



**HAL**  
open science

## Using Convex Combinations of Spatial Weights in Spatial Autoregressive Models

Nicolas Debarsy, James P Lesage

► **To cite this version:**

Nicolas Debarsy, James P Lesage. Using Convex Combinations of Spatial Weights in Spatial Autoregressive Models. Handbook of Regional Science, Springer Berlin Heidelberg, pp.1-16, In press, 10.1007/978-3-642-36203-3\_123-1 . halshs-03509810

**HAL Id: halshs-03509810**

**<https://shs.hal.science/halshs-03509810>**

Submitted on 4 Jan 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Using convex combinations of spatial weights in  
spatial autoregressive models

Nicolas Debarsy  
CNRS- LEM UMR 9221  
Université de Lille 1  
Bat. SH2 F-59655 Villeneuve-d'Ascq  
nicolas.debarsy@cnrs.fr

James P. LeSage  
Texas State University-San Marcos  
Department of Finance & Economics  
601 University Drive  
San Marcos, TX 78666, USA  
james.lesage@txstate.edu

April 8, 2019

## Abstract

Spatial regression models rely on simultaneous autoregressive processes that model spatial or cross-sectional dependence between cross-sectional observations using a weight matrix. A criticism of applied spatial regression methods is that reliance on geographic proximity of observations to form the weight matrix that specifies the structure of cross-sectional dependence might be unrealistic in some applied modeling situations. In cases where the structure of dependence or connectivity between (cross-sectional) observations arises from non-spatial relationships, spatial weights are theoretically unjustifiable. Some literature addresses the structure of dependence between observations by introducing geographic proximity as well as other types of non-spatial proximity, resulting in a model that utilizes multiple weight matrices. Each set of weights reflect a different type of dependence specified using linear combinations of observations defined by alternative characteristics. The multiple weight matrix approach results in a simultaneous autoregressive process that poses a number of challenges for parameter estimation and model interpretation.

We focus on literature that relies on a single connectivity matrix constructed from a convex combination of multiple matrices, each of which reflects a different type of dependence or interaction structure. The advantage of this approach is that the resulting simultaneous autoregressive process is amenable to conventional spatial regression estimation algorithms as well as methods developed for interpretation of estimates from these models. Estimates of the scalar parameters used to form the convex combination of the weight matrices can be used to produce an inference regarding the relative importance of each type of dependence.

KEYWORDS: Markov Chain Monte Carlo estimation, Taylor series approximation, log-marginal likelihood, multiple weight matrices.

## 1 Introduction

Spatial regression models typically rely on spatial proximity or Euclidean distance between observations to specify the structure of *simultaneous dependence* between observations. For example, neighboring regions that have common borders with each region are defined as those on which each observation is dependent. Alternatively, dependent observations for each region can be defined as a function of inverse distance to other observations, perhaps up to some cut-off distance. However, this is often viewed as restrictive

and inappropriate in applications where spatial proximity or distance is theoretically unjustifiable as a basis for the underlying dependence structure of the *simultaneous autoregressive process* (Corrado and Fingleton, 2012). This criticism is most relevant when spatial regression models are used in broader contexts to model spatial phenomena involving interregional flows of goods or population, knowledge transmission between regions, or non-spatial phenomena such as student peer groups or social networks. In these scenarios, the spatial location of observations may not be theoretically appealing as a basis for the underlying structure of dependence between observations.

A *simultaneous autoregressive process* models variation in an  $N \times 1$  variable vector  $y$  using other observations in the vector  $y$ . An  $N \times N$  matrix  $W$  is used to define specific observations on which each observation is dependent. The matrix  $W$  has zeros on the diagonal to prevent an observation from being dependent on itself, and is typically row-normalized to have row-sums of one. This type of normalization aids with interpretation of the model, allowing us to compare the magnitude of scalar dependence parameter estimates arising from models specified with alternative definitions of the matrix  $W$ . A simultaneous autoregressive process specifies a relationship between the vector of observations in  $y$  and the  $N \times 1$  vector  $Wy$ , resulting from the product of the  $N \times N$  matrix with the  $N \times 1$  vector  $y$ . The relationship takes the form:  $y = \rho Wy + \varepsilon$ , where the scalar parameter  $\rho$  defines the strength of dependence between the  $N \times 1$  vector  $y$  and  $N \times 1$  vector  $Wy$ . In the case of row-normalized  $W$ , the relationship is between observations in the vector  $y$  and a linear combination of observations on which each observation is dependent. The  $N \times 1$  vector  $\varepsilon$  is a stochastic disturbance term, typically assumed to be normally distributed with zero mean and a scalar variance,  $\sigma^2$ . We focus on maximum likelihood-based estimation procedures where normality is important, but alternative estimation methods, such as the generalized method of moments do not require normally distributed disturbances. In situations where the dependence structure in the matrix  $W$  is defined based on spatial aspects of a sample of observations located in space, we can label this a *spatial autoregressive process*.

One can also define a simultaneous autoregressive process involving the model disturbances, for example,  $y = \mu \iota_N + u$ ,  $u = \lambda Wu + \varepsilon$  where the scalar  $\mu = E(y)$ ,  $\iota_N$  is an  $N \times 1$  vector of ones, and  $u$  is a  $N \times 1$  set of disturbances. Here, the disturbances exhibit dependence on disturbances from other observations identified by the matrix  $W$ , with the strength of dependence defined by the scalar parameter  $\lambda$ . Generally, simultaneous autoregressive processes model outcomes ( $y$ ) or disturbances ( $u$ ) as related to a linear combination of other observations defined by the dependence structure. This could involve

a scenario where each observation is possibly dependent on a different number of observations in the *dependence set*, with possibly different weights assigned to each of these sets of dependent observations.

