Parsimonious Modeling in Spatial Statistics and Spatial Econometrics

Information of Statistics and Mathematics Daisuke Murakami
University of Tsukuba Morito Tsutsumi

1. Introduction

Statistics for spatial data have been studied in geostatistics and spatial econometrics for decades. Geostatistics, which originates from natural science, focuses on spatial interpolation, mapping, and other data driven analysis. By contrast, spatial econometrics, which stemmed from regional science, emphasizes on model-driven analysis such as hypothesis testing on direct/indirect spatial effects.

Despite such difference, they share the same modeling interest: how to describe spatial dependence. Usual modeling approach based on a spatial connectivity matrix has the following drawbacks: (i) it tends to be computationally expensive; (ii) consideration of spatial dependence can lead a confounding problem, which makes the parameter estimation unstable. Problems (i) and (ii) get severe especially in non-linear/non-Gaussian, and other flexible models. Recently, parsimonious approach, which reduces redundancy of the model, has been suggested to be a useful approach to mitigate these problems.

Based on such a background, this study develops a parsimonious spatial modeling approach that is computationally efficient, copes with the confounding problem by balancing the bias-variance trade-off, and is directly related to models in geostatistics and spatial econometrics.

2. Model

In the basic linear case, our model yields

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\mathbf{V}(\boldsymbol{\theta})\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \mathbf{u} \sim N(\mathbf{0}_L, \sigma^2\mathbf{I}_L), \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{y}$ is a vector of explained variables, $\mathbf{X}$ is a matrix of explanatory variables, $\mathbf{0}_L$ and $\mathbf{0}$ are vector of zeros, and $\mathbf{I}_L$ and $\mathbf{I}$ are identity matrices with appropriate size. $\boldsymbol{\beta}$ is a vector of coefficients, and σ^2 is a variance parameter. $\mathbf{E}$ is a matrix of L eigenvectors of a spatial connectivity matrix, $\mathbf{V}(\boldsymbol{\theta})$ is a diagonal matrix whose elements depend on the L eigenvalues and parameters, $\boldsymbol{\theta}$. The term $\mathbf{E}\mathbf{V}(\boldsymbol{\theta})\mathbf{u}$ captures spatial dependence. This model is capable of reducing the redundancy of the model by selecting L and $\boldsymbol{\theta}$ appropriately.

This specification allows us applying a fast Type II restricted maximum likelihood (or h-likelihood/empirical Bayes) approach that maximizes Eq.(2):

$$L_R(\mathbf{y} \mid \boldsymbol{\theta}) = \log \int\int p(\mathbf{y} \mid \mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\mathbf{u} \mid \boldsymbol{\theta}) d\mathbf{u} d\boldsymbol{\beta}. \tag{2}$$

Advantages of this estimation approach are as follows: (a) computational efficient; (b) it furnishes shrinkage estimators, which are based on a balancing of a bias-variance trade-off; (c) it approximates both geostatistical and spatial econometric models; (d) (a), (b), and (c) hold even after the model Eq.(1) is extended to some other non-Gaussian/non-linear models.