Dimension Reduction and Alleviation of Spatial Confounding for Spatial Generalized Linear Mixed Models

John Hughes [1] and **Murali Haran** [2]

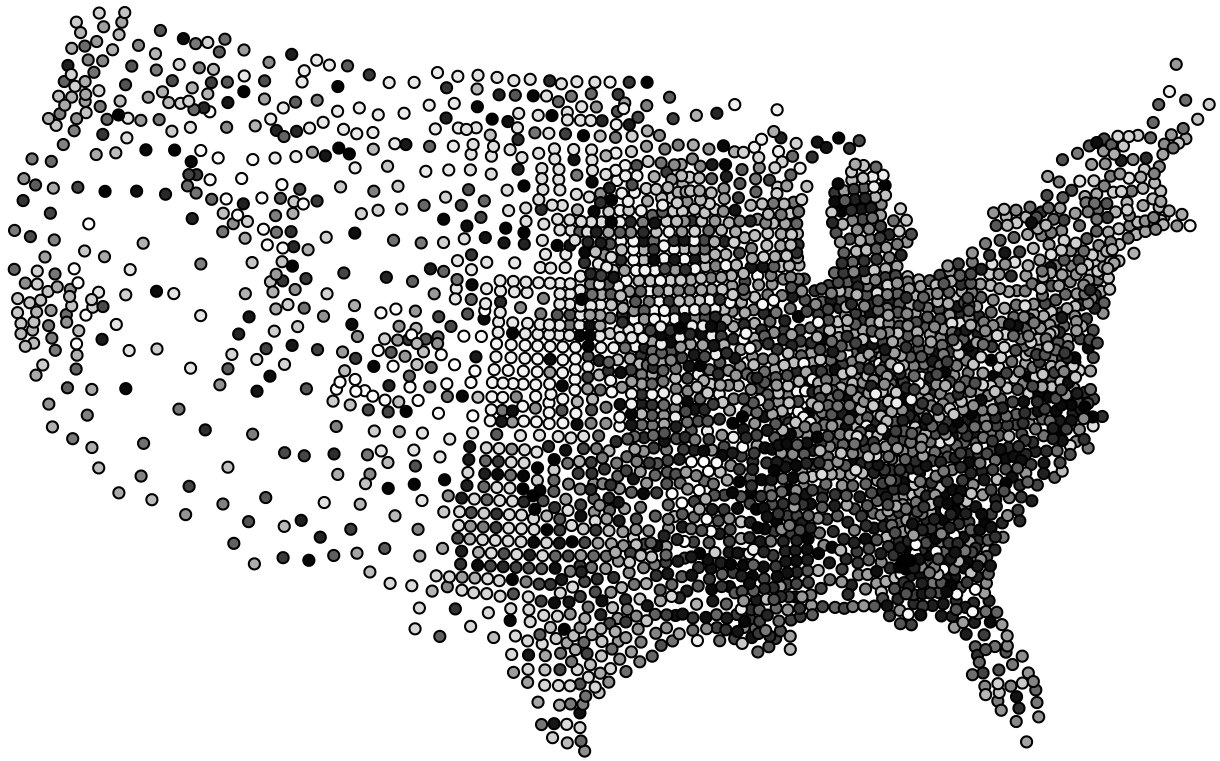
[1] Department of Biostatistics, University of Minnesota

[2] Department of Statistics, Penn State University

The Eighth International Purdue Symposium on Statistics

June 2012

Non-Gaussian Spatial Data



- Non-Gaussian spatial data are very common and appear in a large number of disciplines. E.g. ecology, epidemiology
- Figure: U.S. infant mortality data by county. $n = 3071$
- Common lattice data: normal, binary, count, zero-inflated

Spatial Generalized Linear Mixed Models

- Lattice model: $G = (V, E)$ is underlying graph
- Stage 1: model for spatial data Z at location $\mathbf{s}_i$
  - $f(Z(\mathbf{s}_i)|\beta, \Theta, W(\mathbf{s}_i)), i = 1, \ldots, n$, conditionally independent
  - $g(E(Z(\mathbf{s}_i))) = X(\mathbf{s}_i)\beta + W(\mathbf{s}_i)$
- Stage 2: $\mathbf{W} = (W(\mathbf{s}_1), \ldots, W(\mathbf{s}_n))^T$

  $\mathbf{W}$: Gaussian Markov random field or Gaussian process
- Stage 3: Priors for Θ, β
- Inference based on $\pi(\Theta, \beta, \mathbf{W} \mid \mathbf{Z})$