Simultaneous autoregressive processes can be used to extend ordinary regression models, producing specifications such as:  $y = \rho W y + X\beta + \varepsilon$ , which has been labeled a *spatial autoregressive model* (SAR), where  $X$  is an  $N \times K$  matrix of explanatory variables and  $\varepsilon$  an  $N \times 1$  vector of normally distributed zero mean, constant variance disturbances. Of course, this type of model specification can be viewed more generally as a *simultaneous autoregressive model* when the application involves non-spatial observations. Another spatial regression specification is the spatial error model (SEM),  $y = X\beta + u, u = \lambda W u + \varepsilon$ , where the  $N \times 1$  vector of disturbances  $u$  follow a spatial autoregressive process. Matrix-vector products such as  $W y, W u$  are referred to as *spatial lags*, and we can also use a matrix-matrix product  $W X$  to create spatial lags of the  $K$  explanatory variables of the model. In the case of non-spatial  $W$  matrices, we label these *simultaneous lags*, where  $W y, W u, W X$  simply reflect linear combinations of observations from the *dependence set* of observations. These reflect a linear combination of explanatory variables from dependent observations specified by the matrix  $W$ . These additional explanatory variables can enter the spatial regression model:  $y = \rho W y + X\beta + W X\theta + \varepsilon$ , which has been labeled a spatial Durbin model (SDM). In a spatial setting where the dependence structure reflects nearby or neighboring regions, this model indicates that explanatory variable values of neighboring regions may help explain variation in own-region outcomes reflected in the  $y$ -vector. More generally, in situations where the matrix  $W$  reflects non-spatial dependence relations, the set of explanatory variables defined by  $W X$  have been labeled *contextual variables*, reflecting the notion that a linear combination of explanatory variables from observations in the *dependence set* describe a *context* that may impact outcomes in the vector  $y$ .

One important issue in the spatial regression literature concerns the appropriateness of the connectivity matrices  $W$  to define the underlying simultaneous autoregressive process. An intuitive motivation for reliance on non-spatial weight matrices arises naturally when one considers that spatial dependence relies on the notion of Euclidean distance or spatial proximity to define observations in the dependence set. Euclidean distance can be generalized to define a dependence set of observations based on other metrics. For example, Pace et al. (2000) propose a model for prices of homes sold that occur at irregular points in space and time, generalizing distance to include relative locations in time. Related work by Pace et al. (2002) relied

on generalized distances that considered the number of bedrooms and bathrooms (of nearby homes) to specify the structure of dependence between the selling prices of homes, with the motivation that appraisers determine selling price estimates based on homes comparable in these two metrics.

Once we open the door to non-spatial metrics as a way to specify dependence between cross-sectional observations, a host of issues arise, which are discussed in LeSage and Pace (2011), and Debarsy and LeSage (2018a). Blankmeyer et al (2011, p.94) point out that “*a single weight matrix, based on a multivariate similarity criterion (generalized distance) requires a norm to prevent scale differences from influencing the weight placed on the various measures of similarity. (This is unlike the case of spatial proximity where Euclidian distance provides a natural scaling)*”.

## 1.1 Multiple connectivity matrices

An approach that avoids relative scaling issues involves extensions of the SAR specification to include more than one spatial lag (see, among others Badinger and Egger, 2011; Lacombe, 2004; Lee and Liu, 2010; Elhorst et al., 2011).

These models extend our SAR specification to include multiple spatial or simultaneous lags, each one based on a different connectivity matrix, in an effort to capture different types of dependence between cross-sectional observations using the simultaneous autoregressive specification. The additional simultaneous lags of the dependent variable vector  $y$  have been labeled ‘higher-order’ terms, with this type of extended SAR model shown in Eq. (1), where  $L$  simultaneous lag terms are introduced.

$$y = \left( \sum_{\ell=1}^L \rho_{\ell} W_{\ell} \right) y + X\beta + \varepsilon \quad (1)$$

In Eq. (1),  $y$  is an  $N \times 1$  vector of dependent variable outcomes,  $X$  is an exogenous  $N \times K$  explanatory variables matrix, with  $\beta$  the associated  $K \times 1$  parameters,  $\rho_{\ell}, \ell = 1, \dots, L$  are scalar dependence parameters measuring the strength of each type of dependence modeled by the  $N \times N$  dependence matrices  $W_{\ell}, \ell = 1, \dots, L$ .

## 1.2 Convex combinations of multiple weight matrices

LeSage and Pace (2011) point notably to a number of estimation issues that arise for higher-order models of the type in Eq. (1). In addition, Elhorst et al. (2011) point to estimation complications that arise relating to

the parameter space for the dependence parameters  $\rho_\ell, \ell = 1, \dots, L$ . The parameter space for  $\rho_1$  depends on values taken by the other parameters  $\rho_j, j = 2, \dots, L$ , and similarly, the parameter space for  $\rho_2$  depends on values taken by  $\rho_1, \rho_j, j = 3, \dots, L$ , and so on. They provide a solution based on a trigonometric approach to this problem for the case of two  $W$ -matrices, that would allow maximum likelihood estimation to be carried out. However, they note that extension of their approach to models containing more than two weight matrices becomes increasingly difficult in situations where the number of weight matrices increases.

