Bayesian estimation of spatial filters with Moran's eigenvectors and hierarchical shrinkage priors

Connor Donegan[1]

Geospatial Information Sciences, University of Texas at Dallas
Population and Data Sciences, University of Texas Southwestern Medical Center

Abstract

This paper proposes a Bayesian method for spatial regression using eigenvector spatial filtering (ESF) and Piironen and Vehtari's (2017) regularized horseshoe (RHS) prior. ESF models are most often estimated using variable selection procedures such as stepwise selection, but in the absence of a Bayesian model averaging procedure variable selection methods cannot properly account for parameter uncertainty. Hierarchical shrinkage priors such as the RHS address the foregoing concerns in a computationally efficient manner by encoding prior information about spatial filters into an adaptive prior distribution which shrinks posterior estimates towards zero in the absence of a strong signal while only minimally regularizing coefficients of important eigenvectors. This paper presents results from a large simulation study which compares the performance of the proposed Bayesian model (RHS-ESF) to alternative spatial models under a variety of spatial autocorrelation scenarios and model specifications. The RHS-ESF model performance matched that of the SAR model from which the data was generated. The study highlights that reliable uncertainty estimates require attention to spatial autocorrelation in covariates. A demonstration analysis of 2016 U.S. Presidential election results further evidences robustness of results to hyper-prior specifications as well as the advantages of estimating spatial models using the Stan probabilistic programming language.

Email address: Connor.Donegan@UTDallas.edu (Connor Donegan)

[1]This research was supported in part by the computational resources provided by the BioHPC supercomputing facility located in the Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, TX. URL: https://portal.biohpc.swmed.edu. The author is also grateful for comments received from Yongwan Chun and Daniel Griffith.

Keywords: Spatial regression, spatial filtering, Bayesian inference, Stan.

1. Introduction

The development of spatial regression models is driven by two primary concerns. The first is that spatially referenced data, like time series data, violate the assumption of exchangeability of observations embedded in classical regression techniques as indicated by the presence of spatial autocorrelation (SA) in model residuals. Correlated observations provide less information than would the same number of uncorrelated observations, resulting in higher observed variance and greater uncertainty of parameter estimates (Griffith, 2011). A second motivation, particularly in ecology, geostatistics generally and public health research, is that spatial patterns in outcome variables may themselves be of primary interest. By incorporating evidence of spatial structure in the outcome variable into the model, spatial regression can identify areas of unexplained and elevated (or depressed) probability of event occurrence which may inform subsequent investigations.

Eigenvector spatial filtering (ESF) models incorporate spatial structure into the linear predictor of any generalized linear model (GLM) with a linear combination of the eigenvectors from a spatial connectivity matrix. Roughly concurrent with similar advances in philogenetic and ecological modeling (Diniz-Filho et al., 1998; Borcard and Legendre, 2002), Griffith (2000) introduced ESF in part as a means to circumvent the difficulties pertaining to the estimation of non-normal probability models with positive SA. ESF methodology has since been extended to address problems of network autocorrelation (Chun, 2008), multi-level models (Park and Kim, 2014; Hu et al., 2018), spatially-varying coefficients (Murakami et al., 2017), spatio-temporal autocorrelation (Griffith et al., 2019) and a variety of problems in ecological modeling (see Griffith et al., 2019, xiii–xiv).

However, the estimation of spatial filters still poses significant challenges. Spatial filters are most often estimated by means of stepwise selection procedures (Tiefelsdorf and Griffith, 2007; Chun et al., 2016). While the vicissitudes of stepwise selection are attenuated by the fact that all candidate eigenvectors are mutually orthogonal (Griffith, 2012), absent a model averaging procedure variable selection methods are unable to account for parameter uncertainty deriving from uncertainty in model selection (LeSage and Pace 2009, Ch. 6; Gargallo et al. 2018). In a simulation study comparing spatially

varying coefficient models, Oshan and Fotheringham (2018) raise concerns with multiple hypothesis testing and find that ESF estimated by stepwise selection overfit the data. Alternative methods for estimating spatial filters which seek to address some or all of these concerns include the lasso (Seya et al., 2015), an empirical Bayes random effects approach (Murakami and Griffith, 2015) and a Bayesian adaptive sampling algorithm (Gargallo et al., 2018).

This article draws on methods for high-dimensional signal detection to propose a Bayesian estimation procedure that avoids fitting multiple model specifications. Hierarchical shrinkage priors are normal distributions with scale parameters that adapt to the data such that posterior distributions of strong signals float towards their maximum likelihood estimate while weak signals are subject to aggressive shrinkage but without ever collapsing onto zero (see Polson and Scott, 2010). This article aims to demonstrate that Piironen and Vehtari's (2017) regularized horseshoe (RHS) prior is a computationally efficient and high performing prior distribution for the coefficients of the spatial filter. This article first explains why the calculation of uncertainty remains a key challenge facing ESF and then introduces the proposed Bayesian model (RHS-ESF). Results from a large simulation study are then presented as arbiter of conceptual disputes found in the literature, particularly regarding the interpretation of variance inflation caused by eigenvectors which are correlated with covariates. The study is designed to evaluate the reliability of the RHS-ESF model relative to alternative ESF methods as well as the restricted spatial regression (RSR) framework (Reich et al., 2006; Hughes and Haran, 2013) and the simultaneous autoregressive (SAR) error model under a variety of SA conditions. Finally an analysis of 2016 Presidential election data demonstrates implementation of the RHS-ESF model while drawing on novel insights derived from the simulation study.

2. ESF regression: current methods and limitations

2.1. ESF theory

ESF regression incorporates spatial autocorrelation into the linear predictor of a GLM with a set of spatially-structured, synthetic control variables from a spectral decomposition of the spatial connectivity matrix. If y_i is a normally distributed set of spatially autocorrelated observations taken at locations $i = 1, 2, ...n$, then following Griffith (2011, Theorem 1-a) we may view y_i as a mixture of normal distributions $y_i = y_i^* + \phi_i$ where $y_i^* \sim N(\mu_i, \sigma^2)$ is a

spatially unstructured (exchangeable) process and $\phi_i \sim N(\theta_i, \omega^2)$, $\frac{1}{n}\sum_i^n \theta_i = 0$ is a spatially structured (ordered) process. Then

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I} + \omega^2\boldsymbol{I}). \tag{1}$$

This conceptualization highlights how SA inflates observed variance and, while the expectation of the mean is unaffected by SA, failure to model spatial structure in any finite sample will tend to produce larger estimation errors. To estimate θ_i and recover y_i*, ESF regression takes the eigenfunctions of a transformed spatial connectivity matrix such as $\boldsymbol{C}$, where $c_{ij} = 1$ if polygons i and j are neighbors and all $C_{ii} = 0$. With projection matrix $\boldsymbol{M} = (\boldsymbol{I} - \mathbf{1}\mathbf{1}'/n)$, where $\boldsymbol{I}$ is the identity matrix and $\mathbf{1}$ an n-by-one vector of ones, the eigenfunction decomposition of the matrix $\boldsymbol{MCM}$ (which appears in the numerator of the Moran coefficient (MC) (Griffith, 2012, 5)) produces n mutually orthogonal, zero mean eigenvectors $\boldsymbol{E}$ and associated eigenvalues $\boldsymbol{\Lambda}$. Each E_i represents a distinct pattern of potential SA, with degree of SA indexed by its eigenvalue λ_i. Figure 1 maps a selection of these eigenvectors based on a connectivity matrix representing the 100 counties of North Carolina.

