A Projection-based Approach for Spatial Generalized Linear Mixed Models

Based on joint work with

Yawen Guan, SAMSI/NC State

John Hughes, University of Colorado-Denver

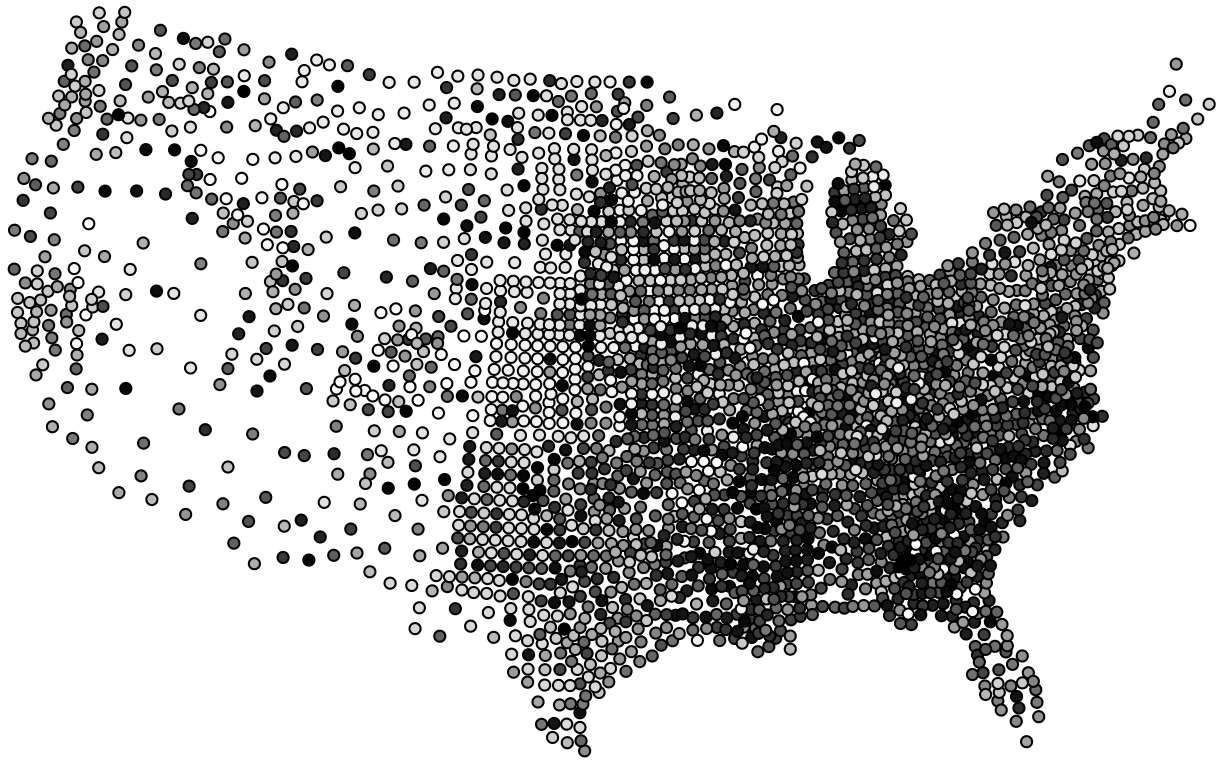
Banff International Research Station, Canada

December 2017

Murali Haran

Department of Statistics, Penn State University

US Infant Mortality Data by County



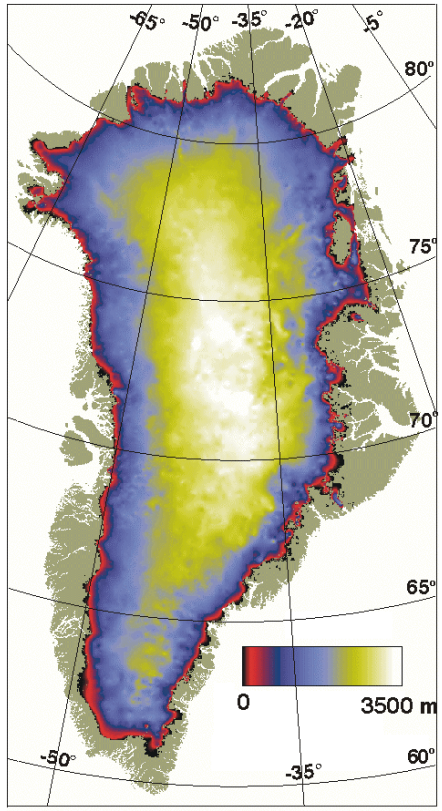
Ratio of deaths to births, each averaged over 2002-2004.

