$, where $\bar{\boldsymbol{\gamma}}$ lies between $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}_0$ elementwise. Then, the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ follows by applying the CLT in Lemma 4 in the Appendix to $\sqrt{n}g(\boldsymbol{\gamma}_0)$ and showing that $\frac{\partial g(\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}'} - \mathrm{E}\left(\frac{\partial g(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}\right) = o_p(1)$. Thus, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty}(\mathbf{G}'\boldsymbol{\Phi}\mathbf{G})^{-1}\mathbf{G}'\boldsymbol{\Phi}\mathbf{H}\boldsymbol{\Phi}\mathbf{G}(\mathbf{G}'\boldsymbol{\Phi}\mathbf{G})^{-1}\right). \tag{6.2}$$

From the expression for the variance-covariance matrix of IGMME, we can see that the precision of the estimator can be improved by replacing the arbitrary weighing matrix $\boldsymbol{\Phi}$ in the objective function with $\mathbf{H}^{-1}$. The resulting GMME is called the optimal GMME (Hansen, 1982). However, this estimator is not feasible as $\mathbf{H}^{-1}$ is unknown. To make it feasible, a plug-in estimator $\hat{\mathbf{H}} \equiv \mathbf{H}(\hat{\boldsymbol{\gamma}})$ based on the initial GMME $\hat{\boldsymbol{\gamma}}$ can be formulated. Then, the feasible optimal GMME is defined by

$$\hat{\boldsymbol{\gamma}}_o = \arg\min_{\boldsymbol{\gamma}} g'(\boldsymbol{\gamma})\hat{\mathbf{H}}^{-1}g(\boldsymbol{\gamma}). \tag{6.3}$$

Under some conditions, Debarsy et al. (2015) show that

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}}_o - \boldsymbol{\gamma}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty}(\mathbf{G}'\mathbf{H}^{-1}\mathbf{G})^{-1}\right). \tag{6.4}$$

[3]The low level assumptions guaranteeing the large sample properties are provided in Debarsy et al. (2015).

Debarsy et al. (2015) determine the best set of moment functions that provide the most efficient GMME for the MESS(1,1) under homoskedasticity. Their idea is to decompose the components of the inverse of the variance-covariance matrix of the optimal GMME, and then use the Cauchy-Schwarz inequality in such a way that an upper bound on the inverse of the variance-covariance matrix that is free of arbitrary pieces of the moment functions (free of $\mathbf{P}_i$'s and $\mathbf{F}$) can be attained. The resulting GMME is termed as the best GMME (BGMME). When the disturbance terms are normally distributed, the BGMME turns out to be asymptotically as efficient as the QMLE. However, when the disturbance terms are not normally distributed, and $\mathbf{W}$ and $\mathbf{M}$ do not commute, the BGMME can be asymptotically more efficient than the QMLE. The best set of moment functions is

$$g^*(\boldsymbol{\gamma}) = \frac{1}{n}\left(\mathbf{V}'(\boldsymbol{\gamma})\mathbf{P}_1^*\mathbf{V}(\boldsymbol{\gamma}), \ldots, \mathbf{V}'(\boldsymbol{\gamma})\mathbf{P}_{k^*+4}^*\mathbf{V}(\boldsymbol{\gamma}), \mathbf{V}'(\boldsymbol{\gamma})\mathbf{F}^*\right)', \tag{6.5}$$

where $\mathbf{P}_1^* = \mathbb{W}$, $\mathbf{P}_2^* = \text{Diag}(\mathbb{W})$, $\mathbf{P}_3^* = \text{Diag}(e^{\rho_0\mathbf{M}}\mathbf{W}\mathbf{X}\boldsymbol{\beta}_0)^{(t)}$, $\mathbf{P}_4^* = \mathbf{M}$, $\mathbf{P}_{m+4}^* = \text{Diag}(e^{\rho_0\mathbf{M}}\mathbf{X}_m)^{(t)}$ for $m = 1, ..., k^*$, and $\mathbf{F}^* = (\mathbf{F}_1^*, \mathbf{F}_2^*, \mathbf{F}_3^*, \mathbf{F}_4^*)$ with $\mathbf{F}_1^* = e^{\rho_0\mathbf{M}}\mathbf{X}^*$, $\mathbf{F}_2^* = e^{\rho_0\mathbf{M}}\mathbf{W}\mathbf{X}\boldsymbol{\beta}_0$, $\mathbf{F}_3^* = \boldsymbol{l}$, $\mathbf{F}_4^* = \text{vec}_D(\mathbb{W})$, where $\mathbf{X}^*$ excludes the intercept term in $\mathbf{X}$ if $\mathbf{M}$ is row-normalized so that $\mathbf{F}_1^*$ does not contain the intercept term generated in $e^{\rho_0\mathbf{M}}\mathbf{X}$, $k^*$ is the number of columns in $\mathbf{X}^*$, $\mathbf{A}^{(t)} = \mathbf{A} - \mathbf{I}_n\text{tr}(\mathbf{A})/n$ for any $n \times n$ matrix $\mathbf{A}$ and $\boldsymbol{l}$ is an $n \times 1$ vector of ones.

## 6.2 Estimation under heteroskedasticity

In this subsection, we consider the GMM estimation of MESS(1,1) under Assumption 2. Recall that $\boldsymbol{\Sigma}$ denotes the variance-covariance matrix of the error terms, i.e., $\boldsymbol{\Sigma} = \text{Diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Similar to the homoskedastic case, we again employ the following vector of moment functions consisting of $k_p$ quadratic moment functions and $k_f$ linear moment functions:

$$g(\boldsymbol{\gamma}) = \frac{1}{n}(\mathbf{V}'(\boldsymbol{\gamma})\mathbf{P}_1\mathbf{V}(\boldsymbol{\gamma}), \ldots, \mathbf{V}'(\boldsymbol{\gamma})\mathbf{P}_{k_p}\mathbf{V}(\boldsymbol{\gamma}), \mathbf{V}'(\boldsymbol{\gamma})\mathbf{F})'.$$

At $\boldsymbol{\gamma}_0$, we have $\text{E}\left(\mathbf{V}'\mathbf{P}_m\mathbf{V}\right) = \text{tr}\left(\mathbf{P}_m\boldsymbol{\Sigma}\right) = \text{tr}\left(\boldsymbol{\Sigma}\,\text{Diag}(\mathbf{P}_i)\right)$, which is equal to zero if the diagonal elements of $\mathbf{P}_m$ are zeros. Hence, in the heteroskedastic case, we require that the diagonal elements of $\mathbf{P}_m$ are zeros, i.e., $\text{Diag}(\mathbf{P}_m) = \mathbf{0}$ for $m = 1, 2, \ldots, k_p$. Then, an initial GMME based on an arbitrary symmetric weighting matrix $\boldsymbol{\Phi}$, with rank greater than or equal to $k + 2$, can be defined as

$$\hat{\boldsymbol{\gamma}} = \arg\min g'(\boldsymbol{\gamma})\boldsymbol{\Phi}g(\boldsymbol{\gamma}). \tag{6.6}$$

Let $\mathbf{G} = \mathrm{E}\left(\frac{\partial g(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}\right)$ and $\mathbf{H} = n\mathrm{E}\left(g\left(\boldsymbol{\gamma}_0\right) g'\left(\boldsymbol{\gamma}_0\right)\right)$. By Lemma 2 in Appendix A, we can show that

$$\mathbf{H} = \frac{1}{n}\begin{pmatrix} \frac{1}{2}\boldsymbol{\omega}'\boldsymbol{\omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}'\boldsymbol{\Sigma}\mathbf{F} \end{pmatrix},$$

and

$$\mathbf{G} = \frac{1}{n}\begin{pmatrix} \mathbf{0} & \frac{1}{2}\boldsymbol{\omega}'\mathrm{vec}\left(\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{-1}\mathbb{W}\right)^s\boldsymbol{\Sigma}^{1/2}\right) & \frac{1}{2}\boldsymbol{\omega}'\,\mathrm{vec}\left(\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{-1}\mathbf{M}\right)^s\boldsymbol{\Sigma}^{1/2}\right) \\ -\mathbf{F}'e^{\tau_0\mathbf{M}}\mathbf{X} & \mathbf{F}'\mathbb{W}e^{\tau_0\mathbf{M}}\mathbf{X}\boldsymbol{\beta}_0 & \mathbf{0} \end{pmatrix},$$

where $\boldsymbol{\omega} = \mathrm{vec}(\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1^s\boldsymbol{\Sigma}^{1/2}, \ldots, \boldsymbol{\Sigma}^{1/2}\mathbf{P}_{k_p}^s\boldsymbol{\Sigma}^{1/2})$. It follows again that

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty}(\mathbf{G}'\boldsymbol{\Phi}\mathbf{G})^{-1}\mathbf{G}'\boldsymbol{\Phi}\mathbf{H}\boldsymbol{\Phi}\mathbf{G}(\mathbf{G}'\boldsymbol{\Phi}\mathbf{G})^{-1}\right). \tag{6.7}$$

Note that $\mathbf{G}$ and $\mathbf{H}$ involve the unknown diagonal matrix $\boldsymbol{\Sigma}$. These terms can be consistently estimated by replacing $\boldsymbol{\Sigma}$ with $\mathrm{Diag}(v_1^2(\hat{\boldsymbol{\gamma}}), \ldots, v_n^2(\hat{\boldsymbol{\gamma}}))$. Let $\hat{\mathbf{H}}$ be the plug-in estimator of $\mathbf{H}$ based on the initial GMME $\hat{\boldsymbol{\gamma}}$. Then, a feasible optimal robust GMME (RGMME) can be obtained as

$$\hat{\boldsymbol{\gamma}}_o = \arg\min g'(\boldsymbol{\gamma})\hat{\mathbf{H}}^{-1}g(\boldsymbol{\gamma}). \tag{6.8}$$

It can be shown that

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}}_o - \boldsymbol{\gamma}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty}(\mathbf{G}'\mathbf{H}^{-1}\mathbf{G})^{-1}\right). \tag{6.9}$$

In the heteroskedastic case, the best set of moment functions is not feasible because the moment functions involve the unknown $\boldsymbol{\Sigma}$, which cannot be consistently estimated. In practice, we can formulate the RGMME based on the following vector of moment functions (Debarsy et al., 2015):

$$g^*(\boldsymbol{\gamma}) = \frac{1}{n}\left(\mathbf{V}'(\boldsymbol{\gamma})(\hat{\mathbb{W}} - \mathrm{Diag}(\hat{\mathbb{W}}))\mathbf{V}(\boldsymbol{\gamma}), \mathbf{V}'(\boldsymbol{\gamma})\mathbf{M}\mathbf{V}(\boldsymbol{\gamma}), \mathbf{V}'(\boldsymbol{\gamma})\mathbf{F}\right)', \tag{6.10}$$

where $\mathbf{F} = (\hat{\mathbb{W}}e^{\hat{\tau}\mathbf{M}}\mathbf{X}\hat{\boldsymbol{\beta}}, e^{\hat{\tau}\mathbf{M}}\mathbf{X})$ with $\hat{\mathbb{W}} = e^{\hat{\tau}\mathbf{M}}\mathbf{W}e^{-\hat{\tau}\mathbf{M}}$.

# 7 Bayesian estimation approach

In this section, we provide a comprehensive review of Bayesian estimation methods for the MESS model under both homoskedastic and heteroskedastic errors (LeSage and Pace, 2007, 2009; Yang et al., 2021; Doğan et al., 2023).

### 7.1 Estimation under homoskedasticity

Following LeSage and Pace (2007), we assume the following independent prior distributions: $\lambda \sim N(\mu_\lambda, V_\lambda)$, $\rho \sim N(\mu_\rho, V_\rho)$, $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$, and $\sigma^2 \sim IG(a, b)$, where $IG$ denotes the inverse-gamma distribution. Under these prior distributions, the posterior distribution of parameters can be expressed as[4]

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

where $p(\boldsymbol{\theta})$ is the joint prior distribution of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2, \lambda, \rho)'$ and $p(\mathbf{Y}|\boldsymbol{\theta})$ is the likelihood function given as

$$p(\mathbf{Y}|\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' e^{\rho\mathbf{M}'} e^{\rho\mathbf{M}} (e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right).$$

Algorithm 1 describes a Gibbs sampler that can be used to generate random draws from $p(\boldsymbol{\theta}|\mathbf{Y})$.

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

1. *Sampling step for $\boldsymbol{\beta}$:*

$$\boldsymbol{\beta}|\mathbf{Y}, \lambda, \rho, \sigma^2 \sim N(\hat{\boldsymbol{\beta}}, \mathbf{K}_{\boldsymbol{\beta}}),$$

*where* $\mathbf{K}_{\boldsymbol{\beta}} = (\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \sigma^{-2}\mathbf{X}' e^{\rho\mathbf{M}'} e^{\rho\mathbf{M}}\mathbf{X})^{-1}$ *and* $\hat{\boldsymbol{\beta}} = \mathbf{K}_{\boldsymbol{\beta}}(\sigma^{-2}\mathbf{X}' e^{\rho\mathbf{M}'} e^{\rho\mathbf{M}} e^{\lambda\mathbf{W}}\mathbf{Y} + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}})$.

2. *Sampling step for $\sigma^2$:*

$$\sigma^2|\mathbf{Y}, \lambda, \rho, \boldsymbol{\beta} \sim IG(\hat{\sigma}^2, K_{\sigma^2}),$$

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

3. *Sampling step for $\lambda$:*

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

*Generate a candidate value $\lambda^{new}$ according to*

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

[4] We use $p(\cdot)$ to denote the relevant density functions, and ignore $\mathbf{X}$ in the conditional sets for the sake of simplicity.

*where $c_\lambda$ is a tuning parameter.*[5] *Then, accept the candidate value $\lambda^{new}$ with probability*

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

*4. Sampling step for $\rho$:*

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

*Use the random-walk Metropolis-Hastings algorithm described in Step 3 to generate random draws from $p(\rho|\mathbf{Y}, \boldsymbol{\beta}, \lambda, \sigma^2)$.*

In Algorithm 1, the conditional posterior distributions of $\boldsymbol{\beta}$ and $\sigma^2$ are determined from $p(\boldsymbol{\beta}|\mathbf{Y}, \lambda, \rho, \sigma^2) \propto p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\beta})$ and $p(\sigma^2|\mathbf{Y}, \lambda, \rho, \boldsymbol{\beta}) \propto p(\mathbf{Y}|\boldsymbol{\theta})p(\sigma^2)$, respectively. Since we assume conjugate priors for $\boldsymbol{\beta}$ and $\sigma^2$, these conditional posterior distributions take known forms as shown in Algorithm 1. The Bayesian argument used to determine these conditional posterior distributions is analogous to the one used for a linear regression model. On the other hand, the conditional posterior distributions of spatial parameters are non-standard because the likelihood function is non-linear in terms of these parameters. To sample these parameters, we use the random walk Metropolis-Hastings algorithm suggested by LeSage and Pace (2009).

### 7.2 Estimation under heteroskedasticity

Following Lesage (1997) and LeSage and Pace (2009), we assume that the disturbance terms have a scale mixture of normal distributions such that the scale mixture variables generate different distributions with distinct variance terms. Thus, we have $v_i|\eta_i \sim N(0, \eta_i\sigma^2)$, where $\eta_i$'s are independent scale mixture variables with $\eta_i \sim \text{IG}(\nu/2, \nu/2)$ for $i = 1, \ldots, n$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2, \lambda, \rho, \nu)'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$, and $\mathbf{H}(\boldsymbol{\eta}) = \text{Diag}(\eta_1, \ldots, \eta_n)$ be the $n \times n$ diagonal matrix with the $i$th diagonal element $\eta_i$. Then, we can derive the conditional likelihood function $p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\eta})$ as

$$p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\eta}) = (2\pi\sigma^2)^{-n/2}\left(\prod_{i=1}^{n}\eta_i\right)^{-1/2} \times \exp\left(-\frac{1}{2\sigma^2}\left(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)' e^{\rho\mathbf{M}'}\mathbf{H}^{-1}(\boldsymbol{\eta})e^{\rho\mathbf{M}}\left(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)\right). \tag{7.1}$$

To introduce a Bayesian estimation approach, we adopt the prior distributions assumed in Section 6.1 for $\boldsymbol{\beta}$, $\lambda$, $\rho$, and $\sigma^2$. In the heteroskedastic case, we also need to determine a prior distribution

[5] The tuning parameter is determined during the estimation such that the acceptance rate falls between 40% and 60% (LeSage and Pace, 2009).

for $\nu$. To that end, we note that the marginal distribution of $v_i$ is a $t$ distribution with mean zero, scale parameter $\sigma^2$ and $\nu$ degrees of freedom, i.e., $v_i \sim t_\nu(0, \sigma^2)$. Thus, we assume the following prior $\nu \sim \text{Uniform}(2, \bar{\nu})$, where $\text{Uniform}(a, b)$ denotes the uniform distribution over the interval $(a, b)$, and $\bar{\nu}$ is a known positive number. This prior distribution ensures that the variance of $v_i$ exists because $\nu > 2$. Also, we can set $\bar{\nu}$ to a large positive number so that the $t$ distribution is allowed to approximate the normal distribution well-enough.

The posterior distribution of parameters then takes the following form:

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

where $p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\eta})$ is the conditional likelihood function stated in (7.1) and $p(\boldsymbol{\theta}, \boldsymbol{\eta})$ is the joint prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Algorithm 2 describes a Gibbs sampler that can be used to generate random draws from $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{Y})$.

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

1. *Sampling step for $\boldsymbol{\beta}$:*

$$\boldsymbol{\beta}|\mathbf{Y}, \lambda, \rho, \sigma^2, \boldsymbol{\eta} \sim N(\hat{\boldsymbol{\beta}}, \mathbf{K}_{\boldsymbol{\beta}}),$$

*where* $\mathbf{K}_{\boldsymbol{\beta}} = (\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \sigma^{-2}\mathbf{X}'e^{\rho\mathbf{M}'}\mathbf{H}^{-1}(\boldsymbol{\eta})e^{\rho\mathbf{M}}\mathbf{X})^{-1}$, $\mathbf{H}(\boldsymbol{\eta}) = \text{Diag}(\eta_1, \ldots, \eta_n)$ *and* $\hat{\boldsymbol{\beta}} = \mathbf{K}_{\boldsymbol{\beta}}(\sigma^{-2}\mathbf{X}'e^{\rho\mathbf{M}'}\mathbf{H}^{-1}(\boldsymbol{\eta})e^{\rho\mathbf{M}}e^{\lambda\mathbf{W}}\mathbf{Y} + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}})$.

2. *Sampling step for $\sigma^2$:*

$$\sigma^2|\mathbf{Y}, \lambda, \rho, \boldsymbol{\beta}, \boldsymbol{\eta} \sim IG(\hat{\sigma}^2, K_{\sigma^2}),$$

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

3. *Sampling step for $\lambda$:*

$$p(\lambda|\mathbf{Y}, \boldsymbol{\beta}, \rho, \sigma^2, \boldsymbol{\eta}) \propto \exp\left(-\frac{1}{2}\left(\sigma^{-2}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'e^{\rho\mathbf{M}'}\mathbf{H}^{-1}(\boldsymbol{\eta})e^{\rho\mathbf{M}}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \mathbf{V}_\lambda^{-1}(\lambda^2 - 2\mu_\lambda\lambda)\right)\right).$$

*Use the random-walk Metropolis-Hastings algorithm described in Step 3 of Algorithm 1 to sample this parameter.*

4. *Sampling step for $\rho$:*

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

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

5. *Sampling step for* $\boldsymbol{\eta}$*:*

$$\eta_i|\mathbf{Y}, \lambda, \rho, \boldsymbol{\beta}, \sigma^2, \nu \sim IG\left(\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{Y_i^2(\boldsymbol{\gamma})}{2\sigma^2}\right) \quad for \quad i = 1, 2, \ldots, n,$$

*where* $Y_i(\boldsymbol{\gamma})$ *is the* $i$*th element of* $\mathbf{Y}(\boldsymbol{\gamma}) = e^{\rho\mathbf{M}}\left(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)$.

6. *Sampling step for* $\nu$*:*

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

*Use the Griddy-Gibbs sampler to sample this parameter.*

The conditional posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\sigma^2$ take known forms as shown in Algorithm 2. In the case of spatial parameters, we again resort to the random walk Metropolis-Hastings algorithm suggested by LeSage and Pace (2009). The conditional posterior distribution of $\nu$ is determined from $p(\nu|\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\eta}, \lambda, \rho, \sigma^2) = p(\nu|\boldsymbol{\eta}) \propto p(\boldsymbol{\eta}|\nu)p(\nu)$. However, this distribution does not take a known form. Since $\nu$ has support over $(2, \bar{\nu})$, we suggest using a Griddy-Gibbs sampler to sample this parameter. Algorithm 3 describes this Griddy-Gibbs sampler.

**Algorithm 3** (The Griddy-Gibbs sampler for $\nu$)**.**

1. *Construct a grid of points* $\nu_1, \ldots, \nu_m$ *from the interval* $(2, \bar{\nu})$.
2. *Compute* $p_i = \sum_{j=1}^{i} p(\nu_j|\boldsymbol{\eta})$ *for* $i = 1, \ldots, m$*, and generate* $u$ *from Uniform*$(0, 1)$.
3. *Determine the smallest* $k$ *such that* $p_k \geq u$ *and return* $\nu = \nu_k$.

## 8 Estimation in the presence of endogenous and Durbin regressors

The preceding sections consider a regression model with spatial dependence specified by the MESS, where no endogenous regressors are included. In this section, we consider a MESS model with endogenous and Durbin regressors. The popular nonlinear two-stage least squares (N2SLS) estimation method in such a setting can have some irregular features (Jin and Lee, 2018).

Consider the following model:

$$e^{\lambda_0\mathbf{W}}\mathbf{Y} = \mathbf{X}^*\boldsymbol{\beta}_{10} + \mathbf{W}\boldsymbol{l}\boldsymbol{\beta}_{20} + \mathbf{W}\mathbf{X}_1\boldsymbol{\beta}_{30} + \mathbf{Z}\boldsymbol{\beta}_{40} + \mathbf{V}, \tag{8.1}$$

where $\boldsymbol{l}$ is an $n \times 1$ vector of ones, $\mathbf{X}_1$ excludes the intercept term from the exogenous variable matrix $\mathbf{X}$, $\mathbf{Z}$ is an $n \times k_z$ matrix of endogenous regressors, and $\mathbf{X}^* = \mathbf{X} = [\boldsymbol{l}, \mathbf{X}_1]$ if $\mathbf{W}$ is not row-normalized to have row sums equal to one, and $\mathbf{X}^* = \mathbf{X}_1$ otherwise. The $\boldsymbol{\beta}_{10}$, $\boldsymbol{\beta}_{20}$, $\boldsymbol{\beta}_{30}$ and $\boldsymbol{\beta}_{40}$ are conformable true parameters, and $\mathbf{W}$, $\mathbf{Y}$ and $\mathbf{V}$ have the same meanings as those in (2.1). The Durbin regressors $\mathbf{W}\mathbf{X}_1$ are neighbors' characteristics and capture exogenous externalities. When $\mathbf{W}$ is row-normalized, $\mathbf{W}\boldsymbol{l} = \boldsymbol{l}$ is the intercept term; when $\mathbf{W}$ is not row-normalized, $\mathbf{W}\boldsymbol{l}$ is also a Durbin regressor. In particular, if $\mathbf{W}$ is not row-normalized and has binary elements, $\mathbf{W}\boldsymbol{l}$ is a vector of out-degrees that measure the overall numbers of links for each spatial unit. Model (8.1) includes Durbin regressors explicitly since the MESS structure and the Durbin regressors lead to some irregular features of the N2SLS estimator. Model (2.1) has not considered Durbin regressors explicitly but can allow for that, where the theoretical analysis will not be affected although the related expressions for estimators need to be modified accordingly. To focus on the N2SLS estimation, a MESS process for the disturbances is not considered in (8.1).[6]

Let $\mathbf{F}$ be an $n \times k_f$ full rank IV matrix for the N2SLS estimation, where $k_f$ is not smaller than the total number of parameters in $\boldsymbol{\phi} = (\lambda, \boldsymbol{\beta}')'$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2, \boldsymbol{\beta}_3', \boldsymbol{\beta}_4')'$. For example, $\mathbf{F}$ can be the matrix formed by the independent columns of $[\boldsymbol{l}, \mathbf{X}_1, \mathbf{W}\boldsymbol{l}, \mathbf{W}\mathbf{X}_1, \mathbf{W}^2\boldsymbol{l}, \mathbf{W}^2\mathbf{X}_1, \bar{\mathbf{Z}}]$, where $\bar{\mathbf{Z}}$ is the IV matrix for $\mathbf{Z}$.[7] Assume that the elements of $\mathbf{V}$ are independent conditional on $\mathbf{F}$ but can have different conditional variances so that $\boldsymbol{\Sigma} = \mathrm{E}(\mathbf{V}\mathbf{V}'|\mathbf{F})$ is a diagonal matrix of conditional variances. Denote $\mathbf{D} = [\mathbf{X}^*, \mathbf{W}\boldsymbol{l}, \mathbf{W}\mathbf{X}_1, \mathbf{Z}]$ and $\boldsymbol{\Pi} = \mathbf{F}'\boldsymbol{\Sigma}\mathbf{F}$. The infeasible N2SLS estimation, as if $\boldsymbol{\Sigma}$ were known, has the objective function

$$Q(\boldsymbol{\phi}) = (e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{D}\boldsymbol{\beta})'\mathbf{F}\boldsymbol{\Pi}^{-1}\mathbf{F}'(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{D}\boldsymbol{\beta}). \tag{8.2}$$

The N2SLS estimator $\hat{\boldsymbol{\phi}}$ derived by minimizing $Q(\boldsymbol{\phi})$ is consistent under regularity conditions.