Griffith and Chun (2014, 1482) summarize the interpretation of $\boldsymbol{E}$:

> The first eigenvector, $\boldsymbol{E}_1$, is the set of real numbers that has the largest MC value achievable by any set of real numbers for the spatial arrangement defined by the spatial weight matrix $\boldsymbol{C}$; the second eigenvector, $\boldsymbol{E}_2$, is the set of real numbers that has the largest achievable MC value by any set that is uncorrelated with $\boldsymbol{E}_1$; the third eigenvector, $\boldsymbol{E}_3$, is the set of real numbers that has the largest achievable MC value by any set that is uncorrelated with both $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ and so on through $\boldsymbol{E}_n$, each eigenvector achieving the largest MC that remains uncorrelated with the preceding $n-1$ eigenvectors.

As illustrated in Figure 1, the first eigenvector with the largest eigenvalue represents a strong, large-scale pattern of SA while eigenvectors corresponding to lower levels of SA depict smaller-scale map patterns. If the spatially structured component $\boldsymbol{\theta}$ can be estimated through a linear combination of these eigenvectors $\boldsymbol{E}\boldsymbol{\beta}_{\boldsymbol{E}}$, where $\boldsymbol{\beta}_{\boldsymbol{E}}$ is a vector of n unknown coefficients, then the data can be modeled as

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu} + \boldsymbol{E}\boldsymbol{\beta}_{\boldsymbol{E}}, \sigma^2\boldsymbol{I}). \tag{2}$$

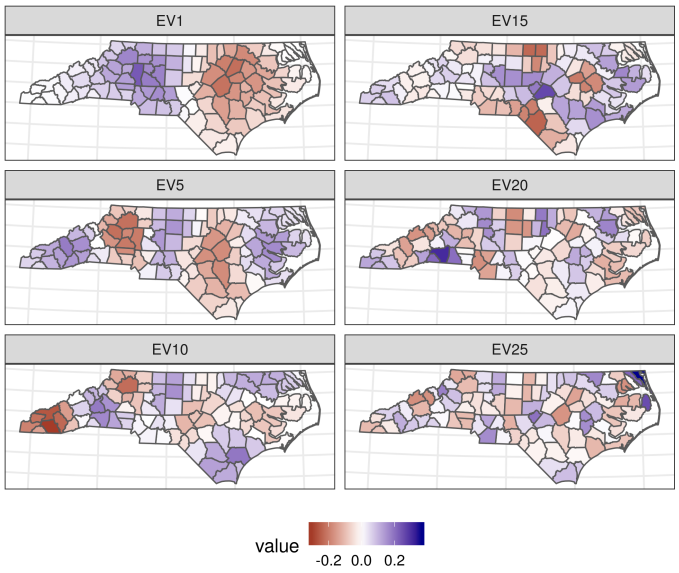

Figure 1: A selection of eigenvectors from the transformed spatial connectivity matrix of 100 counties of North Carolina.

2.2. Current methods for estimating spatial filters

In practice, strong prior information on SA and spatial filters is used to reduce the dimensionality of $\boldsymbol{\beta_E}$. Since most applications are concerned with positive SA, the eigenvectors representing negative SA (whose eigenvalues are negative) are not considered and their coefficients are set to zero. The problem can be simplified further by the exclusion of eigenvectors that represent trace amounts of SA, such as by dropping all eigenvectors for which $\lambda_i/\lambda_{max} < T$, with threshold T set at or below 0.25. For the counties of North Carolina, for example, this practice leaves twenty-four eigenvectors with coefficients to estimate. The most common estimation procedure takes the remaining eigenvectors to be a candidate set and applies a stepwise variable selection procedure to identify a final subset of eigenvectors to include and accepts the maximum likelihood estimate of their coefficients. The stepwise selection procedures may take information criteria (AIC or BIC), coefficient p-values, residual sum of squares or residual SA as their objective function (Chun et al., 2016; Tiefelsdorf and Griffith, 2007).Alternatively, Pace et al.

(2013) argue that a fixed number of eigenvectors may be included based only on tessellation size. Outside of the variable selection paradigm, Murakami and Griffith (2015) apply residual maximum likelihood (REML) to estimate the candidate eigenvector coefficients as a set of random effects (RE-ESF).

2.3. Limitations of current methods

Concerns with automatic variable selection procedures are well known (Chatfield, 1995) and generally apply here. Of particular concern is that all candidate eigenvectors coefficients have non-zero prior probability for inclusion in the model but all that fail to meet the threshold for inclusion are given coefficients of $\hat{\beta}_{Ei} = var(\hat{\beta}_{Ei}) = 0$ which self-evidently eliminates some amount of uncertainty that remains after seeing the data. A proper accounting for uncertainty in the spatial process is important not just in its own right, for the purposes of prediction and possibly cluster detection, but also is necessary for an accurate estimation of the marginal effects of covariates of interest. This section argues that a primary challenge for ESF estimation is the proper calculation of uncertainty across the high-dimensional parameter space of $\boldsymbol{C}$'s spectral decomposition.

Recall that within the frequentist paradigm the variance of a multiple regression coefficient estimate $\hat{\beta}_j$

$$var(\hat{\beta}_j) = \frac{\sigma^2}{var(x_j)(1 - R_j^2)} \tag{3}$$

where σ^2 is the residual variance and R_j^2 is the R^2 statistic from a regression of x_j on the remaining covariates in the design matrix $\boldsymbol{X}$ (Gujarati et al., 2012, 347). The precision of $\hat{\beta}_j$ increases with the variance of x_j and decreases with σ^2 and R_j^2. If $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{B}_{\boldsymbol{x}}$ with $\boldsymbol{\beta}_{\boldsymbol{x}}$ a vector of unknown coefficients then any non-zero pairwise correlation between covariates and eigenvectors necessarily induces additional uncertainty into the estimate of $\boldsymbol{\beta}_{\boldsymbol{X}}$. At the same time as variable selection procedures drop some eigenvectors with non-zero correlations from the model, the (unpenalized) addition of those eigenvectors most correlated with the outcome variable will decrease the residual variance of the regression. Automatic variable selection procedures thus work systematically to minimize the standard error of regression coefficients but without accounting for all sources of uncertainty. The challenge is obviously that a high-variance model with a predetermined number of up to n estimated eigenvector coefficients likewise fails to meaningfully represent the

researchers state of knowledge since it ignores a great deal of prior information regarding the degree of complexity in most spatial processes and may over or under-correct for SA (Chun et al., 2016).

Such correlations between eigenvectors of a spatial connectivity matrix and covariates has been coined 'spatial confounding' (Reich et al., 2006; Hodges and Reich, 2010; Hughes and Haran, 2013). In contrast to the perspective just outlined, Hodges and Reich (2010) argue that the variance inflation caused by correlations between spatial random effects and covariates does not reflect any legitimate inferential uncertainty and can also "mess up" the fixed effects estimates obtained from a non-spatial linear model. This view motivated the search for a model in which "sample size can be discounted without distorting the fixed effect estimate" (Hodges and Reich, 2010, 331). To this end Hodges and Reich propose restricted spatial regression (RSR), a spatial filtering method which introduces the eigenvectors of $\boldsymbol{MCM}$ where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$, so that the eigenvectors are restricted to the space orthogonal to $\boldsymbol{X}$ (Hughes and Haran, 2013).[2] RSR is designed such that no correlated eigenvectors are allowed to 'steal' from the explanatory power that an OLS regression would apportion exclusively to the covariates.