An alternative approach that relies on convex combinations of different dependence matrices to form a single weight matrix, was first explored by Pace and LeSage (2002). They proposed use of a model shown in Eq. (2), where a convex combination of multiple  $N \times N$  primitive spatial basis matrices  $B^{(i)}, i = 1 \dots p$ , each based on a differing sequence of individual *doubly stochastic* nearest neighbor weight matrices is used to form a single weight matrix,  $W_c = \sum_{i=1}^p a_i B^{(i)}$ . The model parameters are subject to restrictions set forth in (3) and (4). (A *doubly stochastic* matrix is one whose rows and columns sum to unity.) This leads to a spatial autoregressive model in Eq. (2).

$$y = \rho W_c y + X\beta + \varepsilon \quad (2)$$

$$W_c = \sum_{i=1}^p a_i B^{(i)} \quad (3)$$

$$0 \leq a_i \leq 1, i = 1 \dots p$$

$$\sum_{i=1}^p a_i = 1 \quad (4)$$

The model specification in Eq. (2) overcomes the interpretation problem noted for the *higher-order* spatial autoregressive model because the estimates for the parameters  $\sum_{i=1}^p a_i = 1$  can be used to draw inferences regarding the relative importance of the different spatial basis matrices  $B^{(i)}$  in the model. It also remedies the problem regarding the parameter space for  $\rho$  because  $W_c$  is a row-normalized weight matrix when individual weight matrices  $B^{(i)}$  are row-normalized, which is of course true of the *doubly stochastic* weights  $B^{(i)}$ . We note that a convex combination of row-normalized weight matrices is also row-normalized. This means the parameter space for  $\rho$  would obey the standard lower and upper bounds of  $1/\mu_{\min}$  and 1, where  $\mu_{\min}$  is the smallest eigenvalue of the matrix  $W_c$ . In typical estimation algorithms

for spatial autoregressive models the value of  $1/\mu_{\min}$  is set to  $-1$ , since this ensures that the matrix inverse:  $(I_N - \rho W_c)^{-1}$  exists, and avoids the need to calculate the minimum eigenvalue of the  $N \times N$  weight matrix  $W_c$ .

Hazir et al. (2018) extend this approach to the more general case of a convex combination of  $L$  row-normalized  $N \times N$  weight matrices based on alternative underlying dependence sets for the  $N$  observations reflected by underlying matrices  $W_\ell$ , so that  $W_c(\Gamma) = \sum_{\ell=1}^L \gamma_\ell W_\ell$ , with  $0 \leq \gamma_\ell \leq 1, \ell = 1 \dots L, \sum_{\ell=1}^L \gamma_\ell = 1$  and  $\Gamma = (\gamma_1, \dots, \gamma_L)'$ . The resulting convex combination of underlying weight structures,  $W_c(\Gamma)$ , can be used to specify dependence between  $N$  observations based on a convex combination of  $L$  different types of dependence between observations. The scalar parameters  $\gamma_\ell$  indicate the relative importance assigned to each type of dependence in the cross-sectional dependence scheme. The two sets of constraints imposed reflect the fact that this approach relies on a convex combination. When each  $W_\ell, \ell = 1, \dots, L$ , is row-normalized, then  $W_c(\Gamma)$  obeys the conventional row-normalization, avoiding the need to re-normalize the matrix for new values of  $\Gamma$ . We note that the approach based on a convex combination of weight matrices also avoids the issue of scaling for different metrics that arises for generalized measures of distance by casting the problem as one of relative distance/proximity, inherent in conventional spatial regression weight matrices.

To estimate the model parameters  $\rho, \beta, \Gamma, \sigma^2$ , Hazir et al. (2018) propose evaluating the likelihood numerous times over a grid of values for the parameters  $\gamma_\ell, \ell = 1 \dots L$ , and using the vector of parameters (that we denote  $\hat{\Gamma}$ ) associated with the highest likelihood function value as point estimates for these parameters. For example, in the case of two weight matrices where we have values of  $\gamma_1$  and  $\gamma_2 = (1 - \gamma_1)$ , they would define a grid of values for  $\gamma_1 = 0, 0.1, 0.2, \dots, 1$ , and perhaps refine the grid based on an initial estimate from the course grid of (say) 0.2 to be a grid defined over 0.19, 0.20, 0.21,  $\dots$ , 0.29. The value of  $\hat{\gamma}_1$  from this grid of values would be used in the second step of the estimation procedure. The second step would produce estimates for the other parameters  $\hat{\rho}, \hat{\beta}, \hat{\sigma}^2$  along with measures of dispersion for these parameters by simply fixing the parameters  $\hat{\Gamma}$  and generating  $W_c(\hat{\Gamma}) = \hat{\gamma}_1 W_1 + \hat{\gamma}_2 W_2 + \dots + \hat{\gamma}_L W_L$ . Treating  $W_c(\hat{\Gamma})$  as fixed, conventional maximum likelihood algorithms are used to estimate  $\rho, \beta, \sigma^2$  and associated measures of dispersion. This approach has the virtue of simplicity, since it allows use of conventional maximum likelihood estimation software to estimate the parameters  $\hat{\Gamma}$  in the first step, and the parameters  $\rho, \beta, \sigma^2$  in the second step. An undesirable aspect of this approach is the



lack of dispersion estimates for the point estimates of the parameters  $\hat{\Gamma}$  on which to draw inferences regarding whether individual  $\gamma_\ell$  parameters are significantly different from zero.

### 1.3 Bayesian estimation of convex combinations of weight matrix models

To overcome the problem of inference regarding individual parameters  $\gamma_\ell$ , Debarsy and LeSage (2018a) propose *Markov Chain Monte Carlo* (MCMC) estimation of the SAR model with convex combination of connectivity matrices. MCMC estimation involves sampling for the complete sequence of conditional distributions of the model parameters. A large sample of draws from the conditional distribution of each parameter is used to construct the *joint posterior distribution* for the model parameters. In the case of only two weight matrices,  $\Gamma = (\gamma_1, (1 - \gamma_1))'$ , this involves sampling from the following sequence of conditional distributions:  $(\rho|\beta, \sigma^2, \gamma_1)$ ,  $(\beta|\rho, \sigma^2, \gamma_1)$ ,  $(\sigma^2|\rho, \beta, \gamma_1)$ , and  $(\gamma_1|\rho, \beta, \sigma^2)$ . A virtue of MCMC sampling is that the conditional distributions for each parameter given values of all other model parameters typically takes a simple form that is computationally easy to sample from. As an example, the conditional distribution for the  $K \times 1$  parameter vector  $\beta|\rho, \sigma^2, \gamma_1$  takes the form of a  $K$ -variate normal distribution with a mean vector and variance-covariance matrix shown in Eqs. (5) and (6) (when no prior distributions are assigned to the model parameters). When we assign no prior distributions to the model parameters, MCMC sampling serves as simply an estimation method for producing parameter estimates that will be equivalent to those from maximum likelihood estimation. The motivation for use of MCMC estimation in the case of no Bayesian prior information is simply to exploit the computationally simple forms taken by the conditional distributions to make estimation a simpler task than that of optimizing a likelihood function with respect to a large number of parameters. This approach is sometimes labeled *simulated maximum likelihood* estimation.

