The Statistical Distribution of Coefficients for Constructing Eigenvector Spatial Filters

P. Sinha[1], M. Lee[2], Y. Chun[3], D. A. Griffith[4]

The University of Texas at Dallas, 800 W. Campbell Rd., Richardson, Texas, USA
Telephone: 1-972-883-4950
Email[1]:pnsinha@utk.edu
Email[2]:mxl120631@utdallas.edu
Email[3]:ywchun@utdallas.edu
Email[4]:dagriffith@utdallas.edu

1. Introduction

This paper presents an exploratory simulation experiment to investigate the distribution of coefficients that are used to construct eigenvector spatial filters (ESFs) (Griffith 2000; Griffith 2003). The fundamental idea of ESF 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. The spatially structured component is modeled for with a linear combination of synthetic proxy variables, which are extracted as eigenvectors from a spatial weights matrix that ties geographic objects together in space; it is added as a control variable to a spatial model specification. This control variable identifies and isolates the stochastic spatial dependencies among the georeferenced observations, thus allowing model building to proceed as if the observations are independent. Identifying the distribution of coefficients used to construct this linear combination may support simulation experiments employing the generation of spatially autocorrelated random numbers using ESF.

2. Eigenvector Spatial Filtering

The ESF methodology utilizes the properties of eigenvectors and their corresponding eigenvalues of the transformed spatial weights matrix $(\mathbf{I} - \mathbf{11}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{11}^T/n)$, where $\mathbf{I}$ is an n-by-n identity matrix, $\mathbf{1}$ is an n-by-1 vector of ones, $\mathbf{C}$ is an n-by-n spatial weights matrix, and superscript T denotes the matrix transpose operator. Studies, including Tiefelsdorf and Boots (1995) and Griffith (1996), show that its n mutually orthogonal and uncorrelated (Griffith 2000) eigenvectors, $\mathbf{E}=\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n\}$, and n corresponding eigenvalues, $\boldsymbol{\lambda}=\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, relate to spatial autocorrelation (SA). Important properties of these vectors include: 1) they furnish distinct map pattern descriptions visualizing latent SA in georeferenced variables, and 2) the eigenvalues index the level of SA of a map pattern that is generated when the corresponding eigenvector is mapped on the given tessellation. That is, the Moran Coefficient (MC) of the map pattern produced by eigenfvector $\mathbf{E}_j$ is $\mathrm{MC}_j = \lambda_j\ n/\mathbf{1}^T\mathbf{C1}$.

The ESF linear regression model specification can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_k\boldsymbol{\beta}_E + \boldsymbol{\varepsilon}$, where $\mathbf{E}_k$ is an n-by-K matrix containing K eigenvectors, $\boldsymbol{\beta}_E$ is the corresponding vector of regression parameters, $\mathbf{E}_k\boldsymbol{\beta}_E$ is the ESF, and $\boldsymbol{\varepsilon}\sim\mathrm{N}\left(\mathbf{0},\mathbf{I}\sigma_\varepsilon^2\right)$ 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. Methodology

Random numbers using a simultaneous autoregressive (SAR) data generating mechanism can be simulated with $\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$, where matrix $\mathbf{W}$ is the row standardized geographic connectivity matrix , and ρ is a parameter indicating level of SA. Theoretically, $(\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$ can be decomposed into a linear combination, $\mathbf{E}_k$ $\boldsymbol{\beta}_{E.}$ (Griffith 2003). This property leads the quest for finding the distribution of ESF coefficients $\boldsymbol{\beta}_E$ while controlling the level of autocorrelation. The corresponding coefficients can be used to generate spatially auto-correlated random numbers using a combination of eigenvectors as shown in equation $\mathbf{Y} = \mathbf{E}_k\boldsymbol{\beta}_E + \boldsymbol{\varepsilon}$.

The candidate set of $\mathbf{E}_k$ varies based on the magnitude of the SA parameter ρ. The number of selected eigenvectors and coefficients $\boldsymbol{\beta}_E$ depend on the result of a stepwise regression of the SAR generated random numbers with the candidate set of eigenvectors. Visualization of results from a simulation experiment suggests that the frequency distribution of the elements of vector $\boldsymbol{\beta}_E$ is similar to the gamma distribution with its scale and shape parameters varying with n and the level of SA. The selection of a candidate set of eigenvectors also depends on the level of SA. For higher levels of SA, eigenvectors with the largest eigenvalues were selected more often; for lower levels of SA, the probability of selecting eigenvectors appears to be nearly uniform.

