A Projection-based Inferential Approach for Spatial Generalized Linear Mixed Models

Based on joint work with Yawen Guan and John Hughes

Department of Statistics, Purdue University, Spring 2017.

Murali Haran

Department of Statistics, Penn State University

Two Manuscripts

1. Dimension-reduced spatial generalized linear mixed models (SGLMMs) for lattice data

   Hughes, J. and Haran, M. (2013) Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models," *Journal of the Royal Statistical Society Series B*

2. Dimension-reduced spatial generalized linear mixed models (SGLMMs) for continuous domain *and* lattice data

   Guan, Y. and Haran, M. (2016) A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models arXiv

What This Talk is About

- Modeling non-Gaussian spatial data is challenging.
- Spatial generalized linear mixed models (SGLMMs) provide a general framework. They are widely used.
- Shortcomings of SGLMMs: (1) Inference presents difficult computational issues. (2) Parameter interpretation is generally misleading.
- I will describe methods that simultaneously resolve both these issues.
- The methods are projection-based.

Outline

- Spatial linear models
- Spatial generalized linear mixed models (SGLMMs)
  - lattice and continuous-domain spatial data
- Shortcomings of SGLMMs
  - interpretation
  - computing issues
- Projection-based approach to these problems.

Modeling Spatial Data

- Gaussian and non-Gaussian spatial data are very common and appear in a large number of disciplines.
- Common: gaussian, binary, count, zero-inflated
- Purpose of the model
  1. regression while adjusting for residual spatial dependence
  2. smoothing the spatial field and "borrowing strength"
- These models are used widely and have become particularly important in disease epidemiology and ecology.

Infant Mortality

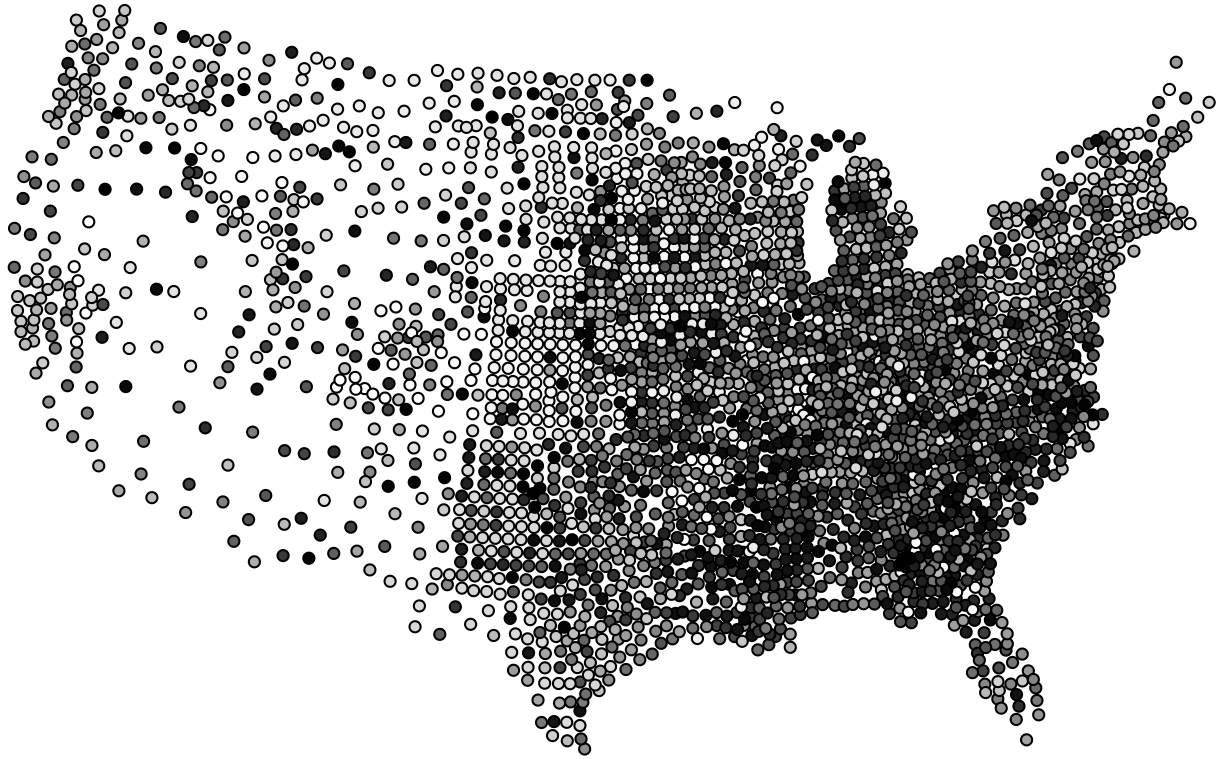


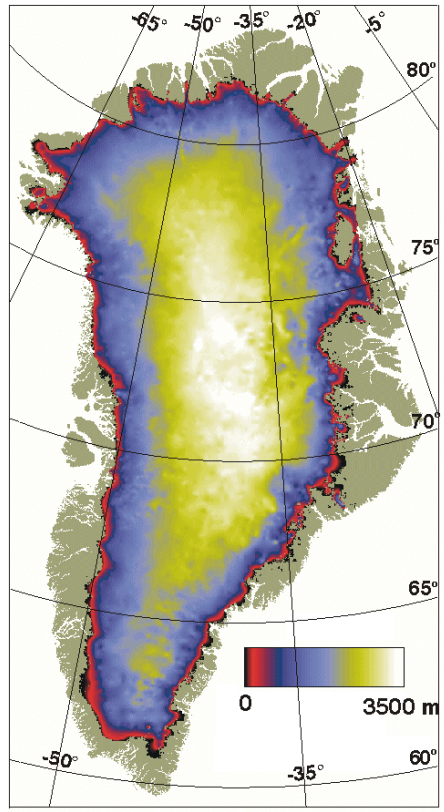
Figure: U.S. infant mortality data by county. $n = 3071$

Ratio of deaths to births, each averaged over 2002-2004.

Darker indicates higher rate.