Originally Besag et al. (1991), Diggle et al. (1998)

SGLMMs: Challenges

SGLMMs have become very popular even outside mainstream statistics. Flexible models but some drawbacks:

(1) Confounding between spatial random effects and "fixed effects" (covariates)

(2) Computational challenges due to high dimensional spatial random effects. Two-pronged issue:

- MCMC is slow per iteration due to high dimensionality
- Markov chain is slow mixing due to strong cross-correlations

Spatial Confounding in SGLMMs

- $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, orthogonal projection onto $\mathrm{span}(\mathbf{X})$
- $\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P}$, orthogonal projection onto $\mathrm{span}(\mathbf{X})$'s orthogonal complement
- Spectral decomposition to acquire orthogonal bases, $\mathbf{K}_{n\times p}$ and $\mathbf{L}_{n\times(n-p)}$, for $\mathrm{span}(\mathbf{X})$ and $\mathrm{span}(\mathbf{X})^{\perp}$. Rewrite:

$$g(\mathbb{E}(Z_i \mid \beta, W_i)) = \mathbf{X}_i\beta + W_i = \mathbf{X}_i\beta + \mathbf{K}_i\gamma + \mathbf{L}_i\delta.$$

$\mathbf{K}$ is collinear with $\mathbf{X}$.

This is the source of confounding. Appears to cause variance inflation.

Spatial Confounding: Reparameterization Solution

- Reich, Hodges and Zadnik (2006) propose solution: since **K** have no scientific meaning, delete them from the model.
- $g(\mathbb{E}(Z_i \mid \beta, \delta)) = \mathbf{X}_i\beta + \mathbf{L}_i\delta$. Prior for random effects δ now

$$p(\delta \mid \tau) \propto \tau^{(n-p)/2} \exp\left(-\frac{\tau}{2}\delta'\mathbf{Q}^*\delta\right),$$

  where $\mathbf{Q}^* = \mathbf{L}'\mathbf{Q}\mathbf{L}$.
- Corrects issues due to confounding
- # of parameters reduced (only slightly) from $n + p + 1$ to $n + 1$. Computational challenge remains.
- RHZ approach ignores underlying graph

Our Sparse Reparameterization

- Represent graph $G = (V, E)$ using $\mathbf{A}$, $n \times n$ adjacency matrix with entries $\mathrm{diag}(\mathbf{A}) = \mathbf{0}$ and $\mathbf{A}_{ij} = 1\{(i,j) \in E, i \neq j\}$, with $1\{\cdot\}$ an indicator function
- Basic idea inspired by Griffith (2003): augment a generalized linear model with selected eigenvectors of $(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{A}(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$. This appears in Moran's I statistic (nonparametric measure of spatial dependence),

$$I(\mathbf{A}) \propto \frac{\mathbf{Z}'(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{A}(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{Z}}{\mathbf{Z}'(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{Z}},$$

Background for Sparse Reparameterization

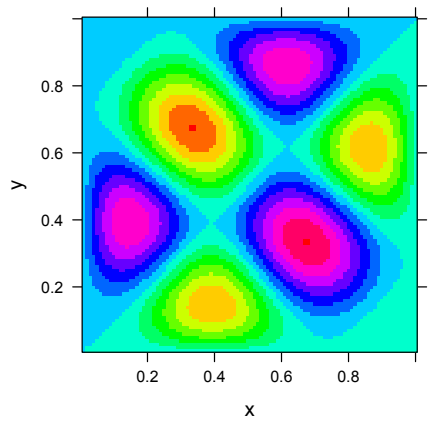
- Griffith's goal: reveal the structure of missing spatial covariates. Our goal: smoothing orthogonal to $\mathbf{X}$
- Hence, we replace $\mathbf{I} - \mathbf{1}\mathbf{1}'/n$ with $\mathbf{P}^{\perp}$
- $\mathbf{M}_{\mathbf{X}}(\mathbf{A}) = \mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}$, Moran operator for $\mathbf{X}$ with respect to the graph G, appears in numerator of generalized Moran's I:

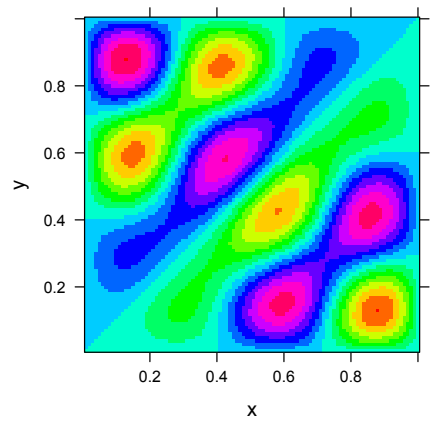
$$I_{\mathbf{X}}(\mathbf{A}) \propto \frac{\mathbf{Z}'\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}\mathbf{Z}}{\mathbf{Z}'\mathbf{P}^{\perp}\mathbf{Z}}.$$