Let $\boldsymbol{\delta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2)'$ and $\boldsymbol{\xi} = (\boldsymbol{\beta}_3', \boldsymbol{\beta}_4')'$ when $\mathbf{W}$ is row-normalized, and let $\boldsymbol{\delta} = \boldsymbol{\beta}_1$ and $\boldsymbol{\xi} = (\boldsymbol{\beta}_2, \boldsymbol{\beta}_3', \boldsymbol{\beta}_4')'$ when $\mathbf{W}$ is not row-normalized. Then, $\boldsymbol{\xi}$ contains the coefficients for the Durbin and endogenous regressors. When $\boldsymbol{\xi}_0 \neq 0$, all components of $\hat{\boldsymbol{\phi}}$ are $\sqrt{n}$-consistent and $\hat{\boldsymbol{\phi}}$ has the asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \xrightarrow{d} N\Big(0, \lim_{n\to\infty}\Big\{\frac{1}{n}\,\mathrm{E}[(-\mathbf{W}\mathbf{D}\boldsymbol{\beta}_0, \mathbf{D})'\mathbf{F}]\bar{\boldsymbol{\Pi}}^{-1}\,\mathrm{E}[\mathbf{F}'(-\mathbf{W}\mathbf{D}\boldsymbol{\beta}_0, \mathbf{D})]\Big\}^{-1}\Big). \tag{8.3}$$

However, some components of $\hat{\boldsymbol{\phi}}$ have a rate of convergence slower than $\sqrt{n}$ and are not asymptotically normal in the case that $\boldsymbol{\xi}_0 = 0$, i.e., the Durbin and endogenous regressors are irrelevant, which is unknown when estimation is considered.

[6] If there is a MESS process for the disturbances, then as in Jin and Wang (2022), the GMM estimation with both linear and quadratic moments can be considered, since instrumental variables alone are not enough to identify parameters for the disturbance process.

[7] If $\mathbf{W}$ is row-normalized, then $\mathbf{W}\boldsymbol{l}$ and $\mathbf{W}^2\boldsymbol{l}$ are redundant.

When $\boldsymbol{\xi}_0 = 0$, we have

$$\frac{1}{\sqrt{n}}\frac{\partial Q(\boldsymbol{\phi}_0)}{\partial \lambda} + \frac{1}{\sqrt{n}}\frac{\partial Q(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\beta}'}(0_{1\times k^*}, \boldsymbol{\delta}_{20}, \boldsymbol{\delta}_{10}', 0_{1\times k_z})' = o_p(1),$$

where $k^*$ is the number of columns in $\mathbf{X}^*$, $\boldsymbol{\delta}_{20}$ is the last element of $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_{10}$ contains the remaining elements. Thus, $\frac{1}{\sqrt{n}}\frac{\partial Q(\boldsymbol{\phi}_0)}{\partial \lambda}$ and $\frac{1}{\sqrt{n}}\frac{\partial Q(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\beta}}$ are linearly dependent with probability approaching one (w.p.a.1.). As a result, $\frac{1}{n}\frac{\partial Q(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\theta}}\frac{\partial Q(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\theta}'}$ is singular w.p.a.1. In addition, we can show that $\frac{1}{n}\frac{\partial^2 Q(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}$ is also singular for large $n$. Hence, the usual method of deriving the asymptotic distribution of an estimator based on the mean value theorem expansion of the first order condition will not work.

The asymptotic distribution of $\hat{\boldsymbol{\phi}}$ in the case with $\boldsymbol{\xi}_0 = 0$ can be derived by first reparameterizing the model so that the derivative of the new N2SLS objective function with respect to a new parameter is exactly zero and then investigating a third order Taylor expansion of the first order condition at the true parameter vector. Let $\bar{\boldsymbol{\Pi}} = \mathrm{E}(\boldsymbol{\Pi})$, $k_d$ be the number of columns in $\mathbf{D}$, $J$ be a random vector that follows the normal distribution $N(0, \boldsymbol{\Delta})$, where

$$\boldsymbol{\Delta} = \lim_{n\to\infty}\begin{pmatrix} 2 & 0 \\ 0 & \mathbf{I}_{k_d} \end{pmatrix}\left(\frac{1}{n}\mathrm{E}[(-\mathbf{W}^2\mathbf{X}\boldsymbol{\delta}_0, \mathbf{D})'\mathbf{F}]\bar{\boldsymbol{\Pi}}^{-1}\,\mathrm{E}[\mathbf{F}'(-\mathbf{W}^2\mathbf{X}\boldsymbol{\delta}_0, \mathbf{D})]\right)^{-1}\begin{pmatrix} 2 & 0 \\ 0 & \mathbf{I}_{k_d} \end{pmatrix},$$

and $L = J_2 - \lim_{n\to\infty}[\frac{2}{n}\,\mathrm{E}(\mathbf{D}'\mathbf{F})\bar{\boldsymbol{\Pi}}^{-1}\,\mathrm{E}(\mathbf{F}'\mathbf{D})]^{-1}\frac{1}{n}\,\mathrm{E}(\mathbf{D}'\mathbf{F})\bar{\boldsymbol{\Pi}}^{-1}\,\mathrm{E}(\mathbf{F}'\mathbf{W}^2\mathbf{X})\boldsymbol{\delta}_0 J_1$, where $J_1$ is the first element of $J$ and $J_2$ contains the remaining elements of $J$. Then, in the case with $\boldsymbol{\xi}_0 = 0$, the N2SLS estimator $\hat{\boldsymbol{\phi}} = (\hat{\lambda}, \hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3', \hat{\boldsymbol{\beta}}_4')'$ has the asymptotic distribution

$$\begin{pmatrix} n^{1/4}(\hat{\lambda} - \lambda_0) \\ n^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \\ n^{1/4}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20}) \\ n^{1/4}(\hat{\boldsymbol{\beta}}_3 - \boldsymbol{\beta}_{30}) \\ n^{1/2}(\hat{\boldsymbol{\beta}}_4 - \boldsymbol{\beta}_{40}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} (-1)^B J_1^{1/2} \\ J_{2x^*} \\ (-1)^B \boldsymbol{\delta}_{20} J_1^{1/2} \\ (-1)^B \boldsymbol{\delta}_{10} J_1^{1/2} \\ J_{2z} \end{pmatrix} I(J_1 > 0) + \begin{pmatrix} 0 \\ L_{x^*} \\ 0_{k\times 1} \\ L_z \end{pmatrix} I(J_1 < 0), \tag{8.4}$$

where $I(\cdot)$ denotes the indicator function, $J_{2x^*}$ and $L_{x^*}$ are vectors consisting of the first $k^*$ elements of $J_2$ and $L$ respectively, $J_{2z}$ and $L_z$ are vectors consisting of the last $k_z$ elements of $J_2$ and $L$ respectively, and $B$ is a Bernoulli random variable with success probability described in Jin and Lee (2018). Thus, only $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_4$ are $\sqrt{n}$-consistent, and the remaining components of $\hat{\boldsymbol{\phi}}$ have a slow rate $n^{1/4}$ of convergence and follow non-standard asymptotic distributions.

The above N2SLS estimator is an infeasible estimator as $\boldsymbol{\Pi}$ is unknown. A feasible N2SLS estimator can be derived as follows. We may first derive an initial consistent but inefficient N2SLS estimator, e.g., the minimizer $\check{\boldsymbol{\phi}}$ of $(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{D}\boldsymbol{\beta})'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{D}\boldsymbol{\beta})$, and then consider the feasible N2SLS estimation with the objective function $\check{Q}(\boldsymbol{\phi}) = (e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{D}\boldsymbol{\beta})'\mathbf{F}(\mathbf{F}'\check{\boldsymbol{\Sigma}}\mathbf{F})^{-1}\mathbf{F}'(e^{\lambda\mathbf{W}}\mathbf{Y} -$

$\mathbf{D}\boldsymbol{\beta}$), where $\check{\boldsymbol{\Sigma}} = \text{Diag}(\check{v}_1^2, \cdots, \check{v}_n^2)$ with $\check{v}_i$ the $i$th element of $e^{\check{\lambda}\mathbf{W}}\mathbf{Y} - \mathbf{D}\check{\boldsymbol{\beta}}$. The feasible N2SLS estimator $\tilde{\boldsymbol{\phi}}$ has the same asymptotic distribution as the infeasible estimator $\hat{\boldsymbol{\phi}}$.

As $\boldsymbol{\xi}_0 = 0$ and $\boldsymbol{\xi}_0 \neq 0$ lead to different asymptotic distributions of $\hat{\boldsymbol{\theta}}$, Jin and Lee (2018) propose several tests for the hypothesis that $\boldsymbol{\xi}_0 = 0$. Depending on whether $\boldsymbol{\xi}_0 = 0$ is rejected or not, inference can be based on (8.3) or (8.4). Consider the case with $\boldsymbol{\xi}_0 \neq 0$ as an example. By (8.3), the variance of $\tilde{\boldsymbol{\phi}}$ can be estimated by $[(-\mathbf{W}\mathbf{D}\tilde{\boldsymbol{\beta}}, \mathbf{D})'\mathbf{F}(\mathbf{F}'\tilde{\boldsymbol{\Sigma}}\mathbf{F})^{-1}\mathbf{F}'(-\mathbf{W}\mathbf{D}\tilde{\boldsymbol{\beta}}, \mathbf{D})]^{-1}$, where $\tilde{\boldsymbol{\Sigma}} = \text{Diag}(\tilde{v}_1^2, \cdots, \tilde{v}_n^2)$ with $\tilde{v}_i$ the $i$th element of $e^{\tilde{\lambda}\mathbf{W}}\mathbf{Y} - \mathbf{D}\tilde{\boldsymbol{\beta}}$.

An interesting alternative estimation method is the adaptive group LASSO (AGLASSO), which can implement model selection and estimation simultaneously. The resulting estimator has the oracle properties (Fan and Li, 2001), so that the true model can be selected w.p.a.1. and the estimator always has the $\sqrt{n}$-rate of convergence and asymptotic normal distribution. The AGLASSO objective function to be minimized is

$$\frac{1}{n}\check{Q}(\boldsymbol{\phi}) + \alpha_n \|\check{\boldsymbol{\xi}}\|^{-\mu}\|\boldsymbol{\xi}\|, \tag{8.5}$$

where $\alpha_n$ is a tuning parameter that is positive and converges to zero, $\check{\boldsymbol{\xi}}$ is an initial consistent estimator, and $\mu$ is some positive number such as 1 or 2. Under regularity conditions, the AGLASSO estimator $\dot{\boldsymbol{\phi}}$ is consistent. In the case that $\boldsymbol{\xi}_0 = 0$, the probability that $\dot{\boldsymbol{\xi}} = 0$ goes to one as $n$ goes to infinity, that is, $\dot{\boldsymbol{\phi}}$ has the sparsity property, and for the remaining parameters $\boldsymbol{\psi} = (\lambda, \boldsymbol{\delta}')'$, the AGLASSO estimator has an asymptotic normal distribution as if $\boldsymbol{\xi}_0$ were known:

$$\sqrt{n}(\dot{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \xrightarrow{d} N\Big(0, \lim_{n\to\infty}\frac{1}{n}\big\{\text{E}[(-\mathbf{W}\mathbf{X}\boldsymbol{\delta}_0, \mathbf{X})'\mathbf{F}]\bar{\boldsymbol{\Pi}}^{-1}\,\text{E}[\mathbf{F}'(-\mathbf{W}\mathbf{X}\boldsymbol{\delta}_0, \mathbf{X})]\big\}^{-1}\Big),$$

in the case that $\boldsymbol{\xi}_0 \neq 0$, under the condition that $\alpha_n = o(n^{-1/2})$ and other regularity conditions, $\dot{\boldsymbol{\phi}}$ has the same asymptotic normal distribution as that stated in (8.3). Similar to the variance estimation of $\tilde{\boldsymbol{\phi}}$, the variance of $\dot{\boldsymbol{\psi}}$ for the case with $\boldsymbol{\xi}_0 = 0$ can be estimated by $[(-\mathbf{W}\mathbf{X}\dot{\boldsymbol{\delta}}, \mathbf{X})'\mathbf{F}(\mathbf{F}'\dot{\boldsymbol{\Sigma}}\mathbf{F})^{-1}\mathbf{F}'(-\mathbf{W}\mathbf{X}\dot{\boldsymbol{\delta}}, \mathbf{X})]^{-1}$, where $\dot{\boldsymbol{\Sigma}}$ is defined similarly to $\tilde{\boldsymbol{\Sigma}}$.

A practical question for the AGLASSO estimator is the selection of the tuning parameter $\alpha_n$. We can use an information criterion to choose $\alpha_n$. To make the dependence of $\dot{\boldsymbol{\phi}}$ on $\alpha_n$ explicit, denote the minimizer of $\frac{1}{n}\check{Q}(\boldsymbol{\phi}) + \alpha\|\tilde{\boldsymbol{\xi}}\|^{-\mu}\|\boldsymbol{\xi}\|$ by $\dot{\boldsymbol{\phi}}_\alpha$. Correspondingly, the AGLASSO estimator of $\boldsymbol{\xi}$ is $\dot{\boldsymbol{\xi}}_\alpha$. Consider the following information criterion:

$$h_n(\alpha) = \frac{1}{n}\check{Q}(\dot{\boldsymbol{\phi}}_\alpha) - I(\dot{\boldsymbol{\xi}}_\alpha = 0)\Gamma_n,$$

where $\Gamma_n > 0$ satisfies $\Gamma_n \to 0$ and $n^{1/2}\Gamma_n \to \infty$ as $n \to \infty$. For example, we may take $\Gamma_n = O(n^{-1/4})$. The tuning parameter chosen by minimizing $h_n(\alpha)$ can achieve model selection consistency.

The Monte Carlo results presented in Jin and Lee (2018) show that the N2SLS and AGLASSO

estimators have similar performance in the regular case with $\boldsymbol{\xi}_0 \neq 0$, but the AGLASSO estimator performs significantly better in the irregular case with $\boldsymbol{\xi}_0 = 0$. Thus, we suggest the use of the AGLASSO estimator.

# 9 Impact measures

In empirical applications, practitioners are often interested in quantifying the marginal effect of an explanatory variable on an outcome variable. In spatial econometric models, due to transmission channels, calculation of marginal effects and their interpretation become less straightforward. In this section, we review the summary measures suggested in the literature for the interpretation and presentation of marginal effects in a MESS(1,1) model.

From the model definition in (2.1), the marginal effect of a change in the $k$th explanatory variable $\mathbf{X}_k$ on $\mathrm{E}(\mathbf{Y})$ is given by $e^{-\lambda_0 \mathbf{W}}\beta_{0k}$, where $\beta_{0k}$ is the $k$th element of the true coefficient vector $\boldsymbol{\beta}_0$. LeSage and Pace (2009) propose three scalar measures for the marginal effect to ease the interpretation and presentation of this marginal effect:

1. Average Direct Impact (ADI): $\frac{1}{n}\mathrm{tr}(e^{-\lambda_0 \mathbf{W}}\beta_{0k})$,
2. Average Indirect Impact (AII): $\frac{1}{n}\left(\beta_{0k}\boldsymbol{l}' e^{-\lambda_0 \mathbf{W}}\boldsymbol{l} - \mathrm{tr}(e^{-\lambda_0 \mathbf{W}}\beta_{0k})\right)$,
3. Average Total Impact (ATI): $\frac{1}{n}\beta_{0k}\boldsymbol{l}' e^{-\lambda_0 \mathbf{W}}\boldsymbol{l}$.

The ADI, AII and ATI are, respectively, the average of the main diagonal elements of $e^{-\lambda_0 \mathbf{W}}\beta_{0k}$, the average of the off-diagonal elements of $e^{-\lambda_0 \mathbf{W}}\beta_{0k}$, and the average of all the elements of $e^{-\lambda_0 \mathbf{W}}\beta_{0k}$. In empirical studies, the ADI, ATI, and AII can be interpreted as the average own response, the average total response, and the average others' response of $\mathbf{Y}$ to a change in $\mathbf{X}_k$, respectively.

There are alternative ways that can be used to determine the dispersion of these scalar impact measures (Arbia et al., 2020). In the Bayesian estimation approach, a sequence of random draws for each impact measure can be obtained by using the posterior draws. Then, the mean and the standard deviation calculated from each sequence of impact measures can be used for inference.

In the classical estimation approaches, the delta method can be used to determine the asymptotic distributions of impact measure estimators. Applying the mean value theorem to ADI estimator $\frac{1}{n}\mathrm{tr}(e^{-\hat{\lambda}\mathbf{W}}\hat{\beta}_k)$ yields

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(\mathrm{tr}(e^{-\hat{\lambda}\mathbf{W}}\hat{\beta}_k) - \mathrm{tr}(e^{-\lambda_0 \mathbf{W}}\beta_{0k})\right) &= \frac{1}{\sqrt{n}}\left(-\mathrm{tr}(e^{-\hat{\lambda}\mathbf{W}}\mathbf{W}\hat{\beta}_k)(\hat{\lambda}-\lambda_0) + \mathrm{tr}(e^{-\hat{\lambda}\mathbf{W}})(\hat{\beta}_k - \beta_{0k})\right) + o_p(1) \\ &= \mathbf{A}_1 \times \sqrt{n}(\hat{\lambda}-\lambda_0, \hat{\beta}_k - \beta_{0k})' + o_p(1) \xrightarrow{d} N\Big(0, \lim_{n\to\infty} \mathbf{A}_1 \mathbf{B} \mathbf{A}_1'\Big),\end{aligned}$$

where $\mathbf{A}_1 = \left(-\frac{1}{n}\text{tr}(e^{-\lambda_0\mathbf{W}}\mathbf{W}\beta_{0k}), \frac{1}{n}\text{tr}(e^{-\lambda_0\mathbf{W}})\right)$ and $\mathbf{B}$ is the asymptotic covariance of $\sqrt{n}(\hat{\lambda}-\lambda_0, \hat{\beta}_k - \beta_{0k})$. Thus, we can estimate the asymptotic variance of the direct impact as $\frac{1}{n}\hat{\mathbf{A}}_1\hat{\mathbf{B}}\hat{\mathbf{A}}_1'$, where $\hat{\mathbf{A}}_1 = \left(-\frac{1}{n}\text{tr}(e^{-\hat{\lambda}\mathbf{W}}\mathbf{W}\hat{\beta}_k), \frac{1}{n}\text{tr}(e^{-\hat{\lambda}\mathbf{W}})\right)$, and $\hat{\mathbf{B}}$ is the estimated asymptotic covariance of $\sqrt{n}(\hat{\lambda}-\lambda_0, \hat{\beta}_k-\beta_{0k})$.

Applying the mean value theorem to ATI estimator $\frac{1}{n}\hat{\beta}_k\boldsymbol{l}'e^{-\hat{\lambda}\mathbf{W}}\boldsymbol{l}$, we obtain

$$\frac{1}{\sqrt{n}}\left(\hat{\beta}_k\boldsymbol{l}'e^{-\hat{\lambda}\mathbf{W}}\boldsymbol{l} - \beta_{0k}\boldsymbol{l}'e^{-\lambda_0\mathbf{W}}\boldsymbol{l}\right) = \mathbf{A}_2 \times \sqrt{n}(\hat{\lambda}-\lambda_0, \hat{\beta}_k-\beta_{0k})' + o_p(1) \xrightarrow{d} N\left(0, \lim_{n\to\infty}\mathbf{A}_2\mathbf{B}\mathbf{A}_2'\right),$$

where $\mathbf{A}_2 = \left(-\frac{1}{n}\beta_k\boldsymbol{l}'e^{-\lambda_0\mathbf{W}}\mathbf{W}\boldsymbol{l}, \frac{1}{n}\boldsymbol{l}'e^{-\lambda_0\mathbf{W}}\boldsymbol{l}\right)$. Thus, $\text{Var}(\frac{1}{n}\hat{\beta}_k\boldsymbol{l}'e^{-\hat{\lambda}\mathbf{W}}\boldsymbol{l})$ can be estimated by $\frac{1}{n}\hat{\mathbf{A}}_2\hat{\mathbf{B}}\hat{\mathbf{A}}_2'$, where $\hat{\mathbf{A}}_2 = \left(-\frac{1}{n}\hat{\beta}_k\boldsymbol{l}'e^{-\hat{\lambda}\mathbf{W}}\mathbf{W}\boldsymbol{l}, \frac{1}{n}\boldsymbol{l}'e^{-\hat{\lambda}\mathbf{W}}\boldsymbol{l}\right)$.

Finally, applying the mean value theorem to the estimator of AII, we obtain

$$\begin{aligned}&\frac{1}{\sqrt{n}}\left(\left(\hat{\beta}_k\boldsymbol{l}'e^{-\hat{\lambda}\mathbf{W}}\boldsymbol{l} - \text{tr}(e^{-\hat{\lambda}\mathbf{W}}\hat{\beta}_k)\right) - \left(\beta_{0k}\boldsymbol{l}'e^{-\hat{\lambda}_0\mathbf{W}}\boldsymbol{l} - \text{tr}(e^{-\hat{\lambda}_0\mathbf{W}}\beta_{0k})\right)\right)\\ &\quad = (\mathbf{A}_2-\mathbf{A}_1) \times \sqrt{n}(\hat{\lambda}-\lambda_0, \hat{\beta}_k-\beta_{0k})' + o_p(1) \xrightarrow{d} N\left(0, \lim_{n\to\infty}(\mathbf{A}_2-\mathbf{A}_1)\mathbf{B}(\mathbf{A}_2-\mathbf{A}_1)'\right).\end{aligned}$$

Then, an estimate of $\text{Var}\left(\frac{1}{n}\left(\hat{\beta}_k\boldsymbol{l}'e^{-\hat{\lambda}\mathbf{W}}\boldsymbol{l} - \text{tr}(e^{-\hat{\lambda}\mathbf{W}}\hat{\beta}_k)\right)\right)$ is given by $\frac{1}{n}(\hat{\mathbf{A}}_2-\hat{\mathbf{A}}_1)\hat{\mathbf{B}}(\hat{\mathbf{A}}_2-\hat{\mathbf{A}}_1)'$.

To illustrate these summary impact measures, we consider two examples. The first example follows from the empirical application in Debarsy et al. (2015). They consider a modified gravity equation for explaining Belgium's outward FDI, which takes the following form:

$$\begin{aligned}&e^{\lambda\mathbf{W}}\text{LFDI} = \beta_0\mathbf{1}_n + \beta_1\text{LGDP} + \beta_2\text{LPOP} + \beta_3\text{OECD} + \beta_4\text{LDIS} + \beta_5\text{MP} + \mathbf{U},\\ &e^{\rho\mathbf{W}}\mathbf{U} = \mathbf{V},\end{aligned} \tag{9.1}$$

where LFDI is the $n \times 1$ vector of the logarithm of Belgium's outward FDI stock to host countries, LGDP is the $n \times 1$ vector of the logarithm of host countries' GDPs, LPOP is the $n \times 1$ vector of the logarithm of host countries' populations, OECD is a dummy variable indicating whether the host country is an OECD country, LDIS is the $n \times 1$ vector of the logarithm of bilateral distance between Belgium and host countries, and the last variable MP is the surrounding-market potential variable constructed by following Blonigen et al. (2007). The sample data is described in detail in Debarsy et al. (2015) and includes data on Belgium's outward FDI stock in 35 host countries in 2009, which constitute 94% of Belgium's total outward FDI stock. The estimation results reported in Table 7 of Debarsy et al. (2015) are included here as Table 1 for easy reference. The results show that the QMLE and the GMME produce almost identical point estimates: the estimates of $\lambda_0$, $\beta_{20}$, $\beta_{30}$, $\beta_{40}$ and $\beta_{50}$ are statistically significant and have expected signs while the estimates of $\beta_{60}$ and $\rho_0$ are statistically insignificant. Since the estimate of $\lambda_0$ is negative and statistically significant, Debarsy et al. (2015) conclude that the vertical FDI mode is the dominant type of outward FDI for Belgium.

They also estimate the SARAR(1,1) version of (9.1) by both the QMLE and the GMME. Note that these estimates of spatial parameters are not directly comparable between MESS and SAR. However, when the spatial weights matrix is row normalized, Debarsy et al. (2015) suggest a relation using ATI: $\lambda_{\text{SAR}} = 1 - e^{\lambda_{\text{MESS}}}$ for the spatial parameter $\lambda$ in $\mathbf{Y}$. In this respect, the positive estimates for $\lambda_{\text{SAR}}$ in columns (2), (3), (5) and (6) and negative estimates for $\lambda_{\text{MESS}}$ in columns (1) and (3) are compatible with this relation.

