Spatial Weights Matrix Selection and Model Averaging for Spatial Autoregressive Models

Xinyu Zhang
Chinese Academy of Sciences
Beijing, China
xinyu@amss.ac.cn

Jihai Yu
Peking University
Beijing, China
jihai.yu@gmail.com

September 27, 2014

Abstract

Spatial econometrics relies on spatial weights matrix to specify the cross sectional dependence, which might not be unique. This paper proposes a model selection procedure to choose an optimal weights matrix from several candidates by using a Mallows type criterion. It also proposes a model averaging procedure to reduce the squared loss. We prove that these procedures are asymptotically optimal in the sense of minimizing the squared loss. Monte Carlo experiments show that proposed procedures have satisfactory finite sample performances. We apply the model selection and model averaging procedures to study the market integration in China using historical rice price.

JEL classification: C13; C23; R15
Keywords: Model Selection, Model Averaging, Spatial Econometrics, Spatial Autoregressive

1 Introduction

Spatial econometrics study autocorrelation among cross sectional units and have wide applications.[1] Among them, the spatial autoregressive (SAR) model proposed by Cliff and Ord (1973) has received the most attention. To estimate the SAR model, various estimation methods have been developed.[2] These estimation methods treat the spatial weights matrix as given and implicitly assume that they have the correct specification of the spatial weights matrix. However, currently there is no concise theory on how to specify the "correct" spatial weights matrix. The purpose of the current paper is to propose a model selection procedure to choose an optimal spatial weights matrix from several candidates, and also propose a model averaging procedure to reduce squared loss.

For selecting a "correct" spatial weights matrix, Kelejian (2008) suggested a J-test for testing a null SAR model against a set of alternative models with different spatial weights matrices. His J-test is based on whether or not predictions based on alternative models add significantly to the explanatory power of the null model. Kelejian and Piras (2011) suggest a modification of Kelejian's J-test which uses available information in a more efficient way. Both Kelejian (2008) and Kelejian and Piras (2011) have a shortcoming. If we reject the null model (using one of the spatial weights matrices) with more than one alternative specifications, there is no formal procedure proposed on selecting alternatives. In the current paper, we propose a model selection procedure where a Mallows type criterion (Mallows, 1973) is used. Commonly used information criteria such as AIC and BIC cannot be used to select spatial weights matrix, because in two models with different spatial weights matrices, the numbers of known parameters are the same. In the current paper, we propose a spatial weights matrix selection method. Unlike studies featuring the identification of weights matrix, our focus here is on the true conditional mean of dependent variables given independent variables. As a step further, we also propose a model averaging to reduce estimation

[1] Early development in estimation and testing can be found in Anselin (1988, 1992), Kelejian and Robinson (1993), Cressie (1993), Anselin and Florax (1995), Anselin and Rey (1997), and Anselin and Bera (1998), among others.

[2] Kelejian and Prucha (1998) proposed a two stage least squares (2SLS) method which uses instrumental variables (IVs) constructed from exogenous variables and the spatial weights matrix . Lee (2003) chose some specific IVs and got the best two stage least square (B2SLS) estimators. Lee (2004) studies the asymptotic properties of the maximum likelihood (ML) and quasi-maximum likelihood (QML) estimators of the SAR model. For the SAR model, the spatial correlation can provide nonlinear moment conditions in addition to linear moments of IV's in the general method of moment (GMM) setting. Lee (2007) established asymptotic properties of GMM estimators, which can be more efficient than the 2SLS estimators. The best GMM can be as efficient as the ML estimators (MLEs) when the true disturbances are normal. Liu et al. (2010) showed that carefully designed linear and quadratic moment functions can generate a GMM estimator which is more efficient relative to the QML estimate when the disturbances are not normal.

error, which makes use of all the information available. We show that the proposed selection and averaging procedures are asymptotically optimal in the sense of achieving the lowest possible squared loss in a class of model selection and model average estimators, respectively. We believe that our procedures are useful for empirical researchers who might have several spatial weights matrices available and do not have an explicit guide of which one to use.

The current paper is different from some other model selection problems in spatial econometrics, such as specifying spatial lag and spatial error models, or selecting SAR model or matrix exponential spatial specification (MESS). Debarsy and Ertur (2010) investigate how to distinguish two different spatial specifications in a panel data model setting, i.e., a spatial lag model or a spatial error model. They construct five different test procedures: one joint test, two conditional tests and two marginal tests, for a fixed effects panel model based on Lee and Yu (2010). Han and Lee (2013a, 2013b) consider J-test to distinguish SAR model and MESS model, where the former implies a geometrical decline pattern of spillover effect or externalities, while the latter exhibits an exponential decline pattern of spatial externalities. The current paper will mainly focus on selecting and averaging different spatial weights matrices.

The rest of the paper is organized as follows. Section 2 introduces the model and maximum likelihood estimation (MLE) of SAR models. Section 3 proposes the model selection procedure and its asymptotic optimality is proved. Section 4 studies the model averaging given a set of spatial weights matrices. Section 5 provides Monte Carlo experiments to investigate the finite sample properties of model selection and model averaging procedures. Section 6 investigates the spatial weights matrix selection in studying the market integration using historical data set on Chinese rice prices. Section 7 concludes. Proofs are in Appendices.

2 Model and Estimation

The model considered is a cross sectional SAR

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of dependent variables, $\mathbf{X}$ is an $n \times p$ nonstochastic exogenous variables, $\mathbf{W}$ is a nonstochastic spatial weights matrix and the disturbance ϵ_i, $i = 1, 2, \cdots, n$, of the n-dimensional vector $\boldsymbol{\epsilon}$ are *i.i.d.* $(0, \sigma^2)$. Here, $\mathbf{W}\mathbf{y}$ is usually referred to as a spatial lag of $\mathbf{y}$. We assume that

$$\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{I}_n). \tag{2}$$

The reduced form for the SAR model is

$$\mathbf{y} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}$$

which has a nonstochastic component $(\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}$ and a stochastic component $(\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}$. Thus, the expected value of $\mathbf{y}$ is

$$\boldsymbol{\mu} = E(\mathbf{y}) = (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}$$

and $\boldsymbol{\Omega} = \sigma^2(\mathbf{I}_n - \rho\mathbf{W})^{-1}(\mathbf{I}_n - \rho\mathbf{W}^{\mathrm{T}})^{-1}$ is the variance matrix of $\mathbf{y}$ given $\mathbf{X}$. Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ be the projection matrix and $\mathbf{A} = \mathbf{I}_n - \mathbf{P}$.

Assume we have a weights matrix set $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_S\}$, and thus S candidate models. The log likelihood function for (1) under the s^{th} model is

$$\log(\mathrm{LH}) = -\frac{n}{2}\log(2\pi\sigma^2) + \log|\mathbf{I}_n - \rho\mathbf{W}_s| - \frac{\|(\mathbf{I}_n - \rho\mathbf{W}_s)\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2}. \tag{3}$$

From first order conditions w.r.t. $\boldsymbol{\beta}$ and σ^2, given the $\widehat{\rho}_s$ as the MLE of ρ under the s^{th} candidate model, the MLEs of $\boldsymbol{\beta}$ and σ^2 are

$$\widehat{\boldsymbol{\beta}}_s = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y} \tag{4}$$

and

$$\widehat{\sigma}_s^2 = \|(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_s\|^2/n. \tag{5}$$

So the estimator of $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}}_s = (\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}}_s = (\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{P}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y} = \widetilde{\mathbf{P}}_s\mathbf{y},$$

where $\widetilde{\mathbf{P}}_s = (\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{P}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)$. The $\widehat{\boldsymbol{\mu}}_s$ under the s^{th} model will be used to evaluate the performance of model selection by the difference of $\widehat{\boldsymbol{\mu}}_s$ and the true $\boldsymbol{\mu}$.

For analytical purpose, it can be obtained that

$$\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s} = (\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{P}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s) - (\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{P}\mathbf{W}_s \tag{6}$$

and

$$\begin{aligned}\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}} &= \big[2n\widehat{\rho}_s\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y} - n(\mathbf{W}_s^{\mathrm{T}}\mathbf{A} + \mathbf{A}\mathbf{W}_s)\mathbf{y} \qquad (7)\\ &\quad +2\mathrm{trace}\{(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s\}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{\mathrm{T}}\mathbf{A}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y}\big]\end{aligned}$$

$$
\begin{aligned}
&\times \big[\|\mathbf{A}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y}\|^2 \mathrm{trace}\{(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s\} \\
&\quad - n\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y} \\
&\quad - \mathrm{trace}\{(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s\}(2\widehat{\rho}_s\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{y})\big]^{-1}.
\end{aligned}
$$

Formula (6) is easy to obtain. See Appendix A.1 for the derivation of (7). The (6) and (7) are useful in constructing the model selection criterion below.

3 Spatial Weights Matrix Selection Method and Its Asymptotic Optimality

Given the space of all possible spatial weights matrices, in this section we will select a model by some selection criterion related to the squared length of error vector. As seminal works on model selection and averaging such as Li (1987), Shao (1997) and Hansen (2007), our focus is to reduce the squared loss $L_s = \|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2$. The associated risk is $R_s = E(L_s) = E\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2$. Assuming $\boldsymbol{\Omega}$ is known, we define the following model selection criterion

$$C_s = \|\widetilde{\mathbf{P}}_s\mathbf{y} - \mathbf{y}\|^2 + 2\mathrm{trace}(\widetilde{\mathbf{P}}_s\boldsymbol{\Omega}) + 2\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\boldsymbol{\Omega}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}.$$

The first term of C_s measures the model fit, while the other terms measures the model complexity and serve as penalties. From the normality of $\boldsymbol{\epsilon}$ and Stein's Lemma 1 (Stein, 1981), we have

$$E(C_s) = R_s - \mathrm{trace}(\boldsymbol{\Omega}). \tag{8}$$

See Appendix A.2 for the derivation of (8).

