Eigenvector Selection for Eigenvector Spatial Filtering

P. Sinha, M. Lee, Y. Chun, D. A. Griffith

The University of Texas at Dallas, 800 W. Campbell rd., Richardson, Texas, USA
Telephone: 1-972-883-4950
Fax: 1-972-883-69676
Email: parmanand.sinha@utdallas.edu
Email: mxl120631@utdallas.edu
Email: ywchun@utdallas.edu
Email: dagriffith@utdallas.edu

1. Introduction

Eigenvector spatial filtering (ESF) furnishes a methodology that accounts for spatial dependency in georeferenced data (Griffith 2003), which, to date, has been the domain of spatial autoregressive (SAR) models. Its fundamental idea exploits the decomposition of a spatial variable into the following three components: trend, spatially structured random component (i.e., spatial stochastic signal), and random noise. Its aim is to separate spatially structured random components from both trend and random noise, and, consequently, furnishes a sounder statistical inferential basis and useful visualization. In other words, ESF uses a set of synthetic proxy variables, which are extracted as eigenvectors from a spatial connectivity matrix that ties geographic objects together in space, and then adds these vectors as control variables to a model specification. These control variables identify and isolate the stochastic spatial dependencies among the georeferenced observations, thus allowing model building to proceed as if the observations are independent. Because ESF model specification is flexible, it can be utilized to describe variables following various types of distributions, including the Gaussian, Poisson, and binomial. Different ESF specifications have been compared with other specifications, such as the SAR (Getis and Griffith 2002, Tiefelsdorf and Griffith 2007, Thayn and Siminas 2012), auto-Poisson (Griffith 2002), and auto-logistic (Griffith 2004) ones.

Although ESF has become more popular in addressing spatial autocorrelation (SA) latent in georeferenced data (e.g., Hughes and Haran 2012), it faces two major computational challenges. Both the extraction of eigenvectors from an *n*-by-*n* modified geographic weights matrix, and the selection of a subset from the resulting set of *n* eigenvectors to construct a spatial filter, become increasingly challenging as *n* increases. In this paper, the selection of eigenvectors to construct ESF-based estimators in linear regression is investigated in terms of this latter computational issue.

2. Eigenvector spatial filtering

The ESF methodology utilizes an eigenfunction mathematical decomposition of the transformed spatial weights matrix, $(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)$, which appears in the numerator of the Moran Coefficient (MC), where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is an *n*-by-1 vector of ones, $\mathbf{C}$ is a spatial weights matrix, and superscript T denotes the matrix transpose operator. This decomposition generates *n* eigenvalues, $\boldsymbol{\lambda}=\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, and *n* corresponding mutually orthogonal and uncorrelated (Griffith 2000) eigenvectors, $\mathbf{E}=\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n\}$. Mapping each eigenvector across the original *n* areal units associated

with **C** portrays a distinct map pattern whose MC value is directly associated to its corresponding eigenvalue: i.e., $MC_j = \lambda_j \cdot n/\mathbf{1}^T\mathbf{C}\mathbf{1}$, for $\mathbf{E}_j$ (Tiefelsdorf and Boots 1995, Griffith 1996). Furthermore, the feasible range of MC for a given spatial tessellation is determined by λ_1 and λ_n (de Jong et al., 1984). Hence, the eigenvectors furnish distinct map pattern descriptions of latent SA in georeferenced variables.

The ESF methodology accounts for SA with a linear combination of a subset of the *n* eigenvectors. The pure ESF linear regression model specification may be written as $\mathbf{Y} = \mathbf{E}_k\boldsymbol{\beta}_E + \boldsymbol{\varepsilon}$, where $\mathbf{E}_k$ is an *n*-by-*K* matrix containing *K* eigenvectors, $\boldsymbol{\beta}_E$ is the corresponding vector of regression parameters, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$ is an *n*-by-1 error vector whose elements are iid normal random variates. Because the linear combination of the eigenvectors, $\mathbf{E}_k\boldsymbol{\beta}_E$, accounts for SA, the ESF linear regression specification does not suffer from spatially autocorrelated residuals.

3. Computational intensities in eigenvector spatial filtering

Computational intensity of the ESF methodology arises in two ways. First, the extraction of eigenvectors from an *n*-by-*n* modified geographic weights matrix is computationally intensive. Specially, the extraction of eigenvectors often requires substantial computations as *n* increases, and has an upper limit that depends upon computer resources. For regular square tessellations such as remotely sensed images, eigenvectors can be efficiently generated with equations that require the pixel location in an image (Griffith 2000). Also, an algorithm for sparse matrices can improve the computation of eigenvector generation (Pace et al. 2011).

