On the quality of eigenvector spatial filtering based parameter estimates for the normal probability model: implications about uncertainty and specification error for georeferenced data

Yongwan Chun[1] *and Daniel A. Griffith*[2]

[1] University of Texas at Dallas, 800 W. Campbell rd. Richardson, Texas 75093, USA. ywchun@utdallas.edu
[2] Ashbel Smith Professor, University of Texas at Dallas, 800 W. Campbell rd. Richardson, Texas 75093, USA. dagriffith@utdallas.edu

Abstract

Eigenvector spatial filtering, which introduces a subset of eigenvectors extracted from a spatial weights matrix as synthetic control variables in a regression model specification, furnishes a solution to extraordinarily intricate statistical modeling problems involving spatial dependencies. It accounts for spatial autocorrelation in standard specifications of regression models. But the quality of the resulting regression parameter estimates has yet to be ascertained. The estimator properties to establish include unbiasedness, efficiency and consistency. The purpose of this paper is to demonstrate these estimator properties for linear regression parameters based on eigenvector spatial filtering, including a comparison with the simultaneous autoregressive (SAR) model. Eigenvector spatial filtering methodology requires the judicious selection of eigenvectors, whose number tends to increase with both level of linear regression residual spatial autocorrelation and number of areal units. A logistic regression description of the number of eigenvectors selected in a simulation pilot study suggests estimator consistency.

Keywords: Eigenvector spatial filtering, unbiasedness, efficiency, consistency.

1. Introduction

While classical mathematical statistics avoids correlation amongst observations by assuming that they are independent, spatial autocorrelation (SA) is commonly embedded in georeferenced data. Spatial autoregressive models and geostatistics are popularly applied to capture dependence among observations in space. Eigenvector spatial filtering (ESF) method furnishes another popular approach to take SA into account in a model specification.

An ESF model specification is flexible, and can describe Gaussian, Poisson, and binomial random variables containing positive SA (i.e., similar values tending to cluster together on a map), whereas auto-models like the Poisson, exponential, and gamma schemes have integrability conditions that ensure that they are able to describe only spatial competition between neighboring sites (i.e., negative SA). ESF specifications appear in, for example, Getis and Griffith (2002). Comparisons also exist between the ESF and the auto-logistic and the auto-Poisson (e.g., Griffith,

2004) specifications. Relationship articulations between the ESF and both the AR and SAR specifications appear in Tiefelsdorf and Griffith (2007).

Although ESF has become more popular in addressing SA latent in georeferenced data, the quality of ESF-based estimators has not been thoroughly investigated. The statistical qualities of ESF-based estimators, including unbiasedness, efficiency, and consistency, remain under- or un-explored. Such a quality assessment of ESF-based estimators can bolster the efficacy of ESF methodology, documenting that it furnishes a solid foundation for analyzing georeferenced data in linear and generalized linear regression analysis. The goal of this paper is to summarize selected empirical results in order to begin to fill this gap in the literature. It utilizes a judiciously selected purposeful sample (criterion: span a wide range of n) of readily available geographic landscape datasets for an exploratory pilot study.

2. Background: eigenvector spatial filtering

Eigenvector spatial filtering methodology utilizes an eigenfunction decomposition of the transformed spatial weights matrix, $(\mathbf{I} - \mathbf{1}\mathbf{1}^{T}/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^{T}/n)$, where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is an n-by-1 vector of ones, and T denotes the matrix transpose operator. This decomposition generates n eigenvectors and their associated n eigenvalues. In descending order, the n eigenvalues can be denoted by the set $\{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n\}$. The corresponding n eigenvectors can be denoted as $\mathbf{E} = (\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \ldots, \mathbf{E}_n)$, where each eigenvector, $\mathbf{E}_j$, is an n-by-1 vector. These eigenvectors furnish distinct map pattern descriptions of latent SA in georeferenced variables, because they are mutually both orthogonal and uncorrelated (see Griffith 2003 for details).