In practice, $\boldsymbol{\Omega}$ is unknown. Let $\widehat{\boldsymbol{\Omega}}$ be an estimator of $\boldsymbol{\Omega}$. So a feasible selection criterion is

$$\widehat{C}_s = \|\widetilde{\mathbf{P}}_s\mathbf{y} - \mathbf{y}\|^2 + 2\mathrm{trace}(\widetilde{\mathbf{P}}_s\widehat{\boldsymbol{\Omega}}) + 2\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}. \tag{9}$$

The following Assumptions 2 and 4 contain some conditions on $\widehat{\boldsymbol{\Omega}}$, but the consistency of $\widehat{\boldsymbol{\Omega}}$ is unnecessary, so the model used to estimate $\boldsymbol{\Omega}$ can be misspecified. Following Hansen (2007) and Liu and Ryo (2013), we estimate $\boldsymbol{\Omega}$ using the largest model, which, in the current paper, means the model with the densest spatial weights matrix $\mathbf{W}_s$.

Define $\widehat{s} = \arg\min_{s\in\{1,\ldots,S\}}\widehat{C}_s$, which is the selected model. Next, we build the asymptotic optimality of this model selection procedure. Assume $\widehat{\rho}_s$ has a limit ρ_s^*. Let $\inf_s$ ($\sup_s$) indicate infimum (supremum) over $s \in \{1,\ldots,S\}$, $\overline{\mathbf{P}}_s = \widetilde{\mathbf{P}}_s \mid_{\widehat{\rho}_s=\rho_s^*}$, $R_s^* = E\|\widehat{\boldsymbol{\mu}}_s \mid_{\widehat{\rho}_s=\rho_s^*} - \boldsymbol{\mu}\|^2$, $\xi_n = \inf_s R_s^*$,

p be the number of columns of $\mathbf{X}$, and $\lambda_{\max}(\mathbf{A})$ denote the largest singular values of a matrix $\mathbf{A}$. All the limiting properties here and throughout the text hold under $n \to \infty$.

Assumption 1 *There exists a positive integer G such that $\sum_{s=1}^{S}(R_s^*)^{-G} = o(1)$.*

Assumption 2 $\|\boldsymbol{\mu}\|^2 = O(n)$*,* $\lambda_{\max}(\mathbf{\Omega}) = O(1)$*, and* $\lambda_{\max}(\widehat{\mathbf{\Omega}}) = O_p(1)$*.*

Assumption 3 $\sup_s \lambda_{\max}(\rho_s^* \mathbf{W}_s) = O(1)$ *and* $\sup_s \lambda_{\max}\{(\mathbf{I}_n - \rho_s^* \mathbf{W}_s)^{-1}\} = O(1)$*.*

Assumption 4 $\xi_n^{-1} \sup_s |(\partial\widehat{\rho}_s/\partial \mathbf{y}^{\mathrm{T}})\widehat{\mathbf{\Omega}}(\partial\widetilde{\mathbf{P}}_s/\partial\widehat{\rho}_s)\mathbf{y}| = o_p(1)$*.*

Assumption 5 $\xi_n^{-1}p = o(1)$ *and* $n\xi_n^{-1}\sup_s \lambda_{\max}(\widehat{\mathbf{P}}_s - \overline{\mathbf{P}}_s) = o_p(1)$*.*

Assumption 1 is commonly-used in literature on the optimality of model selection; see, for example, condition (A.3) of Li (1987) and condition (2.6) of Shao (1997). The first part of Assumptions 2, concerning the sum of n elements of $\boldsymbol{\mu}$, is also commonly-used in linear regression models (Liang et al., 2011). The other parts of Assumption 2 and Assumption 3 require the largest singular values of matrices to be bounded. Assumption 4 is a high level assumption. It is seen from (6) and (7) that the expression $(\partial\widehat{\rho}_s/\partial \mathbf{y}^{\mathrm{T}})\widehat{\mathbf{\Omega}}(\partial\widetilde{\mathbf{P}}_s/\partial\widehat{\rho}_s)\mathbf{y}$ will be tedious. We discuss the rationality of Assumption 4 in Appendix A.3. The first part of Assumption 5 allows the number of regressors increase with sample size n, but places a constraint on the growth rate of the number of regressors. Similar assumptions are condition (22) of Zhang et al. (2013) and condition 6 of Liu and Ryo (2013). The second part of Assumption 5 requires that $\widehat{\rho}_s$ converges to ρ_s^* at a rate such that $\xi_n^{-1}\sup_s \lambda_{\max}(\widehat{\mathbf{P}}_s - \overline{\mathbf{P}}_s)$ converges to 0 at a rate quicker than $n \to \infty$ and is similar to condition (A5) of Zhang et al. (2014).

Theorem 1 *Under Assumptions 1-5,*

$$\frac{L_{\widehat{s}}}{\inf_{s\in\{1,\ldots,S\}} L_s} \to 1 \tag{10}$$

in probability as $n \to \infty$.

The result (10) means that our selected estimator $\boldsymbol{\mu}_{\widehat{s}}$ is optimal in the sense that its squared loss is asymptotically identical to that by the infeasible best candidate estimator. Thus, in large sample sense, the model selection criterion (9) can successfully minimize the loss.

4 Model Averaging and Its Asymptotic Optimality

The model selection in Section 3 can help us to pick up a model and we can proceed estimation and testing based on this model. An alternative to model selection is model averaging. Rather than attaching to a single "winning" model, model averaging compromises across a set of candidate models, by which, it provides a kind of insurance against selecting a very poor model and can substantially reduce risk compared to model selection (see Leung and Barron, 2006 and Hansen, 2014).

Let the weight vector $\mathbf{w} = (w_1, ..., w_S)^{\mathrm{T}}$, belonging to the set $\mathcal{H} = \{\mathbf{w} \in [0,1]^S : \sum_{s=1}^S w_s = 1\}$, and $\widetilde{\mathbf{P}}(\mathbf{w}) = \sum_{s=1}^S w_s \widetilde{\mathbf{P}}_s$. The model average estimator of $\boldsymbol{\mu}$ would then be

$$\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum\nolimits_{s=1}^S w_s \widehat{\boldsymbol{\mu}}_s = \sum\nolimits_{s=1}^S w_s \widetilde{\mathbf{P}}_s \mathbf{y} = \widetilde{\mathbf{P}}(\mathbf{w})\mathbf{y}.$$

Let squared loss $L(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$ and associated risk $R(\mathbf{w}) = E\{L(\mathbf{w})\} = E\|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$. Assuming $\boldsymbol{\Omega}$ is known, we define the following weight choice criterion

$$C(\mathbf{w}) = \|\widetilde{\mathbf{P}}(\mathbf{w})\mathbf{y} - \mathbf{y}\|^2 + 2\mathrm{trace}\{\widetilde{\mathbf{P}}(\mathbf{w})\boldsymbol{\Omega}\} + 2\sum\nolimits_{s=1}^S w_s \frac{\partial \widehat{\rho}_s}{\partial \mathbf{y}^{\mathrm{T}}} \boldsymbol{\Omega} \frac{\partial \widetilde{\mathbf{P}}_s}{\partial \widehat{\rho}_s} \mathbf{y}.$$

Using the proof steps in Appendix A.2, we can show that $E\{C(\mathbf{w})\} = R(\mathbf{w}) - \mathrm{trace}(\boldsymbol{\Omega})$. Similar to the previous section, we use $\widehat{\boldsymbol{\Omega}}$ to estimate $\boldsymbol{\Omega}$, so that a feasible weight choice criterion is

$$\widehat{C}(\mathbf{w}) = \|\widetilde{\mathbf{P}}(\mathbf{w})\mathbf{y} - \mathbf{y}\|^2 + 2\mathrm{trace}\{\widetilde{\mathbf{P}}(\mathbf{w})\widehat{\boldsymbol{\Omega}}\} + 2\sum\nolimits_{s=1}^S w_s \frac{\partial \widehat{\rho}_s}{\partial \mathbf{y}^{\mathrm{T}}} \widehat{\boldsymbol{\Omega}} \frac{\partial \widetilde{\mathbf{P}}_s}{\partial \widehat{\rho}_s} \mathbf{y}. \tag{11}$$

The selected weight is then $\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w} \in \mathcal{H}} \widehat{C}(\mathbf{w})$.

Define $\mathbf{H} = (\widetilde{\mathbf{P}}_1\mathbf{y} - \mathbf{y}, \cdots, \widetilde{\mathbf{P}}_S\mathbf{y} - \mathbf{y})$ and

$$\mathbf{h} = \{\mathrm{trace}(\widetilde{\mathbf{P}}_1\widehat{\boldsymbol{\Omega}}) + \frac{\partial \widehat{\rho}_1}{\partial \mathbf{y}^{\mathrm{T}}} \widehat{\boldsymbol{\Omega}} \frac{\partial \widetilde{\mathbf{P}}_1}{\partial \widehat{\rho}_1} \mathbf{y}, \cdots, \mathrm{trace}(\widetilde{\mathbf{P}}_S\widehat{\boldsymbol{\Omega}}) + \frac{\partial \widehat{\rho}_S}{\partial \mathbf{y}^{\mathrm{T}}} \widehat{\boldsymbol{\Omega}} \frac{\partial \widetilde{\mathbf{P}}_S}{\partial \widehat{\rho}_S} \mathbf{y}\}^{\mathrm{T}}.$$

It is straightforward to show that

$$\widehat{C}(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{w} + 2\mathbf{w}^{\mathrm{T}}\mathbf{h},$$

so that $\widehat{C}(\mathbf{w})$ is a quadratic function of $\mathbf{w}$. Numerous software packages are available for obtaining the solution to this problem (e.g., quadprog of Matlab), and they generally work effectively and efficiently even when S is very large.

Define $R^*(\mathbf{w}) = E\| \sum_{s=1}^S w_s \widehat{\boldsymbol{\mu}}_s \mid_{\widehat{\rho}_s = \rho_s^*} - \boldsymbol{\mu}\|^2$ and $\widetilde{\xi}_n = \inf_{\mathbf{w} \in \mathcal{H}} R^*(\mathbf{w})$.

Assumption 6 *$\widetilde{\xi}_n^{-1}\sup_s|(\partial\widehat{\rho}_s/\partial\mathbf{y}^{\mathrm{T}})\widehat{\mathbf{\Omega}}(\partial\widetilde{\mathbf{P}}_s/\partial\widehat{\rho}_s)\mathbf{y}| = o_p(1)$, $\widetilde{\xi}_n^{-1}p = o(1)$, $n\widetilde{\xi}_n^{-1}\sup_s\lambda_{\max}(\widehat{\mathbf{P}}_s - \overline{\mathbf{P}}_s) = o_p(1)$, and there exists a positive integer G such that $S\widetilde{\xi}_n^{-2G}\sum_{s=1}^S(R_s^*)^G = o(1)$.*

The first three parts of Assumption 6 are natural extensions of Assumptions 4-5 in model averaging study. The fourth part of Assumption 6 is widely used and plays a central role in model averaging literature such as Wan et al. (2010), Liu and Ryo (2013), and Ando and Li (2014).