Ice Sheet Thickness

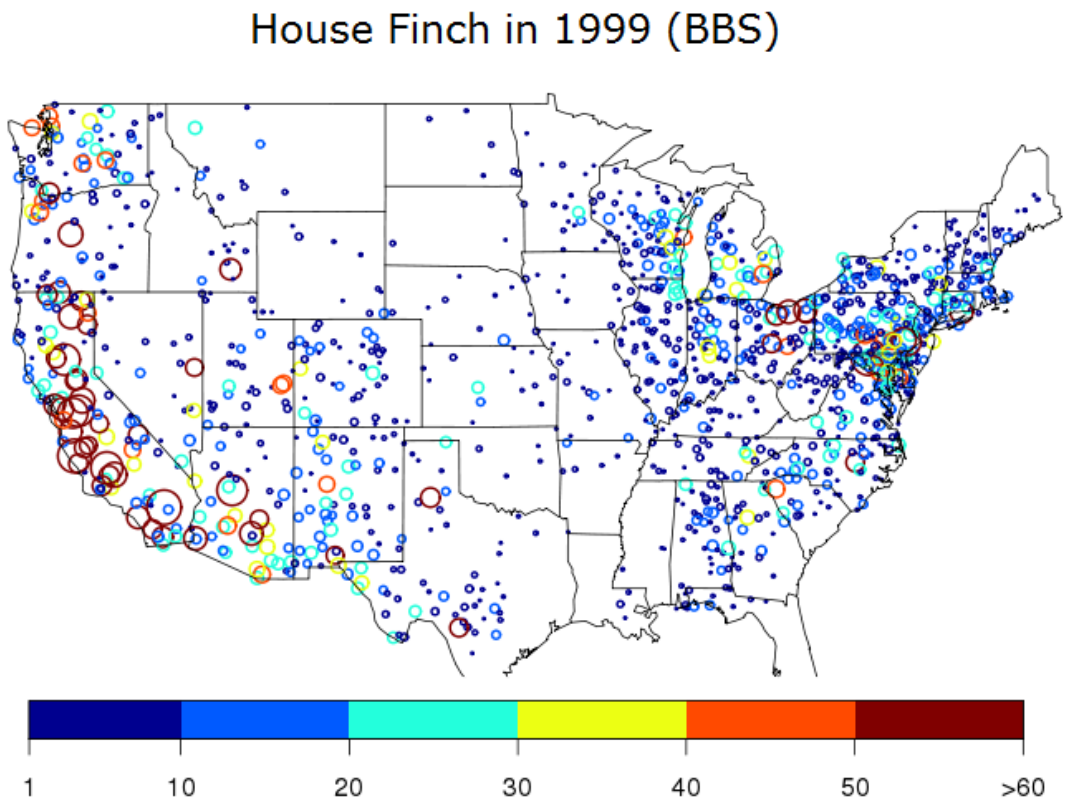
Greenland ice sheet thickness data



Bamber et al. (2001)

Bird Counts

- Data on abundance (counts) of house finch in the US
- Interested in the abundance at unsampled locations.



Pardieck *et al.* 2015. *North American Breeding Bird Survey Dataset 1966 - 2014*

Spatial Linear Mixed Models (SLMMs)

- Spatial process at location $\mathbf{s}$ is $Z(\mathbf{s}) = X(\mathbf{s})\beta + W(\mathbf{s})$.
  - $X(\mathbf{s})$ is covariate at $\mathbf{s}$ and β is a vector of coefficients.
  - Model dependence among spatial random variables by imposing it on $W(\mathbf{s})$s, the random effects.
- Model for $W(\mathbf{s})$s
  - Lattice data: Gaussian Markov Random field (GMRF)
  - Continuous domain: Gaussian process (GP)

SLMMs for Lattice Models

- Gaussian Markov random field

$$W(\mathbf{s}_i) \mid W(\mathbf{s}_{-i}) \sim N\left(\frac{\sum_{j:j\sim i} W(\mathbf{s}_j)}{n_i}, \frac{1}{n_i\tau}\right)$$

  where n_i is number of neighbors of ith region and $j \sim i$ means i, j are neighboring regions

- This specifies $Q(\tau)$, a precision matrix

$$(W(\mathbf{s}_1), \ldots W(\mathbf{s}_n))^T \sim N(0, Q^{-1}(\tau))$$

  $Q = \text{diag}(A\mathbf{1}) - A$, where adjacency matrix A is such that $A_{ij} = 1$ if locations i and j are neighbors, 0 else

SLMMs for Continuous-domain Spatial Data

- Model dependence via a Gaussian process:

$$p((W(\mathbf{s}_1), \ldots W(\mathbf{s}_n))^T \mid \Theta) \sim N(\mathbf{0}, \Sigma(\Theta)),$$

where $\Sigma_{ij} = Cov(W(\mathbf{s}_i), W(\mathbf{s}_j)) = C(||\mathbf{s}_i - \mathbf{s}_j||)$, is specified via a positive definite covariance function with covariance function parameters Θ.

E.g. exponential covariance function with parameters $\Theta = (\sigma^2, \phi, \tau)$.

Spatial Linear Mixed Models: Inference

For both lattice and continuous-domain data:

- Maximum likelihood: maximize $\mathcal{L}(\Theta, \beta; \mathbf{Z})$ w.r.t. Θ, β
- Optimization problem is low-dimensional
- Bayesian inference:
  - Priors for Θ, β
  - Inference based on $\pi(\Theta, \beta \mid \mathbf{Z}) \propto \mathcal{L}(\Theta, \beta; \mathbf{Z}) p(\Theta) p(\beta)$.
- Low-dimensional posterior: use Markov chain Monte Carlo
- Computing: likelihood evaluations involve high-dimensional matrices, n^3 operations
  - GMRFs: sparse matrices $\Rightarrow$ computationally efficient
  - GPs: lots of research, e.g. reduced-rank methods

Spatial Generalized Linear Mixed Models (SGLMMs)

Model for Z at location $\mathbf{s}_i$

1. $Z(\mathbf{s}_i)|\beta, \Theta, W(\mathbf{s}_i), i = 1, \ldots, n$, conditionally independent
   E.g. $Z(\mathbf{s}_i) \mid \beta, W(\mathbf{s}_i) \sim$Poisson$(\mu(\mathbf{s}_i))$
2. Link function $g(\mu(\mathbf{s}_i)) = X(\mathbf{s}_i)\beta + W(\mathbf{s}_i)$
   E.g. $\log(\mu_i) = X(\mathbf{s}_i)\beta + W(\mathbf{s}_i)$
3. Impose dependence: $\mathbf{W} = (W(\mathbf{s}_1), \ldots, W(\mathbf{s}_n))^T$

$$p(\mathbf{W}|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\mathbf{W}'Q\mathbf{W}\right)$$

4. Priors for Θ, β

Inference based on $\pi(\Theta, \beta, \mathbf{W} \mid \mathbf{Z})$

(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

Spatial Confounding in SGLMMs

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

Leads to confounding. This appears to cause variance inflation.