Table 1: Estimation results for the outward FDI example

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Constant | 7.005 | 5.094 | 5.383 | 6.350 | 5.094 | 5.383 |
| | (6.039) | (5.951) | (5.972) | (5.316) | (5.503) | (5.475) |
| LGDP | $1.087^{***}$ | $1.107^{***}$ | $1.104^{***}$ | $1.089^{***}$ | $1.107^{***}$ | $1.103^{***}$ |
| | (0.240) | (0.249) | (0.248) | (0.221) | (0.243) | (0.241) |
| LPOP | $-0.579^{***}$ | $-0.576^{**}$ | $-0.578^{***}$ | $-0.575^{***}$ | $-0.576^{***}$ | $-0.578^{***}$ |
| | (0.241) | (0.249) | (0.249) | (0.241) | (0.254) | (0.252) |
| OECD | $1.064^{*}$ | $1.061^{*}$ | $1.052^{*}$ | $1.115^{*}$ | $1.061^{*}$ | $1.051^{*}$ |
| | (0.549) | (0.557) | (0.559) | (0.592) | (0.615) | (0.616) |
| LDIS | $-1.234^{***}$ | $-1.164^{***}$ | $-1.172^{***}$ | $-1.211^{***}$ | $-1.164^{***}$ | $-1.172^{***}$ |
| | (0.238) | (0.219) | (0.221) | (0.224) | (0.203) | (0.204) |
| MP | 1.062 | 1.080 | 1.049 | 1.101 | 1.080 | 1.049 |
| | (1.094) | (1.107) | (1.131) | (1.177) | (1.205) | (1.205) |
| Spatial parameter in $\mathbf{Y}$ | $-0.329^{**}$ | $0.258^{**}$ | $0.258^{***}$ | $-0.331^{*}$ | $0.258^{**}$ | $0.257^{**}$ |
| | (0.159) | (0.111) | (0.115) | (0.179) | (0.124) | (0.125) |
| Spatial parameter in errors | 0.287 | 0.004 | −0.045 | 0.332 | 0.005 | −0.045 |
| | (0.434) | (0.516) | (0.529) | (0.629) | (0.420) | (0.418) |
| $n$ | 35 | 35 | 35 | 35 | 35 | 35 |

Notes: Standard errors are in parentheses; (1) is homoskedastic SARAR by QML, (2) is homoskedastic MESS(1,1) by QML, (3) is homoskedastic MESS(1,1) by GMM, (4) is heteroskedastic SARAR by GMM, (5) is heteroskedastic MESS(1,1) by QML and (6) is heteroskedastic MESS(1,1) by GMM; *, ** and *** correspond to significance at the 10%, 5% and 1%, respectively. This table is taken from Debarsy et al. (2015).

The average direct effects and average total effects are reported in Table 2. Although the MESS(1,1) model suggests an exponential decay pattern for the influence of high-order neighboring characteristics, while the SARAR(1,1) process indicates a geometric decay, we note that the summary impact measures provided in the table for both models are almost identical.

The second example follows from the empirical application in Pace and Barry (1997) on the US presidential election in 1980. The dataset contains variables on the election results and county characteristics for 3107 US counties. We consider the following MESS(1,1) specification

$$e^{\lambda \mathbf{W}}\text{LPV} = \beta_0 \mathbf{1}_n + \beta_1 \text{EDUC} + \beta_2 \text{LHOW} + \beta_3 \text{LINC} + \mathbf{U}, \quad e^{\rho \mathbf{W}}\mathbf{U} = \mathbf{V}, \tag{9.2}$$

where LPV is the $n \times 1$ vector of log proportion of voting age population that voted in the election,

Table 2: Average direct effects and average total effects for the outward FDI example

| | Average direct effects | | | Average total effects | | |
|---|---|---|---|---|---|---|
| | SARAR | MESS(1,1) | | SARAR | MESS(1,1) | |
| | GMM | QML | GMM | GMM | QML | GMM |
| LGDP | 1.096 | 1.109 | 1.105 | 0.887 | 0.920 | 0.917 |
| LPOP | −0.578 | −0.577 | −0.579 | −0.468 | −0.479 | −0.480 |
| OECD | 1.121 | 1.063 | 1.053 | 0.907 | 0.882 | 0.874 |
| LDIS | −1.218 | −1.165 | −1.174 | −0.986 | −0.968 | −0.975 |

Notes: Effects are computed from estimation results of heteroskedastic SARAR (estimated by GMM) and heteroskedastic MESS(1,1) (estimated by QML and GMM).

EDUC is the $n \times 1$ vector of log percentage of population with a twelfth grade or higher education, LHOW is the $n \times 1$ vector of log percentage of population with home-ownership, and LINC is the $n \times 1$ vector of log per capita income. We consider a contiguity based weights matrix constructed using the latitude and longitude of the counties for this application.

We estimate (9.2) using the QML, GMM and Bayesian methods. We use the `expm` and the matrix-vector product (`mvp`) methods from Section 3 to compute the estimation results and record the corresponding computation times (in seconds). In the case of `mvp` method, the truncation order $q$ is set to 15. For the Bayesian estimator, we set the length of the chain to 1500 draws, with first 500 draws as burn-ins. We also report the results for the SARAR version of (9.2) by using the fast estimation routines available in the Spatial Econometrics Toolbox provided by James LeSage.

Table 3: Presidential election voting example

| | MESS | | | | | | SARAR | | |
|---|---|---|---|---|---|---|---|---|---|
| | QML | | GMM | | Bayesian | | QML | GMM | Bayesian |
| | `expm` | `mvp` | `expm` | `mvp` | `expm` | `mvp` | | | |
| Constant | 0.738*** | 0.738*** | 0.732*** | 0.732*** | 0.734*** | 0.734*** | 0.856*** | 0.662*** | 0.858*** |
| | (0.052) | (0.052) | (0.051) | (0.051) | (0.054) | (0.054) | (0.029) | (0.035) | (0.109) |
| EDUC | 0.316*** | 0.316*** | 0.300*** | 0.300*** | 0.317*** | 0.317*** | 0.161*** | 0.168*** | 0.160*** |
| | (0.021) | (0.021) | (0.020) | (0.020) | (0.020) | (0.020) | (0.009) | (0.013) | (0.020) |
| LHOW | 0.572*** | 0.572*** | 0.571*** | 0.571*** | 0.572*** | 0.572*** | 0.239*** | 0.240*** | 0.237*** |
| | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) | (0.009) | (0.008) | (0.016) |
| LINC | −0.154*** | −0.154*** | −0.144*** | −0.144*** | −0.155*** | −0.155*** | −0.095*** | −0.099*** | −0.094*** |
| | (0.021) | (0.021) | (0.020) | (0.020) | (0.021) | (0.021) | (0.011) | (0.011) | (0.015) |
| $\lambda$ | −0.350*** | −0.350*** | −0.423*** | −0.423*** | −0.337*** | −0.337*** | 0.491*** | 0.468*** | 0.478*** |
| | (0.045) | (0.045) | (0.045) | (0.045) | (0.041) | (0.041) | (0.002) | (0.034) | (0.101) |
| $\rho$ | −0.443*** | −0.443*** | −0.374*** | −0.374*** | −0.458*** | −0.458*** | 0.251*** | 0.250*** | 0.239*** |
| | (0.055) | (0.055) | (0.055) | (0.055) | (0.050) | (0.050) | (0.020) | (0.041) | (0.146) |
| Time (in seconds) | 1072.431 | 6.017 | 4805.841 | 21.103 | 47742.328 | 11.423 | 1.936 | 0.273 | 13.944 |

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

The estimation results are shown in Table 3.[8] Overall, the coefficient estimates of the explanatory variables are consistent across columns in terms of sign and statistical significance. We also note that spatial parameter estimates are statistically significant. The estimates of $\lambda$ range from $-0.42$ to $-0.34$

[8]To estimate the model, we used a MacBook Pro 2016 with a 2.4GHz Intel Core i7 processor and 8 GB 1867 MHz LPDDR3 memory.

in the case of MESS, and from 0.46 to 0.49 in the case of SARAR. The estimates of $\rho$ range from $-0.45$ to $-0.37$ in the case of MESS model, and from 0.23 to 0.25 in the case of SAR model. We note again that the positive estimates for $\lambda_{\text{SARAR}}$ in columns (7) to (9) and negative estimates for $\lambda_{\text{MESS}}$ in columns (1) to (6) are compatible with the relation $\lambda_{\text{SARAR}} = 1 - e^{\lambda_{\text{MESS}}}$ using ATI.

In terms of computational time, the matrix-vector product approach offers significant gains over the scaling and squaring method (combined with a Pad'e approximation) for computing matrix exponential terms. However, the estimation time for the MESS specification using the matrix-vector product approach is slightly higher than that for the SARAR specification, except for the case of Bayesian estimator in this application.

The impact measure estimates are summarized in Table 4. The results show that the corresponding impact measures are very similar across the QML, GMM and Bayesian methods. For example, in the case of MESS, the average direct effect estimate for EDUC is 0.320 using the QML method, 0.305 using the GMM method, and 0.320 for the Bayesian method. In the case of SAR, the average direct effect estimate for EDUC is 0.170 using the QML method, 0.176 using the GMM method, and 0.169 for the Bayesian method.

As a prelude to the next section, we must emphasize that MESS and SAR models are not substitutes for each other, as they are non-nested. In practice, their performance may differ depending on the application. Therefore, the choice between these two models should be guided by formal model selection methods.

## 10 Model selection

There are various approaches in the literature for implementing model selection. In this section, we consider methods based on testing, information criteria, and marginal likelihood.[9]

### 10.1 Testing approach

The classical tests, such as the Wald, Lagrange multiplier (Rao score), and likelihood ratio tests, for inference on spatial parameters can be formulated by using the results on the asymptotic distributions of the estimators (Anselin, 1988; Anselin et al., 1996; Anselin, 2001; LeSage and Pace, 2009; Elhorst, 2014; Doğan et al., 2018; Bera et al., 2018, 2019). In the literature, to test non-nested hypotheses, the Cox statistic and the J statistic are adapted for mainly spatial autoregressive models (Anselin, 1984, 1986; Kelejian, 2008; Kelejian and Piras, 2011; Burridge, 2012; Jin and Lee, 2013). These non-nested

[9]The recent survey by Otto et al. (2024) reviews cross validation methods for geostatistical models. Whether cross validation methods can be applied to spatial econometric models such as the SAR and MESS models we consider here are not obvious, since these models are nonstationary, which can be seen from the distinct means and variances for different spatial units, and splitting the data set into training and test data would destroy the original spatial dependence structure.

Table 4: Impact measures for the presidential election voting example

| | MESS | | | SARAR | | |
|---|---|---|---|---|---|---|
| | QML | GMM | Bayesian | QML | GMM | Bayesian |
| | | | ADI | | | |
| EDUC | $0.320^{***}$ | $0.305^{***}$ | $0.320^{***}$ | $0.170^{***}$ | $0.176^{***}$ | $0.169^{***}$ |
| | (0.020) | (0.020) | (0.020) | (0.009) | (0.006) | (0.017) |
| LHOW | $0.578^{***}$ | $0.580^{***}$ | $0.578^{***}$ | $0.252^{***}$ | $0.251^{***}$ | $0.249^{***}$ |
| | (0.016) | (0.016) | (0.016) | (0.009) | (0.009) | (0.013) |
| LINC | $-0.156^{***}$ | $-0.147^{***}$ | $-0.156^{***}$ | $-0.099^{***}$ | $-0.103^{***}$ | $-0.099^{***}$ |
| | (0.021) | (0.020) | (0.021) | (0.011) | (0.006) | (0.015) |
| | | | AII | | | |
| EDUC | $0.129^{***}$ | $0.153^{***}$ | $0.124^{***}$ | $0.147^{***}$ | $0.140^{***}$ | $0.144^{***}$ |
| | (0.017) | (0.017) | (0.018) | (0.008) | (0.012) | (0.040) |
| LHOW | $0.234^{***}$ | $0.292^{***}$ | $0.224^{***}$ | $0.218^{***}$ | $0.201^{***}$ | $0.217^{***}$ |
| | (0.036) | (0.038) | (0.032) | (0.008) | (0.017) | (0.065) |
| LINC | $-0.063^{***}$ | $-0.074^{***}$ | $-0.060^{***}$ | $-0.086^{***}$ | $-0.083^{***}$ | $-0.084^{***}$ |
| | (0.011) | (0.012) | (0.011) | (0.010) | (0.010) | (0.024) |
| | | | ATI | | | |
| EDUC | $0.449^{***}$ | $0.458^{***}$ | $0.444^{***}$ | $0.317^{***}$ | $0.316^{***}$ | $0.312^{***}$ |
| | (0.027) | (0.027) | (0.029) | (0.018) | (0.016) | (0.034) |
| LHOW | $0.812^{***}$ | $0.872^{***}$ | $0.802^{***}$ | $0.471^{***}$ | $0.452^{***}$ | $0.466^{***}$ |
| | (0.043) | (0.044) | (0.039) | (0.017) | (0.023) | (0.061) |
| LINC | $-0.219^{***}$ | $-0.220^{***}$ | $-0.217^{***}$ | $-0.185^{***}$ | $-0.186^{***}$ | $-0.183^{***}$ |
| | (0.028) | (0.030) | (0.029) | (0.021) | (0.017) | (0.026) |

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

testing approaches can also be used for the model selection problem between the spatial autoregressive models and the MESS models.

In the J-test approach, we augment the null model with the predictor from the alternative model and then check whether the predictor can add significantly to the explanatory power of the augmented model (Davidson and MacKinnon, 1981). Han and Lee (2013b) consider the J-test for the model selection problem between the SARAR(1,0) and MESS (1,0) models. When the SARAR(1,0) model is the null model, we can formulate the null and the alternative hypotheses as

$$H_0 : \mathbf{Y} = \alpha\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{V},$$
$$H_1 : \mathbf{S}^{ex}(\lambda)\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^{ex} + \mathbf{V},$$

where $\mathbf{S}^{ex}(\lambda) = e^{\lambda\mathbf{W}}$ and $\boldsymbol{\beta}^{ex}$ is a conformable parameter vector for $\mathbf{X}$ in the alternative model. As in Kelejian and Piras (2011), Han and Lee (2013b) consider two predictors based on the alternative model. These predictors are $\hat{\mathbf{Y}}_1 = \mathbf{S}^{ex}(\hat{\lambda})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}^{ex}$ and $\hat{\mathbf{Y}}_2 = (\mathbf{I}_n - \mathbf{S}^{ex}(\hat{\lambda}))\mathbf{Y} + \mathbf{X}\hat{\boldsymbol{\beta}}^{ex}$, where $\hat{\lambda}$ and $\hat{\boldsymbol{\beta}}^{ex}$ are the QMLEs of $\lambda$ and $\boldsymbol{\beta}^{ex}$. Note that the first predictor is based on the reduced form of the alternative model while the second predictor is derived from the identity $\mathbf{Y} = (\mathbf{I}_n - \mathbf{S}^{ex}(\lambda))\mathbf{Y} + \mathbf{X}\boldsymbol{\beta}^{ex} + \mathbf{V}$. Then, the null model can be augmented with these predictors to obtain the following testing equation:

$$\mathbf{Y} = \alpha\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{Y}}_{r_1}\delta_{r_1} + \mathbf{V}, \tag{10.1}$$

for $r_1 = 1, 2$. Denote $\mathbf{V}(\boldsymbol{\eta}_{r_1}) = (\mathbf{I}_n - \alpha\mathbf{W})\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \hat{\mathbf{Y}}_{r_1}\delta_{r_1}$, where $\boldsymbol{\eta}_{r_1} = (\alpha, \boldsymbol{\beta}', \delta_{r_1})'$. To estimate the augmented model, Han and Lee (2013b) consider a GMME based on the following vector of linear and quadratic moment functions:

$$g(\boldsymbol{\eta}_{r_1}) = (\mathbf{V}'(\boldsymbol{\eta}_{r_1})\mathbf{P}_1\mathbf{V}(\boldsymbol{\eta}_{r_1}), \ldots, \mathbf{V}'(\boldsymbol{\eta}_{r_1})\mathbf{P}_q\mathbf{V}(\boldsymbol{\eta}_{r_1}), \mathbf{F}'\mathbf{V}(\boldsymbol{\eta}_{r_1})),$$

where $\mathbf{F}$ is a full-column rank matrix of IVs and $\mathbf{P}_m$'s are $n \times n$ matrices of constants with $\mathrm{tr}(\mathbf{P}_m) = 0$ for $m = 1, \ldots, q$. Following Kelejian and Prucha (2010), the IV matrix $\mathbf{F}$ can consist of the linearly independent columns of $(\mathbf{X}, \mathbf{W}\mathbf{X}, \ldots, \mathbf{W}^d\mathbf{X})$, where $d$ is a positive constant. Let $\boldsymbol{\Xi} = \mathrm{E}[g(\boldsymbol{\eta}_{0r_1})g'(\boldsymbol{\eta}_{0r_1})]$, where $\boldsymbol{\eta}_{0r_1} = (\alpha_0, \boldsymbol{\beta}_0', 0)'$ is the true parameter vector under $H_0$. Then, using Lemma 2, it can be shown that

$$\boldsymbol{\Xi} = \begin{pmatrix} (\mu_4 - 3\sigma_0^4)\boldsymbol{\omega}'\boldsymbol{\omega} & \mu_3\boldsymbol{\omega}'\mathbf{F} \\ \mu_3\mathbf{F}'\boldsymbol{\omega} & 0 \end{pmatrix} + \begin{pmatrix} \mathrm{tr}(\mathbf{P}_1\mathbf{P}_1^s) & \ldots & \mathrm{tr}(\mathbf{P}_1\mathbf{P}_q^s) & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{tr}(\mathbf{P}_q\mathbf{P}_1^s) & \ldots & \mathrm{tr}(\mathbf{P}_q\mathbf{P}_q^s) & 0 \\ 0 & \ldots & 0 & \frac{1}{\sigma_0^2}\mathbf{F}'\mathbf{F} \end{pmatrix},$$

where $\boldsymbol{\omega} = [\text{vec}_D(\mathbf{P}_1), \ldots, \text{vec}_D(\mathbf{P}_q)]$. Let $\frac{1}{n}\hat{\boldsymbol{\Xi}}$ be a consistent estimator of $\frac{1}{n}\boldsymbol{\Xi}$. Then, the feasible optimal GMME of $\boldsymbol{\eta}_{0r_1}$ is defined by $\hat{\boldsymbol{\eta}}_{r_1} = \arg\min_{\boldsymbol{\eta}_{r_1}} g^{'}(\boldsymbol{\eta}_{r_1})\hat{\boldsymbol{\Xi}}^{-1}g(\boldsymbol{\eta}_{r_1})$. Let $\lambda^*_{sar}$ and $\boldsymbol{\beta}^{ex*}_{sar}$ be the pseudo true values of $\lambda_0$ and $\boldsymbol{\beta}^{ex}_0$ under the null model, respectively. Define $\mathbf{S}^{ex*}_{sar} = e^{\lambda^*_{sar}\mathbf{W}}$, $\mathbf{S} = \mathbf{I}_n - \alpha_0\mathbf{W}$ and $\mathbf{G} = \mathbf{W}\mathbf{S}^{-1}$. Then, under some assumptions, Han and Lee (2013b) show that

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{r_1} - \boldsymbol{\eta}_{0r_1}) \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty}\left(\mathbf{D}^{'}_{r_1}\boldsymbol{\Xi}^{-1}\mathbf{D}_{r_1}\right)^{-1}\right), \tag{10.2}$$

for $r_1 = 1, 2$, where

$$\mathbf{D}_1 = \begin{pmatrix} \sigma_0^2\text{tr}(\mathbf{P}_1^s\mathbf{G}) & 0 & 0 \\ \vdots & \vdots & \vdots \\ \sigma_0^2\text{tr}(\mathbf{P}_q^s\mathbf{G}) & 0 & 0 \\ \mathbf{F}^{'}\mathbf{G}\mathbf{X}\boldsymbol{\beta}_0 & \mathbf{F}^{'}\mathbf{X} & \mathbf{F}^{'}\mathbf{S}^{ex*-1}_{sar}\mathbf{X}\boldsymbol{\beta}^{ex*}_{sar} \end{pmatrix},$$

$$\mathbf{D}_2 = \begin{pmatrix} \sigma_0^2\text{tr}(\mathbf{P}_1^s\mathbf{G}) & 0 & \sigma_0^2\text{tr}(\mathbf{P}_1^s(\mathbf{I}_n - \mathbf{S}^{ex*}_{sar})\mathbf{S}^{-1}) \\ \vdots & \vdots & \vdots \\ \sigma_0^2\text{tr}(\mathbf{P}_q^s\mathbf{G}) & 0 & \sigma_0^2\text{tr}(\mathbf{P}_q^s(\mathbf{I}_n - \mathbf{S}^{ex*}_{sar})\mathbf{S}^{-1}) \\ \mathbf{F}^{'}\mathbf{G}\mathbf{X}\boldsymbol{\beta}_0 & \mathbf{F}^{'}\mathbf{X} & \mathbf{F}^{'}((\mathbf{I}_n - \mathbf{S}^{ex*}_{sar})\mathbf{S}^{-1}\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta}^{ex*}_{sar}) \end{pmatrix}.$$

We summarize the estimation of $\boldsymbol{\eta}_{r_1}$ in Algorithm 4.

**Algorithm 4** (Estimation of the augmented model in (10.1))**.**

1. *Estimate the alternative model by the QMLE suggested in Section 3.1 and then compute the predictors* $\hat{\mathbf{Y}}_{r_1}$ *for* $r_1 = 1, 2$.
2. *Estimate the null model by one of the methods given in Section 3. Use the estimated values to get a plug-in estimate of* $\boldsymbol{\Xi}$.
3. *Compute* $\hat{\boldsymbol{\eta}}_{r_1} = \arg\min_{\boldsymbol{\eta}_{r_1}} g^{'}(\boldsymbol{\eta}_{r_1})\hat{\boldsymbol{\Xi}}^{-1}g(\boldsymbol{\eta}_{r_1})$.

The result in (10.2) can be used to construct the J statistic in three different ways: (i) the Wald (W) statistic, (ii) the distance difference (DD) statistic, and (iii) the gradient (G) statistic (Newey and West, 1987). Let $\mathbf{R} = (\mathbf{0}_{1\times(k+1)}, 1)$ and $\hat{\mathbf{D}}_{r_1}$ be the plug-in estimator of $\mathbf{D}_{r_1}$ based on $\hat{\boldsymbol{\eta}}_{r_1}$ for $r_1 = 1, 2$. Then, the first two statistics are given as

$$W_{r_1} = (\mathbf{R}\hat{\boldsymbol{\eta}}_{r_1})^{'}\left(\mathbf{R}\left(\hat{\mathbf{D}}^{'}_{r_1}\hat{\boldsymbol{\Xi}}^{-1}\hat{\mathbf{D}}_{r_1}\right)^{-1}\mathbf{R}^{'}\right)^{-1}(\mathbf{R}\hat{\boldsymbol{\eta}}_{r_1}), \tag{10.3}$$

$$\text{DD}_{r_1} = \min_{\{\boldsymbol{\eta}_{r_1}|\delta_{r_1}=0\}} g^{'}(\boldsymbol{\eta}_{r_1})\hat{\boldsymbol{\Xi}}^{-1}g(\boldsymbol{\eta}_{r_1}) - \min_{\boldsymbol{\eta}_{r_1}} g^{'}(\boldsymbol{\eta}_{r_1})\hat{\boldsymbol{\Xi}}^{-1}g(\boldsymbol{\eta}_{r_1}). \tag{10.4}$$

Let $\tilde{\boldsymbol{\eta}}_{r_1} = \arg\min_{\{\boldsymbol{\eta}_{r_1}|\delta_{r_1}=0\}} g^{'}(\boldsymbol{\eta}_{r_1})\hat{\boldsymbol{\Xi}}^{-1}g(\boldsymbol{\eta}_{r_1})$ be the restricted optimal GMME. Then, the gradient

test statistic is defined by

$$\mathrm{G}_{r_1} = g^{'}(\tilde{\boldsymbol{\eta}}_{r_1})\hat{\boldsymbol{\Xi}}^{-1}\tilde{\mathbf{D}}_{r_1}(\tilde{\mathbf{D}}_{r_1}^{'}\hat{\boldsymbol{\Xi}}^{-1}\tilde{\mathbf{D}}_{r_1})^{-1}\tilde{\mathbf{D}}_{r_1}\hat{\boldsymbol{\Xi}}^{-1}g(\tilde{\boldsymbol{\eta}}_{r_1}), \tag{10.5}$$

where $\tilde{\mathbf{D}}_{r_1}$ is the plug-in estimator of $\mathbf{D}_{r_1}$ based on $\tilde{\boldsymbol{\eta}}_{r_1}$ for $r_1 = 1, 2$. Under $H_0$, these statistics have a chi-squared distribution with one degree of freedom. Thus, we will reject $H_0$ at the 5% significance level if the test statistics are larger than 3.84.