The ESF methodology utilizes the eigenvectors to account for SA by adding a linear combination of these eigenvectors. In linear regression, the ESF model specification may be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_k\boldsymbol{\beta}_E + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is an n-by-(p+1) matrix containing covariates (including vector 1 for the intercept term), $\boldsymbol{\beta}$ is the corresponding (p+1)-by-1 vector of regression parameters, $\mathbf{E}_k$ is an n-by-K matrix containing K eigenvectors, $\boldsymbol{\beta}_E$ is the corresponding vector of regression parameters, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$. 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.

Initializing ESF methodology requires the identification of a feasible set of eigenvectors. This procedure involves two steps. In the first step, a candidate set of eigenvectors, which is a noticeably smaller subset (i.e., K << n) of the entire set of eigenvectors, can be demarcated based upon several criteria. Eigenvectors whose MC values are close to the expected value of MC do not explain much spatial variation and can be eliminated from a candidate set. The candidate set can be further restricted to only eigenvectors portraying positive SA (or negative SA), if the MC for a response variable displays positive SA (or negative SA); most empirical datasets display positive SA, and, accordingly, the proposed research would address only this nature of SA. In the second step, a smaller set of eigenvectors can be identified from its candidate set using a stepwise regression selection technique. One way to select an eigenvector is to maximize model fit at each step through statistical significance (e.g., invoking a 10%, 5%, or 1% level). This selection can be implemented easily with conventional stepwise regression procedures. Another

way to select an eigenvector is to minimize residual SA in each stepwise regression iteration. This selection procedure can be repeated until, say, MC ≈ E(MC), the expected value of MC, is achieved.

3. Unbiasedness in the presence of spatial autocorrelation

Using the Frisch-Waugh-Lovell theorem, Pace et al. (2011) confirm that an ESF produces unbiased estimators of covariate regression parameters for data generated with an SAR model specification. Conclusions about biasedness are sensitive to where SA is present in the terms of a model (see Griffith 1976). Nevertheless, Pace et al. conclude that an ESF reduces bias found in ordinary least squares (OLS) estimators, thus improving upon OLS results.

For an SAR model data generating process, the variable Y can be expressed as $\mathbf{Y} = \mathbf{X\beta} + \mathbf{V}^{-1/2}\boldsymbol{\varepsilon}$, where $\mathbf{V}$ is the inverse covariance matrix [e.g.,$(\mathbf{I} - \rho\mathbf{C})^T(\mathbf{I} - \rho\mathbf{C})$ or $(\mathbf{I} - \rho\mathbf{W})^T(\mathbf{I} - \rho\mathbf{W})$ with $\mathbf{W}$ being a row-standardized spatial weights matrix]. The OLS, generalized least squares (GLS), and ESF estimators are unbiased:

$$\text{OLS}: \mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{E}(\mathbf{Y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X\beta} + \mathbf{0}) = \boldsymbol{\beta}$$
$$\text{GLS}: \mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathrm{E}(\mathbf{Y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}(\mathbf{X\beta} + \mathbf{0}) = \boldsymbol{\beta}$$
$$\text{ESF}: \mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}})(\mathbf{X\beta} + \mathbf{0}) = \boldsymbol{\beta}$$

4. Efficiency in the presence of spatial autocorrelation

The degree of sample to sample estimator stability is a quality property often more important than unbiasedness. Relatively efficient estimators have smaller variances, and hence require smaller sample sizes to achieve a given level of precision. The preceding sampling variance of the SAR model specification implies that the expected value of the OLS and GLS constant mean specification variance is

$$\text{OLS}: \mathrm{E}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1/2}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{V}^{-\mathrm{T}/2}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}] = \{\mathbf{1}^{\mathrm{T}}[(\mathbf{I} - \rho\mathbf{W})^{\mathrm{T}}(\mathbf{I} - \rho\mathbf{W})]^{-1}\mathbf{1}/\mathrm{n}^2\}\sigma^2$$
$$\text{GLS}: \mathrm{E}[(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{V}^{-1/2}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{V}^{-\mathrm{T}/2}\mathbf{V}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}] = \sigma^2/[\mathrm{n}(1-\rho)^2]$$

The unbiased OLS variance estimator is $\hat{\sigma}^2/[\mathrm{TR}(\mathbf{V}^{-1})]$, with the denominator term $\mathrm{TR}(\mathbf{V}^{-1})$ adjusting for variance inflation in the presence of positive SA. The GLS estimator is the more efficient of the two (Griffith, 1988; Cordy and Griffith, 1993), in part because the OLS estimator includes a variance inflation factor (VIF) created by positive SA. Most of the efficiency is lost through the variance estimator [i.e., using $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ in place of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$] of the regression coefficients.