Computing for SGLMMs

MCMC algorithms for SGLMMs are challenging to construct:

- Spatial random effects: one random effect for each data point. $n + p + 1$ dimensions where n=size of data, p =number of predictors. MCMC is slow per iteration due to high dimensionality
- Markov chain is slow mixing due to strong cross-correlations among the spatial random effects.

Several attempts to address these issues: Rue and Held (2005), Haran et al. (2003), Haran and Tierney (2010)

Sketch of Solution

- Spatial random effects **W** are the cause of confounding issues as well as computational challenges.
- **W**: just a device to induce dependence.
- Idea: project **W** to lower dimensional random effects δ
  - Preserve spatial dependence implied by original **W**
  - Project orthogonal to space spanned by **X**
- Applies to both Gaussian process and GMRF models
  - GMRF models: projection based on Moran operator which uses neighborhood structure
  - GPs and GMRFs: general approach using random projections

Spatial Confounding: Reparameterization Solution

- Since **K** is collinear, delete it from model
- $g(\mathbb{E}(Z_i \mid \beta, \delta)) = \mathbf{X}_i\beta + \mathbf{L}_i\delta$. Random effects distribution δ

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

  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.

Reich, Hodges, Zadnik (2006)

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:

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

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

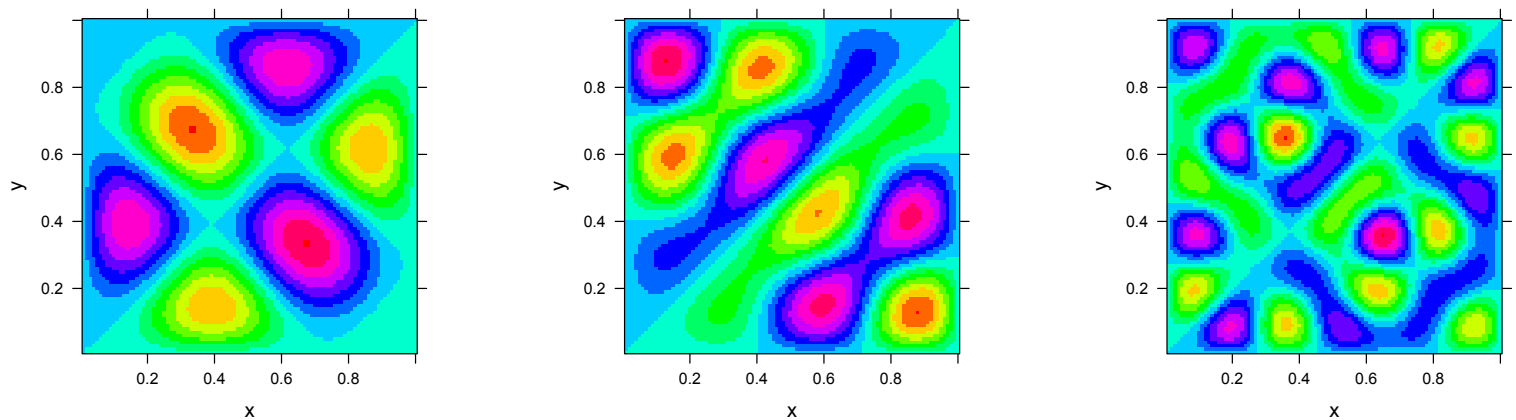
- Corrects issues due to confounding
- Dimension reduction: if $\mathbf{M}_i$ reduced to q dimensions

parameters $q + p + 1 << n + p + 1$ if q is small

Interpreting the Resulting Reparameterization

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

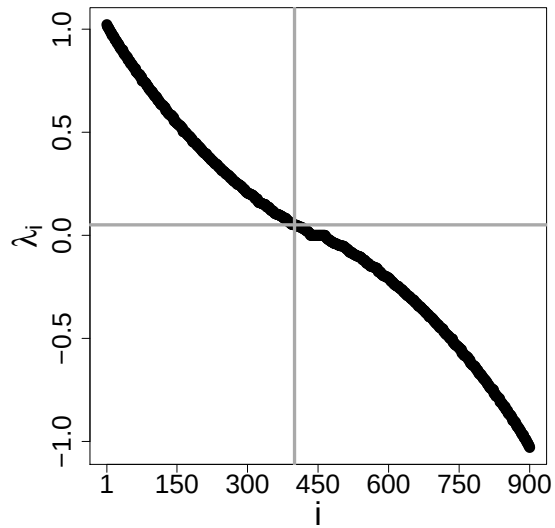
Some selected basis vectors for the 30×30 lattice.