When using the MESS model as the null model, the null and the alternative hypotheses take the following form:

$$H_0 : \mathbf{S}^{ex}(\lambda)\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^{ex} + \mathbf{V},$$
$$H_1 : \mathbf{Y} = \alpha\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{V}.$$

Let $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ be the QML estimates of $\alpha_0$ and $\boldsymbol{\beta}_0$ from the alternative model. Again, we consider two predictors $\hat{\mathbf{Y}}_1 = (\mathbf{I}_n - \hat{\alpha}\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{Y}}_2 = \hat{\alpha}\mathbf{W}\mathbf{Y} + \mathbf{X}\hat{\boldsymbol{\beta}}$. Thus, the augmented model is given by

$$\mathbf{S}^{ex}(\lambda)\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^{ex} + \hat{\mathbf{Y}}_{r_2}\delta_{r_2} + \mathbf{V}, \tag{10.6}$$

for $r_2 = 1, 2$. Let $\boldsymbol{\psi}_{r_2} = (\lambda, \boldsymbol{\beta}^{ex'}, \delta_{r_2})^{'}$, $\boldsymbol{\psi}_{0r_2} = (\lambda_0, \boldsymbol{\beta}^{ex'}, 0)^{'}$ for $r_2 = 1, 2$, and $\alpha_{ex}^*$ and $\boldsymbol{\beta}_{ex}^*$ be the pseudo true values of $\alpha$ and $\boldsymbol{\beta}$ under the null model. Han and Lee (2013b) consider the non-linear 2SLS estimator (N2SLSE) for the estimation of the augmented model. Let $g(\boldsymbol{\psi}_{r_2}) = \mathbf{F}^{'}\mathbf{V}(\boldsymbol{\psi}_{r_2})$ be the vector of linear moment functions, where $\mathbf{V}(\boldsymbol{\psi}_{r_2}) = \mathbf{S}^{ex}(\lambda)\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{ex} + \hat{\mathbf{Y}}_{r_2}\delta_{r_2}$ for $r_2 = 1, 2$. Then, the N2SLSE is defined by

$$\hat{\boldsymbol{\psi}}_{r_2} = \underset{\boldsymbol{\psi}_{r_2}}{\arg\min}\, \mathbf{V}^{'}(\boldsymbol{\psi}_{r_2})\mathbf{F}(\mathbf{F}^{'}\mathbf{F})^{-1}\mathbf{F}^{'}\mathbf{V}(\boldsymbol{\psi}_{r_2}). \tag{10.7}$$

Under some assumptions, it can be shown that

$$\sqrt{n}(\hat{\boldsymbol{\psi}}_{r_2} - \boldsymbol{\psi}_{0r_2}) \xrightarrow{d} N\left(\mathbf{0}, \sigma_0^{ex2}\left(\mathrm{plim}_{n\to\infty}\frac{1}{n}\mathbf{D}_{r_2}^{'}(\mathbf{F}^{'}\mathbf{F})^{-1}\mathbf{D}_{r_2}\right)^{-1}\right), \tag{10.8}$$

where $\mathbf{D}_1 = \mathbf{F}^{'}\left(\mathbf{W}\mathbf{X}\boldsymbol{\beta}_0^{ex}, \mathbf{X}, \mathbf{S}_{ex}^{*-1}\mathbf{X}\boldsymbol{\beta}_{ex}^*\right)$ and $\mathbf{D}_2 = \mathbf{F}^{'}\left(\mathbf{W}\mathbf{X}\boldsymbol{\beta}_0^{ex}, \mathbf{X}, \alpha_{ex}^*\mathbf{W}\mathbf{S}^{ex-1}\mathbf{X}\boldsymbol{\beta}_0^{ex} + \mathbf{X}\boldsymbol{\beta}_{ex}^*\right)$ with $\mathbf{S}_{ex}^* = \mathbf{I}_n - \alpha_{ex}^*\mathbf{W}$ and $\mathbf{S}^{ex} = e^{\lambda_0\mathbf{W}}$. Algorithm 5 summarizes the estimation of the augmented model in (10.6).

**Algorithm 5** (Estimation of the augmented model in (10.6))**.**

1. *Estimate the alternative model by the QMLE and then compute the predictors* $\hat{\mathbf{Y}}_{r_2}$ *for* $r_2 = 1, 2$.
2. *Use* $\mathbf{F} = \left(\mathbf{X}, \mathbf{W}\mathbf{X}, \ldots, \mathbf{W}^d\mathbf{X}\right)$ *to compute* $\hat{\boldsymbol{\psi}}_{r_2} = \arg\min_{\boldsymbol{\psi}_{r_2}} \mathbf{V}^{'}(\boldsymbol{\psi}_{r_2})\mathbf{F}(\mathbf{F}^{'}\mathbf{F})^{-1}\mathbf{F}^{'}\mathbf{V}(\boldsymbol{\psi}_{r_2})$.

Similar to the previous case in which the SARAR(1,0) model was the null model, we can use the result in (10.8) to derive the three test statistics. When the disturbance terms are heteroskedastic, robust methods are necessary to derive consistent estimators. However, the process to derive the three test statistics are similar to the homoskedastic case.

Instead of using the critical value 3.84 from the asymptotic distribution, we can use the bootstrap method to generate the empirical distribution of the test statistics. In this approach, we can report the bootstrapped p-value, which is the percentage of test statistics based on the bootstrapped samples that are greater than the corresponding test statistic obtained from the actual sample, to decide between $H_0$ and $H_1$ (MacKinnon, 2009). The bootstrap procedure for testing $H_0 : \mathbf{Y} = \alpha\mathbf{WY} + \mathbf{X}\boldsymbol{\beta} + \mathbf{V}$ against $H_1 : \mathbf{S}^{ex}(\lambda)\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^{ex} + \mathbf{V}$ is described in Algorithm 6.

**Algorithm 6** (Bootstrap testing procedure)**.**

1. *Compute* $W_{r_1}$, $DD_{r_1}$ *and* $G_{r_1}$ *for* $r_1 = 1, 2$.
2. *Estimate the null model by one of the methods provided in Section 3. Let* $\hat{\mathbf{V}}$ *be the vector of residuals.*
3. *Generate a random sample of size* $n$ *from* $\hat{\mathbf{V}}$ *using sampling with replacement. Denote this re-sampled residual vector by* $\hat{\mathbf{V}}^b$.
4. *Use parameter estimates from Step 2 to compute* $\mathbf{Y}^b = (\mathbf{I}_n - \hat{\lambda}\mathbf{W})^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{V}}^b)$. *Compute the bootstrapped versions of test statistics* $W^b_{r_1}$, $DD^b_{r_1}$ *and* $G^b_{r_1}$ *for* $r_1 = 1, 2$ *by using* $\mathbf{Y}^b$.
5. *Repeat Steps 3–4 for* 99 *times. Then, a test statistic rejects* $H_0$ *if the proportion of its bootstrapped versions that exceed the corresponding one computed in Step 1 is less than* 5%.

In the heteroskedastic case, besides using the heteroskedasticity robust estimation methods, we also need to use a wild bootstrap approach to generate the bootstrapped versions of the test statistics. The details of this approach are summarized in Han and Lee (2013b). The extensive simulation results reported in Han and Lee (2013b) indicate that all versions of the J-statistic can perform satisfactorily when the sample size is large.

Liu and Lee (2019) propose a non-degenerate Vuong-type model selection test for the model selection between the SARAR(1,1) and MESS(1,1) models. The log-likelihood function of the SARAR(1,1) model in (2.2) can be expressed as

$$\ln L_1(\boldsymbol{\theta}_1) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln|\mathbf{I}_n - \alpha\mathbf{W}| + \ln|\mathbf{I}_n - \tau\mathbf{W}| - \frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i(\boldsymbol{\theta}_1)^2,$$

where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}', \alpha, \tau, \sigma^2)'$ and $z_i(\boldsymbol{\theta}_1) = y_i - \alpha\mathbf{W}_{i\cdot}\mathbf{Y} - \tau\mathbf{M}_{i\cdot}\mathbf{Y} + \alpha\tau\sum_{k=1}^{n} m_{ik}\mathbf{W}_{k\cdot}\mathbf{Y} - \mathbf{X}_i\boldsymbol{\beta} + \tau\mathbf{M}_{i\cdot}\mathbf{X}\boldsymbol{\beta}$, with $\mathbf{X}_i$ being the $i$th row of $\mathbf{X}$, $m_{ik}$ being the $(i,k)$th element of $\mathbf{M}$, and $\mathbf{W}_{i\cdot}$ and $\mathbf{M}_{i\cdot}$ being the $i$th row

of $\mathbf{W}$ and $\mathbf{M}$, respectively. Then, we can write the log-likelihood function as $\ln L_1(\boldsymbol{\theta}_1) = \sum_{i=1}^n l_{1i}(\boldsymbol{\theta}_1)$, where $l_{1i}(\boldsymbol{\theta}_1) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 + \frac{1}{n}\ln|\mathbf{I}_n - \alpha\mathbf{W}| + \frac{1}{n}\ln|\mathbf{I}_n - \tau\mathbf{W}| - \frac{1}{2\sigma^2}z_i(\boldsymbol{\theta}_1)^2$. Similarly, we can express the log-likelihood function of the MESS(1,1) as

$$\ln L_2(\boldsymbol{\theta}_2) = \sum_{i=1}^n l_{2i}(\boldsymbol{\theta}_2),$$

where $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}', \lambda, \rho, \sigma^2)'$, $l_{2i}(\boldsymbol{\theta}_2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}h_i(\boldsymbol{\theta}_2)^2$ and $h_i(\boldsymbol{\theta}_2)$ is the $i$th element of $e^{\rho\mathbf{M}}(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Liu and Lee (2019) first show that the QMLEs of both models, where one of the models or both models are possibly misspecified, are consistent estimators of their pseudo-true values and are asymptotically normal.

Liu and Lee (2019) assume that the true data generating process is unknown, and one of the two models or both models might be misspecified. Let $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ be the pseudo-true parameter vectors in the SARAR(1,1) and MESS(1,1) models, respectively. Then, the null hypothesis and alternative hypotheses are given by

$$\begin{aligned}
H_0 &: \lim_{n\to\infty} \frac{1}{\sqrt{n}} \mathrm{E}\left[\ln \frac{L_1(\boldsymbol{\theta}_1^*)}{L_2(\boldsymbol{\theta}_2^*)}\right] = 0 \quad \text{(Models 1 and 2 are asymptotically equivalent)},\\
H_1 &: \lim_{n\to\infty} \frac{1}{\sqrt{n}} \mathrm{E}\left[\ln \frac{L_1(\boldsymbol{\theta}_1^*)}{L_2(\boldsymbol{\theta}_2^*)}\right] = +\infty \quad \text{(Model 1 is asymptotically better than model 2)},\\
H_2 &: \lim_{n\to\infty} \frac{1}{\sqrt{n}} \mathrm{E}\left[\ln \frac{L_1(\boldsymbol{\theta}_1^*)}{L_2(\boldsymbol{\theta}_2^*)}\right] = -\infty \quad \text{(Model 1 is asymptotically worse than model 2)}.
\end{aligned}$$

Let $\mathrm{LR}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \ln L_1(\hat{\boldsymbol{\theta}}) - \ln L_2(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the QMLEs of the two models. Define $\omega^2 = \mathrm{Var}(\frac{1}{\sqrt{n}}\mathrm{LR}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2))$ and $g_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = l_{1i}(\boldsymbol{\theta}_1) - l_{2i}(\boldsymbol{\theta}_2)$. Then, following Hsu and Shi (2017), Liu and Lee (2019) consider the following test statistic:

$$\hat{T} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) + \hat{\sigma}U}{\sqrt{\hat{\omega}^2 + \hat{\sigma}^2}},$$

where $\hat{\sigma}$ is a data-dependent scalar, $\hat{\omega}^2$ is an estimator of $\omega^2$ and $U \sim N(0,1)$. Under some regularity assumptions, it is shown that the test statistic converges to the standard normal distribution under the null hypothesis, i.e., $\hat{T} \xrightarrow{d} N(0,1)$ under $H_0$. Under the alternative hypotheses, they show that $\hat{T} \to +\infty$ under $H_1$, and $\hat{T} \to -\infty$ under $H_2$. In a Monte Carlo study, Liu and Lee (2019) show that the test statistic has good size and power properties.

### 10.2 Information criteria approach

The predictive accuracy of a model is usually measured through an information criterion, which is typically defined based on the deviance term $-2\ln p(\mathbf{Y}|\boldsymbol{\theta})$ (Gelman et al., 2003). The widely used

Akaike information criterion (AIC) takes the following form:

$$\text{AIC} = -2 \ln p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) + 2p, \tag{10.9}$$

where $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ and $p$ is the dimension of $\boldsymbol{\theta}$. In a Bayesian context, Spiegelhalter et al. (2002) suggest another criterion called the deviance information criterion (DIC):

$$\text{DIC} = \bar{D}(\boldsymbol{\theta}) + p_D,$$

where $\bar{D}(\boldsymbol{\theta})$ is called the posterior mean deviance and $p_D$ is a measure of the effective number of parameters in the model. The posterior mean deviance is defined by $\bar{D}(\boldsymbol{\theta}) = -2\,\text{E}\,(\ln p(\mathbf{Y}|\boldsymbol{\theta})|\mathbf{Y})$, where the expectation is taken with respect to the posterior distribution of $\boldsymbol{\theta}$. This term serves as a Bayesian measure of model fit. The effective number of parameters is defined by $p_D = \bar{D}(\boldsymbol{\theta}) - D(\bar{\boldsymbol{\theta}}) = -2\,\text{E}\,(\ln p(\mathbf{Y}|\boldsymbol{\theta})|\mathbf{Y}) + 2 \ln p(\mathbf{Y}|\bar{\boldsymbol{\theta}})$, where $\bar{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$. Thus, the DIC can be written as

$$\text{DIC} = -4\,\text{E}\,(\ln p(\mathbf{Y}|\boldsymbol{\theta})|\mathbf{Y}) + 2 \ln p(\mathbf{Y}|\bar{\boldsymbol{\theta}}).$$

Let $\{\boldsymbol{\theta}^r\}_{r=1}^R$ be a sequence of posterior draws. Then, the first term $\text{E}\,(\ln p(Y|\theta)|Y)$ in the DIC can be computed by $\text{E}\,(\ln p(\mathbf{Y}|\boldsymbol{\theta})|\mathbf{Y}) \approx \frac{1}{R}\sum_{r=1}^R \ln p(\mathbf{Y}|\boldsymbol{\theta}^r)$. The second term $\ln p(\mathbf{Y}|\bar{\boldsymbol{\theta}})$ in the DIC is computed by evaluating the log-likelihood function at the posterior mean $\bar{\boldsymbol{\theta}}$. Using a decision-theoretic perspective, it can be shown that both AIC and DIC choose the model whose predictive distribution is close to the true data generating process (Li et al., 2020).

In our heteroskedastic model considered in Section 7.2, there are alternative likelihood functions: (i) the conditional likelihood function denoted by $p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\eta})$, (ii) the complete-data likelihood function denoted by $p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})$, and (iii) the integrated (or observed) likelihood function denoted by $p(\mathbf{Y}|\boldsymbol{\theta}) = \int p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})\text{d}\boldsymbol{\eta}$. The log-conditional likelihood function is readily available and given by

$$\begin{aligned}\ln p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\eta}) = &-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^n \ln \eta_i \\ &- \frac{1}{2\sigma^2}(e^{\lambda \mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' e^{\rho \mathbf{M}'} \mathbf{H}^{-1}(\boldsymbol{\eta}) e^{\rho \mathbf{M}}(e^{\lambda \mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),\end{aligned} \tag{10.10}$$

where $\mathbf{H}(\boldsymbol{\eta}) = \text{Diag}\,(\eta_1, \ldots, \eta_n)$ is the $n \times n$ diagonal matrix with the $i$th diagonal element $\eta_i$. As shown in Section 6.2, this function facilitates the Bayesian estimation of the heteroskedastic model. Both the conditional likelihood function and the complete-data likelihood function depend on the high-dimensional latent scale mixture variables. Since these high-dimensional variables can not be estimated precisely, the AIC and DIC formulated with the conditional and complete-data likelihood functions

may not perform satisfactorily in model selection exercises (Chan and Grant, 2016). Indeed, the latent variable models violate the conditions of the decision-theoretic perspective, indicating that the AIC and DIC cannot be used as a measure of predictive accuracy (Li et al., 2020). Hopefully, the log-integrated likelihood function can be obtained analytically by integrating out the scale mixture variables $\boldsymbol{\eta}$ from the complete-data likelihood function, i.e., $p(\mathbf{Y}|\boldsymbol{\theta}) = \int p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\eta} = \int p(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\theta})p(\boldsymbol{\eta}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\eta}$. This function can be derived as (Doğan et al., 2023)

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

where $y_i(\boldsymbol{\delta})$ is the $i$th element of $\mathbf{Y}(\boldsymbol{\delta}) = e^{\rho\mathbf{M}}\left(e^{\lambda\mathbf{W}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)$ with $\boldsymbol{\delta} = (\lambda, \rho, \boldsymbol{\beta}')'$. This function can be used to formulate AIC and DIC in the heteroskedastic model.

Another popular criterion is the Bayesian information criterion, which can be derived from a large sample approximation to the log-marginal likelihood of a candidate model. Let $\{M_k\}_{k=1}^K$ be a sequence of candidate models. Then, the marginal likelihood of the model $M_k$ can be expressed as $p(\mathbf{Y}_k|M_k) = \int_{\boldsymbol{\Theta}_k} p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)\mathrm{d}\boldsymbol{\theta}_k$, where $\boldsymbol{\theta}_k$ is the $p_k \times 1$ vector of parameters in $M_k$. Then, the Laplace approximation to $\ln p(\mathbf{Y}_k|M_k)$ yields the following BIC measure (Schwarz, 1978):

$$\mathrm{BIC}_k = -2\ln p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) + 2p\ln(n). \tag{10.11}$$

The Laplace approximation to $\ln p(\mathbf{Y}_k|M_k)$ can also be used to show that (Kass and Raftery, 1995)

$$\lim_{n\to\infty} P\left(\left|\frac{\mathrm{BIC}_k - \mathrm{BIC}_l}{\ln \mathrm{BF}_{kl}} - 1\right| > \epsilon\right) = 0, \tag{10.12}$$

where $\epsilon > 0$ is an arbitrary number and $\mathrm{BF}_{kl} = p(\mathbf{Y}|M_k)/p(\mathbf{Y}|M_l)$ is the Bayes factor of $M_k$ against $M_l$. The result in (10.12) indicates that the BIC is also a consistent model selection criterion like the Bayes factor. Moreover, both BIC and the Bayes factor can be interpreted as the measures of predictive accuracy because the marginal likelihood function can be interpreted as the predictive density evaluated at $\mathbf{Y}$ (Chan and Grant, 2016).

In the Bayesian setting described in Section 7, Doğan et al. (2023) investigate the performance of AIC, DIC and BIC for both nested and non-nested model selection problems through simulations. They consider four popular MESS specifications and aim to see whether the information criteria can select correct model specification and the correct spatial weights matrix from a pool of candidates. Their extensive simulation results show that these criteria perform satisfactorily and can be useful for selecting the correct model in the specification search exercises. In an empirical illustration, they also consider the MESS counterpart of the spatial Durbin model considered in Ertur and Koch (2007)

(EK) under both homoskedasticity and heteroskedasticity. EK incorporate technological interdependence into a neo-classical Solow growth model to explore the impact of technology spillover effects on economic growth. The MESS counterpart of EK's empirical model takes the following form

$$e^{\lambda \mathbf{W}}\text{LOPW} = \beta_0 \mathbf{1}_n + \beta_1 \text{LSAV} + \beta_2 \text{LGLAB} + \beta_3 \mathbf{W} \times \text{LSAV} + \beta_4 \mathbf{W} \times \text{LGLAB} + \mathbf{V}, \tag{10.13}$$

where LOPW is the $n \times 1$ vector of log output per-worker, LSAV is the $n \times 1$ vector of log of fraction of savings, LGLAB the $n \times 1$ vector of growth rate of labor variable. $\mathbf{W}\times$LSAV and $\mathbf{W}\times$LGLAB stand for the Durbin terms. Doğan et al. (2023) consider two specifications for the spatial weights matrix. Let $d_{ij}$ denote the squared great-circle distance between country capitals $i$ and $j$. The first weights matrix is denoted as $\mathbf{W}_1$, and its elements are generated as $d_{ij}^{-2}$. The second one is denoted as $\mathbf{W}_2$ and its elements are generated as $e^{-2d_{ij}}$. Both weights matrices are row normalized.

The sample data come from EK and contain information on 91 countries for the year 1995. Doğan et al. (2023) estimate (10.13) under homoskedastic and heteroskedastic cases using the estimation algorithms described in Section 7. They report the mean and the standard deviation of the posterior draws provided below in Table 5. In columns (1) and (2), EK's results are reproduced for reference. All information criteria are reported in the bottom panel.

Columns (3) and (4) present the estimation results under homoskedasticity. We observe that the estimates for LSAV and LGLAB are close to those from EK. However, while EK report statistically significant negative estimates for $W\times$LSAV, in columns (3) and (4), although the estimates are close, they are no longer statistically significant. The coefficient for $W\times$LGLAB is estimated imprecisely similar to EK.

The estimates of $\lambda$ in columns (3) and (4) are around $-0.85$ and statistically significant. Note again that these estimates are not directly comparable to the estimates of $\lambda$ in columns (1) and (2). However, when the spatial weights matrix is row normalized, the relation $\lambda_{\text{SAR}} = 1 - e^{\lambda_{\text{MESS}}}$ suggested by Debarsy et al. (2015) can be used. We observe that $\lambda_{\text{SAR}} = 0.740$ and $\lambda_{\text{MESS}} = -0.857$ for $\mathbf{W}_1$, and $\lambda_{\text{SAR}} = 0.658$ and $\lambda_{\text{MESS}} = -0.822$ for $\mathbf{W}_2$. These coefficients have opposite signs, and approximately satisfy the relation in this application. For the information criteria, we can see that AIC and BIC have smaller values for $\mathbf{W}_1$, which implies that $\mathbf{W}_1$ is preferred over $\mathbf{W}_2$ for the SAR model. Similarly, for the MESS model, AIC, DIC, and BIC have smaller values for $\mathbf{W}_1$.

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 $\lambda$ in column (6), which is smaller in magnitude. The estimates of the number of degrees of freedom $\nu$ for the $t$ distribution indicate no significant deviations from the normality of the error terms. For the information criteria, we again observe that DIC, AIC

and BIC have smaller values for $\mathbf{W}_1$ compared to $\mathbf{W}_2$. Across columns (3) through (6), the lowest values for all information criteria are observed in column (5).