Meanwhile, the ESF estimator is

$$\text{ESF}: (\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_{\mathrm{k}} - \mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}})\mathbf{V}^{-1} \times$$
$$(\mathbf{I}_{\mathrm{k}} - \mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}})\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2 = (\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2 +$$
$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\textstyle\sum_{\mathrm{h}=1}^{\infty}\rho^{\mathrm{h}}\mathbf{E}_{\mathrm{n-k}}\boldsymbol{\Lambda}_{\mathrm{n-k}}^{\mathrm{h}}\mathbf{E}_{\mathrm{n-k}}^{\mathrm{T}})\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2.$$

If $\mathbf{X}$ is orthogonal to $\mathbf{E}_{\mathrm{n-k}}$, this reduces to $(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{E}_{\mathrm{k}}\mathbf{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2$. The standard ESF variance estimator is $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X\beta} - \mathbf{E}_{\mathrm{k}}\boldsymbol{\beta}_{\mathrm{E}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X\beta} - \mathbf{E}_{\mathrm{k}}\boldsymbol{\beta}_{\mathrm{E}})/(\mathrm{n} - \mathrm{K} - 1)$, implying its unbiased estimator is $\{(\mathrm{n} - \mathrm{K} - 1)/[\mathrm{TR}(\mathbf{V}^{-1}) - \mathrm{TR}(\mathrm{E}_{\mathrm{k}}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{E}_{\mathrm{k}})]\}\ \hat{\sigma}^2$.

Employing the unbiased variance estimators, the constant mean sampling variances become

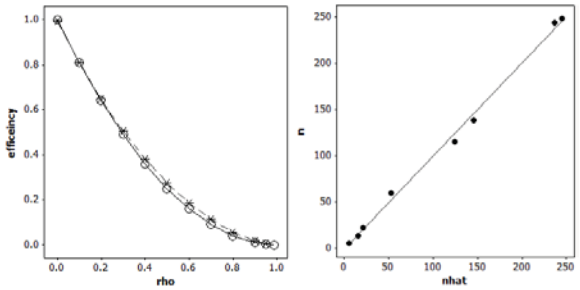
OLS: $(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}/n^2)\ \sigma^2$

GLS for the SAR model: $(n/\mathbf{1}^T\mathbf{V}\mathbf{1})\ \sigma^2 = \sigma^2/(1-\rho)^2$

ESF: $(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{E}_k\mathbf{E}_k^T\mathbf{X})^{-1}\sigma^2 = \sigma^2/n$

The expectation, then, is that the ESF estimator is more efficient than its GLS estimator counterpart (e.g., see Table 1), as well as its OLS estimator counterpart.

Figure 1a summarizes results from a simulation experiment illustrating this improved efficiency, and suggests that $TR(\mathbf{V}^{-1}) - TR(\mathbf{E}_k^T\mathbf{V}^{-1}\mathbf{E}_k) \approx n - K - 1$. Table 1 reports corroborating empirical analysis covariate standard error ratios.



rho	α/K		
	0.01/31	0.10/44	0.10
0.00	0.99469	0.99469	0.99469
0.10	0.81154	0.80557	0.79024
0.20	0.64769	0.64003	0.62360
0.30	0.50538	0.49614	0.47874
0.40	0.38171	0.37125	0.35400
0.50	0.27547	0.26374	0.24739
0.60	0.18556	0.17361	0.15956
0.70	0.11168	0.10□89	0.09067
0.80	0.05490	0.04641	0.04042
0.90	0.01611	0.01209	0.01018
0.95	0.00467	0.00315	0.00259

NOTE: α denotes level of significance

Figure 1: Left (a): Relative efficiency results for the mean: the ESF versus SAR mean estimate; simulation (asterisk) superimposed on theoretical (open circle) results. Right (b): Scatterplot of the predicted [from equation (1)] and the observed number of selected vectors.