$$\beta|\rho, \sigma^2, \gamma_1 \sim \mathcal{N}(\bar{\beta}, \bar{\Sigma}_\beta)$$

$$\bar{\beta} = (X'X)^{-1}X'[I_N - \rho W_c(\gamma_1)]y \quad (5)$$

$$\bar{\Sigma}_\beta = \sigma^2(X'X)^{-1} \quad (6)$$

For the case of only two weight matrices involving the parameter  $\gamma_1$  (and  $\gamma_2 = 1 - \gamma_1$ ), the conditional posterior for  $\gamma_1$  (given  $\beta, \rho, \sigma^2$ ) can be written as in Eq. (7), where  $|A|$  denote the determinant of the matrix  $A$ .

$$\begin{aligned}
p(\gamma_1|\rho, \beta, \sigma^2) &\propto |A|\exp(-1/2\sigma^2)(Ay - X\beta)'(Ay - X\beta) & (7) \\
A &= [I_N - \rho W_c(\gamma_1)]
\end{aligned}$$

A Metropolis-Hastings (M-H) approach can be used to draw samples of the parameter  $\gamma_1$  (see LeSage and Pace, 2009, chap. 5). This involves comparing the conditional distribution in Eq. (7) based on the current value of  $\gamma_1$  to a proposed value, and accepting or rejecting the proposal according to the M-H rules. The restriction that  $0 \leq \gamma_1 \leq 1$  is imposed by rejecting values proposed for  $\gamma_1$  that do not obey the restriction.

A similar conditional distribution arises for the parameter  $\rho$ , shown in Eq. (8), which is also amenable to M-H sampling. Here again, the restriction  $-1 < \rho < 1$  can be imposed by rejecting proposal values that do not obey the restriction.

$$\begin{aligned}
p(\rho|\gamma_1, \beta, \sigma^2) &\propto |A|\exp(-1/2\sigma^2)(Ay - X\beta)'(Ay - X\beta) & (8) \\
A &= [I_N - \rho W_c(\gamma_1)]
\end{aligned}$$

LeSage and Pace (2009) set forth a random-walk procedure for producing M-H proposal values for the parameter  $\rho$ , which can also be used for the parameter  $\gamma_1$ . Debarsy and LeSage (2018a) actually take a different approach to sampling the parameters  $\rho, \gamma_1$ , labeled a *griddy Gibbs* sampler by LeSage and Pace (2009, chap. 5), which should produce equivalent MCMC draws from which the joint posterior estimates for the parameters  $\gamma_1, \rho$  can be constructed.

Simple extension of the approach set forth above to cases involving more than two weight matrices encounters some difficulties. One is that each parameter  $\gamma_i$  must be sampled conditional on all other parameters  $\gamma_j, j \neq i$ . Another is that the adding up constraint  $\sum_{\ell=1}^L \gamma_\ell = 1$  cannot be imposed when we treat  $\gamma_j, j \neq i$  as known and fixed. Quite simply a proposal value for  $\gamma_i$  requires that we would need to adjust  $\gamma_j, j \neq i$  to be consistent with the summing up constraint, which rules out treating these as fixed. Of course, this issue of the adding up constraint does not arise in the case of two weight matrices where  $\gamma_2 = (1 - \gamma_1)$ , since the constraint holds for all values  $0 \leq \gamma_1 \leq 1$ . Further, the Jacobian of the transformation to be assessed at each pass through the sampler cannot be pre-computed for a grid of values for  $\rho$  and  $\Gamma$ .

Debarsy and LeSage (2018b) propose an approach that generalizes to specifications involving more than two matrices. Their approach involves

rewriting the convex combination SAR model presented in Eq. (2) as in Eq. (9), which they argue has a number of computational advantages. One advantage is that the parameter vector  $\omega$  isolates the parameters  $\rho, \gamma_\ell, \ell = 1, \dots, L$  from sample data matrices. Separating model parameters from vectors and matrices that contain  $N$  sample data observations allows calculation of matrices such as the  $N \times (L + 1)$  matrix  $\tilde{y}$  once, prior to beginning the MCMC sampling loop. More regarding this later.

$$\begin{aligned} \tilde{y}\omega &= X\beta + \varepsilon \\ \tilde{y} &= (y, W_1y, W_2y, \dots, W_Ly) \\ \omega &= \begin{pmatrix} 1 \\ -\rho\gamma_1 \\ -\rho\gamma_2 \\ \vdots \\ -\rho\gamma_L \end{pmatrix} = \begin{pmatrix} 1 \\ -\rho\Gamma \end{pmatrix}, \Gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_L \end{pmatrix}, \quad 0 \leq \gamma_\ell \leq 1, \ell = 1 \dots L \end{aligned} \tag{9}$$

## 2 Computational issues in estimating convex combination of weights models

Debarsy and LeSage (2018b) point to three computational challenges that arise for the convex combination of weight matrices model. One is the need to impose the adding up constraint and the non-negativity requirement for the parameters  $\sum_{\ell=1}^L \gamma_\ell = 1$ , which would require use of a constrained optimization algorithm. In the case of MCMC sampling, there is the issue of sampling a single parameter  $\gamma_i$  conditional on other parameters  $\gamma_j, j \neq i$ , while maintaining the adding up constraint and the non-negativity requirement. A second is the need to evaluate a log-determinant term that arises in the likelihood and the conditional distributions for the spatial dependence parameters  $(\rho, \Gamma)$  of the model. The log-determinant depends on both  $\rho$  as well as the vector of parameters  $\Gamma$ , so changes in the values of these require calculation of new values for the log-determinant  $|A| = |I_N - \rho W_c(\Gamma)|$  on each parameter draw from the conditional distributions of  $\rho, \Gamma$ . A third problem they note, involves calculating simulated (empirical) measures of dispersion for the partial derivatives  $\partial y / \partial x$  used to interpret estimates from the model. We discuss the nature of these problems in more detail along with solutions proposed by Debarsy and LeSage (2018b) in the next three sections.