4. A Simulation Experiment

A simulation experiment has been conducted using 10-by-10, 25-by-25, 50-by-50, 75-by-75, and 100-by-100 hexagon tessellations and three levels of SA, with 10,000 replications. The following are the steps involved:

1. Generate 10,000 spatially autocorrelated random variables, Y, for each tessellation using a SAR data generating mechanism with ρ = 0.1, 0.5 and 0.9.
2. Calculate the candidate set of eigenvectors for each tessellation and each level of SA using the criterion

   $$\frac{n_+}{1 + e^{[2.1480 - 6.1808[(z_{MC} + 0.6)^{0.1742}] / n_+^{0.1298} + 3.3534/(z_{MC} + 0.6)^{0.1742}]}}$$

   where, z_{MC} is the Moran coefficient z-score of Y, n_+ is the number of eigenvectors with positive eigenvalues, and e is an exponential function.
3. Select the significant eigenvectors, using the Y as dependent variables in stepwise regressions with the candidate set of eigenvectors and selection based upon an AIC criterion maximizing goodness-of-fit.
4. Calculate the probability of selection of each significant eigenvectors in 10,000 cases, retaining those with a minimal empirical probability of 0.01 for further analysis. Eigenvectors selected lesser than 100 times out of 10,000 simulations are assumed to be not significant.
5. For the significant eigenvectors coefficients, calculate the shape and scale parameters of a gamma distribution.

5. Results

Table 1 shows the candidate set and selected eigenvectors using a stepwise selection procedure for different tessellations and different levels of SA.

	Expected rho	*Candidate set EVs*	*Selected EVs*
10x10	0.1	9	2
	0.5	25	10
	0.9	33	21
25x25	0.1	44	10
	0.5	145	61
	0.9	207	125
50x50	0.1	44	10
	0.5	145	61
	0.9	207	125
75x75	0.1	373	92
	0.5	1214	529
	0.9	1769	1065
100x100	0.1	670	159
	0.5	2133	941
	0.9	3103	1873

Table 1. selected eigenvectors for different tessellations

Figures 1 to 6 display the distributions of coefficients, $\boldsymbol{\beta}_E$, for randomly selected eigenvectors. Each of these histograms has 10,000 coefficients for a given eigenvector, and has been overlaid with a corresponding gamma function.

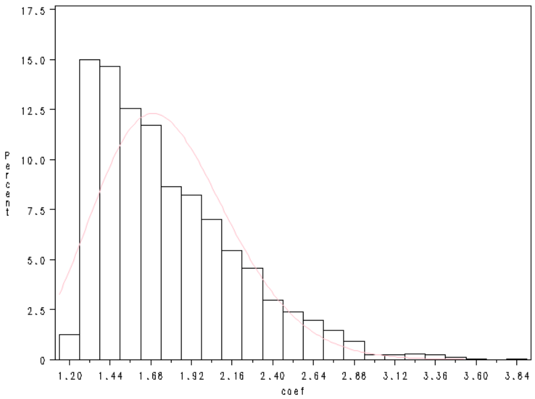

Figure 1. 3rd eigenvector of a 10-by-10 tessellation for ρ = 0.1.

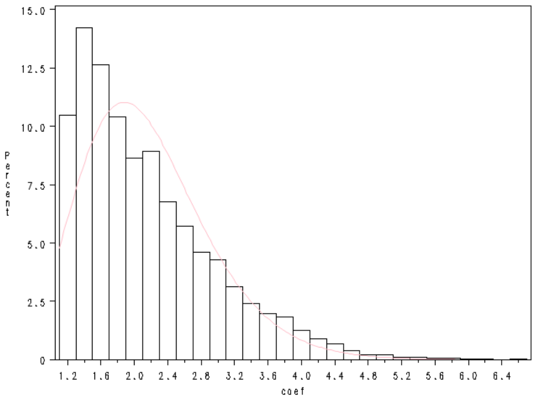

Figure 2. 37th eigenvector of a 25-by-25 tessellation for ρ = 0.9.

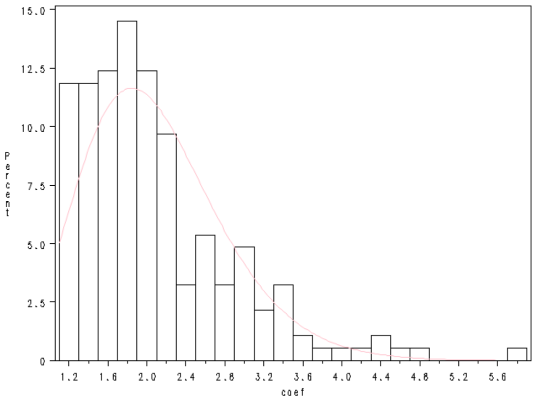

Figure 3. 180th eigenvector of a 50-by-50 tessellation for ρ = 0.5.