Theorem 2 *Under Assumptions 2, 3 and 6,*

$$\frac{L(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}}L(\mathbf{w})} \to 1 \tag{12}$$

in probability as $n \to \infty$.

The result (12) means that the model averaging estimator $\boldsymbol{\mu}_{\widehat{\mathbf{w}}}$ is optimal in the sense that its squared loss is asymptotically identical to that by the infeasible best averaging estimator. Note that $\inf_{\mathbf{w}\in\mathcal{H}}L(\mathbf{w}) \leq \inf_s L_s$, so it is expected that model averaging can reduce estimation error relative to model selection.

Remark: Although the normality assumption of ϵ_i (see (2)) is used in the proofs of Theorems 1 and 2, it is not essential. When we do not assume the normality assumption and assume the following moment condition

$$E(\epsilon_i^{4G}) \leq \kappa < \infty, \tag{13}$$

where κ is a positive constant, the asymptotic optimality shown by Theorems 1 and 2 still hold. See Appendix A.6 for the proof of Theorem 2 without using the normality assumption. The corresponding proof of Theorem 1 without using the normality assumption is simple and is available upon request from authors.

5 Monte Carlo

We conduct a Monte Carlo experiment to evaluate the performance of model selection and model averaging procedures in the current paper.

Denoting $\mathbf{l}$ as $n \times 1$ vector of ones, the data generating process (DGP) is

$$\mathbf{Y} = \rho\mathbf{W}_0\mathbf{Y}_n + \mathbf{l}\beta_0 + \mathbf{X}\beta_1 + \mathbf{V}$$

with $\mathbf{W}_0$ being the true weights matrix, where $\rho = 0.2$, 0.5 or 0.8, $\beta_0 = 1$, $\beta_1 = 1$ and $\mathbf{X}$ and $\mathbf{V}$ are generated from independent standard normal distributions. We use $n = 100$, 400. We have totally four candidate spatial weights matrices $[\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4]$. The $\mathbf{W}_1$ is a square tessellation where each unit only interact with its left neighbor (for the left edge unit, it has the right edge unit as its neighbor). The $\mathbf{W}_2$ is a square tessellation where each unit only interact with its left and right neighbors (for the left and right edge units, they have then only one neighbor). We call this a left-right matrix. $\mathbf{W}_3$ is a rook matrix, which represents a square tessellation with a connectivity of four for the inner fields on the chessboard and two and three for the corner and border fields, respectively. We also specify $\mathbf{W}_4$ as a queen matrix, which presents a square tessellation with a connectivity of eight for the inner fields on the chessboard and three and five for the corner and border fields, respectively. All these weights matrices are row-normalized. From $\mathbf{W}_1$ to $\mathbf{W}_4$, the spatial weights matrices become denser.

For each set of generated sample observations, we obtain the MLE under different spatial weights matrix specifications and the corresponding root mean square error (RMSE) of estimates.[3] Using these estimates, we can compute the $\widehat{C}_s$ in (9) and make the model selection. We can also have model averaging according to $\widehat{C}(\mathbf{w})$ in (11). We do this 1000 times. The average value of the loss $L_s = \|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2$ under each $\mathbf{W}_s$, $s = 1, ..., 4$ is reported, along with average value of the loss under model selection and model averaging. To investigate the accuracy of model selection and performance of model averaging, we also report the selection frequency for each model, and also the weights of model averaging. Results are summarized in Table 1.

From Table 1, we see that the loss of each model is smallest under the correct model specification, although its RMSE of parameter estimates are not necessarily smallest. When $n = 100$, for the model selection, it will pick up the true model with 88.1% probability when $\rho = 0.5$ (84.1% for $\rho = 0.2$ and 85.7% for $\rho = 0.8$), which implies that the model selection procedure in the current paper works. For the model averaging, the true model $\mathbf{W}_1$ is given dominant weights 90.5% on average when $\rho = 0.5$ (85.8% for $\rho = 0.2$ and 87.2% for $\rho = 0.8$). The loss of model averaging is slightly smaller than that of model selection. Comparing the cases of $n = 100$ and $n = 400$, we see that when sample size gets larger, the model selection and model averaging will have a better performance. Comparing different values of ρ, the RMSEs of estimators and the loss become larger when ρ is larger, but the performance of model selection and model averaging procedures are similar.

We also investigate the influence of a denser spatial weights matrix in DGP. Instead of $\mathbf{W}_1$, we

[3] The estimators of the model selection method is based on the selected model in each replication. We did not define the parameter estimates of model averaging, so the following tables do not include RMSE by model averaging.

now use the $\mathbf{W}_4$ as the true spatial weights matrix. Results are in Table 2. We see that the loss is still the smallest for the correct model, along with the RMSE of parameter estimates. However, comparing with the case with $\mathbf{W}_1$ as the true weights matrix, the performance of model selection is poor in picking up the correct model and the model averaging also performs poor in giving weights to the correct model. When sample size is larger, the performances of model selection and model averaging improve to an acceptable degree, but still worse than that of the case with $\mathbf{W}_1$ being the true spatial weights matrix. Also, when ρ becomes larger, the performances of model selection and model averaging become better.

Both Tables 1 and 2 have the true spatial weights matrix in the candidate models. What is the performance of model selection and model averaging when the true model is not in the candidates? In Table 3, we specify the true spatial weights matrix as the summation of $\mathbf{W}_2$ and $\mathbf{W}_4$ (and then row-normalize it). We see that when the true spatial weights matrix is not in the candidate models, the losses are relatively small for $\mathbf{W}_2$, $\mathbf{W}_3$ and $\mathbf{W}_4$. This is understandable because the true model is a weighted average of $\mathbf{W}_2$ and $\mathbf{W}_4$. For the model selection, when sample size or ρ increases, it will pick up the $\mathbf{W}_1$ with a smaller probability. For the model averaging, it also sets smaller weights to $\mathbf{W}_1$ when sample size or ρ increases. When $\rho = 0.5$ or 0.8, the model averaging always has a clearly smaller loss than the model selection and the loss of the model averaging is the smallest among all losses for $n = 400$.

The weighted average of $\mathbf{W}_2$ and $\mathbf{W}_4$, although not in the candidates, has a similar structure to the candidate models. We design a new spatial weights matrix $\mathbf{W}_5$ with an exponential decline pattern of spatial externalities. Each observation is assigned with a positive income generated from a uniform distribution, and elements of $\mathbf{W}_5$ is constructed by $\exp\{-10 \cdot |d_i - d_j|\}$ where $d_i - d_j$ is their income difference. Thus, when their income difference is larger, their economic distance is larger and less correlated. From Table 4, we see that when the spatial weights matrix has some continuous feature with an exponential decline pattern, the model selection will pick up $\mathbf{W}_1$ for most of the time, and the model averaging procedure will also set large weights for $\mathbf{W}_1$. The losses by different weights matrices are very similar and thus the model selection and averaging also lead to similar losses.

The approximation of $\widehat{C}_s$ for the risk (expected value of loss) in the current paper depends on the normal distribution assumption theoretically. We conduct Monte Carlo experiments to evaluate the performance of model selection and model averaging when the disturbances in the DGP are not normally distributed. Table 5 reports the results where the disturbances are $\chi_{(1)}$ distributed (demeaned). Comparing with Table 1, we see that the results are similar so that the performances of model selection and model averaging are still satisfactory even if the disturbances in the DGP are not normally distributed.

6 An Empirical Example

Keller and Shiue (2007) use historical data of the price of rice in China to study the role of spatial features in the expansion of interregional trade and market integration. We have data available for $n = 121$ prefectures (from 10 provinces) and $T = 108$ periods, where months of February and August are recorded from 54 years in the mid-Qing (Qing Dynasty, 1644-1912).[4] The estimation equation is (1)[5] and their reported estimates are the average from 54 years. From Keller and Shiue (2007)'s estimates, the spatial effect is significant. Thus, spatial features are important as the geographical distances influence the trade and possible arbitrage.

In estimating (1) for the rice price arbitrage, different weights matrices could be used. We have distance matrix available $[d_{ij}]_{i,j=1}^{121}$ among these prefecture capitals, where the d_{ij} ranges from 13 to 1854 km. Thus, we can construct one- and two-window distance bands or exponential specifications. For example, the one-window distance band could be (1) $\mathbf{W}_1$, where prefectures are neighbors if the $d_{ij} \leq 300$; and (2) $\mathbf{W}_2$, where prefectures are neighbors if the $d_{ij} \leq 600$. The two-window distance band could be $\mathbf{W}_3$, where $w_{ij}^{(3)} = 1$ if $d_{ij} \leq 300$, $w_{ij}^{(3)} = 0.5$ if $300 < d_{ij} \leq 600$ and $w_{ij}^{(3)} = 0$ if $d_{ij} > 600$. For the exponential specification, we can specify $w_{ij} = \exp\{\theta_d D_{ij}\}$ with $D_{ij} = \frac{d_{ij}}{100}$ and a larger absolute value of a negative θ_d denotes a more rapid decline in the size of the weights when d_{ij} increases. Keller and Shiue (2007) state that the specification with $\theta_d = -1.4$ fits the data well by a limited grid search in terms of likelihood. For distance band specifications, they find that the one-window specification ($\mathbf{W}_1$ and $\mathbf{W}_2$) perform better than the two-window specification ($\mathbf{W}_3$).

Instead of the likelihood approach in choosing the best spatial weights matrix, we use the model selection and model averaging approaches proposed in the current paper. Given each year's data, we run the cross sectional SAR regression using the candidate spatial weights matrix. We have six candidates, where $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ are based on the one- and two- window distance band, and $\mathbf{W}_4$,

[4] We have the minimum price and the maximum price for each prefecture, where the prices are collected from counties of each prefecture. We use the log mid-price for the estimation.

[5] Compared to Keller and Shiue (2007), the weather indicators are not included as exogenous variables due to the data availability. However, as those weather regressors are insignificant in Keller and Shiue (2007), the omission would not be controversial.