Interpreting the Resulting Reparameterization

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

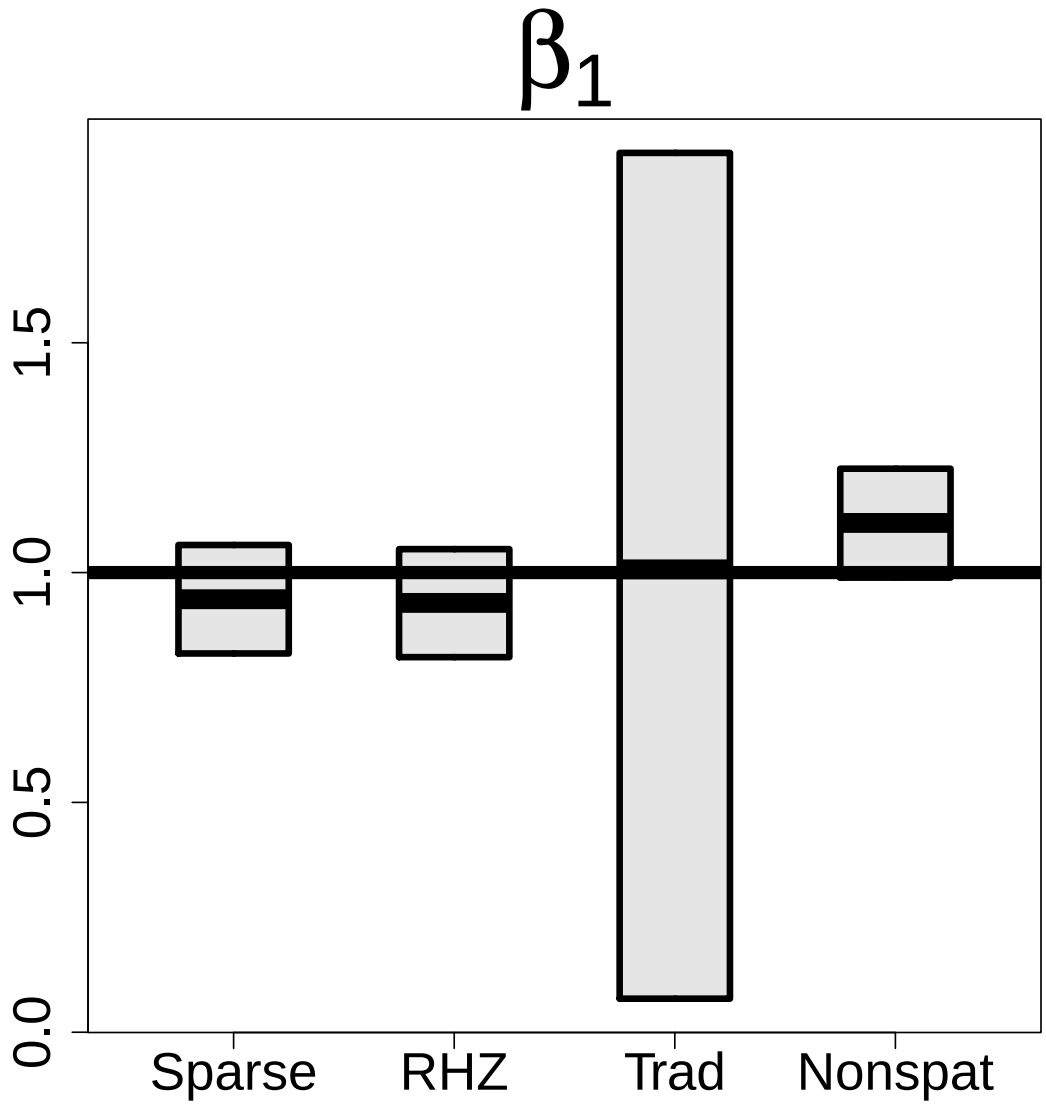
The standardized eigenvalues for the 30×30 lattice.



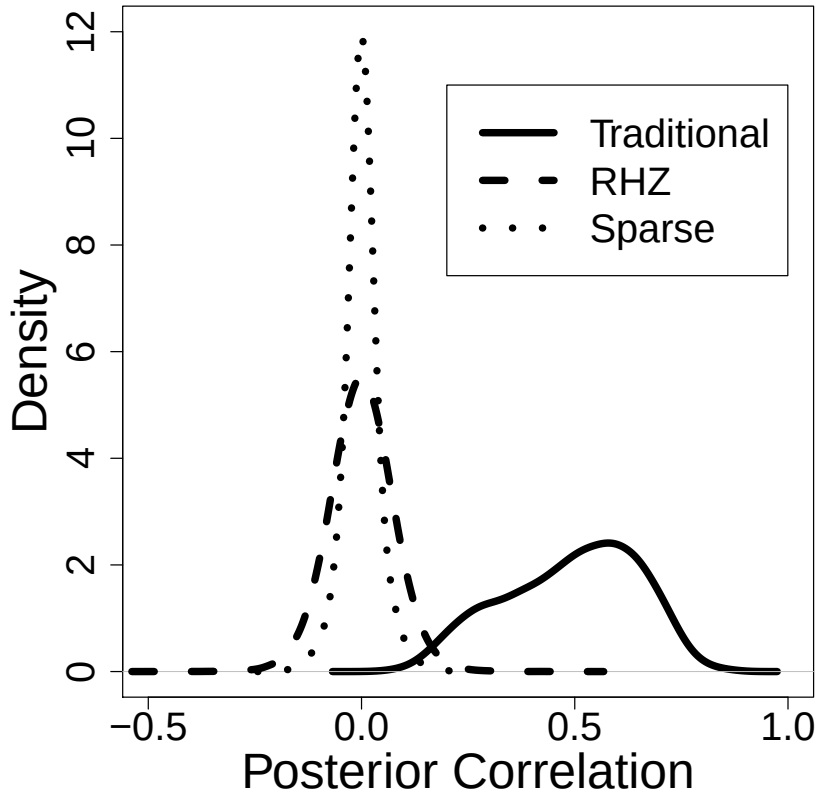
Exploiting the New Parameterization

- If we assume positive spatial dependence, eigenvectors corresponding to negative spatial dependence (negative eigenvalues) should be removed.
- Small eigenvalues may not be meaningful. Remove corresponding eigenvectors.
- Result: much reduced dimensions

Spatial Count Data: Simulation Results



De-correlated Random Effects



Greatly improves efficiency of simple MCMC. No need for elaborate proposals (cf. Held and Rue (2005), Haran et al. (2003), Haran and Tierney (2010)).

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 (we are also being *extremely* careful by running very long chains!)

  Real data example: 14 days (traditional) versus 2-8 hours

Summary for Lattice Models

- SGLMMs provide a very general approach for modeling non-Gaussian spatial data
- Our sparse approach results in more interpretable regression coefficients
- We allow for only meaningful spatial dependence and a natural approach to dimension reduction
- Automated MCMC is computationally efficient, allowing for routine analysis of large data sets

SGLMMs with Latent Gaussian Processes

Recall: example model for count data $Z(\mathbf{s}), s \in \mathcal{D} \subset \mathcal{R}^d$.