## 2.1 MCMC sampling of the parameters $\Gamma$

Debarys and LeSage (2018b) propose block sampling the entire vector of parameters  $\Gamma$ , using a *Metropolis-Hastings* (M-H) approach. This means that a vector of all  $\gamma_\ell$  parameters, labeled  $\Gamma^p$  is proposed and either accepted or rejected using the M-H rule. If rejected, the current vector of all  $\gamma_\ell$  parameters, labeled  $\Gamma^c$ , is retained for the next pass through the MCMC sampler. The virtue of this block sampling approach is that the proposed vector  $\Gamma^p$  can be constructed to ensure the adding up constraint and non-negativity requirements are met.

M-H sampling requires that we evaluate the conditional distribution for the parameters  $\Gamma$  at the current and proposed values for the vectors  $\Gamma^p, \Gamma^c$ . Debarys and LeSage (2018b) show that by integrating out  $\beta$  and  $\sigma^2$ , we can write the conditional distribution of  $\Gamma$  conditioning only on  $\rho$ . As such, the M-H procedure to accept a new block  $\Gamma^p$  uses the conditional distribution of  $\gamma$ , assuming  $\rho$  is given or known. An acceptance probability is calculated based on evaluating the expression shown in Eq. (10) and used to decide if the proposed vector  $\Gamma^p$  is accepted as the new set of parameters.

$$\psi_{M-H}(\Gamma^c, \Gamma^p) = \min(1, \exp[(\log p(\Gamma^p|\rho) - \log p(\Gamma^c|\rho)])] \quad (10)$$

$$\log p(\Gamma|\rho) \propto \log|I_N - \rho W_c(\Gamma)| - \frac{N}{2} \log[\omega' F \omega] \quad (11)$$

$$F = (\tilde{y} - X\beta_d)'(\tilde{y} - X\beta_d)$$

$$\beta_d = (X'X)^{-1}(X'\tilde{y}) \quad (12)$$

We note that  $F$  consists of only sample data, so this expression can be calculated prior to MCMC sampling, leading to a computationally efficient expression reflecting a quadratic form:  $\log(\omega' F \omega)$ , that can be easily evaluated for any vector of dependence parameters  $\omega$ . Since  $\omega$  involves both  $\rho$  and  $\Gamma$ , we can also sample the conditional distribution for  $\rho$  assuming the current vector of values  $\Gamma^c$  is known using the same computationally efficient expression involving the quadratic form.

Of course, the conditional distribution also involves the log-determinant  $\log|I_N - \rho W_c(\Gamma)|$  which depends on both  $\rho$  and the vector  $\Gamma$ , and needs to be calculated for any given values of these parameters in order to evaluate the M-H acceptance probability in Eq. (10), which is the subject of the next section.

## 2.2 An approximation to the log-determinant

Pace and LeSage (2002) set forth a Taylor series approximation for the log-determinant  $\log|I_N - \rho W_c(\Gamma)|$ , which appears in the conditional distribution needed to sample both  $\rho$  and the parameters in the vector  $\Gamma$ . They show that for a nonnegative weight matrix  $W_c(\Gamma)$  with eigenvalues  $\mu_{\min} \geq -1, \mu_{\max} \leq 1$ , and  $1/\mu_{\min} < \rho < 1$ :

$$\log|I_N - \rho W_c(\Gamma)| = - \sum_{j=1}^{\infty} \frac{\rho^j \text{tr} W_c(\Gamma)^j}{j} \quad (13)$$

$$\simeq - \sum_{j=1}^q \frac{\rho^j \text{tr}(W_c(\Gamma)^j)}{j} \quad (14)$$

where  $\text{tr}$  represents the trace operator. Golub and van Loan (1996, p.566) provide the expression in Eq. (13), while Pace and LeSage (2002) note that due to the linearity of the trace operator we have the  $q$ th order approximate expression in Eq. (14). We note that the 1st-order trace,  $\text{tr}(W_c(\Gamma))$ , is zero for any convex combination of weight matrices that have zero diagonal elements. Since the underlying weight matrices  $W_\ell$  have zero diagonal elements, a convex combination of these will also have this property. The second-order trace can be computed as in Eq. (15), and there are similar expressions for higher-order traces shown in Eqs. (16) and (17). Debarsy and LeSage (2018b) carry out Monte Carlo experiments to show that estimates based on  $q = 4$ th order trace approximation have desirable properties, including low bias and good coverage of the true parameter values.

$$\text{tr}(W_c(\Gamma)^2) = \sum_{i=1}^L \sum_{j=1}^L \gamma_i \gamma_j \text{tr}(W_i W_j) \quad (15)$$

$$\text{tr}(W_c(\Gamma)^3) = \sum_{i=1}^L \sum_{j=1}^L \sum_{k=1}^L \gamma_i \gamma_j \gamma_k \text{tr}(W_i W_j W_k) \quad (16)$$

$$\text{tr}(W_c(\Gamma)^4) = \sum_{i=1}^L \sum_{j=1}^L \sum_{k=1}^L \sum_{\ell=1}^L \gamma_i \gamma_j \gamma_k \gamma_\ell \text{tr}(W_i W_j W_k W_\ell) \quad (17)$$

The computationally desirable aspect of these expressions is that the parameters  $\gamma_\ell$  are separated from the trace calculations involving the  $N \times N$  dimensional matrices. This allows the trace calculations to be carried out

prior to MCMC sampling draws. During MCMC sampling, new values for the parameters  $\Gamma$  and  $\rho$  are used in conjunction with the pre-calculated traces to rapidly evaluate the log-determinant term.