$\mathbf{W}_5$ and $\mathbf{W}_6$ use exponential specification $w_{ij} = \exp\{\theta_d D_{ij}\}$ with $\theta_d = -0.8$ for $\mathbf{W}_4$, $\theta_d = -1.4$ for $\mathbf{W}_5$ as in Keller and Shiue (2007), and $\theta_d = -2$ for $\mathbf{W}_6$. For 108 cross sectional regressions, we record the model selection result, and also compute the mean value of model averaging weights. The results are in Table 6. We see that among $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ which use one- and two-window distance band, the model selection has similar frequency for each of them, where the two-window distance band has the highest frequency (0.2406 vs 0.1811 and 0.184). Similar observation is for model averaging. Thus, compared with Keller and Shiue (2007) who use likelihood criterion and prefer one-window distance band, we find that two-window distance band is preferred to the one-window distance band. This seems more consistent with the exponential specifications below. For the exponential specifications, we see that $\theta_d = -1.4$ in Keller and Shiue (2007) has the largest frequency for the model selection and largest weights for model averaging. This implies that $\theta_d = -1.4$ is better than $\theta_d = -0.8$ (which under-values the transportation cost among regions, so that each prefecture has many neighbors effectively) and better than $\theta_d = -2$ (which over-values the transportation cost among regions, so that each prefecture has only neighboring prefectures as his neighbor).

7 Conclusion

This paper proposes a model selection procedure to choose an optimal weights matrix from several candidates by using a Mallows type criterion. This procedure shall be useful for empirical researchers who might have several spatial weights matrices available and do not have an explicit guide of which one to use. It also proposes a model averaging procedure to reduce estimation error. We prove that these procedures can asymptotically minimize the squared loss. Monte Carlo experiments show that proposed procedures have satisfactory finite sample performances, and an empirical example is illustrated.

Appendices

A.1 Derivation of Formula (7)

From the log-likelihood function (3), the concentrated maximum likelihood under the s^{th} candidate model is

$$\log\{\mathrm{LH}(\widehat{\rho}_s)\} \quad = \quad -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\widehat{\sigma}_s^2) + \log|\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s| - \frac{\|(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_s\|^2}{2\widehat{\sigma}_s^2}$$

$$\begin{aligned} &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{\|\mathbf{A}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y}\|^2}{n} + \log|\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s| - \frac{n}{2} \\ &= -\frac{n}{2}\log(2\pi) + \frac{n\log(n)}{2} - \frac{n}{2} - \frac{n}{2}\log(\|\mathbf{A}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y}\|^2) + \log|\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s|. \end{aligned}$$

Thus, $\partial\log\{\mathrm{LH}(\widehat{\rho}_s)\}/\partial\widehat{\rho}_s = 0$ implies

$$-n\widehat{\rho}_s\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y} + n\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{y} - \mathrm{trace}\{(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s\}\|\mathbf{A}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y}\|^2 = 0,$$

where we use the result of Abadir and Magnus (2005) (page 369) to calculate the derivative of determinant. Taking derivative with respect to $\mathbf{y}$ on the above formula, we have

$$\begin{aligned} 0 &= -n\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y}\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}} - 2n\widehat{\rho}_s\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y} + n(\mathbf{W}_s^{\mathrm{T}}\mathbf{A} + \mathbf{A}\mathbf{W}_s)\mathbf{y} \\ &\quad + \|\mathbf{A}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y}\|^2\mathrm{trace}\{(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s\}\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}} \\ &\quad - 2\mathrm{trace}\{(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s\}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{\mathrm{T}}\mathbf{A}(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)\mathbf{y} \\ &\quad - \mathrm{trace}\{(\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)^{-1}\mathbf{W}_s\}(2\widehat{\rho}_s\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{y})\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}}, \end{aligned}$$

which implies (7).

A.2 Derivation of Formula (8)

Define $\mathbf{z} = \mathbf{\Omega}^{-1/2}\mathbf{y}$. Similar to the proof of Theorem 3 of Greven and Kneib (2010), from normality of $\boldsymbol{\epsilon}$ and Stein's Lemma (1981), we have

$$\begin{aligned} E\{(\widetilde{\mathbf{P}}_s\mathbf{y})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\mu})\} &= E\{(\mathbf{\Omega}^{1/2}\widetilde{\mathbf{P}}_s\mathbf{\Omega}^{1/2}\mathbf{z})^{\mathrm{T}}(\mathbf{z} - \mathbf{\Omega}^{-1/2}\boldsymbol{\mu})\} \\ &= E\{\mathrm{trace}\frac{\partial(\mathbf{\Omega}^{1/2}\widetilde{\mathbf{P}}_s\mathbf{\Omega}^{1/2}\mathbf{z})}{\partial\mathbf{z}^{\mathrm{T}}}\} \\ &= E\{\mathrm{trace}(\mathbf{\Omega}^{1/2}\widetilde{\mathbf{P}}_s\mathbf{\Omega}^{1/2}) + E[\mathrm{trace}\{\frac{\partial(\mathbf{\Omega}^{1/2}\widetilde{\mathbf{P}}_s\mathbf{\Omega}^{1/2}\mathbf{z})}{\partial\widehat{\rho}_s}\frac{\partial\widehat{\rho}_s}{\partial\mathbf{z}^{\mathrm{T}}}\}] \\ &= E\{\mathrm{trace}(\widetilde{\mathbf{P}}_s\mathbf{\Omega})\} + E(\frac{\partial\widehat{\rho}_s}{\partial\mathbf{z}^{\mathrm{T}}}\mathbf{\Omega}^{1/2}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}) \\ &= E\{\mathrm{trace}(\widetilde{\mathbf{P}}_s\mathbf{\Omega})\} + E(\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\mathbf{\Omega}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}). \qquad (A.1) \end{aligned}$$

So we have

$$\begin{aligned} R_s &= E\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2 = E\|\widetilde{\mathbf{P}}_s\mathbf{y} - \boldsymbol{\mu}\|^2 = E\|\widetilde{\mathbf{P}}_s\mathbf{y} - \mathbf{y} + \mathbf{y} - \boldsymbol{\mu}\|^2 \\ &= E\|\widetilde{\mathbf{P}}_s\mathbf{y} - \mathbf{y}\|^2 + 2E\{(\widetilde{\mathbf{P}}_s\mathbf{y} - \mathbf{y})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\mu})\} + \mathrm{trace}(\mathbf{\Omega}) \end{aligned}$$

$$
\begin{aligned}
&= E\|\widetilde{\mathbf{P}}_s\mathbf{y}-\mathbf{y}\|^2 + 2E\{(\widetilde{\mathbf{P}}_s\mathbf{y})^{\mathrm{T}}(\mathbf{y}-\boldsymbol{\mu})\} - \mathrm{trace}(\boldsymbol{\Omega}) \\
&= E\|\widetilde{\mathbf{P}}_s\mathbf{y}-\mathbf{y}\|^2 + E\{2\mathrm{trace}(\widetilde{\mathbf{P}}_s\boldsymbol{\Omega})\} + E(2\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\boldsymbol{\Omega}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}) - \mathrm{trace}(\boldsymbol{\Omega}),
\end{aligned}
$$

which implies (8).

A.3 Discussion on Assumption 4

Write $\mathbf{B}_s = (\mathbf{I}_n - \widehat{\rho}_s\mathbf{W}_s)$. From (6), (7) and $\mathbf{A} = \mathbf{I}_n - \mathbf{P}$, we have

$$
\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y} = \frac{\begin{array}{c}\{2n\widehat{\rho}_s\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{W}_s\mathbf{y} - n(\mathbf{W}_s^{\mathrm{T}}\mathbf{A}+\mathbf{A}\mathbf{W}_s)\mathbf{y} + 2\mathrm{trace}(\mathbf{B}_s^{-1}\mathbf{W}_s)\mathbf{B}_s^{\mathrm{T}}\mathbf{A}\mathbf{B}_s\mathbf{y}\}^{\mathrm{T}} \\ \times\widehat{\boldsymbol{\Omega}}(\mathbf{B}_s^{-1}\mathbf{W}_s\mathbf{B}_s^{-1}\mathbf{P}\mathbf{B}_s\mathbf{y} - \mathbf{B}_s^{-1}\mathbf{P}\mathbf{W}_s\mathbf{y})\end{array}}{\begin{array}{c}\|\mathbf{A}\mathbf{B}_s\mathbf{y}\|^2\mathrm{trace}(\mathbf{B}_s^{-1}\mathbf{W}_s\mathbf{B}_s^{-1}\mathbf{W}_s) + n\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{P}\mathbf{W}_s\mathbf{y} \\ +2\mathrm{trace}(\mathbf{B}_s^{-1}\mathbf{W}_s)\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{B}_s\mathbf{y} - n\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{W}_s\mathbf{y}\end{array}}. \quad (A.2)
$$

Assume that $\|\mathbf{y}\|^2$ has order n and uniformly for $s \in \{1,\ldots,S\}$, the smallest singular value of $\mathbf{W}_s^{\mathrm{T}}\mathbf{W}_s$ is bounded away from zero, then $n\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{W}_s\mathbf{y}$ has order n^2 uniformly. We further assume that $\lambda_{\max}(\mathbf{W}_s)$ and $\lambda_{\max}(\mathbf{B}_s^{-1})$ are bounded uniformly. Then, from (A.5) in the following proof and the truth that $\lambda_{\max}(\mathbf{P}) = 1$, we know that the terms $\|\mathbf{A}\mathbf{B}_s\mathbf{y}\|^2$, $\mathrm{trace}(\mathbf{B}_s^{-1}\mathbf{W}\mathbf{B}_{\mathbf{s}}^{-\mathbf{1}}\mathbf{W}_s)$, $\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{P}\mathbf{W}_s\mathbf{y}$, $\mathrm{trace}(\mathbf{B}_{\mathbf{s}}^{-\mathbf{1}}\mathbf{W}_s)$, and $\mathbf{y}^{\mathrm{T}}\mathbf{W}_s^{\mathrm{T}}\mathbf{A}\mathbf{B}_s\mathbf{y}$ are equal to $O(n^2)$ uniformly. So unless there exists a special relationship among these terms, the denominator of the right hand of (A.2) should have order n^2 uniformly.