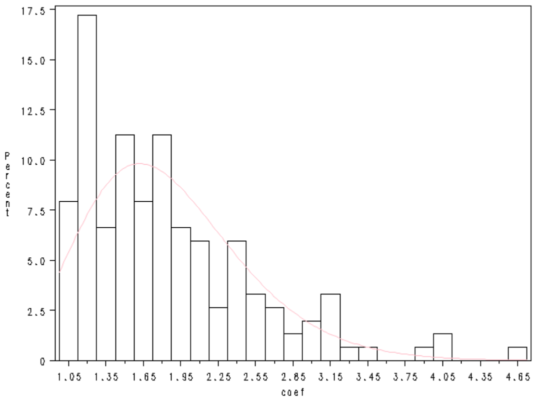

Figure 4. 700th eigenvector of a 50-by-50 tessellation for ρ = 0.9.

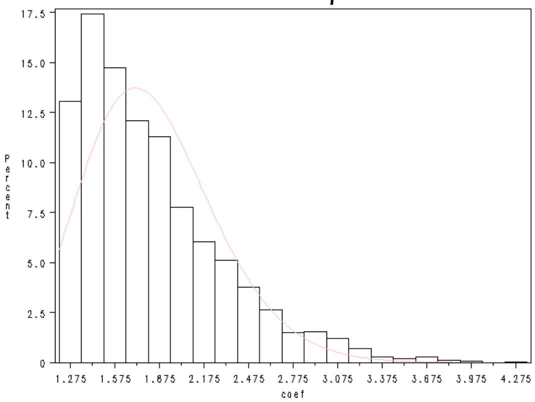

Figure 5. 2nd eigenvector of a 75-by-75 tessellation for ρ = 0.1.

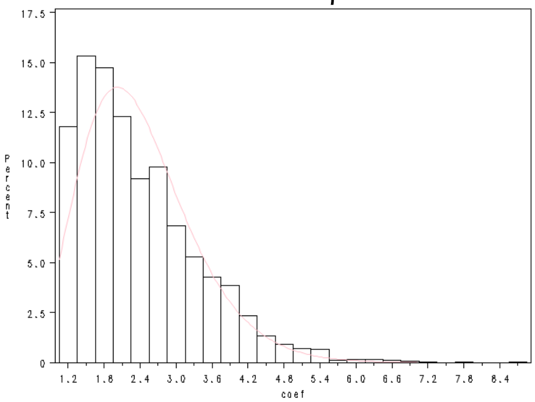

Figure 6. 31st eigenvector of a 100-by-100 tessellation for ρ = 0.5.

Figures 7 to 12 portray the distributions of gamma distribution shape and scale parameters for 25-by-25, 50-by-50, and 75-by-75 tessellations across the number of the eigenvectors. Yellow denotes the trend line. In general, the shape parameter value tends to increase, and the scale parameter tends to decrease, as the number of eigenvectors increases. The shape parameter increment for higher levels of SA tracks an exponential curve, whereas the scale parameter decrement tracks a hyperbolic curve.

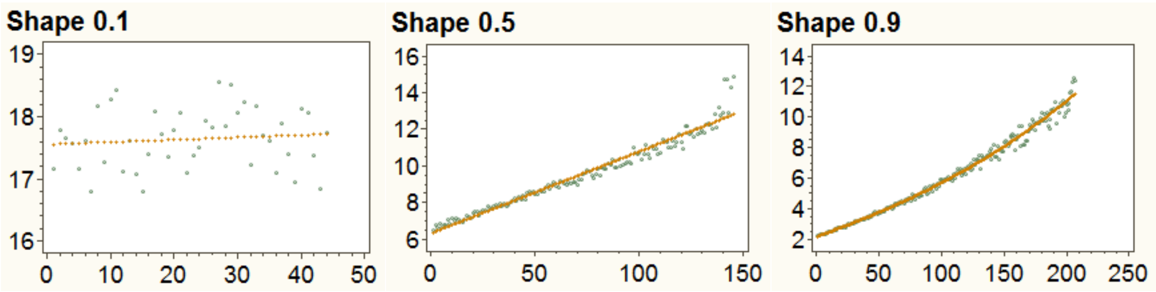

Figure 7. gamma distribution shape parameters for a 25-by-25 tessellation with number of eigenvectors on the horizontal axis. The trends are: (a) uniform for ρ = 0.1; (b) linearly increasing for ρ = 0.5; and, (c) exponentially increasing for ρ = 0.9