Table 5: Estimation results for the MESS spatial growth model

| | SAR | | MESS | | | |
|---|---|---|---|---|---|---|
| | Homoskedasticity | | Homoskedasticity | | Heteroskedasticity | |
| | (1) $\mathbf{W}_1$ | (2) $\mathbf{W}_2$ | (3) $\mathbf{W}_1$ | (4) $\mathbf{W}_2$ | (5) $\mathbf{W}_1$ | (6) $\mathbf{W}_2$ |
| 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) |
| LSAV | 0.825*** | 0.792*** | 0.949*** | 0.893*** | 0.957*** | 0.873*** |
| | (0.000) | (0.000) | (0.116) | (0.121) | (0.116) | (0.117) |
| LGLAB | −1.498*** | −1.451*** | −1.662*** | −1.614*** | −1.673*** | −1.556** |
| | (0.008) | (0.009) | (0.628) | (0.619) | (0.629) | (0.609) |
| $\mathbf{W}\times$LSAV | −0.322*** | −0.372*** | −0.292 | −0.332* | −0.338 | −0.062 |
| | (0.079) | (0.024) | (0.223) | (0.192) | (0.223) | (0.198) |
| $\mathbf{W}\times$LGLAB | 0.571 | 0.137 | 0.149 | −0.050 | 0.189 | −0.345 |
| | (0.501) | (0.863) | (0.842) | (0.788) | (0.844) | (0.787) |
| $\lambda$ | 0.740*** | 0.658*** | −0.857*** | −0.822*** | −0.894*** | −0.581*** |
| | (0.000) | (0.000) | (0.115) | (0.102) | (0.113) | (0.105) |
| $\sigma^2$ | | | 0.334*** | 0.349*** | 0.333*** | 0.355*** |
| | | | (0.052) | (0.054) | (0.052) | (0.056) |
| $\nu$ | | | | | 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 |
| BIC | 156.08 | 168.51 | 182.44 | 186.65 | 177.02 | 183.54 |

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

Yang et al. (2022) suggest using a Mallows $C_p$ type selection criterion for selecting a spatial weights matrix from a pool of candidates. Let $\mathcal{W} = \{(\mathbf{W}_s, \mathbf{M}_s) : s \in \{1, 2, \ldots, S\}\}$ be the pool of spatial weights matrices. The quasi log-likelihood function based on the tuple $(W_s, M_s)$ can be expressed as

$$\ell_s = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\|e^{\tau\mathbf{M}_s}(e^{\alpha\mathbf{W}_s}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\|^2, \tag{10.14}$$

where $\|\cdot\|$ denotes the Euclidean norm. For a given $(\hat{\alpha}_s, \hat{\tau}_s)$ value, the first order conditions of (10.14) with respect to $\boldsymbol{\beta}$ and $\sigma^2$ yield

$$\hat{\boldsymbol{\beta}}_s = \left(\mathbf{X}' e^{\hat{\tau}_s\mathbf{M}_s'} e^{\hat{\tau}_s\mathbf{M}_s}\mathbf{X}\right)^{-1}\mathbf{X}' e^{\hat{\tau}_s\mathbf{M}_s'} e^{\hat{\tau}_s\mathbf{M}_s} e^{\hat{\alpha}_s\mathbf{W}_s}\mathbf{Y}, \tag{10.15}$$

$$\hat{\sigma}_s^2 = \frac{1}{n}\left\|e^{\hat{\tau}_s\mathbf{M}_s}(e^{\hat{\alpha}_s\mathbf{W}_s}\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_s)\right\|^2. \tag{10.16}$$

Let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}) = e^{-\alpha_0 \mathbf{W}} \mathbf{X} \boldsymbol{\beta}_0$. Substituting (10.15) into $\hat{\boldsymbol{\mu}}_s = e^{-\hat{\alpha}_s \mathbf{W}_s} \mathbf{X} \hat{\boldsymbol{\beta}}_s$, we obtain

$$\hat{\boldsymbol{\mu}}_s = e^{-\hat{\alpha}_s \mathbf{W}_s} \mathbf{X} \left( \mathbf{X}' e^{\hat{\tau}_s \mathbf{M}_s'} e^{\hat{\tau}_s \mathbf{M}_s} \mathbf{X} \right)^{-1} \mathbf{X}' e^{\hat{\tau}_s \mathbf{M}_s'} e^{\hat{\tau}_s \mathbf{M}_s} e^{\hat{\alpha}_s \mathbf{W}_s} \mathbf{Y} = \widetilde{\mathbf{P}}_s \mathbf{Y}, \tag{10.17}$$

where $\widetilde{\mathbf{P}}_s = e^{-\hat{\alpha}_s \mathbf{W}_s} e^{-\hat{\tau}_s \mathbf{M}_s} \widehat{\mathbf{P}}_s e^{\hat{\tau}_s \mathbf{M}_s} e^{\hat{\alpha}_s \mathbf{W}_s}$ with $\widehat{\mathbf{P}}_s = e^{\hat{\tau}_s \mathbf{M}_s} \mathbf{X} \left( \mathbf{X}' e^{\hat{\tau}_s \mathbf{M}_s'} e^{\hat{\tau}_s \mathbf{M}_s} \mathbf{X} \right)^{-1} \mathbf{X}' e^{\hat{\tau}_s \mathbf{M}_s'}$. Then, Yang et al. (2022) consider the following selection criterion function:

$$C_s = \left\| \widetilde{\mathbf{P}}_s \mathbf{Y} - \mathbf{Y} \right\|^2 + 2 \left( \mathrm{tr}(\widetilde{\mathbf{P}}_s \boldsymbol{\Omega}) + \frac{\partial \hat{\lambda}_s}{\partial \mathbf{Y}'} \boldsymbol{\Omega} \frac{\partial \widetilde{\mathbf{P}}_s}{\partial \hat{\lambda}} \mathbf{Y} + \frac{\partial \hat{\rho}_s}{\partial \mathbf{Y}'} \boldsymbol{\Omega} \frac{\partial \widetilde{\mathbf{P}}_s}{\partial \hat{\rho}} \mathbf{Y} \right),$$

where $\boldsymbol{\Omega} = \sigma_0^2 e^{-\lambda_0 \mathbf{W}} e^{-\rho_0 \mathbf{M}} e^{-\rho_0 \mathbf{M}'} e^{-\lambda_0 \mathbf{W}'}$ is the variance of $\mathbf{Y}$, and the closed forms of $\frac{\partial \hat{\lambda}_s}{\partial \mathbf{Y}'}$, $\frac{\partial \widetilde{\mathbf{P}}_s}{\partial \hat{\lambda}}$ and $\frac{\partial \hat{\rho}_s}{\partial \mathbf{Y}'}$ can be found in Yang et al. (2022). Given an estimator of $\boldsymbol{\Omega}$, we can compute $C_s$ for each $s$. Thus, the selected model is defined as $\hat{s} = \arg\min_{s \in \{1,\ldots,S\}} C_s$. Under certain assumptions, Yang et al. (2022) show that the selection estimator $\hat{\boldsymbol{\mu}}_{\hat{s}}$ is asymptotically optimal in the sense that it is as efficient as the infeasible estimator that uses the best candidate spatial weights matrix. They also show that the selection procedure is selection consistent in the sense that it chooses the true tuple of weight matrices with probability approaching one as $n \to \infty$.

Instead of selecting the asymptotically optimal model, it is also possible to use a model averaging scheme that compromises across a set of candidate models. Let $\mathbf{w} = (w_1, \ldots, w_S)'$ be a vector of weights, and $\mathcal{N} = \left\{ \mathbf{w} \in [0,1]^S : \sum_{s=1}^S w_s = 1 \right\}$ be the set of model weights vectors. Let $\widetilde{\mathbf{P}}(\mathbf{w}) = \sum_{s=1}^S w_s \widetilde{\mathbf{P}}_s$ be the weighted average of $\left\{ \widetilde{\mathbf{P}}_1, \ldots, \widetilde{\mathbf{P}}_S \right\}$. Then, the model averaging estimator for $\boldsymbol{\mu}$ is given by

$$\hat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{s=1}^S w_s \hat{\boldsymbol{\mu}}_s = \sum_{s=1}^S w_s \widetilde{\mathbf{P}}_s \mathbf{Y} = \widetilde{\mathbf{P}}(\mathbf{w}) \mathbf{Y}. \tag{10.18}$$

Then, Yang et al. (2022) consider the following model weights choice criterion function:

$$C(\mathbf{w}) = \left\| \widetilde{\mathbf{P}}(\mathbf{w}) \mathbf{Y} - \mathbf{Y} \right\|^2 + 2 \left( \mathrm{tr} \left( \widetilde{\mathbf{P}}(\mathbf{w}) \boldsymbol{\Omega} \right) + \sum_{s=1}^S w_s \left( \frac{\partial \hat{\lambda}_s}{\partial \mathbf{Y}'} \boldsymbol{\Omega} \frac{\partial \widetilde{\mathbf{P}}_s}{\partial \hat{\lambda}_s} \mathbf{Y} + \frac{\partial \hat{\rho}_s}{\partial \mathbf{Y}'} \boldsymbol{\Omega} \frac{\partial \widetilde{\mathbf{P}}_s}{\partial \hat{\rho}_s} \mathbf{Y} \right) \right).$$

The optimal model weights vector is thus given by $\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{N}} \widehat{C}(\mathbf{w})$. Similar to the model selection procedure, the model averaging estimator $\hat{\boldsymbol{\mu}}(\hat{\mathbf{w}})$ is also asymptotically optimal.

The selection and averaging estimators can also be considered for the high order MESS models. In the case of heteroskedastic models, Yang et al. (2022) use a heteroskedasticity robust GMM estimator to formulate the selection and model averaging criterion functions. The extensive simulation results in Yang et al. (2022) indicate that the model selection and averaging estimators perform satisfactorily.

### 10.3 Marginal likelihood approach

In the Bayesian approach, the Bayes factor can be used for both nested and non-nested model selection problems. As shown in Section 10.2, the Bayes factor for two models is simply the ratio of the corresponding marginal likelihood functions: $\text{BF}_{kl} = p(\mathbf{Y}|M_k)/p(\mathbf{Y}|M_l)$, where $p(\mathbf{Y}|M_j) = \int_{\boldsymbol{\Theta}_j} p(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j)\text{d}\boldsymbol{\theta}_j$ for $j \in \{k, l\}$. Thus, the Bayes factor chooses $M_k$ if $p(\mathbf{Y}|M_k)$ is larger than $p(\mathbf{Y}|M_l)$. If the data is generated from $M_k$, then the Bayes factor will consistently choose $M_k$ over $M_l$. To see this, consider the expectation of the log-Bayes factor under $p(\mathbf{Y}|M_k)$:

$$\text{E}\left(\log \frac{p(\mathbf{Y}|M_k)}{p(\mathbf{Y}|M_l)}\right) = \int \log \frac{p(\mathbf{Y}|M_k)}{p(\mathbf{Y}|M_l)} p(\mathbf{Y}|M_k)\text{d}\mathbf{Y}, \tag{10.19}$$

which is simply the Kullback-Leibler divergence between $p(\mathbf{Y}|M_k)$ and $p(\mathbf{Y}|M_l)$. Thus, the expectation is strictly positive, unless $p(\mathbf{Y}|M_k) = p(\mathbf{Y}|M_l)$ in which case it is zero.

The Bayes factor reduces to the Savage-Dickey density ratio (SDDR) for the nested model selection problems (Verdinelli and Wasserman, 1995). For example, consider the following null and alternative hypotheses: $H_0 : \lambda = 0$ against $H_1 : \lambda \neq 0$ or $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. Let $M_R$ and $M_U$ be respectively the restricted and the unrestricted model. Then, the Bayes factor in favor of the unrestricted model is

$$\text{BF}_{UR} = \frac{p(\mathbf{Y}|M_U)}{p(\mathbf{Y}|M_R)}, \tag{10.20}$$

where $p(\mathbf{Y}|M_j)$ for $j \in \{U, R\}$ is the corresponding marginal likelihood function. Since our prior distributions are independent, the Bayes factor in (10.20) reduces to the SDDR given by

$$\text{BF}_{UR} = \frac{p(\lambda = 0|M_U)}{p(\lambda = 0|\mathbf{Y}, M_U)}, \tag{10.21}$$

where $p(\lambda = 0|M_U)$ and $p(\lambda = 0|\mathbf{Y}, M_U)$ are respectively the prior and the marginal posterior densities of $\lambda$ evaluated at $\lambda = 0$. The $\text{BF}_{UR}$ indicates that if $\lambda = 0$ is more likely under the prior relative to the marginal posterior, then the $\text{BF}_{UR}$ provides evidence in favor of $H_1$. Under the prior $\lambda \sim N(\mu_\rho, V_\rho)$, we have $p(\lambda = 0|M_U) = (2\pi V_\lambda)^{-1/2} \exp(-\mu_\lambda^2/2V_\lambda)$. Let $\{\boldsymbol{\beta}^r, \lambda^r, \rho^r, \sigma^{2r}\}_{r=1}^R$ be a sequence of posterior draws. Then, one way to estimate the marginal posterior $p(\lambda = 0|\mathbf{Y}, M_U)$ is to use the following Rao-Blackwell estimator (Gelfand and Smith, 1990):

$$\hat{p}(\lambda = 0|\mathbf{Y}, M_U) = \frac{1}{R}\sum_{r=1}^{R} p(\lambda = 0|\mathbf{Y}, \boldsymbol{\beta}^r, \rho^r, \sigma^{2r}). \tag{10.22}$$

This Rao-Blackwell estimator cannot be used in our case because the conditional posterior density of $\lambda$ does not take a standard form. If we assume that the parameter space of $\lambda$ is contained in the

interval $(-\tau, \tau)$, where $\tau$ is a finite positive constant, then we may resort to a Griddy-Gibbs sampler to estimate $p(\lambda = 0|\mathbf{Y}, M_U)$. Algorithm 7 describes how we can use this approach.

**Algorithm 7** (Computing SDDR)**.**

1. *Construct a grid with random points* $\lambda_1, \ldots, \lambda_m$ *from the interval* $(-\tau, \tau)$*. The grid must also include* $\lambda_k = 0$*.*
2. *Compute* $p_r(\lambda_i) = \frac{p(\lambda_i|\mathbf{Y}, \boldsymbol{\beta}^r, \rho^r, \sigma^{2r})}{\sum_{j=1}^{m} p(\lambda_j|\mathbf{Y}, \boldsymbol{\beta}^r, \rho^r, \sigma^{2r})}$ *for* $i = 1, \ldots, m$ *and* $r = 1, \ldots, R$*.*
3. *Compute* $p(\lambda_i) = \sum_{r=1}^{R} p_r(\lambda_i)$ *for* $i = 1, \ldots, m$*.*
4. *Return* $\hat{p}(\lambda = 0|\mathbf{Y}, M_U) = p(\lambda_k)$*.*

The marginal likelihood function of the MESS type models does not take a closed form. There are alternative methods that can be used to estimate or to approximate the marginal likelihood function. In the homoskedastic case, we can analytically integrate out $\boldsymbol{\beta}$ and $\sigma^2$ under the following priors: (i) $\boldsymbol{\beta}|\sigma^2 N(\boldsymbol{\mu}_\beta, \sigma^2 \mathbf{V}_\beta)$ and $\sigma^2 \sim IG(a, b)$ or (ii) $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma$. However, in order to get the marginal likelihood function, we also need to integrate out the spatial parameters, which is not possible analytically. Then, one approach for computing the marginal likelihood function can be based on a numerical integration method (Hepple, 1995; LeSage and Pace, 2009; Han and Lee, 2013a). In the case of MESS(1,1), this approach requires a double numerical integration over the parameter space of $\lambda$ and $\rho$. It is clear that this approach may not be feasible for high order MESS models and for models with heteroskedasticity.

Alternatively, since the conditional posterior distributions of the spatial parameters are in non-standard forms, we may resort to the method suggested by Chib and Jeliazkov (2001) to estimate the marginal likelihood function. This approach is general enough and only requires the MCMC draws of parameters. In the heteroskedastic case, this approach based on the conditional likelihood function $p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\eta})$ provided in Section 7.2 requires the MCMC draws of the high-dimensional scale mixture variables and may therefore not produce precise estimates (Frühwirth-Schnatter and Wagner, 2008). For this reason, we should instead use this method based on the integrated likelihood function $p(\mathbf{Y}|\boldsymbol{\theta})$ given in Section 10.2.

The modified harmonic mean method of Gelfand and Dey (1994) can also be used to estimate the marginal likelihood function. This method requires a probability density function $g$ whose support lies in the support of the posterior distribution. The method produces an approximation based on $\mathrm{E}\left(\frac{g(\boldsymbol{\theta})}{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\middle|\mathbf{Y}\right)$, where the expectation is taken with respect to $p(\boldsymbol{\theta}|\mathbf{Y})$. The expectation gives the

following relationship:

$$\mathrm{E}\left(\frac{g(\boldsymbol{\theta})}{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}|\mathbf{Y}\right) = \int \frac{g(\boldsymbol{\theta})}{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{Y})\mathrm{d}\boldsymbol{\theta} = \int \frac{g(\boldsymbol{\theta})}{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{Y})}\mathrm{d}\boldsymbol{\theta}$$
$$= p^{-1}(\mathbf{Y}) \int g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = p^{-1}(\mathbf{Y}). \tag{10.23}$$

Thus, the marginal likelihood function $p(Y)$ can be estimated by the following estimator:

$$\hat{p}(\mathbf{Y}) = \left(\frac{1}{R}\sum_{r=1}^{R} \frac{g(\boldsymbol{\theta}^r)}{p(\mathbf{Y}|\boldsymbol{\theta}^r)p(\boldsymbol{\theta}^r)}\right)^{-1}, \tag{10.24}$$

where $\{\boldsymbol{\theta}^r\}_{r=1}^R$ is a sequence of the posterior draws from $p(\boldsymbol{\theta}|\mathbf{Y})$. Under the condition that $g(\boldsymbol{\theta})/\left(p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)$ is bounded above over the support of the posterior distribution, it can be shown that this estimator is a simulation consistent estimator when $R$ goes to infinity (Geweke, 1999). To guarantee this boundedness condition, following Geweke (1999), we can consider a truncated multivariate normal density for $g$. Let $A = \{\boldsymbol{\theta} \in \mathbb{R}^p : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) < \chi^2_{\alpha,p}\}$ be the truncation set, where $\hat{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\Omega}}$ is the posterior covariance of $\boldsymbol{\theta}$, and $\chi^2_{\alpha,p}$ is the $(1-\alpha)$ quantile of the $\chi^2_p$ distribution. Then, $g$ takes the following form:

$$g(\boldsymbol{\theta}) = (1-\alpha)^{-1}(2\pi)^{-p/2}\left|\hat{\boldsymbol{\Omega}}\right|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right) \times \mathbf{1}_A(\boldsymbol{\theta}), \tag{10.25}$$

where $\mathbf{1}_A(\boldsymbol{\theta})$ is the indicator function taking value 1 if $\boldsymbol{\theta} \in A$, otherwise 0.

Note that the computation of the modified harmonic mean estimator requires the integrated likelihood function which is available for both homoskedastic and heteroskedastic models. In the context of spatial autoregressive models, Doğan (2023) investigates the finite sample performance of this estimator along with some other popular information criteria for both nested and non-nested model selection problems. His simulation results show that the modified harmonic mean estimator performs satisfactorily, and can be useful for the specification search exercises in spatial econometrics.

# 11 A Monte Carlo study

## 11.1 Design

In this section, we conduct a Monte Carlo study to investigate the finite sample properties of the estimators considered in Sections 4 through 7. To this end, we consider the following data generating process:

$$\mathbf{Y} = e^{-\lambda_0 \mathbf{W}}\mathbf{X}\boldsymbol{\beta}_0 + e^{-\lambda_0 \mathbf{W}}e^{-\rho_0 \mathbf{M}}\mathbf{V},$$

where $\mathbf{X}$ contains two explanatory variables with $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20})' = (1, 1)'$. The observations for the first explanatory variable are drawn independently from the standard normal distribution, while the observations for the second explanatory variable are drawn independently from Uniform$(0, \sqrt{12})$. The spatial parameters $\lambda_0$ and $\rho_0$ can take values from the set $\{(-2, -1), (-2, 1), (0.5, -1), (0.5, 1)\}$. For $\mathbf{V}$, in the homoskedastic scenario, the elements $v_i$ are i.i.d. draws from either (i) the standard normal distribution or (ii) the standardized chi-squared distribution with three degrees of freedom, i.e., $(\chi^2_3 - 3)/\sqrt{6}$. In the heteroskedastic scenario, we set $\mathbf{V} \sim N(\mathbf{0}, \text{Diag}(\gamma_1, \ldots, \gamma_n))$ with either (i) $\gamma_i = 2\vartheta_i/(\sum_{j=1}^n \vartheta_j/n)$, where $\vartheta_i$ is the number of neighbors for unit $i$ using the description of $\mathbf{W}_1$ below, or (ii) $\gamma_i = \exp(0.1 + 0.35X_{2i})$ and $X_{2i}$ is the $i$th element of $X_2$.

For the spatial weights matrix $\mathbf{W}$, we consider the interaction scenario described in Arraiz et al. (2010). To this end, let $n$ entities be distributed across four quadrants of a square grid in such a way that the number of entities in each quadrant can be arranged to allow for sparse or dense quadrants. The location of each entity across the grid is determined by the $xy$-coordinates on the grid. Let $\underline{c}$ and $\bar{c}$ be two integers. The entities in the northeast quadrant of the grid have discrete coordinates satisfying $(\underline{c}+1) \leq x \leq \bar{c}$ and $(\underline{c}+1) \leq y \leq \bar{c}$, with an increment value of 0.5. For the other quadrants, the location coordinates are integers satisfying $1 \leq x \leq \underline{c}$, $1 \leq y \leq \bar{c}$, and $1 \leq x \leq \bar{c}$, $1 \leq y \leq \underline{c}$. The distance $d_{ij}$ between any two entities $i$ and $j$, located respectively at $(x_1, y_1)$ and $(x_2, y_2)$, is measured by the Euclidean distance given by $d_{ij} = \left[(x_1 - x_2)^2 + (y_1 - y_2)^2\right]^{1/2}$. Then, the $(i, j)$th element of $\mathbf{W}$ is set to 1 if $0 \leq d_{ij} \leq 1$, and to 0 otherwise. We then row normalize $\mathbf{W}$. In this scenario, varying the values for $\underline{c}$ and $\bar{c}$ leads to a different sample size and a different share of units in the northeast quadrant. We consider the following two combinations: $(\underline{c}, \bar{c}) = (5, 15)$ and $(\underline{c}, \bar{c}) = (14, 20)$. The first combination produces a sample size of 486 and locates 75 percent of the entities in the northeast quadrant ($\mathbf{W}_1$), whereas the second combination generates a sample size of 485 and locates 25 percent of the entities in the northeast quadrant ($\mathbf{W}_2$).

For the spatial weights matrix $\mathbf{M}$, we consider a nearest neighbors scheme. To this end, using the Euclidean distances ($d_{ij}$'s) from the construction of $\mathbf{W}$ above, we let entity $i$ be dependent on its 5 nearest neighbors so that the weights corresponding to these neighbors in the $i$'th row of $\mathbf{M}$ are set to 1 and the rest are set to zero. Then, $\mathbf{M}$ is row normalized. We use the "makeneighborsw" function from the Spatial Econometrics Toolbox to generate $\mathbf{M}$ (LeSage and Pace, 2009).

We evaluate the performance of the following estimators: (i) the QMLE in (4.4), (ii) the ME in (5.8), (iii) the IGMME in (6.1), (iv) the BGMME in (6.4), (v) the RGMME in (6.8), (vi) the Bayesian estimator (BE) based on Algorithm 1, and (vii) the robust Bayesian estimator (RBE) based on Algorithm 2. For classical estimation methods, we conduct 1000 repetitions. In the case of Bayesian estimation, we choose the following priors: $\lambda \sim N(0, 100)$, $\rho \sim N(0, 100)$, $\boldsymbol{\beta} \sim N(\mathbf{0}, 100\mathbf{I}_2)$ and $\sigma^2 \sim \text{IG}(0.01, 0.01)$. We set the number of repetitions to 100, the number of draws to 1500, and the

burn-ins to 500. For each method, we report bias, root mean squared error (RMSE) and empirical coverage ratio of a 95% confidence interval.[10]

### 11.2 Simulation results

Tables 6 and 7 present the simulation results for two homoskedastic cases: (i) $v_i \sim N(0,1)$ and (ii) $v_i \sim (\chi^2_3 - 3)/\sqrt{6}$. Similarly, Tables 8 and 9 report the simulation results for the heteroskedastic cases. Below, we summarize our main findings from these tables.

1. The results in Tables 6 and 7 demonstrate that all estimators exhibit excellent finite sample performance in terms of bias across all cases. All estimators report negligible bias for all parameters. For instance, in Table 6, when $(\alpha_0, \tau_0, \beta_{10}, \beta_{20})$ is $(-2, -1, 1, 1)$ in the case of $\mathbf{W}_1$, in terms of bias in estimating $\alpha_0$, the QMLE, IGMME, BGMME, ME and BE report $-0.0023$, $-0.0009$, $-0.0026$, $-0.0021$, and $-0.0024$, respectively.

2. In Tables 6 and 7, in terms of finite sample efficiency, the QMLE, BGMME and BE outperform the other estimators and report smaller RMSE when the true disturbance terms are normally distributed. However, when the true disturbance terms are not normally distributed, we observe that the BGMME reports the smallest RMSE. This is not surprising because the theoretical results in Debarsy et al. (2015) show that when the disturbance terms are not normally distributed and $\mathbf{W}$ and $\mathbf{M}$ do not commute, the BGMME can be more efficient than the QMLE. For example, in Table 7, when $(\alpha_0, \tau_0, \beta_{10}, \beta_{20})$ is $(-2, -1, 1, 1)$ in the case of $\mathbf{W}_2$, for $\alpha_0$ the BGMME reports 0.045 for RMSE, whereas the QMLE, IGMME, ME and BE report 0.051, 0.060, 0.052 and 0.061, respectively.

3. In Tables 6 and 7, in terms of finite sample coverage ratios, all estimators perform satisfactorily regardless of the distribution of the true disturbance terms or the denseness of $\mathbf{W}$. There are occasional negligible under coverage cases for the ME and the BE for $\alpha_0$. This is not surprising in the case of ME because it uses the adjusted quasi score (with respect to $\alpha$) that tries to correct the score for the potential heteroskedasticity in the disturbance terms. For example, in Table 6, when $(\alpha_0, \tau_0, \beta_{10}, \beta_{20})$ is $(-2, -1, 1, 1)$ in the case of $\mathbf{W}_1$, for $\alpha_0$, the QMLE, IGMME, BGMME, ME and BE report 94.5%, 93.4%, 94.6%, 92.7%, and 95%, respectively. Overall, all estimators perform satisfactorily.

4. In the heteroskedastic cases, the results in Tables 8 and 9 indicate that all estimators exhibit excellent finite sample performance in terms of bias. For example, in Table 9, when $(\alpha_0, \tau_0, \beta_{10}, \beta_{20})$ is $(0.5, -1, 1, 1)$ in the case of $\mathbf{W}_2$, in terms of bias in estimating $\tau_0$, the QMLE, IGMME, RGMME, ME and BE report 0.0015, 0.0074, 0.0013, 0.0021, and 0.0092, respectively.

[10] An estimation routine written in MATLAB is available in the supplementary online appendix.