Similarly, by the third part of Assumption 2 and the assumptions that $\|\mathbf{y}\|^2$ has order n and $\lambda_{\max}(\mathbf{W}_s)$ and $\lambda_{\max}(\mathbf{B}_s^{-1})$ are bounded uniformly, we know that the nominator of the right hand of (A.2) is $O_p(n^2)$ uniformly. Hence, $(\partial\widehat{\rho}_s/\partial\mathbf{y}^{\mathrm{T}})\widehat{\boldsymbol{\Omega}}(\partial\widetilde{\mathbf{P}}_s/\partial\widehat{\rho}_s)\mathbf{y} = O_p(1)$ uniformly. From Assumption 1, we have $\xi_n \to \infty$. So Assumption 4 is reasonable.

A.4 Proof of Theorem 1

Define $\widetilde{\boldsymbol{\epsilon}} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}$ such that $\boldsymbol{y} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\epsilon}}$. It is seen that

$$
R_s^* = E\|\widehat{\boldsymbol{\mu}}_s\,|_{\widehat{\rho}_s=\rho_s^*} - \boldsymbol{\mu}\|^2 = E\|\overline{\mathbf{P}}_s\mathbf{y}-\boldsymbol{\mu}\|^2 = \|\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu}\|^2 + \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}}) \geq \|\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu}\|^2, \quad (A.3)
$$

by which, we have

$$
\begin{aligned}
|L_s - R_s^*| &= \Big|\|\widetilde{\mathbf{P}}_s\mathbf{y}-\boldsymbol{\mu}\|^2 - \|\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu}\|^2 - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}})\Big| \\
&= \Big|\|(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\boldsymbol{\mu} + (\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\widetilde{\boldsymbol{\epsilon}} + \overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}} + \overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 \\
&\quad -\|\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu}\|^2 - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}})\Big|
\end{aligned}
$$

$$
\begin{aligned}
&= \Big|\|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\mu} + (\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\widetilde{\boldsymbol{\epsilon}} + \overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}\|^2 \\
&\quad +2\{(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\mu} + (\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\widetilde{\boldsymbol{\epsilon}} + \overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}\}^{\mathrm{T}}(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu}) - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}})\Big| \\
&\leq |\|\overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}\|^2 - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}})| + 2|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}\overline{\mathbf{P}}_s\boldsymbol{\epsilon}| + \|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\mu}\|^2 \\
&\quad +\|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\epsilon}\|^2 + 2|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\mu}| + 2|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\widetilde{\boldsymbol{\epsilon}}| \\
&\leq |\|\overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}\|^2 - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}})| + 2|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}\overline{\mathbf{P}}_s\boldsymbol{\epsilon}| + \|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\mu}\|^2 \\
&\quad +\|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\epsilon}\|^2 + 2R_s^{*1/2}\|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\mu}\| + 2R_s^{*1/2}\|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\widetilde{\boldsymbol{\epsilon}}\|.
\end{aligned}
$$

Let $\widehat{C}_s^* = \widehat{C}_s - \|\widetilde{\boldsymbol{\epsilon}}\|^2$ such that $\widehat{s} = \mathrm{argmin}_{s\in\{1,\ldots,S\}}\widehat{C}_s^*$. Then, we have

$$
\begin{aligned}
|\widehat{C}_s^* - L_s| &= \left|\|\widetilde{\mathbf{P}}_s\mathbf{y} - \mathbf{y}\|^2 + 2\mathrm{trace}(\widetilde{\mathbf{P}}_s\widehat{\boldsymbol{\Omega}}) + 2\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y} - \|\widetilde{\boldsymbol{\epsilon}}\|^2 - \|\widetilde{\mathbf{P}}_s\mathbf{y} - \boldsymbol{\mu}\|^2\right| \\
&= \left|\|(\widetilde{\mathbf{P}}_s\mathbf{y} - \boldsymbol{\mu}) - \widetilde{\boldsymbol{\epsilon}}\|^2 + 2\mathrm{trace}(\widetilde{\mathbf{P}}_s\widehat{\boldsymbol{\Omega}}) + 2\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y} - \|\widetilde{\boldsymbol{\epsilon}}\|^2 - \|\widetilde{\mathbf{P}}_s\mathbf{y} - \boldsymbol{\mu}\|^2\right| \\
&= \left|-2(\widetilde{\mathbf{P}}_s\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}} + 2\mathrm{trace}(\widetilde{\mathbf{P}}_s\widehat{\boldsymbol{\Omega}}) + 2\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| \\
&= \left|-2(\widetilde{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}} + 2\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}} + 2\mathrm{trace}(\widetilde{\mathbf{P}}_s\widehat{\boldsymbol{\Omega}}) + 2\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| \\
&\leq 2|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}} - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega})| + 2|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}| + 2\left|\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| \\
&\quad +2|\mathrm{trace}\{(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\widehat{\boldsymbol{\Omega}}\}| + 2|\mathrm{trace}\{\overline{\mathbf{P}}_s(\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}})\}| \\
&\quad +2|\boldsymbol{\mu}^{\mathrm{T}}(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}| + 2|\boldsymbol{\epsilon}^{\mathrm{T}}(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}| \\
&\leq 2|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}} - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega})| + 2|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}| + 2\left|\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| \\
&\quad +2|\mathrm{trace}\{(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\widehat{\boldsymbol{\Omega}}\}| + 2|\mathrm{trace}\{\overline{\mathbf{P}}_s(\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}})\}| \\
&\quad +2\|\boldsymbol{\mu}\|\|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}\| + 2\|\widetilde{\boldsymbol{\epsilon}}\|\|(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}\|.
\end{aligned}
$$

So, as in the proof of Theorem 2.1 in Li (1987), in order to prove (10), we need only to verify that

$$\sup_s R_s^{*-1}\left|\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| = o_p(1), \tag{A.4a}$$

$$\sup_s R_s^{*-1}|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}| = o_p(1), \tag{A.4b}$$

$$\sup_s R_s^{*-1}|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}} - \mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega})| = o_p(1), \tag{A.4c}$$

$$\sup_s R_s^{*-1}|(\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu})^{\mathrm{T}}\overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}| = o_p(1), \tag{A.4d}$$

$$\sup_s R_s^{*-1}|\|\overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}\|^2-\mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}})| = o_p(1), \tag{A.4e}$$

$$\sup_s R_s^{*-1}\|(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\boldsymbol{\mu}\|^2 = o_p(1), \tag{A.4f}$$

$$\sup_s R_s^{*-1}\|(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\widetilde{\boldsymbol{\epsilon}}\|^2 = o_p(1), \tag{A.4g}$$

$$\sup_s R_s^{*-1}\|\boldsymbol{\mu}\|\|(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}\| = o_p(1), \tag{A.4h}$$

$$\sup_s R_s^{*-1}\|\widetilde{\boldsymbol{\epsilon}}\|\|(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}\| = o_p(1), \tag{A.4i}$$

$$\sup_s R_s^{*-1}|\mathrm{trace}\{(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\widehat{\boldsymbol{\Omega}}\}| = o_p(1), \tag{A.4j}$$

and

$$\sup_s R_s^{*-1}|\mathrm{trace}\{\overline{\mathbf{P}}_s(\boldsymbol{\Omega}-\widehat{\boldsymbol{\Omega}})\}| = o_p(1). \tag{A.4k}$$

By Assumptions 4, it is straightforward to show that

$$\sup_s R_s^{*-1}\left|\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| \le \xi_n^{-1}\sup_s\left|\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\boldsymbol{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| = o_p(1),$$

which is (A.4a).

It is well known that for any two $n \times n$ matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ (Li, 1987),

$$\lambda_{\max}(\mathbf{B}_1\mathbf{B}_2) \le \lambda_{\max}(\mathbf{B}_1)\lambda_{\max}(\mathbf{B}_2) \quad \text{and} \quad \lambda_{\max}(\mathbf{B}_1+\mathbf{B}_2) \le \lambda_{\max}(\mathbf{B}_1)+\lambda_{\max}(\mathbf{B}_2). \tag{A.5}$$

From (A.5), Assumption 3, and the truth that $\lambda_{\max}(\mathbf{P}) = 1$, we have

$$\begin{aligned}\sup_s\{\lambda_{\max}(\overline{\mathbf{P}}_s)\} &= \sup_s[\lambda_{\max}\{(\mathbf{I}_n-\rho_s^*\mathbf{W}_s)^{-1}\mathbf{P}(\mathbf{I}_n-\rho_s^*\mathbf{W}_s)\}] \\ &\le \sup_s[\lambda_{\max}\{(\mathbf{I}_n-\rho_s^*\mathbf{W}_s)^{-1}\}][1+\sup_s\{\lambda_{\max}(\rho_s^*\mathbf{W}_s)\}] \\ &= O(1).\end{aligned} \tag{A.6}$$

From the normality of $\widetilde{\boldsymbol{\epsilon}}$, (A.6), and Assumptions 1-2, the equations (A.4b)-(A.4e) can be shown by using the same steps as in the proof of Theorem 2.1 of Li (1987).

From (A.5) and Assumptions 2 and 5, we have

$$\sup_s R_s^{*-1}\|(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\boldsymbol{\mu}\|^2 \le \xi_n^{-1}\|\boldsymbol{\mu}\|^2\sup_s\lambda_{\max}^2(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s) = o_p(1)$$

and

$$\sup_s R_s^{*-1}\|\boldsymbol{\mu}\|\|(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}\| \le \xi_n^{-1}\|\boldsymbol{\mu}\|\sup_s\lambda_{\max}(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\|\widetilde{\boldsymbol{\epsilon}}\| = o_p(1),$$

which are (A.4f) and (A.4i), respectively. Similarly, we can get (A.4g)-(A.4h). Also from (A.5) and Assumptions 2 and 5, we have

$$\begin{aligned}\sup_s R_s^{*-1}|\mathrm{trace}\{(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\widehat{\mathbf{\Omega}}\}| &\le \xi_n^{-1}\sup_s\{\lambda_{\max}(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\lambda_{\max}(\widehat{\mathbf{\Omega}})\mathrm{rank}(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\}\\ &\le 2p\xi_n^{-1}\sup_s\lambda_{\max}(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\lambda_{\max}(\widehat{\mathbf{\Omega}})\\ &= o_p(1), \qquad\qquad (\mathrm{A.7})\end{aligned}$$

which is (A.4j).