1. Data model:

$$Z(\mathbf{s}_i) \mid \beta, W(\mathbf{s}_i) \overset{Indep.}{\sim} \text{Poisson}(\mu(\mathbf{s}_i)), i = 1, \ldots, n$$

$$\log\left(\mu(\mathbf{s}_i)\right) = X(\mathbf{s}_i)\beta + W(\mathbf{s}_i),$$

2. Process model: impose dependence via Gaussian process

$$\mathbf{W} \mid \sigma^2, \phi \sim N\left(\mathbf{0}, \sigma^2 \Sigma_\phi\right)$$

3. Priors for β, σ^2, ϕ

MCMC Inference based on posterior, $\pi(\beta, \sigma^2, \phi, \mathbf{W} \mid \mathbf{Z})$

Reference: [Diggle et al., 1998]

Posterior Distribution

$$\pi(\boldsymbol{\beta}, \sigma^2, \phi, \mathbf{W} \mid \mathbf{Z}) \propto$$

$$\prod_{i}^{n} f\left(Z(\mathbf{s}_i) \mid \boldsymbol{\beta}, W(\mathbf{s}_i)\right) |\sigma^2 \Sigma_\phi|^{-\frac{1}{2}} \exp\left(-\frac{\mathbf{W}' \Sigma_\phi^{-1} \mathbf{W}}{2\sigma^2}\right) p(\boldsymbol{\beta}, \sigma^2, \phi)$$

Outline of Projection-based Approach

1. Fast approximation to the principal components of Σ_ϕ:
   - 1.1 Low-distortion embedding of Σ_ϕ,
   - 1.2 Approximate first m eigenvectors $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and eigenvalues $D_m = \text{diag}(\lambda_1, ..., \lambda_m)$ via Nyström method. [Banerjee et al., 2012] used a similar algorithm to approximate Σ_ϕ in Gaussian process regression
2. Replace n-dimensional $\mathbf{W}$ with $UD_m^{1/2}\delta$

   δ: lower dimensional, components $\approx$independent
3. Project $UD_m^{1/2}\delta$ to $C^\perp(X)$
   - Makes random effects orthogonal to fixed effects
4. Fit the reduced model under Bayesian framework.

Step 1: Approximate Eigenvectors by Nyström's Method

- Partition the positive semi-definite $K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$
- $K_{11} = \Phi^T K \Phi$, where $\Phi = [I_{m \times m}; 0_{m \times (n-m)}]^T$
- Denote V and Λ the eigen vectors and values, respectively.
- Compute exact decomposition for the small $m \times m$ matrix

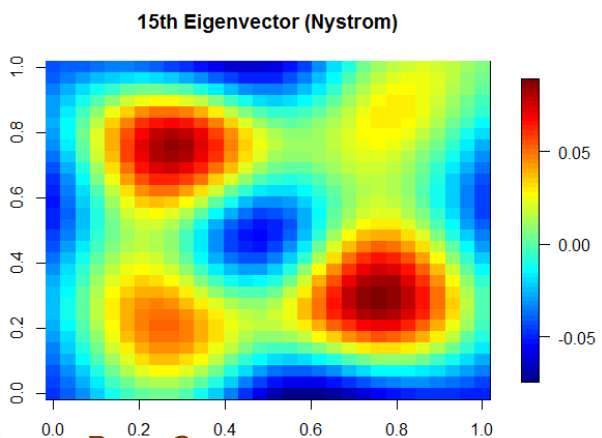
$$K_{11} = V(K_{11}) \Lambda(K_{11}) V(K_{11})^T$$

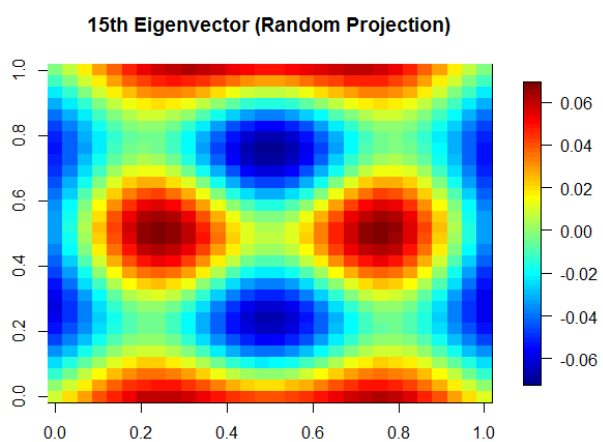
- Approximate $V(K)$ by scaling $V(K_{11})$ up to high dimension [Williams and Seeger, 2001]

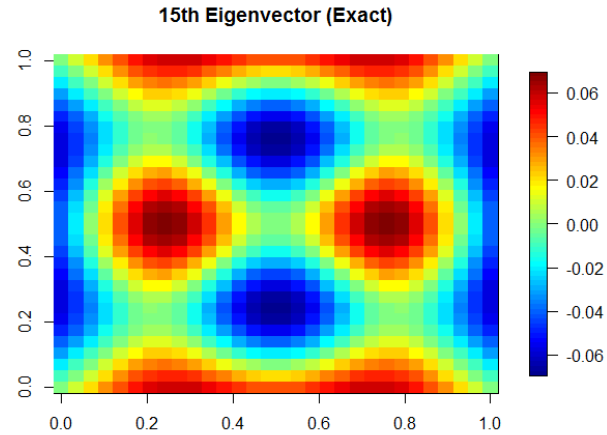
$$\tilde{V}(K) = \sqrt{\frac{m}{n}} \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} V(K_{11}) \Lambda(K_{11})^{-1},$$

Step 1: Approximation using Random Projection

- Originally used to approximate large matrices [Sarló, 2006]
- Efficient Gaussian process regression by approximating the covariance matrix [Banerjee et al., 2012]
- We use random projection to improve approximation of eigen components of Σ_ϕ
- Replace truncation matrix with a random matrix $\Phi = \Omega_{n\times m}$, with $\Omega_{ij} \sim N(0, 1/\sqrt{m})$
- $K_{11} = \Phi^T K \Phi$, and $\tilde{V}(K) = [K\Phi]V(K_{11})\Lambda(K_{11})^{-1}$







Step 2: Random Projection Algorithm Details

Approximates the leading m eigencomponents of the covariance matrix $K = \Sigma_\phi$.