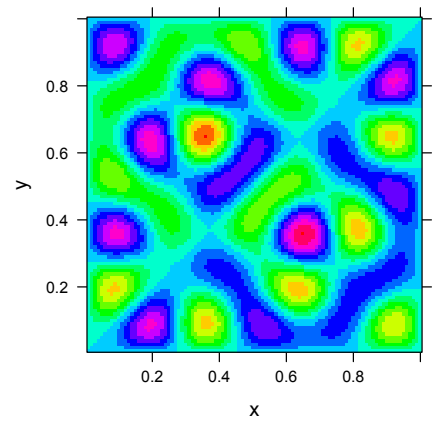
The Resulting Reparameterization: Eigenvectors

- "Tailored" to $\mathbf{X}$ and G: eigenvectors comprise all possible patterns of clustering residual to $\mathbf{X}$ and accounting for G

Some selected eigenvectors for the 30×30 lattice.



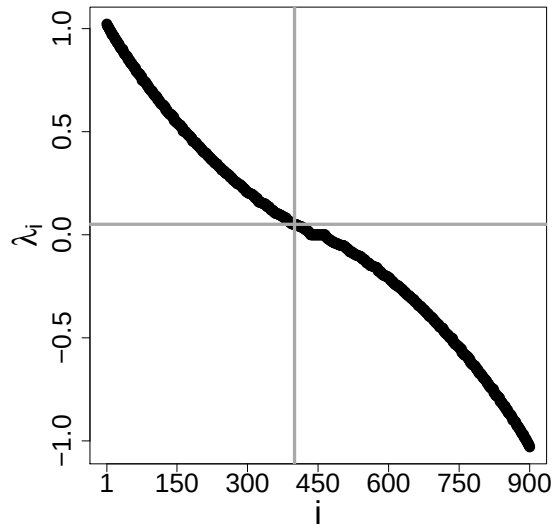




The Resulting Reparameterization: Eigenvalues

- Positive (negative) eigenvalues correspond to varying degrees of positive (negative) spatial dependence (Boots and Tiefelsdorf, 2000)

The standardized eigenvalues for the 30 $\times$ 30 lattice.



Applying the Sparse Reparameterization

- Replacing **L** with **M** in the RHZ model gives

$$g(\mathbb{E}(Z_i \mid \beta, \delta)) = \mathbf{X}_i\beta + \mathbf{M}_i\delta.$$

  And the prior for the random effects is now

$$p(\delta \mid \tau) \propto \tau^{q/2} \exp\left(-\frac{\tau}{2}\delta'\mathbf{Q}^{**}\delta\right),$$

  where $\mathbf{Q}^{**} = \mathbf{M}'\mathbf{Q}\mathbf{M}$.

- Corrects issues due to confounding
- # parameters reduced from $n + p + 1$ to $q + p + 1$. Dramatic speed-up. q is $n/4$ in examples that follow but could be much smaller if model is sparse.

Study: Inference for Spatial Binary

30×30 lattice simulated from RHZ model with $\beta_1 = \beta_2 = 1$.