Second, the selection of a subset from the resulting set of *n* eigenvectors to construct a spatial filter, $\mathbf{E}_k$ is computationally intensive, and involves two steps. In the first step, a candidate set of eigenvectors, which is a noticeably smaller subset (i.e., m << n) of the entire set of eigenvectors, can be demarcated based upon various criteria. One such criterion employs a threshold minimum MC of 0.25, which relates to roughly 5% of the variance in a response variable being attributable to positive SA. Another criterion devised by Griffith and Chun (2009) utilizes the level of SA detected in a response variable, Y, or residuals when a specification contains covariates, and may be summarized as follows:

$$MC_j \geq 2.9970 - 2.8805/(1 + e^{-0.6606 - 0.2525 z_{MC}}), \quad (1)$$

where z_{MC} denotes the z-score of the MC for the response variable Y. Equation (1) indicates that as positive SA decreases, the number of eigenvectors in a candidate set tends to decrease, with the inclusion of fewer and fewer eigenvectors. But equation (1) is based upon only a 20-by-20 square tessellation, for which n = 400.

In the second step, a smaller set of eigenvectors can be identified from a candidate set using an easily implementable standard forward stepwise regression selection technique. One way to select eigenvectors is to maximize model fit at each step employing statistical significance (e.g., invoking a 10% level). Another way to select eigenvectors is to minimize residual SA at each step until the MC $\approx$ E(MC), the expected value of the MC.

4. Simulation experiment results

The simulation experimental design seeks to assess the stepwise selection of eigenvectors to construct an eigenvector spatial filter. The data generating mechanism used is an eigenvector spatial filter model without covariates. The experiment steps are as follows: (1) generate *n* iid random variables; (2) discard any sample whose normality diagnostic statistic suggests rejection of the null hypothesis; (3) randomly permute the *n* generated values until their MC value falls into the interval (-1/(n-1)-0.01, -1/(n-1)+0.01)—i.e., failure to reject the null hypothesis; (4) inbed SA by summing randomly selected eigenvectors with randomly selected coefficients, and adding a iid variable to each sum; and, (5) replicate the experiment 10,000 times for each *n*.

The output of interest is the number of superfluous and missed eigenvectors via the selection process when constructing an eigenvector spatial filter. Additional analyses address the issue of multiple testing associated with forward selection stepwise regression. An initial experiment indicates that in the presence of zero SA, the Bonferroni adjustment holds. The simulation experiments summarized in this paper assess this finding for the full range of positive SA.

5. Implications

The ESF method potentially is plagued by common problems associated with stepwise regression techniques. Is the correct set of eigenvectors selected? Are superfluous eigenvectors selected? Are eigenvectors not selected that should be selected? Are results biased? How does ESF relate to the multiple testing problem? Fortunately, ESF avoids complications affiliated with multicollinearity because the eigenvectors are mutually orthogonal and uncorrelated. This paper contributes to a better understanding of the statistical basis supporting ESF through geo-computation.

6. References

de Jong P, Sprenger C and van Veen F, 1984, On extreme values of Moran's I and Geary's C. *Geographical Analysis*, 16:17-24.

Getis A and Griffith DA, 2002, Comparative spatial filtering in regression analysis. *Geographical Analysis*, 34: 130–140.

Griffith DA, 1996, Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying georeferenced data. *The Canadian Geographer*, 40: 351-367.

Griffith DA, 2000, Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses, *Linear Algebra & Its Applications*, 321:95-112.

Griffith DA, 2002, A spatial filtering specification for the auto-Poisson model. *Statistics & Probability Letters*, 58:245-251.

Griffith DA, 2003, *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization*, Springer-Verlag, Berlin.

Griffith DA, 2004, A spatial filtering specification for the auto-logistic model. *Environment & Planning A*, 36:1791-1811.

Griffith DA and Chun Y, 2009, Eigenvector selection with stepwise regression techniques to construct spatial filters. paper presented at the 105th annual Association of American Geographers meeting, Las Vegas, NV, March 25.

Hughes J and Haran M, 2012, Dimension reduction and alleviation of confounding for spatial generalized linear mixed models, *Journal of the Royal Statistical Society, Series B*, in press, DOI: 10.1111/j.1467-9868.2012.01041.x.

Pace K, LeSage J and Zhu S, 2011, Interpretation and Computation of Estimates from Regression Models using Spatial Filtering., paper presented to the Vth World Conference of the Spatial Econometrics Association, Toulouse, FR, July 6-8.

Thayn, J., and J. Simanis. 2012. Accounting for spatial autocorrelation in linear regression models using spatial filtering with eigenvectors, *Annals of the Association of American Geographers*. 103: 47-66.

Tiefelsdorf M and Boots B, 1995, The exact distribution of Moran's I. *Environment and Planning A*, 27:985-999.

Tiefelsdorf M and Griffith DA, 2007, Semi-parametric filtering of spatial autocorrelation: the eigenvector approach, *Environment & Planning A*, 39:1193-1221.