To the extent that correlations between covariates and eigenvectors produce a challenge for probable inference—particularly that of simultaneously estimating a nonstationary mean and the effects of covariates with limited information—surely it should be addressed as such. One reason to remain skeptical of RSR is its substitution of a deterministic separation procedure designed to purge a source of parameter uncertainty from the model, for established inferential methods. The notion that the spatial trend component of a model should have no impact on non-spatial parameter estimates overlooks the geographic analog to Yule's (1926) "nonsense correlations" which routinely appear in nonstationary time-series data and motivated much foundational work in spatial statistics (see Haining, 1991). This suggests that the *a priori* restriction of the space spanned by the spatial filter is counterproductive in the sense that it ignores potentially useful information. These issues are discussed further in Sections 4.2 and 5.

Lastly, Murakami and Griffith's RE-ESF model, which was an extension

[2]Tiefelsdorf and Griffith (2007) previously introduced this method but for different purposes, noting an algebraic relationship to the spatial lag model specification.

of Hughes and Haran's Bayesian estimation procedure, is a highly promising development for ESF since it produces (fast) penalized estimates of $\boldsymbol{\beta_E}$ with empirical Bayes techniques rather than resorting to variable selection methods. However, calculating the degrees of freedom for REML models remains a challenge. Without a proper solution the $var(\hat{\beta}_j)$ is unreliable.[3]

3. High-dimensional Bayesian inference

3.1. Hierarchical shrinkage priors

Sparsity-imposing priors are designed for high-dimensional regression and classification problems in which the researchers expects only a small proportion of many available covariates to be highly relevant to the outcome variable. Hierarchical shrinkage priors are a class of sparsity priors which model a set of coefficients as a mixture of normal distributions. These "global-local scale mixtures of normals" (Polson and Scott, 2010) estimate a separate scale parameter for each coefficient β_j as a product of a global scale parameter τ and a local scale parameter λ_j:

$$\beta_j|\tau^2, \lambda_j^2 \sim N(0, \tau^2\lambda_j^2). \tag{4}$$

A small value for τ expresses our prior expectation that most of the coefficients are near zero while λ_j learns from the data and enables any given parameter to become large. As Polson and Scott (2010) summarize, "To squelch noise and shrink all of the means toward zero, τ^2 should be small. Yet in order for large signals to override this effect, λ_j^2 must be allowed to be quite large" (5). Carvalho et al.'s (2009) horseshoe prior accomplishes this by placing a positive-valued half-Cauchy prior over λ_j

$$\lambda_j \sim C^+(0, 1) \tag{5}$$

centered on zero with scale parameter of 1. This places most prior probability density very near zero for aggressive shrinkage of weak signals but the fat tail of the distribution enables large values to arise without being subjected to any shrinkage. The horseshoe achieves dimensionality reduction without ever causing the posterior distribution of $\hat{\beta}_j$ to collapse completely on zero, following the rules of probability theory.

[3]The `spmoran` software implementation of RE-ESF (Murakami, 2019) uses equation 15 from Müller et al. (2013) to calculate the effective degrees of freedom (Daisuke Murakami, personal communication).

3.2. The regularized horseshoe prior

Piironen and Vehtari (2017) modify the horseshoe prior to address notable limitations. Their regularized horseshoe (RHS) enables researchers to encode prior information regarding the degree of sparsity into hyperprior distributions for τ while also enabling the specification of weakly informative or regularizing priors over any large $\hat{\beta}_j$ (i.e. any which escape extreme shrinkage). Whereas Carvalho et al. (2009) gave τ a half-Cauchy prior $\tau \sim C^+(0,1)$ and Carvalho et al. (2010) allowed τ to scale with the data, $\tau|\sigma \sim C^+(0,\sigma)$, Piironen and Vehtari (2017) incorporate the scale of the data, the number of observations n and the researchers expectation regarding the number of parameters far from zero, p_0, into the hyper-prior distribution

$$\begin{aligned} \tau|\sigma &\sim C^+(0, s_\tau) \\ s_\tau = \tau_0\sigma &= \frac{p_0}{D - p_0}\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{6}$$

where D is the number of parameters in the design matrix (for ESF this is the number of candidate eigenvectors). This is particularly convenient for ESF regression since we have strong prior information on the number of important eigenvectors (Chun et al., 2016). However, this helpful formula for s_τ only applies for normal probability models. For other models such as Poisson or binomial likelihoods s_τ may need to be set directly using model comparison techniques (see Section 5) and trial and error (with $s_\tau = 1$ as in the original horseshoe a sensible starting point).

Since the original horseshoe prior "resembles a spike-and-slab with an infinitely wide slab", results are not reliable when λ_j is poorly identified by the data (Piironen and Vehtari, 2017). The RHS prior addresses this by regularizing the large $\hat{beta}_j$ with weakly informative Student-t "slabs." Since placing an inverse Gamma prior on the scale parameter of a normal prior distribution results in marginal Student-t priors, λ_j is augmented by an auxiliary parameter c with an inverse Gamma prior:

$$\begin{aligned} \beta_j|\lambda_j, \tau, c &\sim N(0, \tau^2\tilde{\lambda}_j^2), \quad \tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{c^2 + \tau^2\lambda_j^2} \\ \lambda_j &\sim C^+(0,1) \\ c^2|\nu_c, s_c &\sim IG(\alpha, \beta), \quad \alpha = \nu_c/2, \quad \beta = \nu_c s_c^2/2. \end{aligned} \tag{7}$$

When the estimate of λ_j is small, the prior scale parameter for β_j ($\tau^2\tilde{\lambda}_j^2$) is near zero and c remains superfluous. When λ_j is large enough to overwhelm

τ, then the auxiliary parameter acts through $\tilde{\lambda}_j$ as a Student-t prior on β_j with degrees of freedom ν_c and scale s_c^2.

3.3. Bayesian ESF

To estimate a Bayesian ESF regression (RHS-ESF) one can place an RHS prior on the eigenvector coefficients in equation 2:

$$\boldsymbol{\beta_E} \sim RHS(\tau_0, \nu_c, s_c). \tag{8}$$

This requires user specification of three hyper-parameters and is aided by the strong prior information from existing literature on spatial filtering. The scale s_c of the Student-t slab in equation 7 can be set nearly automatically based on the scale of the data, such as one-half the standard deviation of the outcome variable divided by the standard deviation of the eigenvectors, which conveniently all have the same scale. Any value for degrees of freedom ν_c ranging from, say, 10 to 20 is reasonable. Setting τ_0 can be guided by Chun et al.'s (2016) equation for the expected number of non-zero eigenvector coefficients

$$\begin{aligned} p_0 &= \frac{n_{pos}}{1+\gamma} \\ \gamma &= exp[\frac{2.1480 - (6.1808(z_{MC}+0.6)^{0.1742})}{n_{pos}^{0.1298} + 3.3534/(z_{MC}+0.6)^{0.1742}}] \end{aligned} \tag{9}$$

where n_{pos} is the total number of positive-valued eigenvalues and z_{MC} is the residual MC z-score obtained from a non-spatial model. This quantity can be entered into equation 6 if the likelihood model for the data is Gaussian or Student's t. Various specifications may be compared using residual SA and Widely Applicable Information Criteria (Watanabe, 2010) (see Section 5). The model can be estimated (and extended arbitrarily as needed) using the Stan probabilistic programming language (Stan Development Team, 2019) making use of Stan code provided by Piironen and Vehtari for the RHS prior.