5. In Tables 8 and 9, in terms of finite sample efficiency, the QMLE, RGMME, ME and RBE perform similarly. The RGMME and RBE report smaller RMSE values, whereas the IGMME reports the largest RMSE values. For example, in Table 9, when $(\alpha_0, \tau_0, \beta_{10}, \beta_{20})$ is $(0.5, -1, 1, 1)$ in the case of $\mathbf{W}_1$, for $\tau_0$ the RGMME and RBE report 0.094 and 0.091 for RMSE, respectively, whereas the QMLE, IGMME, and ME report 0.093, 0.107 and 0.093, respectively.

6. In Tables 8 and 9, in terms of finite sample coverage ratio, all estimators perform satisfactorily. For example, in Table 9, when $(\alpha_0, \tau_0, \beta_{10}, \beta_{20})$ is $(0.5, 1, 1, 1)$ in the case of $\mathbf{W}_1$, for $\tau_0$, the QMLE, IGMME, RGMME, ME and RBE report 95.0%, 94.6%, 94.9%, 96.8%, and 90.0%, respectively.

Table 6: Estimation results under homoskedasticity with $v_i \sim N(0,1)$

| | QMLE | IGMME | BGMME | ME | BE |
|---|---|---|---|---|---|
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | −0.0023(.043)[.945] | −0.0009(.054)[.934] | −0.0026(.043)[.946] | −0.0021(.045)[.927] | −0.0024(.045)[.950] |
| $\tau_0 = -1$ | 0.0015(.089)[.941] | 0.0041(.100)[.938] | 0.0013(.089)[.939] | 0.0013(.090)[.941] | 0.0141(.089)[.930] |
| $\beta_{10} = 1$ | 0.0034(.040)[.955] | 0.0034(.042)[.960] | 0.0031(.040)[.953] | 0.0030(.040)[.909] | 0.0036(.042)[.970] |
| $\beta_{20} = 1$ | −0.0007(.035)[.943] | −0.0011(.037)[.947] | −0.0009(.035)[.942] | 0.0010(.035)[.976] | 0.0002(.035)[.940] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | −0.0006(.038)[.942] | −0.0005(.044)[.949] | −0.0002(.039)[.939] | −0.0012(.048)[.908] | 0.0037(.036)[.970] |
| $\tau_0 = 1$ | 0.0140(.088)[.942] | 0.0094(.097)[.941] | 0.0115(.088)[.941] | 0.0141(.091)[.954] | 0.0066(.080)[.960] |
| $\beta_{10} = 1$ | 0.0012(.040)[.946] | 0.0010(.043)[.945] | 0.0011(.040)[.946] | 0.0020(.041)[.972] | 0.0039(.040)[.960] |
| $\beta_{20} = 1$ | 0.0000(.038)[.944] | 0.0004(.044)[.938] | 0.0004(.038)[.939] | 0.0000(.047)[.940] | 0.0054(.036)[.960] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | −0.0002(.048)[.939] | 0.0022(.061)[.948] | −0.0004(.048)[.942] | −0.0001(.049)[.883] | 0.0021(.048)[.960] |
| $\tau_0 = -1$ | 0.0032(.092)[.936] | 0.0037(.106)[.949] | 0.0028(.092)[.931] | 0.0031(.092)[.940] | −0.0032(.093)[.940] |
| $\beta_{10} = 1$ | 0.0020(.041)[.944] | 0.0014(.043)[.939] | 0.0017(.041)[.944] | 0.0024(.041)[.890] | −0.0001(.046)[.950] |
| $\beta_{20} = 1$ | 0.0000(.034)[.954] | 0.0001(.034)[.955] | −0.0002(.034)[.951] | 0.0020(.034)[.995] | −0.0026(.034)[.940] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | −0.0009(.041)[.954] | −0.0005(.047)[.951] | −0.0005(.041)[.955] | −0.0034(.054)[.891] | 0.0037(.040)[.930] |
| $\tau_0 = 1$ | 0.0147(.089)[.940] | 0.0094(.096)[.953] | 0.0121(.089)[.942] | 0.0168(.097)[.957] | −0.0073(.078)[.980] |
| $\beta_{10} = 1$ | −0.0017(.042)[.934] | −0.0016(.045)[.936] | −0.0020(.043)[.933] | −0.0006(.043)[.971] | 0.0011(.041)[.950] |
| $\beta_{20} = 1$ | −0.0006(.040)[.947] | 0.0000(.046)[.951] | −0.0002(.040)[.946] | −0.0003(.052)[.955] | 0.0060(.040)[.930] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | −0.0001(.051)[.951] | 0.0003(.061)[.952] | 0.0000(.051)[.952] | −0.0004(.052)[.894] | 0.0057(.053)[.960] |
| $\tau_0 = -1$ | 0.0074(.083)[.944] | 0.0111(.093)[.950] | 0.0065(.084)[.943] | 0.0077(.083)[.948] | 0.0068(.073)[.980] |
| $\beta_{10} = 1$ | 0.0034(.044)[.948] | 0.0036(.046)[.950] | 0.0031(.044)[.948] | 0.0036(.044)[.874] | −0.0049(.044)[.940] |
| $\beta_{20} = 1$ | −0.0006(.034)[.943] | −0.0004(.035)[.951] | −0.0013(.034)[.944] | 0.0005(.034)[.997] | −0.0002(.031)[.970] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | −0.0013(.038)[.959] | −0.0016(.045)[.952] | −0.0012(.038)[.958] | −0.0037(.054)[.845] | 0.0010(.044)[.900] |
| $\tau_0 = 1$ | 0.0143(.083)[.941] | 0.0096(.087)[.950] | 0.0106(.082)[.940] | 0.0143(.086)[.956] | −0.0219(.083)[.950] |
| $\beta_{10} = 1$ | 0.0008(.036)[.948] | 0.0007(.039)[.949] | 0.0003(.036)[.949] | 0.0015(.038)[.985] | −0.0015(.040)[.950] |
| $\beta_{20} = 1$ | −0.0006(.037)[.955] | −0.0006(.044)[.954] | −0.0006(.037)[.957] | 0.0001(.051)[.953] | 0.0011(.043)[.900] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | 0.0025(.050)[.953] | 0.0025(.062)[.933] | 0.0027(.050)[.952] | 0.0025(.052)[.845] | 0.0122(.056)[.880] |
| $\tau_0 = -1$ | −0.0030(.086)[.941] | 0.0033(.096)[.939] | −0.0036(.087)[.937] | −0.0031(.087)[.933] | −0.0014(.089)[.930] |
| $\beta_{10} = 1$ | 0.0001(.038)[.963] | −0.0006(.041)[.947] | −0.0003(.038)[.962] | 0.0008(.038)[.933] | 0.0046(.041)[.930] |
| $\beta_{20} = 1$ | −0.0025(.035)[.951] | −0.0024(.036)[.941] | −0.0030(.035)[.947] | −0.0015(.035)[.998] | 0.0035(.028)[.970] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | −0.0005(.038)[.949] | −0.0008(.044)[.948] | −0.0005(.038)[.948] | 0.0010(.049)[.869] | −0.0010(.040)[.910] |
| $\tau_0 = 1$ | 0.0099(.080)[.950] | 0.0056(.085)[.957] | 0.0063(.080)[.950] | 0.0082(.082)[.960] | −0.0023(.085)[.950] |
| $\beta_{10} = 1$ | 0.0001(.037)[.945] | 0.0001(.040)[.940] | −0.0004(.037)[.943] | 0.0006(.038)[.981] | −0.0036(.034)[.980] |
| $\beta_{20} = 1$ | −0.0001(.038)[.944] | −0.0002(.044)[.949] | −0.0001(.038)[.943] | −0.0002(.049)[.942] | 0.0010(.040)[.900] |

*Notes:* We report the bias (RMSE) [95% coverage ratio].

Table 7: Estimation results under homoskedasticity with $v_i \sim (\chi^2_3 - 3)/\sqrt{6}$

| | QMLE | IGMME | BGMME | ME | BE |
|---|---|---|---|---|---|
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | 0.0012(.045)[.933] | 0.0009(.054)[.946] | −0.0024(.036)[.945] | 0.0010(.045)[.906] | −0.0006(.040)[.950] |
| $\tau_0 = -1$ | 0.0005(.088)[.944] | 0.0037(.099)[.942] | 0.0036(.086)[.939] | 0.0009(.088)[.944] | 0.0037(.074)[.990] |
| $\beta_{10} = 1$ | −0.0016(.043)[.944] | −0.0017(.045)[.943] | −0.0004(.032)[.948] | −0.0008(.043)[.890] | 0.0018(.041)[.950] |
| $\beta_{20} = 1$ | −0.0001(.035)[.948] | 0.0003(.037)[.945] | 0.0006(.027)[.961] | 0.0018(.035)[.981] | 0.0011(.036)[.930] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | 0.0016(.038)[.946] | 0.0020(.043)[.943] | 0.0002(.029)[.948] | 0.0007(.046)[.910] | 0.0022(.041)[.930] |
| $\tau_0 = 1$ | 0.0071(.088)[.936] | 0.0013(.094)[.940] | 0.0057(.086)[.927] | 0.0079(.091)[.943] | −0.0006(.081)[.960] |
| $\beta_{10} = 1$ | 0.0013(.040)[.941] | 0.0016(.042)[.937] | 0.0002(.028)[.957] | 0.0013(.041)[.968] | −0.0027(.043)[.910] |
| $\beta_{20} = 1$ | 0.0027(.038)[.953] | 0.0033(.042)[.951] | 0.0015(.029)[.946] | 0.0006(.044)[.951] | 0.0034(.041)[.940] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | 0.0024(.049)[.947] | 0.0027(.061)[.941] | −0.0003(.041)[.946] | 0.0025(.050)[.891] | 0.0020(.044)[.970] |
| $\tau_0 = -1$ | 0.0027(.092)[.942] | 0.0061(.106)[.953] | 0.0047(.089)[.937] | 0.0027(.092)[.942] | −0.0051(.082)[.950] |
| $\beta_{10} = 1$ | 0.0017(.041)[.941] | 0.0017(.043)[.932] | 0.0005(.030)[.946] | 0.0026(.041)[.890] | −0.0010(.046)[.950] |
| $\beta_{20} = 1$ | 0.0011(.034)[.934] | 0.0010(.035)[.936] | 0.0017(.027)[.951] | 0.0031(.034)[.990] | −0.0046(.036)[.930] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | 0.0009(.042)[.940] | 0.0016(.046)[.945] | −0.0009(.031)[.949] | −0.0022(.055)[.881] | −0.0039(.040)[.940] |
| $\tau_0 = 1$ | 0.0121(.087)[.945] | 0.0050(.096)[.944] | 0.0114(.083)[.948] | 0.0149(.094)[.950] | 0.0010(.089)[.920] |
| $\beta_{10} = 1$ | −0.0003(.042)[.941] | 0.0000(.044)[.943] | −0.0008(.031)[.939] | 0.0003(.043)[.976] | −0.0051(.039)[.930] |
| $\beta_{20} = 1$ | 0.0019(.041)[.944] | 0.0027(.045)[.949] | 0.0002(.030)[.946] | 0.0003(.052)[.948] | −0.0038(.040)[.950] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | 0.0036(.051)[.949] | 0.0025(.060)[.954] | −0.0011(.045)[.946] | 0.0036(.052)[.888] | −0.0059(.061)[.900] |
| $\tau_0 = -1$ | 0.0006(.084)[.939] | 0.0069(.093)[.948] | 0.0028(.083)[.939] | 0.0008(.085)[.941] | −0.0052(.080)[.980] |
| $\beta_{10} = 1$ | −0.0017(.042)[.948] | −0.0017(.044)[.943] | −0.0018(.032)[.956] | −0.0011(.042)[.883] | 0.0014(.041)[.930] |
| $\beta_{20} = 1$ | −0.0013(.033)[.952] | −0.0015(.034)[.953] | −0.0016(.025)[.949] | −0.0002(.033)[.999] | −0.0041(.036)[.950] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | 0.0009(.039)[.939] | 0.0015(.045)[.945] | −0.0001(.030)[.951] | −0.0008(.053)[.847] | −0.0072(.036)[.940] |
| $\tau_0 = 1$ | 0.0102(.075)[.958] | 0.0050(.082)[.959] | 0.0047(.075)[.960] | 0.0102(.079)[.963] | 0.0029(.083)[.920] |
| $\beta_{10} = 1$ | −0.0011(.036)[.943] | −0.0016(.039)[.953] | −0.0010(.026)[.952] | −0.0011(.039)[.983] | −0.0025(.039)[.920] |
| $\beta_{20} = 1$ | 0.0014(.038)[.945] | 0.0023(.044)[.950] | 0.0007(.029)[.949] | 0.0000(.050)[.952] | −0.0057(.036)[.950] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | 0.0019(.050)[.947] | 0.0012(.061)[.949] | −0.0007(.045)[.943] | 0.0023(.051)[.857] | −0.0002(.054)[.930] |
| $\tau_0 = -1$ | 0.0006(.085)[.946] | 0.0077(.094)[.941] | 0.0011(.085)[.936] | 0.0004(.085)[.945] | 0.0099(.091)[.940] |
| $\beta_{10} = 1$ | −0.0033(.037)[.950] | −0.0017(.040)[.945] | −0.0026(.029)[.941] | −0.0027(.038)[.922] | −0.0022(.043)[.940] |
| $\beta_{20} = 1$ | −0.0004(.035)[.959] | −0.0008(.036)[.951] | 0.0003(.026)[.957] | 0.0006(.035)[.997] | −0.0004(.032)[.950] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | 0.0006(.039)[.946] | −0.0006(.045)[.945] | 0.0009(.029)[.949] | −0.0016(.050)[.863] | 0.0124(.048)[.850] |
| $\tau_0 = 1$ | 0.0142(.082)[.940] | 0.0114(.088)[.940] | 0.0093(.081)[.940] | 0.0149(.084)[.943] | 0.0038(.089)[.930] |
| $\beta_{10} = 1$ | −0.0012(.035)[.953] | −0.0022(.038)[.952] | −0.0007(.026)[.958] | −0.0010(.037)[.986] | 0.0090(.041)[.870] |
| $\beta_{20} = 1$ | 0.0008(.039)[.948] | 0.0000(.045)[.947] | 0.0011(.028)[.957] | −0.0004(.049)[.952] | 0.0138(.048)[.880] |

*Notes:* We report the bias (RMSE) [95% coverage ratio].

Table 8: Estimation results under heteroskedasticity with $\gamma_i = 2\vartheta_i/(\sum_{j=1}^{n} \vartheta_j/n)$

| | QMLE | IGMME | RGMME | ME | RBE |
|---|---|---|---|---|---|
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | −0.0014(.057)[.950] | 0.0004(.074)[.945] | −0.0022(.057)[.953] | −0.0011(.059)[.948] | −0.0049(.049)[.960] |
| $\tau_0 = -1$ | 0.0001(.104)[.911] | 0.0043(.112)[.945] | −0.0017(.111)[.893] | −0.0001(.105)[.942] | −0.0012(.103)[.920] |
| $\beta_{10} = 1$ | 0.0044(.057)[.952] | 0.0046(.059)[.954] | 0.0037(.057)[.953] | 0.0044(.052)[.933] | 0.0013(.043)[.960] |
| $\beta_{20} = 1$ | −0.0016(.049)[.943] | −0.0017(.051)[.953] | −0.0028(.049)[.944] | 0.0014(.049)[.970] | 0.0118(.048)[.970] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | 0.0024(.052)[.947] | 0.0000(.057)[.957] | 0.0008(.052)[.948] | −0.0021(.064)[.913] | 0.0104(.059)[.820] |
| $\tau_0 = 1$ | 0.0097(.096)[.938] | 0.0097(.096)[.960] | 0.0068(.101)[.923] | 0.0131(.100)[.962] | −0.0037(.097)[.900] |
| $\beta_{10} = 1$ | 0.0015(.054)[.951] | −0.0002(.056)[.950] | 0.0000(.054)[.947] | 0.0011(.047)[.992] | −0.0035(.042)[.930] |
| $\beta_{20} = 1$ | 0.0035(.052)[.945] | 0.0015(.057)[.961] | 0.0019(.052)[.944] | 0.0001(.013)[.999] | 0.0116(.060)[.840] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | −0.0036(.056)[.965] | −0.0004(.071)[.955] | −0.0042(.056)[.963] | −0.0039(.058)[.958] | 0.0005(.050)[.980] |
| $\tau_0 = -1$ | 0.0071(.104)[.931] | 0.0074(.108)[.962] | 0.0073(.111)[.909] | 0.0075(.104)[.946] | −0.0026(.086)[.970] |
| $\beta_{10} = 1$ | 0.0010(.057)[.948] | 0.0010(.059)[.948] | 0.0003(.057)[.945] | 0.0009(.056)[.891] | 0.0027(.041)[.960] |
| $\beta_{20} = 1$ | −0.0019(.044)[.972] | −0.0016(.045)[.971] | −0.0033(.044)[.970] | 0.0011(.044)[.989] | −0.0061(.050)[.970] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | 0.0016(.058)[.937] | −0.0006(.062)[.963] | 0.0002(.058)[.936] | −0.0111(.095)[.892] | 0.0056(.047)[.900] |
| $\tau_0 = 1$ | 0.0075(.103)[.924] | 0.0060(.102)[.962] | 0.0058(.109)[.907] | 0.0174(.125)[.976] | −0.0143(.098)[.890] |
| $\beta_{10} = 1$ | 0.0002(.060)[.935] | −0.0004(.063)[.939] | −0.0012(.061)[.930] | 0.0002(.054)[.993] | 0.0016(.041)[.920] |
| $\beta_{20} = 1$ | 0.0030(.059)[.937] | 0.0012(.064)[.956] | 0.0016(.059)[.935] | 0.0000(.014)[.999] | 0.0072(.047)[.900] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | 0.0033(.068)[.953] | 0.0027(.091)[.941] | 0.0031(.067)[.954] | 0.0049(.069)[.883] | −0.0096(.061)[.999] |
| $\tau_0 = -1$ | 0.0018(.094)[.935] | 0.0093(.105)[.956] | −0.0015(.095)[.930] | 0.0008(.094)[.945] | −0.0129(.090)[.930] |
| $\beta_{10} = 1$ | 0.0016(.054)[.942] | 0.0019(.057)[.948] | 0.0011(.055)[.945] | 0.0023(.053)[.945] | 0.0051(.039)[.980] |
| $\beta_{20} = 1$ | −0.0020(.048)[.955] | −0.0022(.052)[.947] | −0.0034(.049)[.954] | 0.0000(.048)[.995] | −0.0063(.045)[.960] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | 0.0001(.052)[.958] | −0.0028(.061)[.960] | −0.0029(.053)[.955] | −0.0062(.074)[.900] | −0.0017(.054)[.920] |
| $\tau_0 = 1$ | 0.0117(.091)[.928] | 0.0100(.095)[.946] | 0.0078(.092)[.927] | 0.0147(.097)[.948] | −0.0012(.089)[.940] |
| $\beta_{10} = 1$ | 0.0023(.059)[.939] | 0.0013(.062)[.956] | 0.0010(.060)[.940] | 0.0028(.058)[.992] | 0.0001(.038)[.980] |
| $\beta_{20} = 1$ | 0.0012(.052)[.957] | −0.0011(.061)[.960] | −0.0018(.052)[.954] | −0.0001(.013)[.999] | 0.0009(.054)[.890] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | 0.0069(.074)[.951] | 0.0082(.094)[.940] | 0.0074(.074)[.947] | 0.0086(.076)[.843] | 0.0086(.059)[.980] |
| $\tau_0 = -1$ | −0.0007(.099)[.923] | 0.0043(.105)[.949] | −0.0036(.100)[.918] | −0.0016(.099)[.943] | 0.0074(.089)[.930] |
| $\beta_{10} = 1$ | −0.0024(.056)[.943] | −0.0017(.059)[.955] | −0.0028(.056)[.945] | −0.0011(.055)[.926] | 0.0055(.036)[.970] |
| $\beta_{20} = 1$ | 0.0004(.050)[.954] | 0.0014(.051)[.948] | −0.0008(.050)[.951] | 0.0021(.049)[.999] | −0.0061(.049)[.930] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | 0.0030(.056)[.942] | −0.0039(.065)[.954] | −0.0004(.057)[.945] | 0.0009(.073)[.871] | 0.0068(.059)[.880] |
| $\tau_0 = 1$ | 0.0048(.088)[.947] | 0.0069(.092)[.955] | 0.0009(.089)[.947] | 0.0053(.092)[.963] | 0.0079(.084)[.930] |
| $\beta_{10} = 1$ | 0.0037(.055)[.944] | 0.0030(.058)[.944] | 0.0022(.055)[.942] | 0.0033(.052)[.992] | 0.0019(.034)[.990] |
| $\beta_{20} = 1$ | 0.0042(.056)[.935] | −0.0017(.062)[.951] | 0.0010(.056)[.936] | 0.0001(.013)[.999] | 0.0102(.054)[.880] |

*Notes:* We report the bias (RMSE) [95% coverage ratio].

Table 9: Estimation results under heteroskedasticity with $\gamma_i = \exp(0.1 + 0.35X_{2i})$

| | QMLE | IGMME | RGMME | ME | RBE |
|---|---|---|---|---|---|
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | −0.0023(.050)[.942] | −0.0006(.060)[.958] | −0.0025(.050)[.941] | −0.0019(.051)[.941] | 0.0062(.040)[.960] |
| $\tau_0 = -1$ | 0.0035(.092)[.952] | 0.0056(.106)[.952] | 0.0028(.093)[.950] | 0.0031(.093)[.955] | −0.0031(.094)[.910] |
| $\beta_{10} = 1$ | −0.0001(.050)[.942] | −0.0009(.052)[.947] | −0.0005(.050)[.942] | −0.0004(.047)[.933] | −0.0062(.032)[.990] |
| $\beta_{20} = 1$ | 0.0004(.045)[.933] | 0.0005(.047)[.925] | 0.0000(.045)[.932] | 0.0025(.045)[.969] | 0.0031(.045)[.920] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = -2$ | 0.0008(.046)[.954] | 0.0002(.050)[.955] | 0.0012(.046)[.956] | −0.0016(.056)[.933] | −0.0003(.043)[.940] |
| $\tau_0 = 1$ | 0.0087(.093)[.925] | 0.0063(.098)[.938] | 0.0060(.093)[.932] | 0.0107(.100)[.940] | −0.0071(.078)[.950] |
| $\beta_{10} = 1$ | 0.0000(.048)[.942] | −0.0004(.050)[.945] | 0.0001(.048)[.941] | 0.0005(.041)[.993] | −0.0039(.036)[.920] |
| $\beta_{20} = 1$ | 0.0017(.047)[.945] | 0.0013(.051)[.949] | 0.0021(.047)[.944] | 0.0000(.011)[1.000] | 0.0008(.044)[.950] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | 0.0029(.049)[.960] | 0.0023(.059)[.970] | 0.0028(.049)[.963] | 0.0026(.050)[.948] | −0.0027(.054)[.890] |
| $\tau_0 = -1$ | 0.0015(.093)[.942] | 0.0061(.107)[.948] | 0.0014(.094)[.943] | 0.0019(.093)[.947] | 0.0035(.091)[.930] |
| $\beta_{10} = 1$ | −0.0021(.047)[.958] | −0.0017(.049)[.958] | −0.0027(.047)[.957] | −0.0011(.046)[.924] | −0.0011(.038)[.970] |
| $\beta_{20} = 1$ | 0.0011(.042)[.943] | 0.0012(.043)[.939] | 0.0007(.042)[.944] | 0.0031(.042)[.982] | 0.0061(.043)[.980] |
| | | | $\mathbf{W}_1$ | | |
| $\alpha_0 = 0.5$ | −0.0008(.048)[.948] | −0.0018(.055)[.953] | −0.0003(.048)[.952] | −0.0029(.069)[.897] | −0.0097(.046)[.930] |
| $\tau_0 = 1$ | 0.0087(.090)[.950] | 0.0056(.099)[.946] | 0.0064(.090)[.949] | 0.0102(.102)[.968] | 0.0050(.092)[.900] |
| $\beta_{10} = 1$ | 0.0006(.049)[.956] | 0.0003(.051)[.951] | 0.0006(.049)[.953] | 0.0015(.044)[.994] | −0.0019(.037)[.930] |
| $\beta_{20} = 1$ | 0.0001(.051)[.942] | −0.0006(.057)[.950] | 0.0007(.051)[.940] | −0.0001(.013)[1.000] | −0.0066(.043)[.950] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | 0.0006(.059)[.961] | 0.0008(.073)[.968] | 0.0001(.059)[.962] | 0.0009(.061)[.858] | −0.0029(.064)[.910] |
| $\tau_0 = -1$ | 0.0058(.084)[.953] | 0.0116(.098)[.953] | 0.0054(.084)[.956] | 0.0057(.085)[.962] | −0.0072(.098)[.930] |
| $\beta_{10} = 1$ | 0.0023(.045)[.952] | 0.0025(.047)[.961] | 0.0017(.045)[.951] | 0.0027(.044)[.955] | 0.0062(.034)[.980] |
| $\beta_{20} = 1$ | −0.0003(.047)[.929] | −0.0004(.049)[.934] | −0.0011(.047)[.931] | 0.0013(.047)[.989] | 0.0085(.044)[.950] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = -2$ | −0.0025(.047)[.957] | −0.0029(.055)[.957] | −0.0022(.047)[.954] | −0.0045(.062)[.890] | 0.0039(.044)[.940] |
| $\tau_0 = 1$ | 0.0130(.084)[.943] | 0.0090(.090)[.948] | 0.0093(.083)[.949] | 0.0135(.088)[.948] | −0.0092(.087)[.920] |
| $\beta_{10} = 1$ | −0.0019(.048)[.955] | −0.0031(.051)[.955] | −0.0023(.048)[.953] | −0.0009(.047)[.993] | −0.0045(.036)[.930] |
| $\beta_{20} = 1$ | −0.0010(.048)[.937] | −0.0010(.056)[.947] | −0.0008(.048)[.941] | 0.0004(.011)[1.000] | 0.0056(.046)[.930] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | −0.0005(.059)[.964] | −0.0014(.073)[.961] | −0.0013(.059)[.964] | −0.0013(.061)[.874] | 0.0059(.057)[.950] |
| $\tau_0 = -1$ | 0.0015(.087)[.949] | 0.0074(.098)[.950] | 0.0013(.088)[.946] | 0.0021(.088)[.945] | 0.0092(.093)[.940] |
| $\beta_{10} = 1$ | −0.0017(.048)[.957] | −0.0011(.051)[.953] | −0.0019(.048)[.958] | −0.0015(.047)[.932] | 0.0003(.036)[.940] |
| $\beta_{20} = 1$ | 0.0010(.043)[.946] | 0.0008(.045)[.945] | 0.0001(.043)[.947] | 0.0029(.043)[.995] | 0.0003(.050)[.900] |
| | | | $\mathbf{W}_2$ | | |
| $\alpha_0 = 0.5$ | −0.0012(.048)[.948] | −0.0022(.057)[.953] | −0.0015(.048)[.954] | −0.0037(.068)[.846] | −0.0058(.043)[.940] |
| $\tau_0 = 1$ | 0.0098(.082)[.948] | 0.0056(.087)[.961] | 0.0066(.082)[.951] | 0.0109(.088)[.963] | 0.0054(.078)[.970] |
| $\beta_{10} = 1$ | −0.0004(.046)[.945] | 0.0008(.048)[.955] | −0.0008(.046)[.947] | 0.0005(.043)[.994] | 0.0004(.032)[.970] |
| $\beta_{20} = 1$ | −0.0004(.048)[.948] | −0.0009(.055)[.949] | −0.0008(.047)[.947] | −0.0003(.012)[1.000] | −0.0045(.043)[.910] |

*Notes:* We report the bias (RMSE) [95% coverage ratio].