Table 1. Summary spatial autocorrelation and ESF results for selected geographic landscapes with 2 spatially unautocorrelated covariates.

Landscape	attribute	z_{MC}	n	$n_{selected}$	$n_{candidate}$	s^2_{GLS}/s^2_{OLS}	s^2_{ESF}/s^2_{GLS}
Columbus, OH	crime	6.2507	49	5	12	4.7431	0.0705
North Carolina	SIDS	10.9617	100	13	24	0.4733	0.5574
Murray superfund	arsenic	8.1821	253	22	64	0.7112	0.8889
Mercer-Hall	yield	14.4023	500	60	155	1.0773	0.3284
Toronto	Pop. density	28.9400	731	115	308	3.3166	0.0881
High Peak	biomass	36.3366	900	244	349	0.0613	0.4801
Wiebe	yield	37.6297	1,500	138	462	1.6719	0.1349
China	Pop. density	61.1059	2,379	248	561	0.3613	0.7543

NOTE: s_j denotes the standard error of estimator j; the ratios are for the sums of the variances (see Kramer and Donninger, 1987).

5. Consistency in the presence of spatial autocorrelation

An extremely important quality feature of an estimator is that its sampling distribution becomes increasingly more concentrated near its corresponding parameter value with increasing n. Linear regression parameter estimators are consistent when the covariates are orthogonal and uncorrelated with the error term in a model specification. But the context for this property to hold is that n increases while the number of covariates remains unchanged. An ESF tends to have an increasing number

of eigenvectors with increasing n (Griffith and Chun 2009). Based on a selected set of empirical data analyses—where n ranges from 49 to 2,379 (see Table 2), and model specifications contain two covariates—the following logistic equation describes this increase extremely well (see Figure 1b):

$$n_{selected} \approx n_{candidate} \Big/ (1 + 3.2508\, e^{-0.1995\, z_{MC} + 0.0047\, n}) \qquad (1)$$

where $n_{selected}$ is the number of selected eigenvectors for constructing an ESF, $n_{candidate}$ is the number of candidate eigenvectors for a given surface partitioning [e.g., those with a $MC_j/MC_1 > 0.25$], and z_{MC} is the z-score for the response variable's residual MC. Even when $n_{candidate} = n$, MC = 1.5 (an excessively large value that is quite unlikely) and for a completely connected planar graph [i.e., $\mathbf{1}^T\mathbf{C}\mathbf{1} = 6(n-2)$], the asymptotic limit of equation (1) divided by n (i.e., $n_{selected}/n$) is 0:

$$\lim_{n\to\infty} [n/(1 + 3.2508\, e^{-0.1995\cdot(1.5+1/(n-1))/\sqrt{2/[6(n-2)]} + 0.0047n})] \Big/ n = 0.$$

This result is in keeping with findings reported by Portnoy (1984). In other words, the ESF-based covariate parameter estimates appear to be consistent.

Table 2. Summary standard errors and VIFs for subset of selected geographic landscapes with 2 ESF residual (i.e., spatially unautocorrelated) covariates.

Landscape	Y	estimator	residual MC	b_0 se	X_1		X_2		(pseudo-) R^2
					se	VIF	se	VIF	
Columbus, OH	crime	OLS	0.563	2.138	0.608	1.154	0.163	1.154	0.233
		GLS	***	4.844	0.286	***	0.119	***	0.629
		ESF	***	1.236	0.353	1.174	0.098	1.126	0.712
Toronto	Pop. density	OLS	0.670	0.022	0.007	1.026	0.028	1.026	0.044
		GLS	***	0.064	0.004	***	0.016	***	0.609
		ESF	***	0.012	0.004	1.100	0.015	1.122	0.780
China	Pop. density	OLS	0.753	0.021	0.163	1.193	0.155	1.193	0.086
		GLS	***	0.089	0.074	***	0.071	***	0.800
		ESF	***	0.011	0.085	1.255	0.081	1.277	0.790

6. Implications

ESF-based linear regression estimators are unbiased. Empirical evidence presented in this paper supports the contention that they tend to be more efficient than either their OLS or their GLS counterparts. The presence of SA in covariates, which can cause variance inflation in the ESF estimators, can compromise this efficiency. Figure 2 portrays relationships between VIFs and relative efficiencies. Intercept terms do not have VIFs. In contrast, Figure 2a displays a scatterplot of intercept standard error ratios versus average VIFs for the covariates, and reveals no relationship. Figures 2b and 2c display scatterplots for the covariate standard error ratios versus their corresponding VIFs, and suggest that efficiency calculations may need to be adjusted for VIFs arising from common SA in order to be meaningful. VIFs are not available for the SAR model specification, preventing a similar analysis for the GLS estimators. Finally, the ESF estimators appear to be consistent.