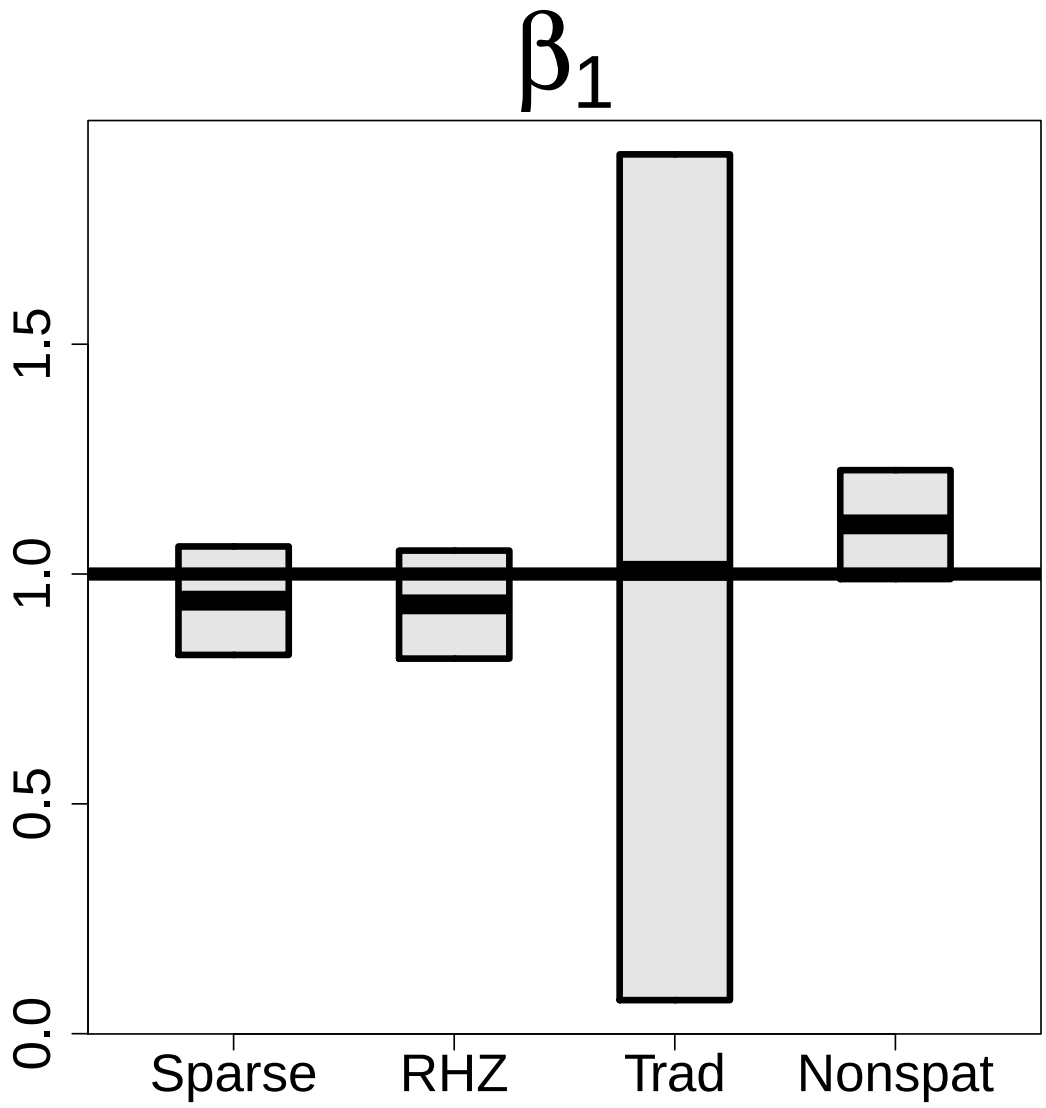
Predictors are the coordinates of unit square.

Model	$\hat{\beta}_1$ CI(β_1)	$\hat{\beta}_2$ CI(β_2)
Sparse	1.080 (0.613, 1.556)	1.130 (0.644, 1.635)
RHZ	1.120 (0.637, 1.606)	1.192 (0.679, 1.713)
Traditional	0.500 (-2.655, 3.616)	-0.605 (-3.698, 2.577)

- Point and interval estimates for Traditional are very poor: 95% interval includes 0
- Sparse and RHZ produce similar (good) results

Similar results for Poisson and Gaussian (linear)

Spatial Count Data: Simulation Results



Spatial Binary: Computational Efficiency

Model	Dimension	Running Time
Sparse	228	2.5 hours
RHZ	901	18.5 hours
Traditional	903	38.5 hours

- MCMC algorithm is
  - faster per iteration (far fewer random effects)
  - mixes faster (random effects are “decorrelated”)
- Far greater speed-ups with much smaller q, e.g. 25-50 is adequate for our examples

Summary

- SGLMMs provide a very general approach for modeling non-Gaussian spatial data
- Our sparse approach results in more interpretable regression coefficients
- We utilize geometry of spatial dependence, a natural approach to dimension reduction
- MCMC easier to construct. Computational efficiency allows for more routine analysis of larger data sets.

References

- Besag, York, Mollie (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*
- Griffith (2003) Spatial Autocorrelation and Spatial Filtering. Springer.
- Reich, Hodges and Zadnik (2006) Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*

Hughes, J. and Haran, M. (2012) "Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models," *Journal of the Royal Statistical Society (B), in press.*

Software: http://www.biostat.umn.edu/∼johnh/software.html