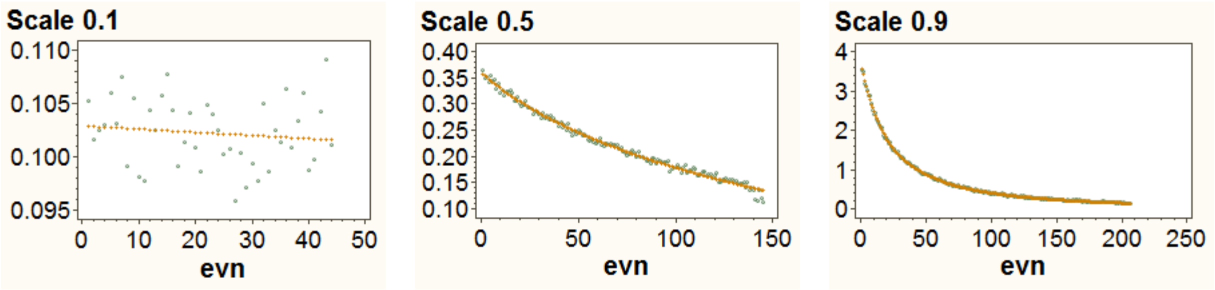

Figure 8. gamma distribution scale parameters for a 25-by-25 tessellation with number of eigenvectors on the horizontal axis. The trends are: (a) uniform for ρ = 0.1; (b) nonlinearly decreasing for ρ = 0.5; and, (c) nonlinearly decreasing for ρ = 0.9

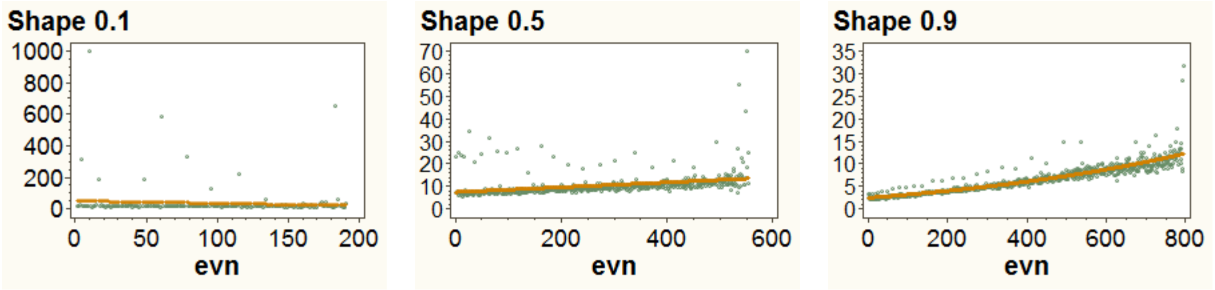

Figure 9. gamma distribution shape parameters for a 50-by-50 tessellation with number of eigenvectors on the horizontal axis. The trends are: (a) uniform for ρ = 0.1; (b) linearly increasing for ρ = 0.5; and, (c) exponentially increasing for ρ = 0.9

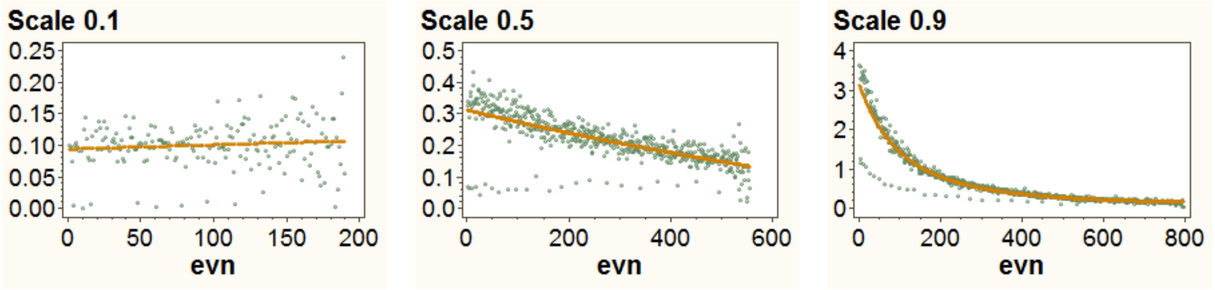

Figure 10. gamma distribution scale parameters for a 50-by-50 tessellation with number of eigenvectors on the horizontal axis. The trends are: (a) uniform for ρ = 0.1; (b) nonlinearly decreasing for ρ = 0.5; and, (c) nonlinearly decreasing for ρ = 0.9