4. Simulation study

4.1. Study aims and design

The purpose of the simulation study is to evaluate the reliability of the RHS-ESF model with a Gaussian likelihood for the purposes of estimating the marginal effects of covariates and to compare results to commonly used

alternative specifications. Spatially autocorrelated data was simulated from a multivariate normal distribution with a simultaneous autoregressive process in the covariance matrix:

$$\begin{aligned} \boldsymbol{y} &\sim N_k(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}) \\ \boldsymbol{Q} &= (\boldsymbol{I} - \rho\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W})'\sigma^2 \end{aligned} \tag{10}$$

where $\boldsymbol{\mu}$ is the length-k vector, $\boldsymbol{Q}$ is a k-by-k dimensional matrix, $\boldsymbol{W}$ is a k-by-k row-standardized spatial connectivity matrix, $\boldsymbol{I}$ is a k-by-k identity matrix, ρ is a scalar providing the degree of spatial autocorrelation and σ is the standard deviation of $\boldsymbol{y}$. All specifications utilized in this study have $\sigma = 1$. The 100 counties of North Carolina and the queen contiguity condition were used to produce the spatial connectivity matrix.

The mean $\boldsymbol{\mu}$ is a linear function of two covariates:

$$\boldsymbol{\mu} = \mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2}, \quad \beta_1 = 0.5, \quad \beta_2 = -0.5. \tag{11}$$

The two covariates consist of random draws from equation 10 but with $\boldsymbol{\mu}$ set to a vector of zeros. The degree of SA in the outcome variable (ρ_y) and covariates (ρ_x), respectively, were systematically varied to capture a range of SA conditions. For each combination of $\rho_y, \rho_x \in \{.1, .3., .5, .7, .9\}$ $1,000$ sets of data were drawn for a total of $25,000$ data sets.

The following models were fit to each data set: ordinary-least squares regression (OLS), ESF regression fit by stepwise selection (STEP-ESF) with p-value of .1 as the objective criteria, Murakami and Griffith's RE-ESF method using the `spmoran` R package (Murakami, 2019), the RHS-ESF model estimated using Stan and a SAR error model estimated using the `spatialreg` R package (Bivand, 2019). In addition, the STEP-ESF and RE-ESF models were each fit a second time using eigenvectors orthogonal to $\boldsymbol{X}$ (with the eigenvectors being specific to each data set) following Hughes and Haran's RSR theory. Finally, each of the preceding models was fit a second time with the mean spatial lag of each covariate (SLX) added to the linear predictor. The SLX term was calculated by WX with W a row-standardized spatial weights matrix.

Prior distributions for the Bayesian ESF models were set to be weakly informative (i.e. they are centered on zero and are flat across the region of potential parameter values given the scale of the data but taper off outside of that region). The RHS prior were given the following hyper-priors: $\tau_0 = .4$, $\nu_c = 15$, $s_c = 4$ which produces a Student-t slab centered on zero with a

scale of four. In practice more than one value for τ_0 should be tested and results generally should not be accepted if the model has not removed SA from the residuals. For this study the model residuals were tested for SA using the MC; if the MC was higher than its expected value[4] then the degree of sparsity was relaxed by setting $\tau_0 = .9$, the model was fit again and the results were accepted unconditionally.

The estimates of $\boldsymbol{\beta_x}$ are evaluated mainly on two criteria: root-mean squared error (RMSE) $\sqrt{\frac{1}{n}\sum_{i=1}^{n=1,000}(\hat{\beta}_{j_i} - \beta_j)^2}$ and interval coverage defined as the proportion of 95% confidence/credible intervals that contain the true parameter value.[5] Interval coverage for a properly specified model is expected to be near 950 out of each $1,000$ models. A secondary criterion is efficiency, calculated as the mean width of the 95% intervals. For the Bayesian model the mean of the posterior distribution is used for $\hat{\beta}_j$ and the quantiles of the posterior distribution are used to construct the credible intervals. Finally, since x_1 and x_2 are uncorrelated with each other the reported RMSE, interval coverage and efficiency scores are the averages across β_1 and β_2.

4.2. Results

Figure 2 reports RMSE of each model specification across all SA conditions. Consider first the upper half of the figure which shows results for specifications without the SLX term. With $\rho_y = .1$ all models have indistinguishable RMSE, though the RMSE of STEP-ESF is slightly above the others. As ρ_y increases the accuracy of the OLS estimates steadily deteriorate relative to all conventional spatial models. Whereas the RMSE of OLS estimates are approximately 50% larger than the SAR model when $\rho_y = .7$, when $\rho_y = .9$ the magnitude of error is double that of the SAR model. Notably, neither of the RSR specifications manages to meaningfully improve upon OLS. The RMSE for conventional RE-ESF, RHS-ESF and SAR specifications are nearly identical for most SA conditions, though when

[4]The formula for the expected value of the residual MC is $-nTR[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}]/[(n-k)'\boldsymbol{C}\mathbf{1}]$ where TR denotes the matrix trace operator, the n x k design matrix $\boldsymbol{X}$ includes a column of ones and $\mathbf{1}$ is an n x 1 vector of ones (Chun and Griffith, 2013, 18).

[5]Mean estimation error $(\frac{1}{n}\sum_{i=1}^{n=1,000}(\hat{\beta}_{j_i} - \beta_j))$ was confirmed generally to be very small for all models ($< .01$), confirming their unbiasedness. The highest mean errors observed were around .02 (or four percent of β_j) for the OLS and the RSR models with SLX when $\rho_y = .9$.

$\rho_y \in \{.5, .7\}$ the SAR model registers a slight advantage. Finally, we notice that estimation error falls slightly when $\rho_x > .5$. This is the result of the increased $var(x_j)$ which SA induces; the higher variance seems to provide additional information content for the model. Yet with very strong SA in the outcome variable ($\rho_y = .9$), this pattern is reversed.