Algorithm 1 Random projection algorithm:

---

1. Low dimensional projection from $R^{n \times n}$ to $R^{n \times k}$, $m < k << n$:
   Form random matrix $\Phi = \Omega_{n \times k}$, or take $\Phi = K^\alpha \Omega_{n \times k}$
2. Nyström's method to approximate eigendecomposition:
   Form $K_{11} = \Phi^T K \Phi$
   SVD for K_{11}: $V(K_{11}) \Lambda(K_{11}) V(K_{11})^T$
   Form Nyström extension $C = [K\Phi][V(K_{11}) \Lambda(K_{11})^{-1/2}]$
   SVD for C: UDV^T
3. Take first m columns of U, and first m diagonal elements of D^2 as our approximation to leading m eigencomponents of K

Replace W with $UD_m^{1/2}\delta$

(Recall) Outline of Projection-based Approach

1. Fast approximation to the principal components of Σ_ϕ:
   - 1.1 Low-distortion embedding of Σ_ϕ,
   - 1.2 Approximate first m eigen-components via Nyström method.

     [Banerjee et al., 2012] used a similar algorithm to approximate Σ_ϕ in Gaussian process regression
2. Replace n-dimensional $\mathbf{W}$ with $UD_m^{1/2}\delta$

   δ: lower dimensional; components $\approx$ independent
3. Project $UD_m^{1/2}\delta$ to $C^\perp(X)$
   - Makes random effects orthogonal to fixed effects
4. Fit the reduced model under Bayesian framework.

Step 3: Orthogonal Projection

- Let $P_{[X]} = X(X^TX)^{-1}X^T$, and $P^{\perp}_{[X]} = I - P_{[X]}$.
- Source of confounding:

$$g\{E(\mathbf{Z} \mid \beta, \mathbf{W}, \sigma^2, \phi)\} = X\beta + \mathbf{W} = X\beta + P_{[X]}(\mathbf{W}) + P^{\perp}_{[X]}(\mathbf{W})$$

- Restricted spatial regression: $P_{[X]}(\mathbf{W})$ is in span of X.
  Remove this to eliminate confounding [Reich et al., 2006].
- Need adjustment for valid inference [Hanks et al., 2015]

$$\beta^{(k)} = \tilde{\beta}^{(k)} - (X^TX)^{-1}X^T\mathbf{W}^{(k)}$$

- Problem: $P^{\perp}_{[X]}(\mathbf{W}) \sim N(\mathbf{0}, P^{\perp}_{[X]}\Sigma P^{\perp}_{[X]})$ is still high-dim.
  If X is nxp input matrix, then $P^{\perp}_{[X]}\Sigma P^{\perp}_{[X]}$ has rank n-p.
- Reduce dimension and confounding by $P^{\perp}_{[X]}UD_m^{1/2}\mathbf{s}\delta$

Step 4: Inference Based on Reparameterizaion

- Spatial generalized linear mixed models:

$$g\{E(Z_i \mid \beta, W_i)\} = X_i\beta + W_i$$

$$\boldsymbol{W} \mid \sigma^2, \phi \sim N_n(\mathbf{0}, \sigma^2\Sigma_\phi)$$

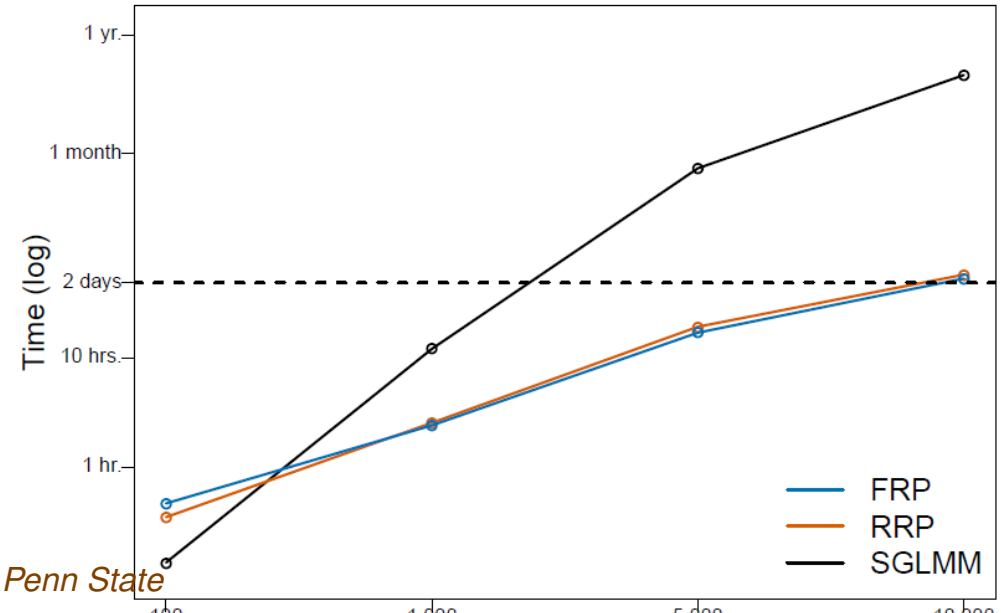
Usual: inference based on $\pi(\beta, \sigma^2, \phi, \mathbf{W} \mid \mathbf{Z})$

- Obtain U, D_m of Σ_ϕ
- D_m is m-dim diagonal matrix with $D_{ii} = i^{th}$ eigenvalue
- FRP: replace $\boldsymbol{W}$ with $UD_m^{1/2}\delta$ to approximate SGLMM or RRP: replace $\boldsymbol{W}$ with $P_{[X]}^{\perp}UD_m^{1/2}\delta$ to approximate restricted spatial model
- Reduced Model:

$$g\{E(Z_i \mid \beta, U, D_m, \delta)\} = X_i\beta + (P_{[X]}^{\perp}UD_m^{1/2})_i\delta$$

Computational Advantages: Reduced Random Effects

- Reduced dimension of random effects δ to $m << n$, e.g. m = 50, n = 1000.
- Restricted spatial generalized linear mixed model with random projections takes about 4 hrs
- Complexity: $O(n^2 m)$ for reparameterization vs $O(n^3)$
- Our approach can fit large data sets within a reasonable amount of time, but fitting the traditional model is infeasible.