The empirical-based pilot study summarized in this paper suggests a number of worthwhile future research projects. First, the efficiency of ESF estimators needs to be established across the possible sources of SA in a linear regression model specification (Griffith 1976). Second, consistency of ESF estimators needs to be established with properly designed simulation experiments. Third, relationships between VIFs and SA need to be articulated. Completion of such research will help build a sound mathematical statistical foundation for ESF methodology.

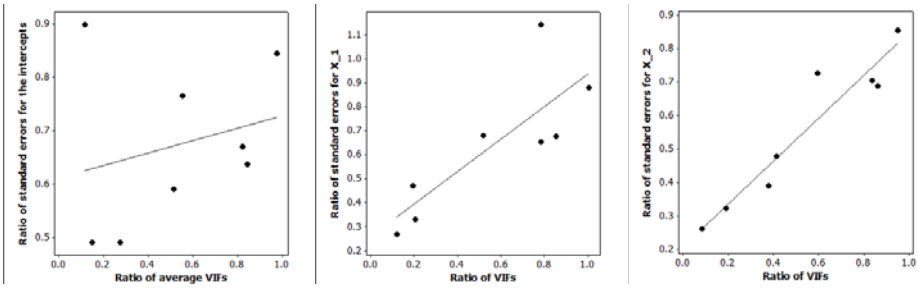


Fig. 2. Scatterplots of the ratio of spatially autocorrelated to spatially unautocorrelated relative efficiency of ESF estimators vis-a-vis OLS estimators versus their corresponding VIF ratios. Left (a): intercepts. Middle (b): 1st covariate regression coefficients. Right (c): 2nd covariate regression coefficients.

References

Cordy, C., Griffith, D.A. (1993), "Efficiency of least squares estimators in the presence of spatial autocorrelation". *Communications in Statistics*, *Series B*, Vol. 22:1161-1179.

Getis, A., Griffith, D.A (2002), "Comparative spatial filtering in regression analysis". *Geographical Analysis*, Vol. 34:130-140.

Griffith, D.A. (1976), "Spatial autocorrelation problems: some preliminary sketches of a structural taxonomy". *The East Lakes Geographer*, Vol. 11:59-68.

Griffith, D.A. (1988), "*Advanced Spatial Statistics"*. Martinus Nijhof, Dordrecht.

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

Griffith, D.A. (2004), "A spatial filtering specification for the auto-logistic model". *Environment & Planning A*, Vol. 36:1791-1811.

Griffith, D.A. (2010), "Spatial filtering". *In:* Getis, A., Fischer, M. (eds.). *Handbook of Applied Spatial Analysis*, Springer-Verlag, Berlin, pp. 301-318.

Griffith, D.A., 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.

Kramer, W., Donninger, C. (1987), "Spatial autocorrelation among errors and the relative efficiency of OLS in the linear regression model". *Journal of the American Statistical Association*, Vol. 82:577-579.

Pace, K., LeSage, J., Zhu, S.(2011), "Interpretation and Computation of Estimates from Regression Models using Spatial Filtering". paper presented to the 10th World Conference of the Spatial Econometrics Association, Toulouse, FR, July 6-8.

Portnoy, S. (1984), "Asymptotic behavior of *M* estimators of *p* regression parameters when p^2/n is large: I. Consistency". *Annals of Statistic,* Vol. 12:1298-1309.

Tiefelsdorf, M., Griffith, D.A. (2007), "Semi-parametric filtering of spatial autocorrelation: the eigenvector approach". *Environment & Planning A*, Vol. 39:1193-1221.