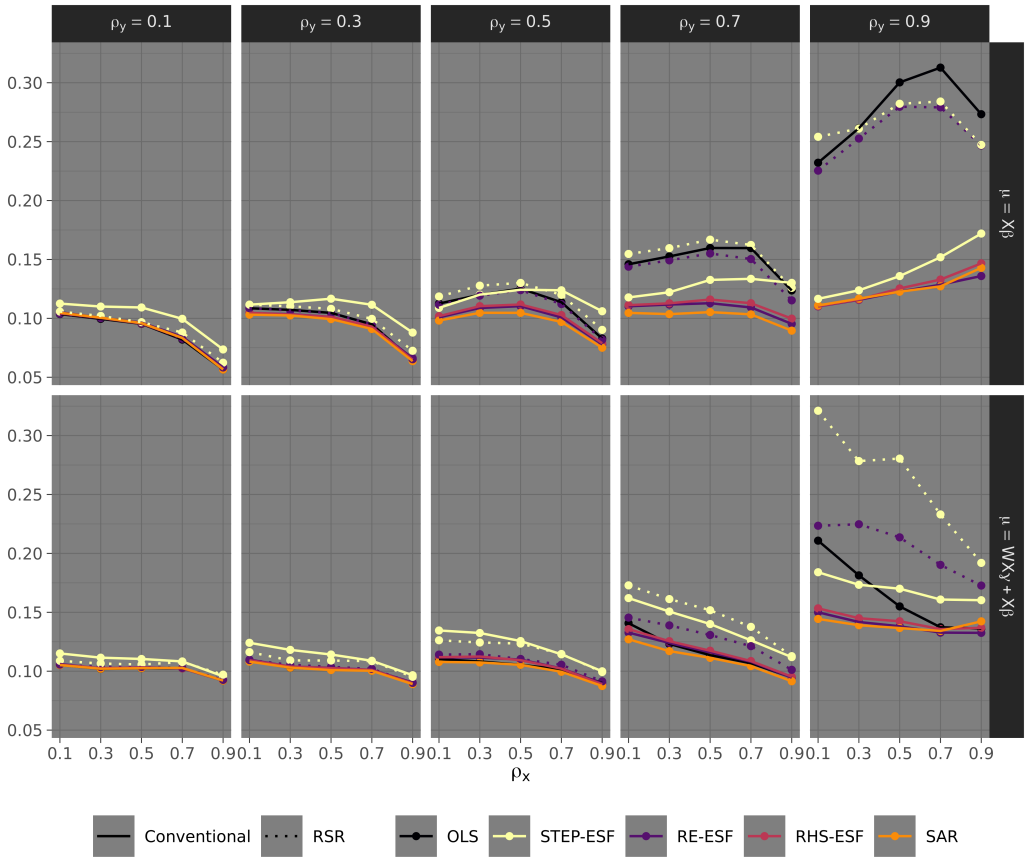

Figure 2: RMSE of parameter estimates.

The bottom panel of Figure 2 reports results for models that include the SLX term. The most notable impact is on the OLS model which produces RMSE comparable to the SAR model in all but a few SA conditions ($\rho_y = .9$ combined with $\rho_x < .7$). Most surprising though is the failure of the RSR specifications to achieve comparable improvement; consequently, OLS outperforms RSR when SLX is included. Interestingly, at lower levels of ρ_y the SLX term flattens out the RMSE trend such that high ρ_x no longer

improves the RMSE; when $\rho_y \in \{.7, .9\}$ the addition of SLX increases error if ρ_x is small but does not cause harm when $\rho_x > .5$. Finally, the small RMSE advantage of the SAR model over RHS-ESF and RE-ESF essentially disappears with the addition of SLX.

Whereas RMSE of the OLS model deteriorates with ρ_y, OLS interval coverage falls as ρ_x rises and the rate at which coverage falls increases with ρ_y (see the top panel of Figure 3). To greatly varying degrees, all of the spatial models exhibit this trend. STEP-ESF and RE-ESF both outperform OLS at high levels of ρ_y but are nonetheless unable to reach 95% coverage even when $\rho_x = .1$. RHS-ESF and SAR specifications share remarkably similar interval coverage: they are both on target when $\rho_x < .5$ and otherwise they are near 2.5% below target. Yet when $\rho_y = \rho_x = .9$ even SAR and RHS-ESF exhibit miserable coverage ($< 82.5\%$). RSR specifications are once again inferior to OLS and often considerably so.

Turning to the bottom panel of Figure 3, the addition of the SLX term has a dramatic impact on interval coverage rates: OLS coverage now exceeds 95% and the trend of declining coverage with rising ρ_x disappears for all ESF specifications (with the exception of $\rho_y = \rho_x = .9$). RHS-ESF maintains coverage of 95% across nearly all SA conditions, SAR coverage is slightly below that of RHS-ESF for $\rho_x > .5$ and the RSR specifications still generally perform more poorly than conventional ESF specifications.

Figure 4 provides some insight into the patterns in coverage rates. It reveals that as ρ_x and thus $var(x)$ increase, all of the intervals become narrower (following the logic of equation 3). Addition of the SLX term counteracts this pattern, dampening but not eliminating the trend. Further, we can see that the SAR and RHS-ESF specifications are far more efficient than OLS. Interval widths of RHS-ESF are slightly wider than those of the SAR model in most conditions, which explains the slight advantage in interval coverage rates apparent in certain SA conditions. The STEP-ESF and RE-ESF specifications, however, produce uncertainty estimates that are consistently even narrower (with and without RSR).

4.3. Discussion

Results from the simulation study confirm the main concerns discussed in the theoretical discussion of Section 2.3, reflect well on the proposed Bayesian ESF model and provide additional insight. As predicted, the STEP-ESF model and the RSR specifications systematically underestimated the degree of uncertainty present in parameter estimates. RE-ESF also underestimated

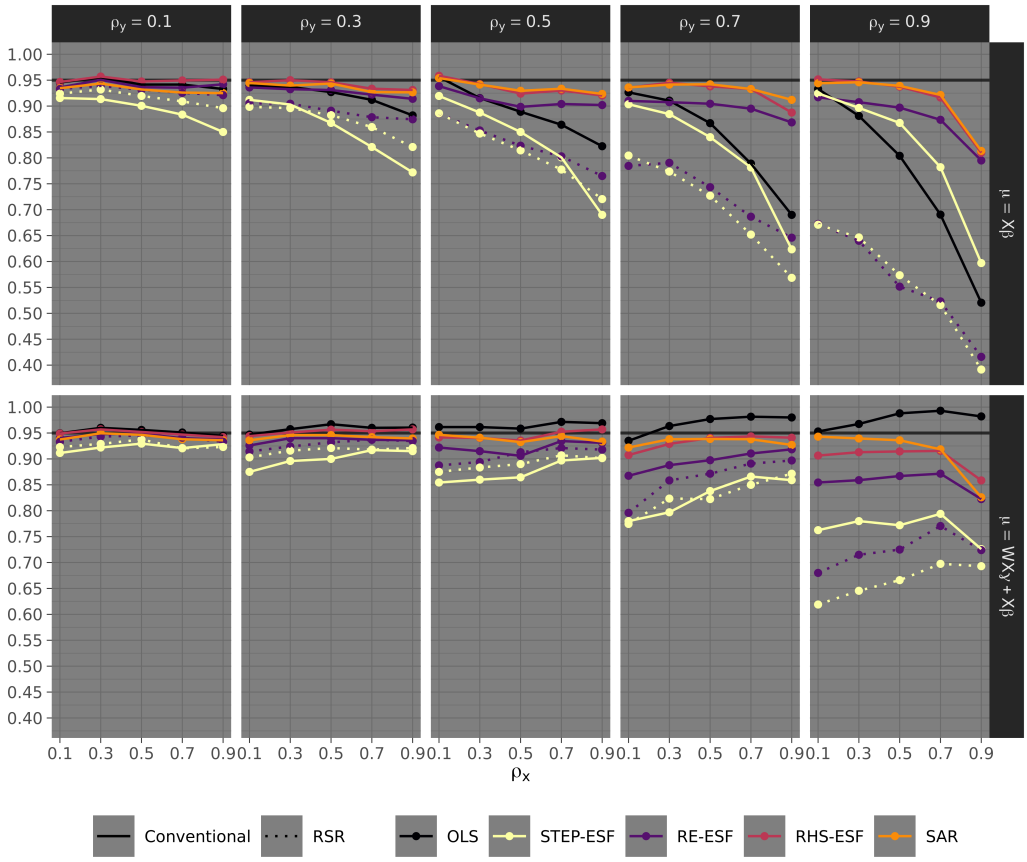

Figure 3: Proportion of 95% confidence/credible intervals containing the true parameter value.

the degree of uncertainty but it is possible that alternative approaches to calculating degrees of freedom would yield satisfactory results. While STEP-ESF and RE-ESF greatly improved upon OLS estimates in terms of RMSE, RSR led to high-error point estimates with overly confident standard errors, a most undesirable combination. This is not surprising given that RSR intends to maintain OLS estimates, which are known to have higher error than spatial models when SA is present. The overly confident uncertainty estimates are simply a result of adding uncorrelated eigenvectors to the linear predictor such that σ^2 is decreased without impacting R_j^2.