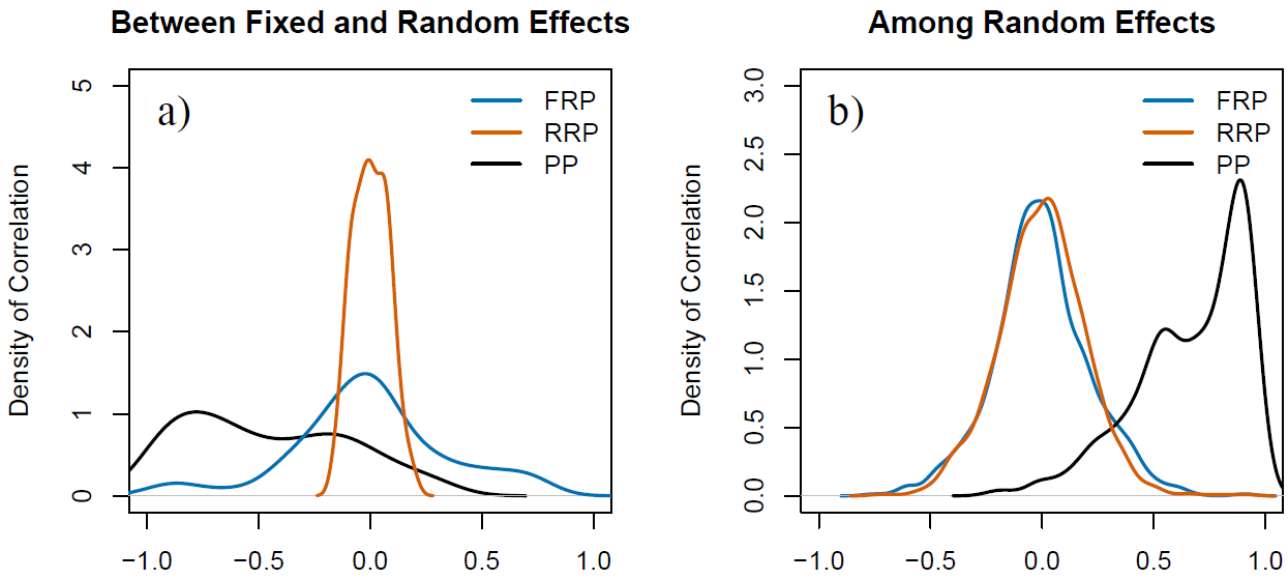
FRP = full model with random projection

RRP = restricted model with random projection

SGLMM = traditional model

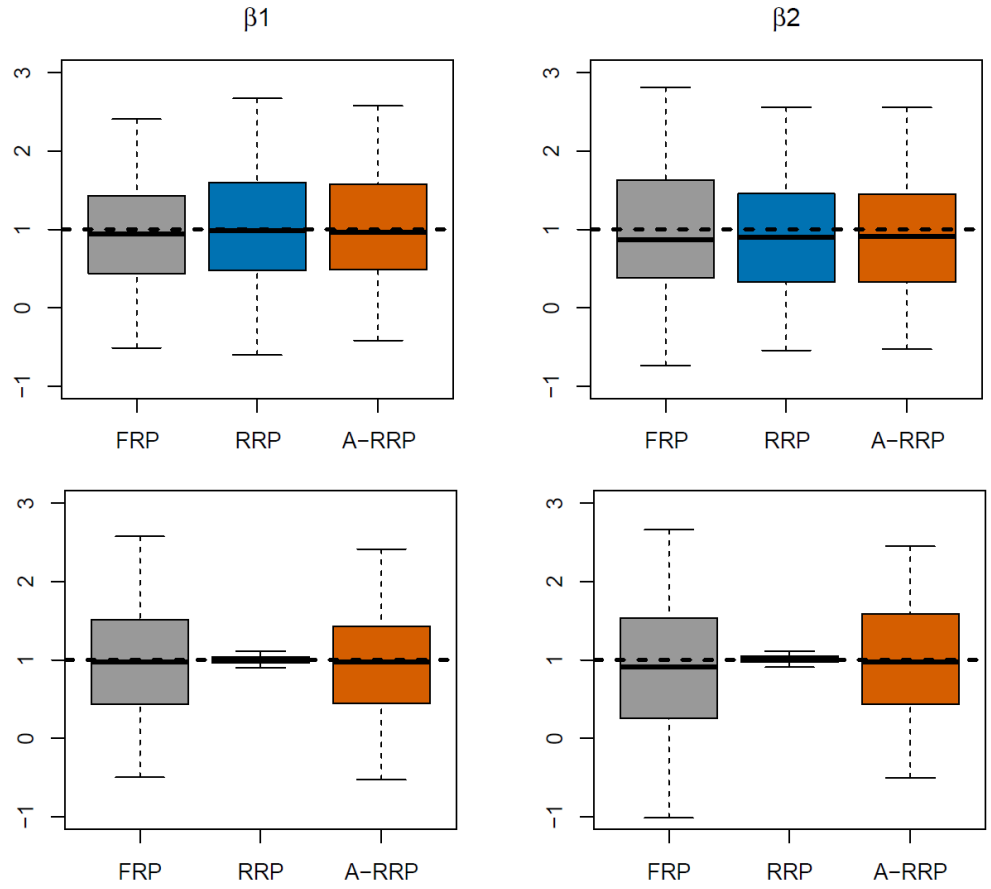
Computational Advantages: Improved MCMC Mixing

- Alleviate confounding between fixed and random effects.
- Reparameterized δ are approximately independent.
- De-correlating random effects results in better MCMC mixing.