## 12 Conclusion and outlook

In this article, we provided an extensive review of cross-sectional MESS models. We mainly focused on the first-order MESS model to discuss specification, estimation, model selection, and interpretation issues. The primary characteristic of a MESS-type model lies in its use of matrix exponential terms to specify spatial dependence in both the dependent variable and the disturbance terms. These models possess several distinctive properties.

- The power series representation of a matrix exponential term indicates an exponential decay of spatial dependence in these models.
- The reduced forms of MESS-type models always exist and do not require any restrictions on the parameter space of spatial parameters. The reduced forms of these models imply an exponential decay for the influence of high-order neighboring characteristics.
- The likelihood functions of these types of models are free of any Jacobian terms that must be computed at each iteration during the estimation process.
- When the spatial weights matrices are commutative, the QMLEs of these types of models can be consistent under an unknown form of heteroskedasticity.
- When the spatial weights matrices are not commutative, the QMLE can be inconsistent under an unknown form of heteroskedasticity. In such cases, a heteroskedasticity-robust estimation is required.
- The MESS and SAR models are not perfect substitutes because these two classes of models are non-nested. In practice, non-nested model selection procedures such as the J-test statistic, the Vuong-type model selection statistic, or Bayesian methods based on the marginal likelihood functions should be used for model selection exercises.

We provided a comprehensive description of various estimation methods, including the QML approach, the M-estimation approach, the GMM approach, and the Bayesian estimation approach. This detailed overview may enable practitioners to select and adapt a method that aligns with their specific needs. Additionally, we addressed estimation in the presence of endogenous explanatory variables and Durbin terms. We also discussed model selection methods based on testing, information criteria, and marginal likelihood approaches.

In future studies, it might be interesting to consider the MESS in a social interactions scenario, and compare its implications with the SAR-type social interaction models. As the QMLE of the MESS can still remain consistent under an unknown form of heteroskedasticity, allowing for such heteroskedasticity in a social interactions model would be a significant contribution. The performance

of model selection procedures such as the J-test statistic, the Vuong-type model selection statistic, the Cox test statistics, and Bayesian methods based on the marginal likelihood functions should be assessed in future studies for non-nested model selection problems between the MESS and SAR models through both simulation studies and empirical applications. We also think that the literature on nonlinear spatial models, such as the spatial extensions of the limited dependent variable data models, still holds some open questions, and estimation strategies for the the MESS-type limited dependent variable data models must be studied carefully. Finally, although the matrix-vector product approach to compute the matrix exponential terms can reduce the computation time significantly, we think that a faster and more reliable computation approach would be a significant contribution to the literature.

# Appendix

## A Useful Lemmas

In this section, we collect some lemmas that are required for our asymptotic results in Theorems 5.1-5.3 on the M-estimator. Lemma 1 can be found in Kelejian and Prucha (1999) and Lee (2002). The homoskedastic and heteroskedastic versions of Lemma 2 can be found in Lee (2007) and Lin and Lee (2010), respectively. Lemma 3 can be found in Debarsy et al. (2015), Lemma 4 gives a CLT result from Kelejian and Prucha (2010), and Lemma 5 is a modified version of Lemma A.4 in Yang (2018).

**Lemma 1.** *Let* $\{\mathbf{A}\}$ *and* $\{\mathbf{B}\}$ *be two sequences of* $n\times n$ *matrices that are uniformly bounded in both row sum and column sum matrix norms. Let* $\{\mathbf{C}\}$ *be a sequence of conformable matrices whose elements are uniformly* $O(h_n^{-1})$*, where the rate sequence* $\{h_n\}$ *can be bounded or divergent. Then,*

*(a) the sequence* $\{\mathbf{AB}\}$ *are uniformly bounded in both row sum and column sum matrix norms,*

*(b) the elements of* $\{\mathbf{A}\}$ *are uniformly bounded and* $tr(\mathbf{A}) = O(n)$*, and*

*(c) the elements of* $\{\mathbf{AC}\}$ *and* $\{\mathbf{CA}\}$ *are uniformly* $O(h_n^{-1})$*.*

**Lemma 2.** *Let* $\{\mathbf{V}\}$ *be a sequence of random* $n\times 1$ *column vectors,* $\boldsymbol{c}$ *be the* $n\times 1$ *vector of constants, and* $\{\mathbf{A}\}$ *and* $\{\mathbf{B}\}$ *be two sequences of* $n\times n$ *matrices of constants. Let* $\mathrm{vec}_D(\mathbf{A})$ *be a column vector formed by the diagonal elements of* $\mathbf{A}$*, and* $\mathbf{A}^s = \mathbf{A} + \mathbf{A}'$*.*

1. ***Homoskedastic case****: Suppose that the elements of* $\mathbf{V}$ *satisfy* $v_i \sim i.i.d.(0,\sigma_0^2)$*. Let* $\mathrm{E}(v_i^3) = \mu_3$ *and* $\mathrm{E}(v_i^4) = \mu_4$*. Then, we have the following results:*

   *(a)* $\mathrm{E}(\mathbf{AV}\cdot\mathbf{V}'\mathbf{BV}) = \mu_3\mathbf{A}vec_D(\mathbf{B})$*,*

   *(b)* $\mathrm{E}(\mathbf{V}'\mathbf{AV}\cdot\mathbf{V}'\mathbf{B}) = \mu_3 vec_D'(\mathbf{A})\mathbf{B}$*, and*

   *(c)* $\mathrm{E}(\mathbf{V}'\mathbf{AV}\cdot\mathbf{V}'\mathbf{BV}) = \left(\mu_4 - 3\sigma_0^4\right)\mathrm{vec}_D'(\mathbf{A})\,\mathrm{vec}_D(\mathbf{B}) + \sigma_0^4\left(\mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B}) + \mathrm{tr}(\mathbf{AB}^s)\right)$*.*

2. ***Heteroskedastic case****: Suppose* $\mathbf{V}$ *has elements that are independently distributed (i.n.i.d.) with* $v_i \sim i.n.i.d.(0,\sigma_i^2)$*. Let* $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_1^2,\ldots,\sigma_n^2)$*. Then, we have the following results:*

   *(a)* $\mathrm{E}(\mathbf{V}'\mathbf{AV}) = \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{A})$*,*

   *(b)* $\mathrm{E}(\mathbf{V}'\mathbf{AV}\cdot\boldsymbol{c}'\mathbf{V}) = \sum_{i=1}^n \mathrm{E}(v_i^3)a_{ii}c_i$*, where* $a_{ii}$ *is the* $(i,i)$*th element of* $\mathbf{A}$ *and* $c_i$ *is the* $i$*th element of* $\boldsymbol{c}$*, and*

   *(c)* $\mathrm{E}(\mathbf{V}'\mathbf{AV}\cdot\mathbf{V}'\mathbf{BV}) = \sum_{i=1}^n a_{ii}b_{ii}(\mathrm{E}(v_i^4) - 3\sigma_i^4) + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{A})\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{B}) + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^s)$*.*

   *If the diagonal elements of* $\mathbf{A}$ *are zeros, then these results take the following form:*

*(a)* $\mathrm{E}(\mathbf{V}'\mathbf{A}\mathbf{V}) = 0$,

*(b)* $\mathrm{E}(\mathbf{V}'\mathbf{A}\mathbf{V} \cdot \boldsymbol{c}'\mathbf{V}) = 0$, *and*

*(c)* $\mathrm{E}(\mathbf{V}'\mathbf{A}\mathbf{V} \cdot \mathbf{V}'\mathbf{B}\mathbf{V}) = \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^s)$.

**Lemma 3.** *Let* $\mathbf{A}$ *be any* $n \times n$ *matrix that is uniformly bounded in row sum and column sum matrix norms, and let* $a = o_p(1)$. *Then,* $\left\|e^{a\mathbf{A}} - \mathbf{I}_n\right\|_\infty = o_p(1)$ *and* $\left\|e^{a\mathbf{A}} - \mathbf{I}_n\right\|_1 = o_p(1)$, *where* $\|\cdot\|_\infty$ *denotes the row sum matrix norm, and* $\|\cdot\|_1$ *denotes the column sum matrix norm.*

**Lemma 4.** *Suppose that* $\{\mathbf{A}\}$ *is a sequence of* $n \times n$ *matrices uniformly bounded in both row sum and column sum matrix norms,* $\{\boldsymbol{c}\}$ *is a sequence of constant column vectors such that* $\sup_n \frac{1}{n}\sum_{i=1}^n |c_i|^{2+\eta_1} < \infty$ *for some* $\eta_1 > 0$, $v_i$ *in* $\mathbf{V}$ *are independent random variables with mean zero and variance* $\sigma_i^2$ *and* $\sup_i \mathrm{E}(|v_i|^{4+\eta_2}) < \infty$ *for some* $\eta_2 > 0$. *Denote* $\sigma_Z^2 = \mathrm{Var}\,(Z)$, *where* $Z = \boldsymbol{c}'\mathbf{V} + \mathbf{V}'\mathbf{A}\mathbf{V} - \mathrm{tr}\,(\mathbf{A}\boldsymbol{\Sigma})$ *where* $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_n^2)$. *Assume that* $\frac{1}{n}\sigma_Z^2$ *is bounded away from zero. Then* $\frac{Z}{\sigma_Z^2} \xrightarrow{d} N(0,1)$.

**Lemma 5.** *Let* $\{\mathbf{A}\}$ *be a sequence of* $n \times n$ *matrices that are bounded in both row sum and column sum matrix norms. Suppose also that the elements of* $\mathbf{A}$ *are* $O\left(h_n^{-1}\right)$, *uniformly. Let* $\boldsymbol{c}$ *be an* $n \times 1$ *vector with elements of the uniform order* $O(h_n^{-1/2})$. *Assume that the elements of the* $n \times 1$ *innovation vector* $\mathbf{V}$ *have zero mean and finite variance, and are mutually independent. Then,*

*(1)* $\mathrm{E}(\mathbf{V}'\mathbf{A}\mathbf{V}) = O\,(n/h_n)$, $\mathrm{Var}(\mathbf{V}'\mathbf{A}\mathbf{V}) = O\,(n/h_n)$,

*(2)* $\mathbf{V}'\mathbf{A}\mathbf{V} = O_p\,(n/h_n)$, $\mathbf{V}'\mathbf{A}\mathbf{V} - \mathrm{E}(\mathbf{V}'\mathbf{A}\mathbf{V}) = O_p\big((n/h_n)^{1/2}\big)$, *and* $\boldsymbol{c}'\mathbf{A}\mathbf{V} = O_p\big((n/h_n)^{1/2}\big)$.

# B Assumptions for the M-Estimator

To investigate the asymptotic properties of $\hat{\boldsymbol{\zeta}}_M$ stated in Theorems 5.1-5.3, we maintain the following assumptions.

**Assumption 3.** *The spatial weights matrices* $\mathbf{W}$ *and* $\mathbf{M}$ *are uniformly bounded in both row sum and column sum matrix norms.*

**Assumption 4.** *There exists a constant* $c > 0$ *such that* $|\lambda| \leq c$ *and* $|\rho| \leq c$, *and the true parameter vector* $\boldsymbol{\zeta}_0$ *lies in the interior of* $\boldsymbol{\Delta} = [-c, c] \times [-c, c]$.

**Assumption 5.** $\mathbf{X}$ *is exogenous, with uniformly bounded elements, and has full column rank. Also,* $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}'e^{\rho\mathbf{M}'}e^{\rho\mathbf{M}}\mathbf{X}$ *exists and is nonsingular, uniformly in* $\rho \in [-c, c]$.

**Assumption 6.** $\inf_{\boldsymbol{\zeta}:\, d(\boldsymbol{\zeta},\boldsymbol{\zeta}_0)\geq\vartheta} \left\|\bar{S}^{*c}(\boldsymbol{\zeta})\right\| > 0$ *for every* $\vartheta > 0$, *where* $d(\boldsymbol{\zeta}, \boldsymbol{\zeta}_0)$ *is a measure of distance between* $\boldsymbol{\zeta}$ *and* $\boldsymbol{\zeta}_0$.

Assumption 3 provides the essential properties of the spatial weights matrices. It ensures that the spatial correlation is limited to a manageable degree (Kelejian and Prucha, 2001, 2010). Assumption 4 requires that the parameter space of the parameters in the matrix exponential terms is compact. Assumption 3 and Assumption 4 imply that the matrix exponential terms are uniformly bounded in both row sum and column sum matrix norms. This can be seen from $\left\|e^{\lambda \mathbf{W}}\right\| = \left\|\sum_{i=0}^{\infty} \lambda^i \mathbf{W}^i / i!\right\| \leq \sum_{i=0}^{\infty} |\lambda|^i \|\mathbf{W}\|^i / i! = e^{|\lambda| \|\mathbf{W}\|}$, which is bounded if $|\lambda|$ and $\|\mathbf{W}\|$ are bounded, where $\|\cdot\|$ is either the row sum or the column sum matrix norm. Assumption 5 provides some regularity conditions and corresponds to Assumption 4 of Debarsy et al. (2015). Assumption 6 is a high-level assumption and ensures the identification of $\boldsymbol{\zeta}_0$. In Appendix C, we provide two low-level conditions that are sufficient for Assumption 6.

# C Proofs of Main Results

In this section, we provide the proofs of the main theorems in Section 5 of the paper.

## C.1 Proof of Theorem 5.1

As discussed in the main paper, we only need to show the consistency of $\hat{\boldsymbol{\zeta}}$. To that end, given Assumption 5, we need to show that $\sup_{\boldsymbol{\zeta} \in \boldsymbol{\Delta}} \frac{1}{n} \left\|S^{*c}(\boldsymbol{\zeta}) - \bar{S}^{*c}(\boldsymbol{\zeta})\right\| \xrightarrow{p} 0$. Note that

$$\hat{\mathbf{V}}(\boldsymbol{\zeta}) = \mathbf{V}(\hat{\boldsymbol{\beta}}_M, \boldsymbol{\zeta}) = e^{\rho \mathbf{M}}(e^{\lambda \mathbf{W}} \mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}_M) = \mathbf{Q}(\rho) e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}} \mathbf{Y}, \tag{C.1}$$

and

$$\bar{\mathbf{V}}(\boldsymbol{\zeta}) = \mathbf{V}(\bar{\boldsymbol{\beta}}_M, \boldsymbol{\zeta}) = e^{\rho \mathbf{M}}(e^{\lambda \mathbf{W}} \mathbf{Y} - \mathbf{X} \bar{\boldsymbol{\beta}}_M) = \mathbf{Q}(\rho) e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}} \mathbf{Y} + \mathbf{P}(\rho) e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}} (\mathbf{Y} - \mathrm{E}(\mathbf{Y})), \tag{C.2}$$

where $\mathbf{P}(\rho)$ is the projection matrix based on $e^{\rho \mathbf{M}} \mathbf{X}$ and $\mathbf{Q}(\rho) = \mathbf{I}_n - \mathbf{P}(\rho)$. Denote $\mathbf{G}(\boldsymbol{\zeta}) = e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}}$ and $\mathbf{G}(\boldsymbol{\zeta}_0) = e^{\rho_0 \mathbf{M}} e^{\lambda_0 \mathbf{W}}$. Substituting (C.1) and (C.2) into $S^{*c}(\boldsymbol{\zeta})$ and $\bar{S}^{*c}(\boldsymbol{\zeta})$, the proof of $\sup_{\boldsymbol{\zeta} \in \boldsymbol{\Delta}} \frac{1}{n} \left\|S^{c*}(\boldsymbol{\zeta}) - \bar{S}^{c*}(\boldsymbol{\zeta})\right\| \xrightarrow{p} 0$ is equivalent to that of the following:

(i) $\sup_{\boldsymbol{\zeta} \in \boldsymbol{\Delta}} \frac{1}{n} \left(\mathbf{Y}' \mathbf{R}_i(\boldsymbol{\zeta}) \mathbf{Y} - \mathrm{E}\left(\mathbf{Y}' \mathbf{R}_i(\boldsymbol{\zeta}) \mathbf{Y}\right)\right) = o_p(1)$, for $i = 1, 2$,

(ii) $\sup_{\boldsymbol{\zeta} \in \boldsymbol{\Delta}} \frac{1}{n} \mathrm{tr}\left(\boldsymbol{\Sigma} \mathbf{G}^{-1'}(\boldsymbol{\zeta}_0) \mathbf{T}_i(\boldsymbol{\zeta}) \mathbf{G}^{-1}(\boldsymbol{\zeta}_0)\right) = o(1)$, for $i = 1, 2, 3$,

where the terms are defined as $\mathbf{R}_1(\boldsymbol{\zeta}) = \mathbf{G}'(\boldsymbol{\zeta}) \mathbb{W}_D(\rho) \mathbf{Q}(\rho) \mathbf{G}(\boldsymbol{\zeta})$, $\mathbf{R}_2(\boldsymbol{\zeta}) = \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{Q}(\rho) \mathbf{M} \mathbf{Q}(\rho) \mathbf{G}(\boldsymbol{\zeta})$, $\mathbf{T}_1(\boldsymbol{\zeta}) = \mathbf{G}'(\boldsymbol{\zeta}) \mathbb{W}_D(\rho) \mathbf{P}(\rho) \mathbf{G}(\boldsymbol{\zeta})$, $\mathbf{T}_2(\boldsymbol{\zeta}) = \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{P}(\rho) \mathbf{G}(\boldsymbol{\zeta})$ and $\mathbf{T}_3(\boldsymbol{\zeta}) = \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{Q}(\rho) \mathbf{M}^s \mathbf{P}(\rho) \mathbf{G}(\boldsymbol{\zeta})$.

**Proof of (i).** Note that (i) follows from the point-wise convergence of $\frac{1}{n}\left(\mathbf{Y}' \mathbf{R}_i(\boldsymbol{\zeta}) \mathbf{Y} - \mathrm{E}(\mathbf{Y}' \mathbf{R}_i(\boldsymbol{\zeta}) \mathbf{Y})\right)$ in each $\boldsymbol{\zeta} \in \boldsymbol{\Delta}$ and stochastic equicontinuity of $\frac{1}{n} \mathbf{Y}' \mathbf{R}_i(\boldsymbol{\zeta}) \mathbf{Y}$ for $i = 1, 2$. To prove the point-wise

convergence, we have

$$\frac{1}{n}\left(\mathbf{Y}^{'}\mathbf{R}_i(\boldsymbol{\zeta})\mathbf{Y}-\mathrm{E}(\mathbf{Y}^{'}\mathbf{R}_i(\boldsymbol{\zeta})\mathbf{Y})\right)$$
$$=\frac{2}{n}\beta_0^{'}\mathbf{X}^{'}e^{\lambda_0\mathbf{W}^{'}}\mathbf{R}_i(\boldsymbol{\zeta})\mathbf{G}^{-1}(\boldsymbol{\zeta}_0)\mathbf{V}+\frac{1}{n}\left(\mathbf{V}^{'}\mathbf{G}^{-1^{'}}(\boldsymbol{\zeta}_0)\mathbf{R}_i(\boldsymbol{\zeta})\mathbf{G}^{-1}(\boldsymbol{\zeta}_0)\mathbf{V}-\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{G}^{-1^{'}}(\boldsymbol{\zeta}_0)\mathbf{R}_i(\boldsymbol{\zeta})\mathbf{G}^{-1}(\boldsymbol{\zeta}_0))\right).$$

Note that $e^{\lambda_0\mathbf{W}^{'}}\mathbf{R}_i(\boldsymbol{\zeta})\mathbf{G}^{-1}(\boldsymbol{\zeta}_0)$ and $\mathbf{G}^{-1^{'}}(\boldsymbol{\zeta}_0)\mathbf{R}_i(\boldsymbol{\zeta})\mathbf{G}^{-1}(\boldsymbol{\zeta}_0)$ are uniformly bounded in row and column sum norms for $i=1,2$ by Lemma 1. Thus, the terms on the r.h.s. are pointwise convergent by Lemma 5(2).

To prove the stochastic equicontinuity, by the mean value theorem, for any two parameter vectors $\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2\in\boldsymbol{\Delta}$, we have

$$\frac{1}{n}\left(\mathbf{Y}^{'}\mathbf{R}_i(\boldsymbol{\zeta}_1)\mathbf{Y}-\mathbf{Y}^{'}\mathbf{R}_i(\boldsymbol{\zeta}_2)\mathbf{Y}\right)=\frac{1}{n}\mathbf{Y}^{'}\frac{\partial\mathbf{R}_i(\bar{\boldsymbol{\zeta}})}{\partial\boldsymbol{\zeta}^{'}}\mathbf{Y}(\boldsymbol{\zeta}_1-\boldsymbol{\zeta}_2),$$

where $\bar{\boldsymbol{\zeta}}$ is between $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ elementwise. Thus we need to prove that $\sup_{\boldsymbol{\zeta}\in\boldsymbol{\Delta}}\frac{1}{n}\mathbf{Y}^{'}\frac{\partial\mathbf{R}_i(\boldsymbol{\zeta})}{\partial\lambda}\mathbf{Y}=O_p(1)$ and $\sup_{\boldsymbol{\zeta}\in\boldsymbol{\Delta}}\frac{1}{n}\mathbf{Y}^{'}\frac{\partial\mathbf{R}_i(\boldsymbol{\zeta})}{\partial\rho}\mathbf{Y}=O_p(1)$ for $i=1,2$. Note that

$$\begin{aligned}
\frac{\partial\mathbf{R}_1(\boldsymbol{\zeta})}{\partial\lambda}&=\mathbf{W}^{'}\mathbf{R}_1(\boldsymbol{\zeta})+\mathbf{R}_1(\boldsymbol{\zeta})\mathbf{W},\quad \frac{\partial\mathbf{R}_2(\boldsymbol{\zeta})}{\partial\lambda}=\mathbf{W}^{'}\mathbf{R}_2(\boldsymbol{\zeta})+\mathbf{R}_2(\boldsymbol{\zeta})\mathbf{W},\\
\frac{\partial\mathbf{R}_1(\boldsymbol{\zeta})}{\partial\rho}&=\mathbf{M}^{'}\mathbf{R}_1(\boldsymbol{\zeta})+\mathbf{G}^{'}(\boldsymbol{\zeta})\dot{\mathbb{W}}_D(\rho)\mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta})+\mathbf{G}^{'}(\boldsymbol{\zeta})\mathbb{W}_D(\rho)\dot{\mathbf{Q}}(\rho)\mathbf{G}(\boldsymbol{\zeta})+\mathbf{R}_1(\boldsymbol{\zeta})\mathbf{M},\\
\frac{\partial\mathbf{R}_2(\boldsymbol{\zeta})}{\partial\rho}&=\mathbf{M}^{'}\mathbf{R}_2(\boldsymbol{\zeta})+\mathbf{G}^{'}(\boldsymbol{\zeta})\dot{\mathbf{Q}}(\rho)\mathbf{M}\mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta})+\mathbf{G}^{'}(\boldsymbol{\zeta})\mathbf{Q}(\rho)\mathbf{M}\dot{\mathbf{Q}}(\rho)\mathbf{G}(\boldsymbol{\zeta})+\mathbf{R}_2(\boldsymbol{\zeta})\mathbf{M},
\end{aligned}$$

where $\dot{\mathbb{W}}_D(\rho)=\frac{\partial\mathbb{W}_D(\rho)}{\partial\rho}=\mathbf{M}\mathbb{W}_D(\rho)-\mathbb{W}_D(\rho)\mathbf{M}-\mathrm{Diag}\left(\mathbf{M}\mathbb{W}_D(\rho)-\mathbb{W}_D(\rho)\mathbf{M}\right)$ and $\dot{\mathbf{Q}}(\rho)=\frac{\partial\mathbf{Q}(\rho)}{\partial\rho}=-(\mathbf{Q}(\rho)\mathbf{M}\mathbf{P}(\rho)+\mathbf{P}(\rho)\mathbf{M}^{'}\mathbf{Q}(\rho))$. By substituting the reduced form $\mathbf{Y}=e^{-\lambda_0\mathbf{W}}(\mathbf{X}\boldsymbol{\beta}_0+e^{-\rho_0\mathbf{M}}\mathbf{V})$ into $\mathbf{Y}^{'}\frac{\partial\mathbf{R}_i(\boldsymbol{\zeta})}{\partial\lambda}\mathbf{Y}$ and $\mathbf{Y}^{'}\frac{\partial\mathbf{R}_i(\boldsymbol{\zeta})}{\partial\rho}\mathbf{Y}$ for $i=1,2$, we get a group of nonstochastic terms and linear and quadratic forms in $\mathbf{V}$. By Lemma 5, $\sup_{\boldsymbol{\zeta}\in\boldsymbol{\Delta}}\frac{1}{n}\mathbf{Y}^{'}\frac{\partial\mathbf{R}_i(\boldsymbol{\zeta})}{\partial\lambda}\mathbf{Y}=O_p(1)$ and $\sup_{\boldsymbol{\zeta}\in\boldsymbol{\Delta}}\frac{1}{n}\mathbf{Y}^{'}\frac{\partial\mathbf{R}_i(\boldsymbol{\zeta})}{\partial\rho}\mathbf{Y}=O_p(1)$ for $i=1,2$.