On the positive side, RHS-ESF matched and sometimes exceeded the performance of the SAR model in terms of RMSE and interval coverage without

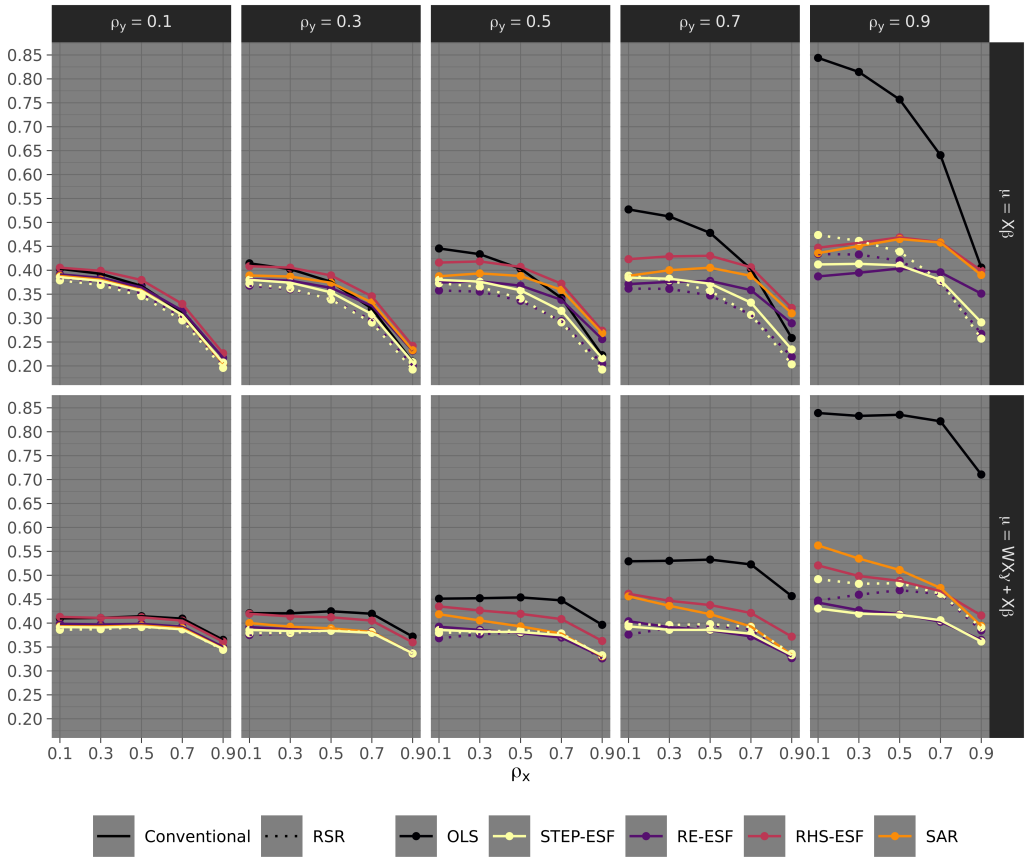

Figure 4: Mean width of 95% confidence/credible intervals.

a notable loss of efficiency. It is important to note that when the covariates were embedded with moderate to strong SA none of the models were able to reach the target interval coverage without the inclusion of the SLX term. Considering both RMSE and interval coverage, the results indicate that when a covariate has $\rho > .5$ it is important to include its mean spatial lag in the model in order to obtain reliable uncertainty estimates. For dependent variables with very strong SA a higher threshold should be set, e.g. $\rho > .65$. This is true even when spatial spillover effects are *not* theorized to be relevant (cf. LeSage and Pace, 2009, 41-2). Indeed, SLX appears nowhere in the data generating model used for this study. The SLX term may be interpreted as a control variable or nuisance parameter that corrects $var(\hat{\beta})$ (or the posterior $p(\beta|y, x, W)$) for the reduced information content of

x relative to an uncorrelated set of observations with the same variance.[6]

5. Demonstration: analyzing unemployment and the Trump vote

This section provides a brief demonstration of the RHS-ESF model through an analysis of 2016 Presidential election results in the state of Ohio. In particular, we will consider the relationship between local unemployment estimates and support for Donald Trump relative to the historic average of the Republican (GOP) vote share (taken from 2004 to 2012). The analysis is relevant for evaluation of competing explanations of the international resurgence of exclusionary populism, often hinging on distinctions between deteriorating economic conditions for low skilled workers and perceived social-status threat among workers of the dominant ethnic group (see Agnew and Shin, 2019).

Table 1 provides summary statistics of the data. Growth in the GOP vote share (`gop_growth`) is defined as Donald Trump's 2016 Presidential vote share minus the historic GOP average (`historic_gop`) (MIT Election Data and Science Lab, 2018). The independent variables included are the natural logarithm of the population size, log of percent of the population age 25 and older with a college degree, percent nonhispanic white (U.S. Census Bureau, 2016; Walker, 2019) and the local area unemployment estimate (Bureau of Labor Statistics, 2016). The variables show moderate to strong SA with the dependent variable and the main explanatory variable of interest having $\rho > .8$.

Figure 5 uses maps and a scatter plot to illustrate the association between GOP growth and unemployment. Both variables share a strong spatial trend, with high values concentrated in Appalchian Ohio (the environmental and cultural region bordering Kentucky, West Virginia and Pennsylvania) and low values surrounding the Cincinnati, Columbus, Toledo and Cleveland metropolitan areas. This is potentially the kind of situation for which RSR was designed with the concern being that the spatial component of the model may arbitrarily detract from the apparently strong relationship (as evidenced by the scatter plot in Figure 5) by removing the shared spatial pattern. A more conventional spatial statistics perspective, buttressed by results in Section 4.2, is that any number of social, economic and environmental variables

[6] As Jaynes (2003) emphasizes, probability theory "is concerned with *logical* connections, which may or may not correspond to causal physical influences" (60). Neighboring values of x may provide information about y independently of any causal relations.

Table 1: Summary statistics for Ohio election data including estimated ρ from an intercept only SAR model.

	Mean	SD	Min	Max	ρ
gop_growth	8.61	6.32	-7.68	25.56	0.82
historic_gop	56.05	9.05	31.44	73.73	0.60
log(population)	4.26	1.00	2.57	7.14	0.64
college_educated	12.28	5.08	5.13	33.59	0.69
white_nonhispanic	90.45	7.51	59.98	97.95	0.32
unemployment	5.50	1.46	3.30	11.10	0.83

surely exhibit a similar spatial pattern. Biomass of standing trees and the beech scale insect invasion are fine examples (see Widmann et al., 2009, 55, 93). Thus a conventional spatial filter will be employed to lessen the risk of such "nonsense correlations" and seek evidence of covariance that exceeds the major SA pattern. SLX terms will be included for all independent variables with $\rho \geq .6$. To ensure the model is robust to the outliers, the model will employ a Student-s t likelihood function with a Gamma prior distribution for the degrees of freedom $\nu \sim G(2, .1)$.