Poisson Model Simulation Study: Point Estimation

- Simulate: $\boldsymbol{\beta} = (1, 1)^T$, and Matérn $(\nu, \phi, \sigma^2) = (2.5, 0.2, 1)$



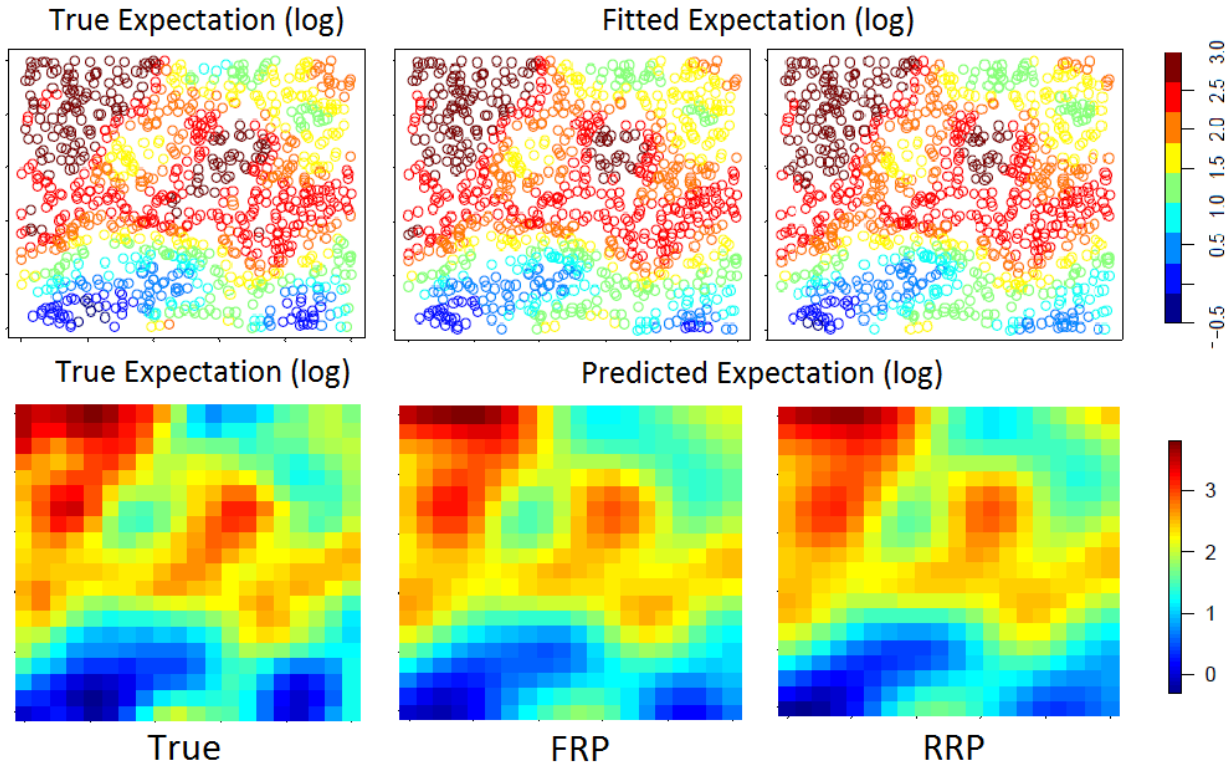
FRP: full model

RRP: restricted model (orthogonalized random effects)

A-RRP: adjusted inference

Poisson Model Prediction Performance

- Simulate n = 1000 spatial count data
- Prediction on 20 x 20 grid using rank = 50



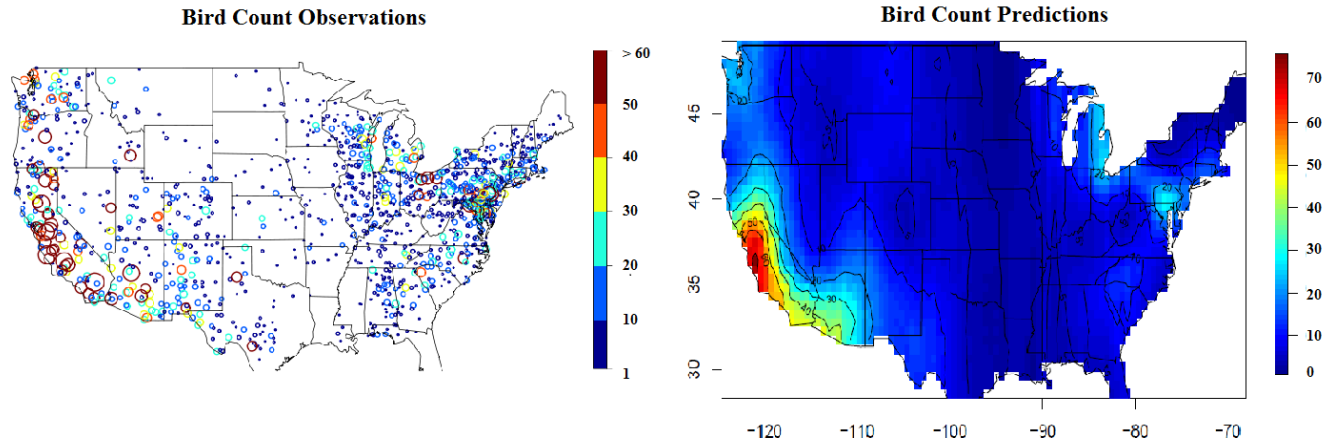
FRP: full model

RRP: restricted model (orthogonalized random effects)

(All utilize random projections)

Interpolated Bird Counts

- Approximate the SGLMM with only the intercept term.
- Computation time is about 7 hours,
- Small bird counts in the center and most of the East Coast
- Large counts centered near New York area and the West



Pardieck *et al.* 2015. *North American Breeding Bird Survey Dataset 1966 - 2014*

Summary

- Projection-based approaches for spatial data
  1. reduces dimensions of posterior distribution
  2. reparameterization improves mixing of MCMC algorithm
  3. adjust for spatial confounding
  4. principled method for determining rank select rank
- We recommend fitting restricted model
- Can adjust for confounding *a posteriori*
- Simulations: good inference and prediction performance
- Caveat: our approach is faster than existing approaches but does not scale to larger data (n > 5,000 may be problematic)

Gaussian Process for Dependence and Interpolation

- A Gaussian process is an infinite-dimensional random process, any finite-dimension of which is a multivariate normal.
- Matérn covariance function describes dependence, e.g.

$\nu = 0.5, \quad C(h) = \sigma^2 \exp(-\frac{|h|}{\phi})$ (Exponential)

$\nu = 2.5, \quad C(h) = \sigma^2 \left(1 + \frac{\sqrt{5}|h|}{\phi} + \frac{5|h|^2}{3\phi^2}\right) \exp(-\frac{\sqrt{5}|h|}{\phi})$

$\nu = \infty, \quad C(h) = \sigma^2 \exp(-\frac{|h|^2}{2\phi^2})$ (Square exponential)