From (A.5)-(A.6) and Assumptions 2 and 5, we have

$$\begin{aligned}\sup_s R_s^{*-1}|\mathrm{trace}\{\overline{\mathbf{P}}_s(\mathbf{\Omega}-\widehat{\mathbf{\Omega}})\}| &\le \xi_n^{-1}\sup_s\{\lambda_{\max}(\mathbf{\Omega}-\widehat{\mathbf{\Omega}})\lambda_{\max}(\overline{\mathbf{P}}_s)\mathrm{rank}(\overline{\mathbf{P}}_s)\}\\ &\le \xi_n^{-1}\{\lambda_{\max}(\mathbf{\Omega})+\lambda_{\max}(\widehat{\mathbf{\Omega}})\}\sup_s\lambda_{\max}(\overline{\mathbf{P}}_s)p\\ &= o_p(1),\end{aligned}$$

which is (A.4k). This completes the proof.

A.5 Proof of Theorem 2

Let $\sup_{\mathbf{w}}$ indicate supremum over $\mathbf{w}\in\mathcal{H}$. Following the steps in the proof of Theorem 1, we need only to verify that

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}\left|\sum\nolimits_{s=1}^S w_s\frac{\partial\widehat{\rho}_s}{\partial\mathbf{y}^{\mathrm{T}}}\widehat{\mathbf{\Omega}}\frac{\partial\widetilde{\mathbf{P}}_s}{\partial\widehat{\rho}_s}\mathbf{y}\right| = o_p(1), \qquad (\mathrm{A.8a})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\{\overline{\mathbf{P}}(\mathbf{w})\boldsymbol{\mu}-\boldsymbol{\mu}\}^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}| = o_p(1), \qquad (\mathrm{A.8b})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}-\mathrm{trace}\{\overline{\mathbf{P}}(\mathbf{w})\mathbf{\Omega}\}| = o_p(1), \qquad (\mathrm{A.8c})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\{\overline{\mathbf{P}}(\mathbf{w})\boldsymbol{\mu}-\boldsymbol{\mu}\}^{\mathrm{T}}\overline{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}| = o_p(1), \qquad (\mathrm{A.8d})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\|\overline{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}\|^2-\mathrm{trace}\{\overline{\mathbf{P}}(\mathbf{w})\mathbf{\Omega}\overline{\mathbf{P}}(\mathbf{w})^{\mathrm{T}}\}| = o_p(1), \qquad (\mathrm{A.8e})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}\|\sum\nolimits_{s=1}^S w_s(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\boldsymbol{\mu}\|^2 = o_p(1), \qquad (\mathrm{A.8f})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}\|\sum\nolimits_{s=1}^S w_s(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)\widetilde{\boldsymbol{\epsilon}}\|^2 = o_p(1), \qquad (\mathrm{A.8g})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}\|\boldsymbol{\mu}\|\|\sum\nolimits_{s=1}^S w_s(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}\| = o_p(1), \qquad (\mathrm{A.8h})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}\|\widetilde{\boldsymbol{\epsilon}}\|\|\sum\nolimits_{s=1}^S w_s(\widetilde{\mathbf{P}}_s-\overline{\mathbf{P}}_s)^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}\| = o_p(1), \qquad (\mathrm{A.8i})$$

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\mathrm{trace}\{\textstyle\sum_{s=1}^{S} w_s(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\widehat{\boldsymbol{\Omega}}\}| = o_p(1), \tag{A.8j}$$

and

$$\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\mathrm{trace}\{\overline{\mathbf{P}}(\mathbf{w})(\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}})\}| = o_p(1). \tag{A.8k}$$

Following the first part of Assumption 6, we can obtain (A.8a). From the normality of $\widetilde{\boldsymbol{\epsilon}}$, (A.6), Assumption 2, and the fourth part of Assumption 6, we can obtain (A.8b)-(A.8e) by using the same steps as in the proof of Theorem 1' of Wan et al. (2010). From (A.5), and Assumption 2, and the third part of Assumption 6, we have

$$\begin{aligned}
&\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}\|\textstyle\sum_{s=1}^{S} w_s(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\boldsymbol{\mu}\|^2 \\
&\leq \widetilde{\xi}_n^{-1}\|\boldsymbol{\mu}\|^2\lambda_{\max}\{\textstyle\sum_{s=1}^{S} w_s(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)^{\mathrm{T}}\sum_{s=1}^{S} w_s(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\} \\
&\leq \widetilde{\xi}_n^{-1}\|\boldsymbol{\mu}\|^2\lambda_{\max}^2\{\textstyle\sum_{s=1}^{S} w_s(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\} \\
&\leq \widetilde{\xi}_n^{-1}\|\boldsymbol{\mu}\|^2\{\textstyle\sum_{s=1}^{S} w_s\lambda_{\max}(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s)\}^2 \\
&\leq \widetilde{\xi}_n^{-1}\|\boldsymbol{\mu}\|^2\sup_s\lambda_{\max}^2(\widetilde{\mathbf{P}}_s - \overline{\mathbf{P}}_s) \\
&= o_p(1),
\end{aligned}$$

which is (A.8f). Next, following the steps in the proofs of (A.4g)-(A.4k), we can obtain (A.8g)-(A.8k).

A.6 Proof of Theorem 2 without using the normality assumption

Seeing Appendix A.5, the proofs of (A.8a) and (A.8f)-(A.8k) do not depend on the normality assumption of $\boldsymbol{\epsilon}$. Thus, we need only to reprove (A.8b)-(A.8e).

From (A.3), (A.5), (A.6), the moment condition (13), Conditions 2 and 6, and Theorem 2 of Whittle (1960), we have that for any $\delta > 0$,

$$\begin{aligned}
&\Pr[\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\{\overline{\mathbf{P}}(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu}\}^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}| > \delta] \\
&\leq \Pr\{\sup_s |(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}| > \delta\widetilde{\xi}_n\} \\
&\leq \textstyle\sum_{s=1}^{S}\Pr\{|(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}| > \delta\widetilde{\xi}_n\} \\
&\leq \delta^{-2G}\widetilde{\xi}_n^{-2G}\textstyle\sum_{s=1}^{S} E\{(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}\}^{2G} \\
&\leq C_1\delta^{-2G}\widetilde{\xi}_n^{-2G}\textstyle\sum_{s=1}^{S}\|(\mathbf{I}_n - \rho\mathbf{W}^{\mathrm{T}})^{-1}(\overline{\mathbf{P}}_s\boldsymbol{\mu} - \boldsymbol{\mu})\|^{2G}
\end{aligned}$$

$$
\begin{aligned}
&\le C_1\delta^{-2G}\lambda_{\max}^{G}(\boldsymbol{\Omega})\widetilde{\xi}_n^{-2G}\sum\nolimits_{s=1}^{S}(R_s^*)^G\\
&= o_p(1),
\end{aligned}
$$

$$
\begin{aligned}
&\Pr[\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}-\mathrm{trace}\{\overline{\mathbf{P}}(\mathbf{w})\boldsymbol{\Omega}\}|>\delta]\\
&\le \Pr\{\sup_{s}|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}-\mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega})|>\delta\widetilde{\xi}_n\}\\
&\le \sum\nolimits_{s=1}^{S}\Pr\{|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s\widetilde{\boldsymbol{\epsilon}}-\mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega})|>\delta\widetilde{\xi}_n\}\\
&= \sum\nolimits_{s=1}^{S}\Pr\{|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\rho\mathbf{W}^{\mathrm{T}})^{-1}\overline{\mathbf{P}}_s(\mathbf{I}_n-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}-\mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega})|>\delta\widetilde{\xi}_n\}\\
&\le C_2\delta^{-2G}\widetilde{\xi}_n^{-2G}\sum\nolimits_{s=1}^{S}\mathrm{trace}^G\{(\mathbf{I}_n-\rho\mathbf{W}^{\mathrm{T}})^{-1}\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_s^{\mathrm{T}}(\mathbf{I}_n-\rho\mathbf{W})^{-1}\}\\
&\le C_2\delta^{-2G}\lambda_{\max}^{G}(\boldsymbol{\Omega})\widetilde{\xi}_n^{-2G}\sum\nolimits_{s=1}^{S}\mathrm{trace}^G(\overline{\mathbf{P}}_s^{\mathrm{T}}\boldsymbol{\Omega}\overline{\mathbf{P}}_s)\\
&\le C_2\delta^{-2G}\lambda_{\max}^{G}(\boldsymbol{\Omega})\widetilde{\xi}_n^{-2G}\sum\nolimits_{s=1}^{S}(R_s^*)^G\\
&= o_p(1),
\end{aligned}
$$

$$
\begin{aligned}
&\Pr[\sup_{\mathbf{w}} R^*(\mathbf{w})^{-1}|\{\overline{\mathbf{P}}(\mathbf{w})\boldsymbol{\mu}-\boldsymbol{\mu}\}^{\mathrm{T}}\overline{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}|>\delta]\\
&\le \Pr\{\sup_{t}\sup_{s}|(\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu})^{\mathrm{T}}\overline{\mathbf{P}}_t(\mathbf{I}_n-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}|>\delta\widetilde{\xi}_n\}\\
&\le \sum\nolimits_{t=1}^{S}\sum\nolimits_{s=1}^{S}\Pr\{|(\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu})^{\mathrm{T}}\overline{\mathbf{P}}_t(\mathbf{I}_n-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}|>\delta\widetilde{\xi}_n\}\\
&\le \delta^{-2G}\widetilde{\xi}_n^{-2G}\sum\nolimits_{t=1}^{S}\sum\nolimits_{s=1}^{S}E\{(\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu})^{\mathrm{T}}\overline{\mathbf{P}}_t(\mathbf{I}_n-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}\}^{2G}\\
&\le C_3\delta^{-2G}\widetilde{\xi}_n^{-2G}\sum\nolimits_{t=1}^{S}\sum\nolimits_{s=1}^{S}\|(\mathbf{I}_n-\rho\mathbf{W}^{\mathrm{T}})^{-1}\overline{\mathbf{P}}_t^{\mathrm{T}}(\overline{\mathbf{P}}_s\boldsymbol{\mu}-\boldsymbol{\mu})\|^{2G}\\
&\le C_3\delta^{-2G}\lambda_{\max}^{G}(\boldsymbol{\Omega})\widetilde{\xi}_n^{-2G}\sum\nolimits_{t=1}^{S}\sum\nolimits_{s=1}^{S}\lambda_{\max}^{2G}(\overline{\mathbf{P}}_t)\,(R_s^*)^G\\
&= o_p(1),
\end{aligned}
$$