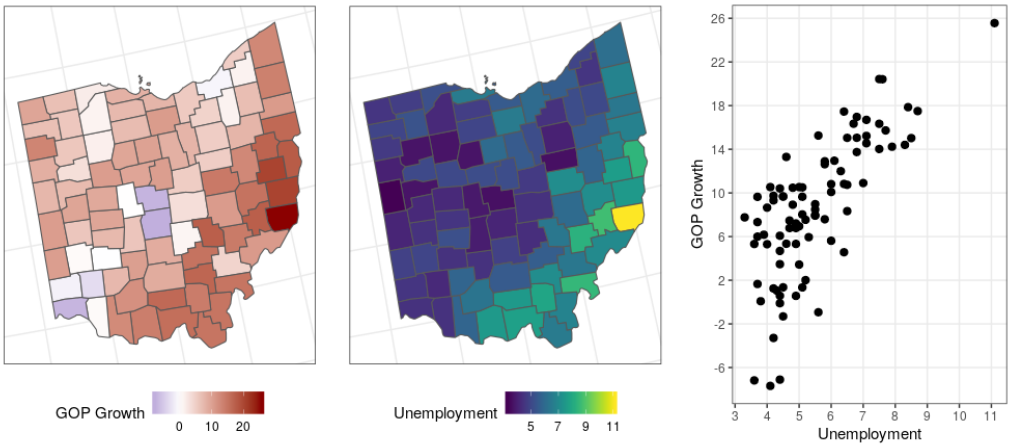

Figure 5: Maps and scatter plot of the unemployment rate and 2016 growth in the GOP Presidential vote share over the historic average.

A potential concern with the RHS-ESF model is its reliance on hyper-prior specifications. If posterior distributions tend to be highly sensitive to

hyper-prior specifications then reliability and ease of use will suffer. Results from the simulation study already suggest this is not the case, as favorable estimates were obtained when hyper-priors were set in crude fashion. Figure 6 reinforces this conclusion. The left panel illustrates residual SA for 31 different values of τ_0 ranging from .01 to 3.01 in the Ohio election model; the middle panel reports how WAIC varies with τ_0.[7] Spatial filters may not be particularly sparse in the sense that $0 \ll \frac{p0}{D} < 1$ and the only *necessary* constraint on the global scale parameter is $\tau > 0$. Lower values of WAIC indicate improved model fit. Across the entire range of τ_0 values considered, residual MC ranges from a maximum of $-.04$ to a minimum near $-.09$, all of which are acceptable. Aiming for a model which first has minimum WAIC and, secondly, has an MC z-score near zero to avoid over-correcting for SA, an optimal specification is somewhere near $\tau_0 = .6$ with similar results obtained within the range $.5 \leq \tau_0 \leq 1$. Equation 9 produces a value of $p0 = 19$ as the expected number of non-zero eigenvector coefficients; plugging this into equation 6, $\tau_0 = .51$, a fine choice. As the right-hand panel of Figure 6 shows, the posterior distribution of β_{unemp} is highly robust even to large changes in τ_0.

Results from the Student-s t RHS-ESF model with $\tau_0 = .51$, $\nu_c = 15$ and $s_c = 5$ and weakly informative Student's t priors on each coefficient are reported in Table 2. Five McMC chains each with $10,000$ samples, the first $3,000$ of each being discarded as warm-up, were combined to construct the posterior distributions. The Rhat statistics of 1.00 indicate convergence across chains. Sampling is highly efficient as the effective number of independent samples in the posterior distributions `n_eff` are all $> 10,000$. Whereas OLS produces $\hat{\beta}_{unemp} = 1.21$ with 95% confidence interval $[.60, 1.83]$, the SAR error model $\hat{\beta}_{unemp} = 1.18$ $[.70, 1.68]$ and the Gaussian RHS-ESF $\hat{\beta}_{unemp} = 1.16$ $[.63, 1.69]$, the robust RHS-ESF estimate is substantially lower, $\hat{\beta}_{unemp} = .85$ $[.41, 1.31]$. Preference for the robust model is supported by WAIC which improved from 365 to 340 simply by replacing the Gaussian likelihood with Student's t. Estimates are thus far more sensitive to a poorly chosen data model than to the hyper-prior settings but currently only Bayesian spatial models can adequately address this. The final results are consistent with the view that poor economic prospects among peripherally located

[7] For each value of τ_0, a single McMC chain was estimated with $14,000$ samples, $4,000$ of them discarded as warm-up.

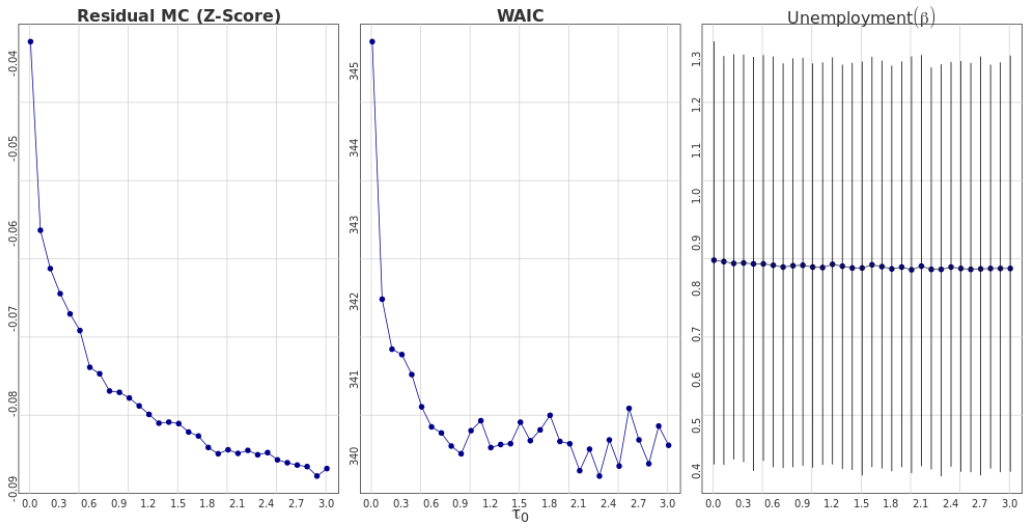

Figure 6: Diagnostics from the Ohio election model for various τ_0 hyper-prior specifications with posterior mean and 95% credible interval of the marginal effect of unemployment.

whites contributed to the collective appeal of the Trump campaign in (what *was*) a key Midwest swing state. Any further substantive interpretation of the results is not appropriate without additional contextual and theoretical discussion which is beyond the scope of this demonstration. For a more thorough analysis the Stan model could be extended to consider measurement error or the entire Midwest with varying intercepts and/or slopes by state, though computation time would no longer be trivial.

6. Conclusion

This study has argued that key challenge for spatial filtering-based regression is the accurate quantification of uncertainty in the spatial component of the model and, by extension, other model parameters such as the coefficients of independent variables. The paper introduced Piironen and Vehtari's RHS prior as a remedy for this challenge since it accomplishes dimensionality reduction without violating the rules of probability theory. The RHS-ESF model was then tested in a simulation study designed to ascertain the reliability of parameter estimates under a wide range of SA conditions. The Bayesian ESF model performed as well as the SAR model from which the data was generated while alternative specifications provided overly confident standard errors. These results reinforce the importance of building models

Table 2: Posterior summary of Student's t RHS-ESF regression on growth in the GOP county-level vote share, from the historic average to 2016. The intercept represents the expected GOP growth when all covariates are at their mean value.