### 2.3 Efficient calculation of partial derivatives for the model

In addition to producing estimates for the underlying model parameters  $\rho, \beta, \Gamma, \sigma^2$ , we need to compute the reduced form of the model to compute the partial derivatives of  $y$  with the relevant exogenous determinant to interpret its effect. For the standard simultaneous autoregressive model, LeSage and Pace (2009) have derived these partial derivatives and further develop scalar summary measures to interpret them. For the simultaneous autoregressive model with a convex combination of connectivity matrices, the partial derivative of the outcome variable with respect to the  $r^{th}$  determinant is shown in Eq. (18) while the scalar summary measures, which represent own- and cross-partial derivatives that they label Eq. *direct* and Eq. *indirect* effects, are shown in Eqs. (19) and (21).

$$\partial y / \partial x_r = S_r(W_c(\Gamma)) \quad (18)$$

$$\begin{aligned} S_r(W_c(\Gamma)) &= (I_N - \rho W_c(\Gamma))^{-1} \beta_r \\ &= I_N \beta_r + \rho W_c(\Gamma) \beta_r + \rho^2 W_c(\Gamma)^2 \beta_r + \dots \end{aligned} \quad (19)$$

$$\bar{M}(r)_{direct} = N^{-1} tr(S_r(W_c(\Gamma))) \quad (20)$$

$$\bar{M}(r)_{total} = n^{-1} \iota'_N S_r(W_c(\Gamma)) \iota_N \quad (21)$$

$$\bar{M}(r)_{indirect} = \bar{M}(r)_{total} - \bar{M}(r)_{direct}$$

$$W_c(\Gamma) = \sum_{\ell=1}^L \gamma_\ell W_\ell$$

While expressions in Eqs. (19), (20) and (21) produce point estimates for the scalar summary measures of effects (own- and cross-partial derivatives) used to interpret the impact of changes in explanatory variables on dependent variable outcomes, we also require measures of dispersion for the purpose of statistical tests regarding the significance of these effects. Use of an empirical distribution constructed by simulating the non-linear expressions in Eq. (18) using (say 1,000) draws from the posterior distribution of the underlying parameters  $\rho, \beta_r, \gamma_\ell, \ell = 1, \dots, L$  is suggested by LeSage and Pace (2009).

A naive approach to such a simulation-based empirical distribution would require calculation of the  $N \times N$  matrix inverse  $S_r(W_c(\Gamma))$  a large number

(say 1,000) times, for varying values of the parameters  $\rho, \beta_r, \gamma_\ell, \ell = 1, \dots, L$ , which would be very compute intensive. For the SAR model with only one connectivity matrix,  $W$ , LeSage and Pace (2009) show that the required quantity for constructing the empirical distribution of the effects is  $\text{tr}(S_r(W))$ , which can be estimated without a great deal of computational effort. Debarsy and LeSage (2018b) extend this approach using estimated traces in conjunction with the four first orders traces evaluated at each pass through the sampler to approximate the log-determinant term. This allows the simulation-based empirical distribution of the direct, indirect and total effects estimates to account for uncertainty in  $\Gamma$ , the parameters of the convex combination. It is important to take into account the stochastic uncertainty that arises as a result of these parameters and the additional variation/uncertainty they contribute to the *effects estimates* because ignoring it leads to under-coverage of the other parameters (too narrow standard errors).

### 3 Bayesian model averaging for multiple models

The convex combination of weights model specification raises questions regarding which weight matrices should be used in these models. For a modeling situation where  $L$  weight matrices are considered as candidates for the model specification, there are  $M = 2^L - L - 1$  possible ways to employ *two or more* of the  $L$  weight matrices in alternative model specifications. When  $L = 5$ , we have  $M = 26$  possible models involving two or more weight matrices, and for  $L = 10$ ,  $M = 1,013$ . This makes a method for comparing model estimates based on alternative combinations of weight matrices an important issue. Debarsy and LeSage (2018b) take a Bayesian approach to deriving the log-marginal likelihood for these models. Given the log-marginal likelihood for a model  $M_i$ , that we denote  $\text{Log}M_i$ , we can calculate its associated posterior probability, namely  $\text{prob}(M_i) = \exp(\text{Log}M_i / \sum_{i=1}^Q \text{Log}M_i)$  (in the case of  $Q$  different models). The highest  $\text{prob}(M_i)$  answers the question — which model is most consistent with the sample data  $(y, X, W_1, W_2, \dots, W_L)$ ? — *unconditional* on any parameter value/estimates. The answer is unconditional with respect to the parameter values because the parameters have been integrated out of the joint likelihood to produce the log-marginal likelihood and associated model probabilities.

Debarsy and LeSage (2018b) combine analytical and numerical strategy to produce *estimates of the log-marginal likelihood*. This involves analytically integrating out the parameters  $\beta, \sigma^2$  from the joint likelihood to produce a

joint posterior distribution for the dependence parameters  $\rho, \Gamma$ . They then rely on Metropolis-Hastings guided Monte Carlo integration during MCMC estimation of the models to produce estimates of log-marginal likelihoods for each model.

The Bayesian solution to the problem of uncertainty regarding which model specification is correct involves model averaging, where estimates from all candidate models are weighted by the posterior model probability. This means that model averaged estimates incorporate uncertainty regarding the model parameters as well as uncertainty regarding the model specification.