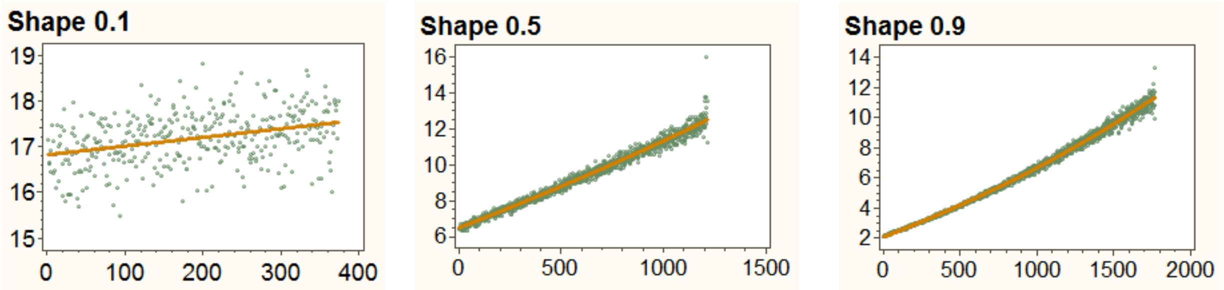

Figure 11. gamma distribution shape parameters for a 75-by-75 tessellation with number of eigenvectors on the horizontal axis. The trends are: (a) uniform for ρ = 0.1; (b) linearly increasing for ρ = 0.5; and, (c) exponentially increasing for ρ = 0.9

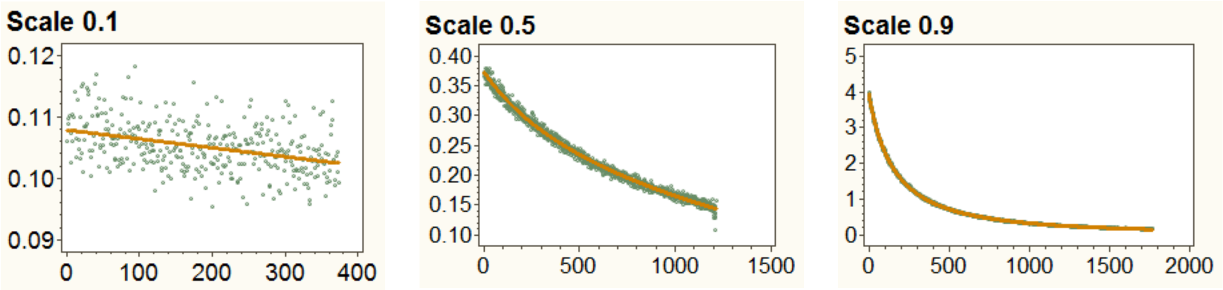

Figure 12. gamma distribution scale parameters for a 75-by-75 tessellation with number of eigenvectors on the horizontal axis. The trends are: (a) uniform for ρ = 0.1; (b) nonlinearly decreasing for ρ = 0.5; and, (c) nonlinearly decreasing for ρ = 0.9

6. Implications and Future Research

Correctly identifying the statistical distribution of coefficients will support generating spatially autocorrelated random numbers using ESF while controlling for the level of autocorrelation. Theoretically, in the generation of spatially autocorrelated random numbers, a SAR mechanism uses all of the eigenvectors, whereas with an ESF, the stepwise selection is done with eigenvectors reflecting the nature of SA. The statistical distribution of the coefficients would be useful for designing Monte Carlo simulation studies using ESF. An ESF could allow for more experimental control in a simulation experiment. Surfaces beyond 100-by-100 will be included in the future research.

7. Acknowledgements

This research was conducted with supported from the U.S. National Science Foundation, grant BCS-1229223; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

8. References

Griffith, Daniel A. 1996. “Spatial Autocorrelation and Eigenfunctions of the Geographic Weights Matrix Accompanying Geo-Referenced Data.” *The Canadian Geographer/Le G éographe Canadien* 40 (4). 351–67.

———. 2000. “Eigenfunction Properties and Approximations of Selected Incidence Matrices Employed in Spatial Analyses.” *Linear Algebra and Its Applications* 321 (1-3): 95–112.

———. 2003. *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*. Springer-Verlag, Berlin.

Tiefelsdorf, Michael, and Barry Boots. 1995. “The Exact Distribution of Moran’s I.” *Environment and Planning A* 27. 985-999.