	Mean	SD	2.5%	97.5%	n_eff	Rhat
Intercept	8.68	0.14	8.40	8.97	21034.32	1.00
b_w.historic_gop	0.13	0.06	0.01	0.26	11101.93	1.00
b_w.log.population	1.13	0.82	-0.52	2.67	9896.77	1.00
b_w.college_educated	-0.24	0.15	-0.54	0.06	12363.49	1.00
b_w.unemployment	0.83	0.38	0.05	1.56	13987.27	1.00
b_historic_gop	-0.28	0.04	-0.37	-0.19	13205.62	1.00
b_log(population)	-0.89	0.40	-1.67	-0.09	13938.34	1.00
b_college_educated	-0.47	0.06	-0.59	-0.36	14747.60	1.00
b_white_nonhispanic	0.36	0.05	0.26	0.46	15739.78	1.00
b_unemployment	0.85	0.23	0.41	1.31	17203.00	1.00

that incorporate all sources of uncertainty and also provided insight into the challenge of estimating marginal effects when covariates exhibit strong SA. The results lead to the recommendation that the mean-spatial lag of *high-SA* covariates be included in the model absent some compelling reason not to. Since the spatial lag term provides useful information about the outcome and corrects the estimated uncertainty of the coefficient, including this term is more than appropriate even when no spatial spillover (causal) effect is theorized to occur. However, the simulation study revealed that highly correlated data ($\rho_x, \rho_y > .9$) was exceedingly difficult for all of the models to process, leading to useless uncertainty estimates. Lastly, a demonstration analysis of election data showed that the model is highly robust to hyper-prior specifications while Chun et al.'s formula for the expected number of non-zero eigenvector coefficients appears to be a reliable guide for setting the prior degree of sparsity of the model.

Avenues for further research include extending RHS-ESF to Binomial and Poisson models and particularly to disease mapping applications. At the same time, the simulation study could be extended to consider variation across other important parameters such as tessellation size and shape, the signal-to-noise ration and correlation among covariates, all of which were held constant in this study. While the RHS prior performs exceedingly well, it would be wise to consider alternative sparsity priors as well.

References

Agnew, J., Shin, M., 2019. Mapping Populism: Taking Politics to the People. Rowman & Littlefield Publishers.

Bivand, R., 2019. spatialreg: spatial models estimation and testing. R package version 1.1-3.

Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological modelling 153, 51–68.

Bureau of Labor Statistics, 2016. Local area unemployment statistics. https://www.bls.gov/lau/#cntyaa.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2009. Handling sparsity via the horseshoe, in: Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009, pp. 73–80.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society: Series A (Statistics in Society) 158, 419–444.

Chun, Y., 2008. Modeling network autocorrelation within migration flows by eigenvector spatial filtering. Journal of Geographical Systems 10, 317–344.

Chun, Y., Griffith, D.A., 2013. Spatial statistics and geostatistics: theory and applications for geographic information science and technology. Sage, Los Angeles.

Chun, Y., Griffith, D.A., Lee, M., Sinha, P., 2016. Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. Journal of Geographical Systems 18, 67–85.

Diniz-Filho, J.A.F., de Sant'Ana, C.E.R., Bini, L.M., 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52, 1247–1262.

Gargallo, P., Miguel, J.A., Salvador, M.J., 2018. Bayesian spatial filtering for hedonic models: An application for the real estate market. Geographical Analysis 50, 247–279.

Griffith, D., Chun, Y., 2014. Spatial Autocorrelation and Spatial Filtering, in: Fischer, M.M., Nijkamp, P. (Eds.), Handbook of Regional Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1477–1507.

Griffith, D., Chun, Y., Li, B., 2019. Spatial Regression Analysis Using Eigenvector Spatial Filtering. Academic Press, London.

Griffith, D.A., 2000. A linear regression solution to the spatial autocorrelation problem. Journal of Geographical Systems 2, 141–156.

Griffith, D.A., 2011. Positive spatial autocorrelation, mixture distributions, and geospatial data histograms, in: IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, pp. 1–6.

Griffith, D.A., 2012. Spatial statistics: A quantitative geographer's perspective. Spatial Statistics 1, 3–15.

Gujarati, D.N., Porter, D.C., Gunasekar, S., 2012. Basic Econometrics. McGraw Hill Education, Chennai.

Haining, R., 1991. Bivariate correlation with spatial data. Geographical Analysis 23, 210–227.

Hodges, J.S., Reich, B.J., 2010. Adding spatially-correlated errors can mess up the fixed effect you love. The American Statistician 64, 325–334.

Hu, L., Griffith, D., Chun, Y., 2018. Space-time statistical insights about geographic variation in lung cancer incidence rates: Florida, usa, 2000–2011. International journal of environmental research and public health 15, 2406.

Hughes, J., Haran, M., 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75, 139–159.

Jaynes, E.T., 2003. Probability theory: the logic of science. Cambridge University Press, Cambridge.

LeSage, J., Pace, R.K., 2009. Introduction to spatial econometrics. CRC Press, Boca Raton.

MIT Election Data and Science Lab, 2018. County Presidential Election Returns 2000-2016. https://doi.org/10.7910/DVN/VOQCHQ.

Müller, S., Scealy, J.L., Welsh, A.H., et al., 2013. Model selection in linear mixed models. Statistical Science 28, 135–167.

Murakami, D., 2019. spmoran: Moran's Eigenvector-Based Spatial Regression Models. R package version 0.1.6.2.

Murakami, D., Griffith, D.A., 2015. Random effects specifications in eigenvector spatial filtering: a simulation study. Journal of Geographical Systems 17, 311–331.

Murakami, D., Yoshida, T., Seya, H., Griffith, D.A., Yamagata, Y., 2017. A moran coefficient-based mixed effects approach to investigate spatially varying relationships. Spatial Statistics 19, 68–89.

Oshan, T.M., Fotheringham, A.S., 2018. A comparison of spatially varying regression coefficient estimates using geographically weighted and spatial-filter-based techniques. Geographical Analysis 50, 53–75.

Pace, R.K., LeSage, J.P., Zhu, S., 2013. Interpretation and computation of estimates from regression models using spatial filtering. Spatial Economic Analysis 8, 352–369.

Park, Y.M., Kim, Y., 2014. A spatially filtered multilevel model to account for spatial dependency: application to self-rated health status in south korea. International journal of health geographics 13, 6.

Piironen, J., Vehtari, A., 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics 11, 5018–5051.

Polson, N.G., Scott, J.G., 2010. Shrink globally, act locally: Sparse bayesian regularization and prediction. Bayesian statistics 9, 501–538.

Reich, B.J., Hodges, J.S., Zadnik, V., 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics 62, 1197–1206.

Seya, H., Murakami, D., Tsutsumi, M., Yamagata, Y., 2015. Application of lasso to the eigenvector selection problem in eigenvector-based spatial filtering. Geographical Analysis 47, 284–299.

Stan Development Team, 2019. RStan: the R interface to Stan. R package version 2.19.2.

Tiefelsdorf, M., Griffith, D.A., 2007. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. Environment and Planning A 39, 1193–1221.

U.S. Census Bureau, 2016. American Community Survey 5-year estimates b01003_001, b15003_001, b15003_022, b03002_003. Accessed from the tidycensus R package, Nov. 6, 2019.

Walker, K., 2019. tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 0.9.2.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571–3594.

Widmann, R.H., Balser, D., Barnett, C., Butler, B.J., Griffith, D.M., Lister, T.W., Moser, K., Perry, C.H., Riemann, R., Woodall, C.W., 2009. Ohio Forests 2006. Technical Report NRS-36. U.S. Department of Agriculture. Accessed January 5, 2020.

Yule, G.U., 1926. Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. Journal of the royal statistical society 89, 1–63.