To illustrate this approach, we use a sample of 856 nursing homes located in Texas taken from Blankmeyer et al. (2011). They argue that CEO compensation depends on that of Chief Executive Officers (CEO's) from peer nursing homes, where peers can be defined using alternative weight matrices. Specifically, we consider a spatial weight matrix based on 3-nearest spatial neighbors ( $W_{space}$ ), a size similarity matrix based on three nursing homes with the most similar square foot area ( $W_{size}$ ), and a payroll similarity matrix constructed from the three nursing homes with the most similar payroll magnitudes ( $W_{pay}$ ). We note that peer nursing homes defined using the size and payroll measures could be located anywhere in the state of Texas, which comprises the sample of 856 nursing homes.

The explanatory variables are the share of medicaid and private pay patients, the occupancy rate, and dummy variables indicating for-profit nursing homes and nursing homes that are part of a chain operation. Blankmeyer et al (2011) indicate these are criterion on which the executive board evaluates the CEOs managerial performance.

Given  $L = 3$  different candidate weight matrices, we have  $M = 4$  possible models involving two or more weight matrices. The log-marginal likelihoods, model probabilities and estimates for the parameters  $\rho, \gamma_1, \gamma_2, \gamma_3$  are shown in Table 1 for the  $M = 4$  possible models, along with Bayesian model averaged estimates (BMA).

The highest model probability of 0.659 points to *Model 3* that includes  $W_{payroll}$  (three most similar payroll peers) and  $W_{size}$  (square foot area size-based peers), placing no weight on the model that includes  $W_{space}$  (the 3-nearest spatial neighboring nursing homes). The second highest log-marginal likelihood value is associated with *Model 1* that has a model probability of 0.196 and includes  $W_{space}$  and  $W_{payroll}$  only, excluding the area-based size peers  $W_{size}$ . The BMA estimates will be based on MCMC draws from *Model 1*, *Model 3*, and *Model 4* which have non-zero posterior probabilities (0.196, 0.659, 0.145), respectively. These are reported in Table 2.



Table 1: Model probabilities for models involving two or more weight matrices

Models	LogM	Prob	$\rho$	$\gamma_1, W_{space}$	$\gamma_2, W_{payroll}$	$\gamma_3, W_{size}$
Model 1	-455.822	0.196	0.506	0.102	0.898	—
Model 2	-560.499	0.000	0.144	0.363	—	0.637
Model 3	-454.611	0.659	0.518	—	0.871	0.129
Model 4	-456.126	0.145	0.567	0.090	0.792	0.117
BMA	-455.068	1.000	0.523	0.033	0.865	0.102

From the table, we see evidence of significant dependence on peer institutions in explaining variation in CEO compensation, the dependent variable, as evidenced by the estimate for  $\rho = 0.5225$ , which is significant. The most important peer nursing homes for determining CEO compensation are other nursing homes with similar payroll magnitudes indicated by the estimate for  $\gamma_2 = 0.8644$ . This indicates that management of staff and nursing personnel is viewed as the most important basis for determining a peer institution by the executive board that evaluates and determines the CEO's compensation. Larger payrolls point to a larger CEO role in managing personnel at the nursing home.

The estimate of  $\gamma_3 = 0.1031$  for the weight matrix associated with the matrix  $W_{size}$  based on the physical size of the nursing home is the next most important basis for determining a peer institution, with a very small the role played by spatial proximity, since  $\gamma_1 = 0.0329$ , with a lower 0.01 credible interval of 0.0074, very close to zero.

Turning to the *effects estimates*, we see that a larger number of Medicaid and private pay patients in the nursing home (which are considered less desirable than Medicare patients) has a negative *direct impact* on CEO compensation. By direct impact, we mean that the own-institution characteristic impacts the own-institution CEO compensation, in a negative way for this variable. Higher occupancy rates have a positive impact on CEO compensation, and Blankmeyer et al (2011) point to low occupancy rates as a problem for Texas nursing homes. The dummy variable selecting *For profit* institutions has a positive and significant impact on CEO compensation, indicating that these institutions have higher compensation levels than non-profit institutions, many of which are affiliated with religious institutions. Finally, the dummy variable selecting nursing homes that are part of a chain of homes administered by a single company, has a positive direct

Table 2: Model averaged estimates

Variables	Lower 0.01	Lower 0.05	Median	Upper 0.95	Upper 0.99
Constant	3.8810	4.0613	4.6649	5.2588	5.4205
Medicaid+Private	-0.2191	-0.1978	-0.1178	-0.0417	-0.0180
Occupancy rate	0.7193	0.7494	0.8234	0.9009	0.9224
For profit	0.0342	0.0452	0.0807	0.1185	0.1296
Chain	0.0044	0.0130	0.0460	0.0813	0.0933
$\rho$	0.4544	0.4696	0.5225	0.5797	0.5973
$\gamma_1, W_{space}$	0.0074	0.0110	0.0329	0.0557	0.0629
$\gamma_2, W_{payroll}$	0.7773	0.7965	0.8644	0.9366	0.9496
$\gamma_3, W_{size}$	0.0215	0.0355	0.1031	0.1658	0.1847
Direct	Lower 0.01	Lower 0.05	Median	Upper 0.95	Upper 0.99
Medicaid+Private	-0.2366	-0.2131	-0.1276	-0.0454	-0.0196
Occupancy rate	0.7839	0.8117	0.8904	0.9711	1.0014
For profit	0.0373	0.0489	0.0871	0.1283	0.1401
Chain	0.0047	0.0138	0.0499	0.0878	0.1016
Indirect	Lower 0.01	Lower 0.05	Median	Upper 0.95	Upper 0.99
Medicaid+Private	-0.2344	-0.2039	-0.1192	-0.0434	-0.0150
Occupancy rate	0.6346	0.6804	0.8416	1.0602	1.1607
For profit	0.0362	0.0462	0.0825	0.1297	0.1466
Chain	0.0057	0.0135	0.0477	0.0880	0.1006
Total	Lower 0.01	Lower 0.05	Median	Upper 0.95	Upper 0.99
Medicaid+Private	-0.4571	-0.4141	-0.2471	-0.0907	-0.0337
Occupancy rate	1.4662	1.5341	1.7364	1.9897	2.1080
For profit	0.0764	0.0960	0.1709	0.2569	0.2808
Chain	0.0113	0.0272	0.0982	0.1735	0.1973

impact on CEO compensation.