**Proof of (ii).** Under Assumption 5, Lemma 1 ensures that $\frac{1}{n}\mathrm{tr}\left(\boldsymbol{\Sigma}\mathbf{G}^{-1^{'}}(\boldsymbol{\zeta}_0)\mathbf{T}_i(\boldsymbol{\zeta})\mathbf{G}^{-1}(\boldsymbol{\zeta}_0)\right)=o(1)$, for $i=1,2,3$.

## C.2 Proof of Theorem 5.2

By the mean value theorem, $\sqrt{n}(\hat{\boldsymbol{\gamma}}_M-\boldsymbol{\gamma}_0)=-\left(\frac{1}{n}\frac{\partial S^*(\overline{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}^{'}}\right)^{-1}\frac{1}{\sqrt{n}}S^*(\boldsymbol{\gamma}_0)$, where $\overline{\boldsymbol{\gamma}}$ is between $\hat{\boldsymbol{\gamma}}_M$ and $\boldsymbol{\gamma}_0$ elementwise. Thus we need to prove:

(i) $\frac{1}{\sqrt{n}}S^*(\boldsymbol{\gamma}_0)\xrightarrow{d}N\left(0,\lim_{N\to\infty}\boldsymbol{\Omega}(\boldsymbol{\gamma}_0)\right)$,

(ii) $\frac{1}{n}\left(\frac{\partial S^*(\overline{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}^{'}}-\frac{\partial S^*(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^{'}}\right)=o_p(1)$, and

(iii) $\frac{1}{n}\left(\frac{\partial S^*(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^{'}}-\mathrm{E}\left(\frac{\partial S^*(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^{'}}\right)\right)=o_p(1)$.

**Proof of (i).** Note that the elements of $S^*(\boldsymbol{\gamma}_0)$ are linear-quadratic forms of $\mathbf{V}$ as shown in (5.10). Then we can construct an $(k+2)\times 1$ vector $\mathbf{a} = (\mathbf{a}_1', a_2, a_3)'$ such that $\mathbf{a}'S^*(\boldsymbol{\gamma}_0) = \mathbf{b}'\mathbf{V} + \mathbf{V}'\mathbf{B}\mathbf{V}$, where $\mathbf{b}' = \mathbf{a}_1'\mathbf{X}'e^{\rho_0\mathbf{M}'} - a_2\boldsymbol{\beta}_0'\mathbf{X}'e^{\rho_0\mathbf{M}'}\mathbb{W}_D$ and $\mathbf{B} = -a_2\mathbb{W}_D - a_3\mathbf{M}$. Then, $\frac{1}{n}\mathbf{a}'S^*(\boldsymbol{\gamma}_0)$ is asymptotically normal by Lemma 4. Then, the Cramer-Wold device leads to (i).

**Proof of (ii).** Let $\boldsymbol{\Pi}(\boldsymbol{\gamma}) = -\frac{1}{n}\frac{\partial S^*(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}$. Since $\mathbf{Y} = e^{-\lambda_0\mathbf{W}}(\mathbf{X}\boldsymbol{\beta}_0 + e^{-\rho_0\mathbf{M}}\mathbf{V})$, all terms in the Hessian matrix can be written in forms of functions in Lemma 5. By Lemma 5, we know that $\frac{1}{n}\boldsymbol{\Pi}(\boldsymbol{\gamma}_0) = O_p(1)$, which implies $\frac{1}{n}\boldsymbol{\Pi}(\bar{\boldsymbol{\gamma}}) = O_p(1)$. We can write $e^{\bar{\lambda}\mathbf{W}} = (e^{\bar{\lambda}\mathbf{W}} - e^{\lambda_0\mathbf{W}}) + e^{\lambda_0\mathbf{W}}$, $e^{\bar{\rho}\mathbf{M}} = (e^{\bar{\rho}\mathbf{M}} - e^{\rho_0\mathbf{M}}) + e^{\rho_0\mathbf{M}}$ and $\bar{\boldsymbol{\beta}} = (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \boldsymbol{\beta}_0$, and then expand the terms in $\frac{1}{n}\boldsymbol{\Pi}(\boldsymbol{\gamma})$. By Lemma 5 and the reduced form of $\mathbf{Y}$, $\frac{1}{n}\mathbf{Y}'\mathbf{A}\mathbf{Y} = O_p(1)$ and $\frac{1}{n}\mathbf{X}'\mathbf{A}\mathbf{Y} = O_p(1)$, where $A$ is an $n\times n$ matrix that is bounded in both row and column sum matrix norms. Also note $\left\|e^{\bar{\lambda}\mathbf{W}} - e^{\lambda\mathbf{W}}\right\|_\infty = \left\|(e^{(\bar{\lambda}-\lambda_0)\mathbf{W}} - I_n)e^{\lambda_0\mathbf{W}}\right\|_\infty \leq \left\|(e^{(\bar{\lambda}-\lambda_0)\mathbf{W}} - I_n\right\|_\infty \left\|e^{\lambda_0\mathbf{W}}\right\|_\infty = o_p(1)$ by Lemma 3, and similarly $\left\|e^{\bar{\rho}\mathbf{M}} - e^{\rho_0\mathbf{M}}\right\|_\infty = o_p(1)$. Then the expanded forms of $\frac{1}{n}\left(\boldsymbol{\Pi}(\bar{\boldsymbol{\gamma}}) - \boldsymbol{\Pi}(\boldsymbol{\gamma}_0)\right)$ imply that it is $o_p(1)$.

**Proof of (iii).** Substituting the reduced form of $\mathbf{Y}$ into $\frac{1}{n}\left(\frac{\partial S^*(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'} - \mathrm{E}\left(\frac{\partial S^*(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}\right)\right)$, we know that each element is a linear or quadratic function of $\mathbf{V}$. For example, for $\boldsymbol{\Pi}^*_{\lambda\rho}(\boldsymbol{\gamma}_0)$,

$$\begin{aligned}
&\boldsymbol{\Pi}^*_{\lambda\rho}(\boldsymbol{\gamma}_0) - \mathrm{E}\left(\boldsymbol{\Pi}^*_{\lambda\rho}(\boldsymbol{\gamma}_0)\right) = -\boldsymbol{\beta}_0'\mathbf{X}'e^{\rho_0\mathbf{M}'}\mathbf{M}'\mathbb{W}_D(\rho)\mathbf{V} - \mathbf{V}'\mathbf{M}'\mathbb{W}_D(\rho)\mathbf{V} + \mathrm{E}\left(\mathbf{V}'\mathbf{M}'\mathbb{W}_D(\rho)\mathbf{V}\right)\\
&- \boldsymbol{\beta}_0'\mathbf{X}'e^{\rho_0\mathbf{M}'}\dot{\mathbb{W}}_D(\rho)\mathbf{V} - \mathbf{V}'\dot{\mathbb{W}}_D(\rho)\mathbf{V} + \mathrm{E}\left(\mathbf{V}'\mathbf{M}'\dot{\mathbb{W}}_D(\rho)\mathbf{V}\right) - \boldsymbol{\beta}_0'\mathbf{X}'e^{\rho_0\mathbf{M}'}\mathbb{W}_D(\rho)\mathbf{M}\mathbf{V}\\
&- \mathbf{V}'\mathbb{W}_D(\rho)\mathbf{M}\mathbf{V} + \mathrm{E}\left(\mathbf{V}'\mathbb{W}_D(\rho)\mathbf{M}\mathbf{V}\right).
\end{aligned}$$

By Lemma 5, $\frac{1}{n}\left(\boldsymbol{\Pi}^*_{\lambda\rho}(\boldsymbol{\gamma}_0) - \mathrm{E}\left(\boldsymbol{\Pi}^*_{\lambda\rho}(\boldsymbol{\gamma}_0)\right)\right) = o_p(1)$. The proof for the rest of the terms are similar to that for $\boldsymbol{\Pi}^*_{\lambda\rho}(\boldsymbol{\gamma}_0)$ and thus are omitted.

### C.3 Proof of Theorem 5.3

Since the terms in $\boldsymbol{\Omega}(\boldsymbol{\gamma}_0)$ are similar to those in Proposition 5 in Debarsy et al. (2015), the proof is similar to that of Proposition 5 and thus is omitted.

## D Details of the Identification Conditions

The identification condition in Assumption 6 under the heteroskedastic error terms is a high level assumption. In this section, we derive low level conditions for the identification of $\boldsymbol{\zeta}_0$. Note that the identification of $\boldsymbol{\zeta}_0$ requires that $\bar{\mathbf{S}}^{c*}(\boldsymbol{\zeta}) \neq 0$ for $\boldsymbol{\zeta} \neq \boldsymbol{\zeta}_0$ under the exact identification case, similar to the method of moment approach. Also recall that the population counterpart of the concentrated

adjusted score function is given by

$$\bar{S}^{c*}(\boldsymbol{\zeta}) = \begin{cases} \lambda: & -\mathrm{E}\left(\mathbf{Y}' e^{\lambda \mathbf{W}'} e^{\rho \mathbf{M}'} \mathbb{W}_D(\rho) \bar{\mathbf{V}}(\boldsymbol{\zeta})\right), \\ \rho: & -\mathrm{E}\left(\bar{\mathbf{V}}'(\boldsymbol{\zeta}) \mathbf{M} \bar{\mathbf{V}}(\boldsymbol{\zeta})\right), \end{cases}$$

where $\bar{\mathbf{V}}(\boldsymbol{\zeta}) = \mathbf{V}(\bar{\boldsymbol{\beta}}_M(\boldsymbol{\zeta}), \boldsymbol{\zeta})$. Also recall that (C.2) implies that

$$\bar{\mathbf{V}}(\boldsymbol{\zeta}) = \mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta})\mathbf{Y} + \mathbf{P}(\rho)\mathbf{G}(\boldsymbol{\zeta})(\mathbf{Y} - \mathrm{E}(\mathbf{Y})),$$

where $\mathbf{G}(\boldsymbol{\zeta}) = e^{\rho \mathbf{M}} e^{\lambda \mathbf{W}}$. Denote $\mathbf{G} = \mathbf{G}(\boldsymbol{\zeta}_0)$. From the reduced form $\mathbf{Y} = e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{G}^{-1}\mathbf{V}$, we know $\mathbf{Y} - \mathrm{E}(\mathbf{Y}) = \mathbf{G}^{-1}\mathbf{V}$. Then,

$$\begin{aligned} \bar{\mathbf{V}}(\boldsymbol{\zeta}) &= \mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta})\mathbf{Y} + \mathbf{P}(\rho)\mathbf{G}(\boldsymbol{\zeta})\mathbf{G}^{-1}\mathbf{V} \\ &= \mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta})(e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{G}^{-1}\mathbf{V}) + \mathbf{P}(\rho)\mathbf{G}(\boldsymbol{\zeta})\mathbf{G}^{-1}\mathbf{V} \\ &= \mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{G}(\boldsymbol{\zeta})\mathbf{G}^{-1}\mathbf{V}. \end{aligned}$$

Then we have

$$\begin{aligned} &\mathrm{E}(\bar{\mathbf{V}}'(\boldsymbol{\zeta})\mathbb{W}_D(\rho)\bar{\mathbf{V}}(\boldsymbol{\zeta})) \\ &= \mathrm{E}\left((\mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{G}(\boldsymbol{\zeta})\mathbf{G}^{-1}\mathbf{V})' \mathbb{W}_D(\rho)(\mathbf{Q}(\rho)\mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{G}(\boldsymbol{\zeta})\mathbf{G}^{-1}\mathbf{V})\right) \\ &= \boldsymbol{\beta}_0' \mathbf{X}' e^{-\lambda_0 \mathbf{W}'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{Q}(\rho) \mathbb{W}_D \mathbf{Q}(\rho) \mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathrm{E}(\mathbf{V}' \mathbf{G}^{-1'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbb{W}_D(\rho) \mathbf{G}(\boldsymbol{\zeta}) \mathbf{G}^{-1} \mathbf{V}) \\ &= \boldsymbol{\beta}_0' \mathbf{X}' e^{-\lambda_0 \mathbf{W}'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{Q}(\rho) \mathbb{W}_D \mathbf{Q}(\rho) \mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathrm{tr}(\boldsymbol{\Sigma} \mathbf{G}^{-1'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbb{W}_D(\rho) \mathbf{G}(\boldsymbol{\zeta}) \mathbf{G}^{-1}). \end{aligned}$$

Similarly $\mathrm{E}(\bar{\mathbf{V}}'(\boldsymbol{\zeta})\mathbf{M}\bar{\mathbf{V}}(\boldsymbol{\zeta}))$ can be expressed as

$$\begin{aligned} &\mathrm{E}(\bar{\mathbf{V}}'(\boldsymbol{\zeta})\mathbf{M}(\rho)\bar{\mathbf{V}}(\boldsymbol{\zeta})) \\ &= \boldsymbol{\beta}_0' \mathbf{X}' e^{-\lambda_0 \mathbf{W}'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{Q}(\rho) \mathbf{M} \mathbf{Q}(\rho) \mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathrm{tr}(\boldsymbol{\Sigma} \mathbf{G}^{-1'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{M} \mathbf{G}(\boldsymbol{\zeta}) \mathbf{G}^{-1}). \end{aligned}$$

Thus, the identification of $\boldsymbol{\zeta}_0$ follows, if $\boldsymbol{\zeta} \neq \boldsymbol{\zeta}_0$, one of the following conditions holds:

(i) $\lim_{n\to\infty} \frac{1}{n}[\boldsymbol{\beta}_0' \mathbf{X}' e^{-\lambda_0 \mathbf{W}'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{Q}(\rho) \mathbb{W}_D \mathbf{Q}(\rho) \mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathrm{tr}(\boldsymbol{\Sigma} \mathbf{G}^{-1'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbb{W}_D(\rho) \mathbf{G}(\boldsymbol{\zeta}) \mathbf{G}^{-1})] \neq 0,$

(ii) $\lim_{n\to\infty} \frac{1}{n}[\boldsymbol{\beta}_0' \mathbf{X}' e^{-\lambda_0 \mathbf{W}'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{Q}(\rho) \mathbf{M} \mathbf{Q}(\rho) \mathbf{G}(\boldsymbol{\zeta}) e^{-\lambda_0 \mathbf{W}} \mathbf{X}\boldsymbol{\beta}_0 + \mathrm{tr}(\boldsymbol{\Sigma} \mathbf{G}^{-1'} \mathbf{G}'(\boldsymbol{\zeta}) \mathbf{M} \mathbf{G}(\boldsymbol{\zeta}) \mathbf{G}^{-1})] \neq 0.$

**Acknowledgements** Ye Yang gratefully acknowledges the financial support from the Beijing Natural Science Foundation (9242005) and the research fund for new professors at Capital University of Economics and Business (XRZ2023042).

## References


Anselin, L. (1984). Specification tests on the structure of interaction in spatial econometric models. *Papers of the Regional Science Association*, 54(1):165–182.

Anselin, L. (1986). Non-nested tests on the weight structure in spatial autoregressive models: Some monte carlo results. *Journal of Regional Science*, 26(2):267–284.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Springer, New York.

Anselin, L. (2001). Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference*, 97(1):113–139.

Anselin, L., Bera, A. K., Florax, R., and Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26(1):77–104.

Arbia, G., Bera, A. K., Doğan, O., and Taşpınar, S. (2020). Testing impact measures in spatial autoregressive models. *International Regional Science Review*, 43(1-2):40–75.

Arraiz, I., Drukker, D. M., Kelejian, H. H., and Prucha, I. R. (2010). A spatial Cliff-Ord-Type model with heteroskedastic innovations: Small and large sample results. *Journal of Regional Science*, 50(2):592–614.

Bera, A. K., Doğan, O., and Taşpınar, S. (2018). Simple tests for endogeneity of spatial weights matrices. *Regional Science and Urban Economics*, 69:130–142.

Bera, A. K., Doğan, O., and Taşpınar, S. (2019). Testing spatial dependence in spatial models with endogenous weights matrices. *Journal of Econometric Methods*, 8(1):20170015.

Blonigen, B. A., Davies, R. B., Waddell, G. R., and Naughton, H. T. (2007). Fdi in space: Spatial autoregressive relationships in foreign direct investment. *European Economic Review*, 51(5):1303–1325.

Burridge, P. (2012). Improving the J test in the SARAR model by likelihood-based estimation. *Spatial Economic Analysis*, 7(1):75–107.

Chan, J. C. C. and Grant, A. L. (2016). On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics*, 14(4):772–802.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

Cliff, A. D. and Ord, J. (1969). The problem of spatial autocorrelation. In Scott, A. J., editor, *London Papers in Regional Science 1 Studies in Regional Science*, pages 22–55. Pion, London.

Cliff, A. D. and Ord, J. (1973). *Spatial autocorrelation*. Pion, London.

Davidson, R. and MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, 49(3):781–793.

Debarsy, N., Jin, F., and Lee, L.-F. (2015). Large sample properties of the matrix exponential spatial specification with an application to FDI. *Journal of Econometrics*, 188(1):1–21.

Doğan, O. (2023). Modified harmonic mean method for spatial autoregressive models. *Economics Letters*, 223:110978.

Doğan, O., Taşpınar, S., and Bera, A. K. (2018). Simple tests for social interaction models with network structures. *Spatial Economic Analysis*, 13(2):212–246.

Doğan, O., Yang, Y., and Taşpınar, S. (2023). Information criteria for matrix exponential spatial specifications. *Spatial Statistics*, 57:100776.

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

Ertur, C. and Koch, W. (2007). Growth, technological interdependence and spatial externalities: theory and evidence. *Journal of Applied Econometrics*, 22(6):1033–1062.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Frühwirth-Schnatter, S. and Wagner, H. (2008). Marginal likelihoods for non-gaussian models using auxiliary mixture sampling. *Computational Statistics & Data Analysis*, 52(10):4608–4624.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, New York, third edition edition.

Geweke, J. (1999). Using simulation methods for bayesian econometric models: Inference, development,and communication. *Econometric Reviews*, 18(1):1–73.

Han, X. and Lee, L.-f. (2013a). Bayesian estimation and model selection for spatial Durbin error model with finite distributed lags. *Regional Science and Urban Economics*, 43(5):816–837.

Han, X. and Lee, L.-f. (2013b). Model selection using J-test for the spatial autoregressive model vs. the matrix exponential spatial model. *Regional Science and Urban Economics*, 43(2):250–271.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hepple, W. L. (1995). Bayesian techniques in spatial and network econometrics: 1. model comparison and posterior odds. *Environment and Planning A: Economy and Space*, 27(3):447–469.

Hsu, Y. and Shi, X. (2017). Model-selection tests for conditional moment restriction models. *The Econometrics Journal*, 20(1):52–85.

Jin, F. and Lee, L.-f. (2013). Cox-type tests for competing spatial autoregressive models with spatial autoregressive disturbances. *Regional Science and Urban Economics*, 43(4):590–616.

Jin, F. and Lee, L.-F. (2018). Irregular N2SLS and LASSO estimation of the matrix exponential spatial specification model. *Journal of Econometrics*, 206(2):336–358.

Jin, F. and Wang, Y. (2022). GMM estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. *Econometric Reviews*, 41(6):652–674.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kelejian, H. H. (2008). A spatial J-test for model specification against a single or a set of non-nested alternatives. *Letters in Spatial and Resource Sciences*, 1:3–11.

Kelejian, H. H. and Piras, G. (2011). An extension of kelejian's J-test for non-nested spatial models. *Regional Science and Urban Economics*, 41(3):281–292.

Kelejian, H. H. and Prucha, I. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, 17:99–121.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40(2):509–533.

Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distribution of the moran I test statistic with applications. *Journal of Econometrics*, 104(2):219–257.

Lee, L.-f. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory*, 18(2):252–277.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925.

Lee, L.-f. (2007). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514.

LeSage, J. and Chih, Y.-Y. (2018). A matrix exponential spatial panel model with heterogeneous coefficients. *Geographical Analysis*, 50(4):422–453.

Lesage, J. P. (1997). Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, 20(1-2):113–129.

LeSage, J. P. and Pace, R. K. (2007). A matrix exponential spatial specification. *Journal of Econometrics*, 140(1):190–214.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC, London.

LeSage, J. P. and Parent, O. (2007). Bayesian model averaging for spatial econometric models. *Geographical Analysis*, 39(3):241–267.

Li, Y., Yu, J., and Zeng, T. (2020). Deviance information criterion for latent variable models and misspecified models. *Journal of Econometrics*, 216(2):450–493.

Lin, X. and Lee, L.-f. (2010). GMM estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics*, 157(1):34–52.

Liu, T. and Lee, L.-f. (2019). A likelihood ratio test for spatial model selection. *Journal of Econometrics*, 213(2):434–458.

MacKinnon, J. G. (2009). Bootstrap hypothesis testing. In Belsley, D. A. and Kontoghiorghes, E. J., editors, *Handbook of Computational Econometrics*, pages 183–210. John Wiley & Sons Ltd, West Sussex.

Moler, C. and Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20(4):801–836.

Moler, C. and Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49.

Newey, W. K. and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 28(3):777–787.

Otto, P., Fassò, A., and Maranzano, P. (2024). A review of regularised estimation methods and cross-validation in spatiotemporal statistics. *https://arxiv.org/abs/2402.00183*.

Otto, P. and Sibbertsen, P. (2023). Spatial autoregressive fractionally integrated moving average model. *https://arxiv.org/abs/2309.06880*.

Pace, R. K. and Barry, R. (1997). Quick computation of spatial autoregressive estimators. *Geographical Analysis*, 29(3):232–247.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461 – 464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618.

White, H. (1994). *Estimation, inference and specification analysis*. Econometric Society Monographs.

White, H. G. (1980). A heteroskedasticity-consistent covariance matrix estimator a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41(3/4):434–449.

Yang, Y. (2022). Unified M-estimation of matrix exponential spatial dynamic panel specification. *Econometric Reviews*, 41(7):729–748.

Yang, Y., Doğan, O., and Taşpınar, S. (2021). Fast estimation of matrix exponential spatial models. *Journal of Spatial Econometrics*, 2(9).

Yang, Y., Doğan, O., and Taşpınar, S. (2022). Model selection and model averaging for matrix exponential spatial models. *Econometric Reviews*, 41(8):827–858.

Yang, Y., Doğan, O., and Taşpınar, S. (2024). Estimation of matrix exponential unbalanced panel data models with fixed effects: an application to US outward FDI stock. *Journal of Business & Economic Statistics*, 42(2):469–484.

Yang, Z. (2018). Unified M-estimation of fixed-effects spatial dynamic models with short panels. *Journal of Econometrics*, 205(2):423–447.

Zhang, Y., Feng, S., and Jin, F. (2019). QML estimation of the matrix exponential spatial specification panel data model with fixed effects and heteroskedasticity. *Economics Letters*, 180:1–5.