and

$$
\begin{aligned}
&\Pr[R^*(\mathbf{w})^{-1}|\|\overline{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}\|^2-\mathrm{trace}\{\overline{\mathbf{P}}(\mathbf{w})\boldsymbol{\Omega}\overline{\mathbf{P}}(\mathbf{w})^{\mathrm{T}}\}|>\delta]\\
&\le \Pr\{\sup_{t}\sup_{s}|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s\overline{\mathbf{P}}_t^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}-\mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_t^{\mathrm{T}})|>\delta\widetilde{\xi}_n\}\\
&\le \sum\nolimits_{t=1}^{S}\sum\nolimits_{s=1}^{S}\Pr\{|\widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\overline{\mathbf{P}}_s\overline{\mathbf{P}}_t^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}-\mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_t^{\mathrm{T}})|>\delta\widetilde{\xi}_n\}\\
&= \sum\nolimits_{t=1}^{S}\sum\nolimits_{s=1}^{S}\Pr\{|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\rho\mathbf{W}^{\mathrm{T}})^{-1}\overline{\mathbf{P}}_s\overline{\mathbf{P}}_t^{\mathrm{T}}(\mathbf{I}_n-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}-\mathrm{trace}(\overline{\mathbf{P}}_s\boldsymbol{\Omega}\overline{\mathbf{P}}_t^{\mathrm{T}})|>\delta\widetilde{\xi}_n\}\\
&\le C_4\delta^{-2G}\widetilde{\xi}_n^{-2G}\sum\nolimits_{t=1}^{S}\sum\nolimits_{s=1}^{S}\mathrm{trace}^G\{(\mathbf{I}_n-\rho\mathbf{W}^{\mathrm{T}})^{-1}\overline{\mathbf{P}}_s\overline{\mathbf{P}}_t^{\mathrm{T}}\boldsymbol{\Omega}\overline{\mathbf{P}}_t\overline{\mathbf{P}}_s^{\mathrm{T}}(\mathbf{I}_n-\rho\mathbf{W})^{-1}\}
\end{aligned}
$$

$$
\begin{aligned}
&\leq C_4 \delta^{-2G} \lambda^{G}_{\max}(\mathbf{\Omega}) \widetilde{\xi}_n^{-2G} \sum\nolimits_{t=1}^{S} \sum\nolimits_{s=1}^{S} \lambda^{2G}_{\max}(\overline{\mathbf{P}}_t) \mathrm{trace}^{G}(\overline{\mathbf{P}}_s^{\mathrm{T}} \mathbf{\Omega} \overline{\mathbf{P}}_s) \\
&\leq C_4 \delta^{-2G} \lambda^{G}_{\max}(\mathbf{\Omega}) \widetilde{\xi}_n^{-2G} \sum\nolimits_{t=1}^{S} \sum\nolimits_{s=1}^{S} \lambda^{2G}_{\max}(\overline{\mathbf{P}}_t) \left(R^{*}_s\right)^{G} \\
&= o_p(1),
\end{aligned}
$$

where C_1, C_2, C_3 and C_4 are positive constants. There results imply (A.8b)-(A.8e), respectively.

Reference

Abadir, K.M. & J. R. Magnus (2005). *Matrix Algebra.* Cambridge University Press.

Ando, T. & K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* 109, 254-265.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer Academic, The Netherlands.

Anselin, L. (1992). Space and applied econometrics, Anselin (ed.), Special Issue. *Regional Science and Urban Economics* 22.

Anselin, L. & A.K. Bera (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics, A. Ullah and D.E.A. Giles (eds.), *Handbook of Applied Economics Statistics.* Marcel Dekker, New York.

Anselin, L. & R. Florax (1995). *New Directions in Spatial Econometrics.* Springer-Verlag, Berlin.

Anselin, L. & S. Rey (1997). Spatial econometrics, Anselin, L. and S. Rey (eds.), Special Issue, *International Regional Science Review* 20.

Cliff, A.D. & J.K. Ord (1973). Spatial Autocorrelation. London: Pion Ltd.

Cressie, N. (1993). Statistics for Spatial Data. Wiley, New York.

Debarsy, N. & C. Ertur (2010). Testing for spatial autocorrelation in a fixed effects panel data model. *Regional Science and Urban Economics* 40, 453-470.

Greven, S. & T. Kneib (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* 97, 773-789.

Kelejian H.H. (2008). A spatial J-test for model specification agaist a single or a set of nonnested alternatives. *Letters in Spatial and Rescources Science* 1, 3-11.

Kelejian, H.H. & I.R. Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbance. *Journal of Real Estate Finance and Economics* 17:1, 99-121.

Kelejian, H.H. & D. Robinson (1993). A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science* 72, 297-312.

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

Keller, W. & C.H. Shiue (2007). The origin of spatial interaction. *Journal of Econometrics* 140, 304-332.

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

Han, X. & L.F. Lee (2013b). Bayesian estimation and model selection for spatial Durbin error model with finite distributed lags. *Regional Science and Urban Economics* 43, 250-271.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica* 75, 1175-1189.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics* forthcoming.

Lee, L.F. (2003). Best spatial two-stage least squares estimator for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews* 22, No.4, 307-335.

Lee, L.F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica* 72, No.6, 1899-1925.

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

Lee, L.F. & J. Yu (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* 154, 165-185.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized crossvalidation: Discrete index set. *Annals of Statistics* 15, 958-975.

Liang H., G. Zou, A.T.K. Wan & X. Zhang (2011). Optimal weight choice for frequentist model average estimators. *Journal of American Statistical Association* 106, 1053-1066.

Liu, X., L.F. Lee & C.R. Bollinger (2010). Improved efficient quasi maximum likelihood estimator of spatial autoregressive models. *Journal of Econometrics* 159, 303-319.

Liu, Q. & R. Okui (2013). Heteroskedasticity-robust C_p model averaging. *Econometrics Journal* 16, 462-473.

Leung, G. & A.R. Barron (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* 52, 3396-3410.

Mallows, C.L. (1973). Some Comments on C_p. *Technometrics* 15, 661-675.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). textitStatistica Sinica 7, 221-264.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* 153, 1135-1151.

Wan, A. T. K., X. Zhang & G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156, 277-283.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and Its Applications* 5, 302-305.

Zhang, X., A. T. K. Wan & G. Zou (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* 174, 82-94.

Zhang, X., G. Zou & H. Liang (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika* 101, 205-218.

Table 1: True DGP is $\mathbf{W}_1$

			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.2	$n = 100$	RMSE of ρ	0.084	0.077	0.101	0.136	0.086	
		RMSE of β_0	0.152	0.140	0.165	0.202	0.151	
		RMSE of β_1	0.101	0.103	0.103	0.105	0.101	
		Loss	0.033	0.050	0.059	0.064	0.036	0.035
		MS accuracy	0.841	0.095	0.030	0.033		
		MA weights	0.858	0.089	0.023	0.030		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.070	0.039	0.055	0.086	0.067	
		RMSE of β_0	0.105	0.070	0.085	0.121	0.102	
		RMSE of β_1	0.050	0.054	0.052	0.052	0.051	
		Loss	0.012	0.028	0.039	0.044	0.013	0.013
		MS accuracy	0.903	0.072	0.016	0.009		
		MA weights	0.926	0.061	0.009	0.004		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.5	$n = 100$	RMSE of ρ	0.163	0.072	0.108	0.264	0.159	
		RMSE of β_0	0.353	0.180	0.245	0.553	0.345	
		RMSE of β_1	0.104	0.157	0.131	0.130	0.110	
		Loss	0.101	0.236	0.318	0.348	0.120	0.119
		MS accuracy	0.881	0.044	0.061	0.015		
		MA weights	0.905	0.023	0.059	0.013		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.146	0.036	0.054	0.195	0.146	
		RMSE of β_0	0.303	0.091	0.123	0.400	0.303	
		RMSE of β_1	0.052	0.133	0.084	0.067	0.052	
		Loss	0.058	0.194	0.281	0.312	0.059	0.060
		MS accuracy	0.984	0.011	0.005	0.000		
		MA weights	0.996	0.000	0.004	0.000		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.8	$n = 100$	RMSE of ρ	0.178	0.077	0.129	0.406	0.171	
		RMSE of β_0	0.915	0.411	0.665	2.065	0.881	
		RMSE of β_1	0.108	0.321	0.228	0.204	0.145	
		Loss	0.587	1.379	1.710	1.777	0.702	0.701
		MS accuracy	0.857	0.046	0.065	0.031		
		MA weights	0.872	0.043	0.058	0.027		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.144	0.046	0.089	0.390	0.144	
		RMSE of β_0	0.731	0.242	0.454	1.966	0.729	
		RMSE of β_1	0.053	0.312	0.173	0.113	0.055	
		Loss	0.366	1.211	1.639	1.715	0.373	0.374
		MS accuracy	0.992	0.002	0.007	0.000		
		MA weights	0.994	0.001	0.005	0.000		

Note: $(\beta_0, \beta_1) = (1, 1)$.