The indirect effects in this model reflect *peer effects*, such that compensation of CEO's at nursing home  $i$  is influenced by compensation levels for CEOs at peer nursing homes  $j \neq i$ . The median estimates for indirect effects are similar in magnitude to the direct effects, indicating that compensation of other nursing home CEOs plays an equally important role as the

own-institution characteristics in determining compensation for the typical CEO in our sample. The total effects are of course the sum of direct plus peer/indirect effects.

## 4 Conclusion

The simultaneous autoregressive model specified using a convex combination of dependence matrices represents one solution to the important problem of selecting an appropriate *dependence set* of observations for use in modeling cross-sectional dependence. Using a specification based on a convex combination of dependence matrices allows formal Bayesian tests for the relative importance of alternative definitions of the dependence set of observations. The specification also allows a combination of several weight matrices that avoids scaling issues that arise in specifications based on multivariate similarity criterion such as those set forth in Blankmeyer et al (2011). Another advantage is avoidance of complex parameter space constraints inherent in higher order simultaneous autoregressive models.

Debarsy and LeSage (2018b) propose a solution to the computational challenges inherent in simultaneous autoregressive model specifications based on convex combinations of weight matrices and further develop a Bayesian model averaging approach that allows for model uncertainty regarding alternative definitions of the dependence set. For example, one can draw inferences regarding which matrices should enter the specification and which should be excluded.

This approach should be useful in applied research where there is little theoretical guidance regarding an appropriate dependence set and associated weight matrix. An important caveat is that practitioners should attempt to consider definitions of dependence sets that convey differing types of potential cross-sectional dependence relationships. Use of alternative weight matrices that convey similar cross-sectional dependence relationships can encounter difficulties when simultaneous lags of the dependent variable vector  $(W_1y, W_2y, \dots, W_Ly)$ , disturbances  $(W_1u, W_2u, \dots, W_Lu)$ , or explanatory variables  $(W_1X, W_2X, \dots, W_LX)$  exhibit high levels of correlation.

LeSage and Pace (2014) discuss the issue of highly correlated spatial lags that can arise in spatial autoregressive model specifications, and difficulties that arise in distinguishing between models based on similar definitions of dependence sets of observations. They label the mistaken belief that small changes in the definition of dependence sets can lead to large changes in model estimates and inferences as the *the biggest myth in spatial econo-*

*metrics*. The same caveat applies to simultaneous autoregressive model specifications described here. Practitioners should strive for inclusion of alternative weight matrices in the convex combination of weights that reflect distinctly different types of dependence. Use of specifications where multiple weight matrices result in similar dependence sets of observations will lead to poor estimates and inferences regarding the relative importance of each set of weights.

## Cross References

LeGallo, J. , Chapter 1. Cross-Section Spatial Regression Models.

Mills, JA and Parent O. Chapter 3. Bayesian MCMC Estimation.

LeSage, JP and Pace, RK Chapter 8. Interpreting Spatial Econometric Models.

## References

Badinger, H., P. Egger (2011), Estimation of higher-order spatial autoregressive cross-section models with heteroscedastic disturbances, *Papers in Regional Science*, 90(1), 213–235.

Blankmeyer, E., LeSage, J. P., Stutzman, J. R., Knox, K. J., Pace, R. K. (2011), Peer-group dependence in salary benchmarking: A statistical model, *Management and Decision Economics*, 32(2), 91–104.

Debarsy N., J. P. LeSage (2018a), Flexible dependence modeling using convex combinations of different types of connectivity structures, *Regional Science and Urban Economics*, 69, 48–68.

Debarsy N., J. P. LeSage (2018b), Bayesian model averaging for spatial autoregressive models based on convex combinations of different types of connectivity matrices, LEM Working Paper, DP2018-22, Université de Lille.

Elhorst, J.P., D.J. Lacombe, G. Piras (2011), On model specification and parameter space definitions in higher order spatial econometric models, *Regional Science and Urban Economics*, 42, 211–220.

Golub, G, C. van Loan (1996), *Matrix Computation, Third Edition*, The John Hopkins University Press, Baltimore.

- Hazir, C.S., J.P. LeSage, C. Autant-Bernard (2018), The role of R&D collaboration networks on regional innovation performance, *Papers in Regional Science*, 97(3), 549–567.
- Lacombe, D.J., 2004, Does econometric methodology matter? An analysis of public policy using spatial econometric techniques. *Geographical Analysis*, 36(2), 105–118.
- Lee, L.-F., X. Liu, 2010, Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances, *Econometric Theory* 26, 187–230.
- LeSage, J. P., R. K. Pace (2009), *Introduction to Spatial Econometrics*, Taylor and Francis/CRC, Boca Raton.
- LeSage, J. P., R. K. Pace, (2011), Pitfalls in higher order model extensions of basic regression methodology, *Review of Regional Studies*, 41(1), 13–26.
- LeSage, J. P., R. K. Pace, (2014) The biggest myth in spatial econometrics, *Econometrics*, 2(4), 217–249.
- Pace, R. K., R. Barry, O.W. Gilley, C.F. Sirmans (2000), A method for spatial-temporal forecasting with an application to real estate prices, *International Journal of Forecasting*, 16(2), 229–246.
- Pace, R.K., J.P. LeSage (2002), Semiparametric maximum likelihood estimates of spatial dependence, *Geographical Analysis*, 34(1), 75–90.
- Pace R.K., C.F. Sirmans, V.C. Slawson (2002), Automated valuation models in Wang K. and Wolverton M.L. (Eds), *Real Estate valuation theory*, Vol. 8, chap 7, 133–156, Springer.