Table 2: True DGP is $\mathbf{W}_4$

			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.2	$n = 100$	RMSE of ρ	0.168	0.148	0.132	0.128	0.156	
		RMSE of β_0	0.244	0.221	0.202	0.192	0.227	
		RMSE of β_1	0.102	0.102	0.102	0.102	0.102	
		Loss	0.032	0.032	0.030	0.028	0.032	0.031
		MS accuracy	0.733	0.046	0.092	0.129		
		MA weights	0.735	0.045	0.092	0.128		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.170	0.143	0.105	0.073	0.121	
		RMSE of β_0	0.224	0.192	0.145	0.105	0.163	
		RMSE of β_1	0.051	0.051	0.050	0.050	0.050	
		Loss	0.012	0.012	0.010	0.007	0.010	0.009
		MS accuracy	0.412	0.051	0.148	0.388		
		MA weights	0.400	0.047	0.168	0.385		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.5	$n = 100$	RMSE of ρ	0.406	0.320	0.224	0.129	0.262	
		RMSE of β_0	0.836	0.665	0.477	0.283	0.549	
		RMSE of β_1	0.114	0.110	0.106	0.103	0.103	
		Loss	0.105	0.098	0.073	0.057	0.075	0.072
		MS accuracy	0.345	0.047	0.229	0.379		
		MA weights	0.311	0.040	0.269	0.380		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.397	0.301	0.182	0.060	0.120	
		RMSE of β_0	0.805	0.612	0.375	0.131	0.246	
		RMSE of β_1	0.063	0.058	0.053	0.050	0.050	
		Loss	0.068	0.060	0.033	0.015	0.020	0.019
		MS accuracy	0.095	0.034	0.133	0.738		
		MA weights	0.035	0.015	0.163	0.787		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.8	$n = 100$	RMSE of ρ	0.565	0.379	0.224	0.092	0.235	
		RMSE of β_0	2.866	1.935	1.154	0.484	1.212	
		RMSE of β_1	0.169	0.137	0.115	0.103	0.106	
		Loss	0.622	0.579	0.364	0.284	0.347	0.342
		MS accuracy	0.139	0.019	0.393	0.449		
		MA weights	0.087	0.013	0.461	0.439		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.531	0.336	0.167	0.037	0.081	
		RMSE of β_0	2.675	1.696	0.848	0.195	0.413	
		RMSE of β_1	0.124	0.085	0.059	0.050	0.051	
		Loss	0.471	0.419	0.168	0.075	0.092	0.090
		MS accuracy	0.048	0.007	0.162	0.783		
		MA weights	0.000	0.000	0.187	0.813		

Note: $(\beta_0, \beta_1) = (1, 1)$.

Table 3: True DGP is not in the candidates: $\mathbf{W}_0 = \mathbf{W}_2 + \mathbf{W}_4$

			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.2	$n = 100$	RMSE of ρ	0.144	0.107	0.113	0.130	0.131	
		RMSE of β_0	0.216	0.175	0.180	0.194	0.198	
		RMSE of β_1	0.102	0.102	0.102	0.103	0.102	
		Loss	0.032	0.030	0.031	0.032	0.032	0.031
		MS accuracy	0.651	0.156	0.094	0.099		
		MA weights	0.663	0.152	0.091	0.094		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.137	0.082	0.071	0.075	0.093	
		RMSE of β_0	0.184	0.118	0.105	0.107	0.130	
		RMSE of β_1	0.051	0.050	0.050	0.051	0.050	
		Loss	0.013	0.009	0.010	0.011	0.011	0.010
		MS accuracy	0.325	0.322	0.165	0.187		
		MA weights	0.325	0.330	0.166	0.179		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.5	$n = 100$	RMSE of ρ	0.336	0.188	0.147	0.144	0.217	
		RMSE of β_0	0.697	0.404	0.323	0.312	0.459	
		RMSE of β_1	0.113	0.104	0.106	0.112	0.105	
		Loss	0.105	0.074	0.073	0.080	0.084	0.075
		MS accuracy	0.305	0.313	0.201	0.180		
		MA weights	0.304	0.318	0.221	0.157		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.324	0.166	0.097	0.070	0.140	
		RMSE of β_0	0.659	0.343	0.205	0.151	0.289	
		RMSE of β_1	0.063	0.051	0.054	0.059	0.052	
		Loss	0.068	0.035	0.032	0.037	0.035	0.025
		MS accuracy	0.076	0.433	0.216	0.275		
		MA weights	0.045	0.454	0.309	0.192		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.8	$n = 100$	RMSE of ρ	0.470	0.227	0.133	0.113	0.221	
		RMSE of β_0	2.383	1.167	0.69	0.584	1.130	
		RMSE of β_1	0.165	0.110	0.117	0.142	0.119	
		Loss	0.646	0.485	0.356	0.375	0.417	0.386
		MS accuracy	0.174	0.226	0.348	0.253		
		MA weights	0.133	0.207	0.431	0.229		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.439	0.195	0.079	0.049	0.095	
		RMSE of β_0	2.216	0.985	0.404	0.252	0.481	
		RMSE of β_1	0.121	0.054	0.063	0.092	0.068	
		Loss	0.498	0.313	0.145	0.156	0.160	0.131
		MS accuracy	0.032	0.260	0.354	0.354		
		MA weights	0.003	0.087	0.675	0.235		

Note: $(\beta_0, \beta_1) = (1, 1)$.

Table 4: True DGP is not in the candidates: $\mathbf{W}_0$ has exponential decline pattern

			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.2	$n = 100$	RMSE of ρ	0.185	0.177	0.173	0.174	0.18	
		RMSE of β_0	0.263	0.254	0.250	0.249	0.258	
		RMSE of β_1	0.101	0.101	0.101	0.101	0.101	
		Loss	0.027	0.027	0.027	0.027	0.028	0.027
		MS accuracy	0.914	0.030	0.029	0.027		
		MA weights	0.917	0.030	0.028	0.025		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.191	0.186	0.181	0.179	0.186	
		RMSE of β_0	0.250	0.244	0.237	0.234	0.244	
		RMSE of β_1	0.050	0.050	0.050	0.050	0.050	
		Loss	0.007	0.007	0.007	0.007	0.007	0.007
		MS accuracy	0.897	0.029	0.042	0.033		
		MA weights	0.897	0.031	0.042	0.030		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.5	$n = 100$	RMSE of ρ	0.484	0.475	0.468	0.464	0.477	
		RMSE of β_0	0.993	0.974	0.960	0.952	0.978	
		RMSE of β_1	0.102	0.102	0.103	0.103	0.103	
		Loss	0.061	0.062	0.062	0.062	0.062	0.062
		MS accuracy	0.915	0.028	0.030	0.027		
		MA weights	0.915	0.028	0.029	0.028		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.491	0.485	0.479	0.476	0.485	
		RMSE of β_0	0.995	0.983	0.970	0.964	0.983	
		RMSE of β_1	0.051	0.051	0.051	0.050	0.051	
		Loss	0.016	0.016	0.016	0.016	0.016	0.016
		MS accuracy	0.898	0.027	0.041	0.033		
		MA weights	0.898	0.030	0.040	0.032		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.8	$n = 100$	RMSE of ρ	0.786	0.778	0.771	0.766	0.781	
		RMSE of β_0	3.977	3.936	3.899	3.875	3.948	
		RMSE of β_1	0.115	0.114	0.114	0.115	0.115	
		Loss	0.341	0.341	0.341	0.341	0.341	0.341
		MS accuracy	0.927	0.016	0.025	0.032		
		MA weights	0.929	0.016	0.025	0.030		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.791	0.786	0.779	0.776	0.785	
		RMSE of β_0	3.988	3.960	3.925	3.910	3.958	
		RMSE of β_1	0.053	0.053	0.053	0.052	0.053	
		Loss	0.088	0.088	0.088	0.088	0.088	0.088
		MS accuracy	0.907	0.025	0.040	0.028		
		MA weights	0.908	0.026	0.039	0.027		

Note: $(\beta_0, \beta_1) = (1, 1)$.

Table 5: True DGP is not normal

			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.2	$n = 100$	RMSE of ρ	0.086	0.081	0.106	0.136	0.087	
		RMSE of β_0	0.162	0.152	0.175	0.207	0.162	
		RMSE of β_1	0.104	0.105	0.107	0.107	0.104	
		Loss	0.034	0.051	0.061	0.065	0.037	0.037
		MS accuracy	0.861	0.085	0.026	0.029		
		MA weights	0.875	0.078	0.023	0.024		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.071	0.04	0.055	0.087	0.069	
		RMSE of β_0	0.110	0.074	0.09	0.127	0.107	
		RMSE of β_1	0.049	0.052	0.051	0.050	0.049	
		Loss	0.012	0.029	0.039	0.044	0.014	0.013
		MS accuracy	0.907	0.066	0.018	0.009		
		MA weights	0.934	0.053	0.010	0.003		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.5	$n = 100$	RMSE of ρ	0.161	0.075	0.110	0.267	0.156	
		RMSE of β_0	0.362	0.193	0.253	0.558	0.347	
		RMSE of β_1	0.106	0.159	0.133	0.130	0.112	
		Loss	0.103	0.238	0.319	0.349	0.119	0.118
		MS accuracy	0.898	0.040	0.055	0.007		
		MA weights	0.914	0.029	0.052	0.005		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.147	0.037	0.055	0.196	0.147	
		RMSE of β_0	0.308	0.094	0.127	0.404	0.307	
		RMSE of β_1	0.050	0.129	0.082	0.064	0.050	
		Loss	0.060	0.195	0.281	0.312	0.060	0.061
		MS accuracy	0.987	0.009	0.004	0.000		
		MA weights	0.997	0.000	0.003	0.000		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
ρ= 0.8	$n = 100$	RMSE of ρ	0.177	0.078	0.131	0.413	0.173	
		RMSE of β_0	0.929	0.426	0.682	2.095	0.909	
		RMSE of β_1	0.109	0.323	0.231	0.204	0.148	
		Loss	0.594	1.384	1.695	1.763	0.706	0.706
		MS accuracy	0.860	0.050	0.064	0.026		
		MA weights	0.878	0.047	0.053	0.022		
			$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	MS	MA
	$n = 400$	RMSE of ρ	0.146	0.047	0.089	0.390	0.146	
		RMSE of β_0	0.743	0.245	0.458	1.968	0.743	
		RMSE of β_1	0.051	0.308	0.170	0.111	0.052	
		Loss	0.377	1.215	1.639	1.716	0.383	0.385
		MS accuracy	0.992	0.000	0.004	0.003		
		MA weights	0.995	0.000	0.002	0.003		

Note: $(\beta_0, \beta_1) = (1, 1)$.

Table 6: Spatial Effect in Rice Prices

		$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	$\mathbf{W}_5$	$\mathbf{W}_6$
$n = 121$	Estimate of ρ	0.9272	0.9765	0.9478	0.9673	0.9120	0.8637
	Estimate of β_0	0.0209	0.0125	0.0098	0.0113	0.0270	0.0411
	MS result	0.1811	0.1840	0.2406	0.1615	0.2115	0.0213
	MA weight	0.1389	0.1667	0.2222	0.2037	0.2500	0.0185