Darker indicates higher rate. $n = 3071$

Question (regression): which factors impact infant mortality?

(Yang, Haran, Matthews, 2008)

Greenland Ice Sheet Thickness

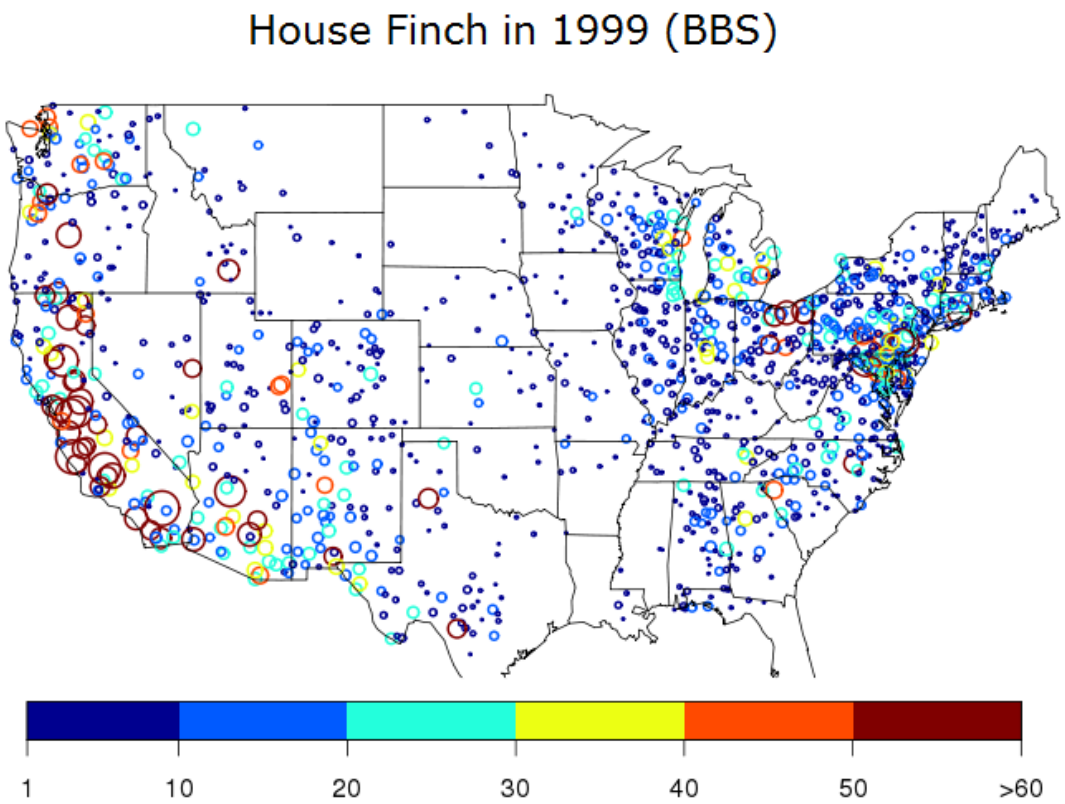


Bamber et al. (2001)

Question: How to interpolate this surface?

How to calibrate (infer parameters for) ice sheet model based on these data? (Chang, Haran, Applegate, Pollard, 2016a,b,c)

House Finch Abundances



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

Question (interpolation): Abundance at unsampled location?

Talk Summary

- Spatial data are common in environmental science: disease modeling, ecology, climate...
- Spatial generalized linear mixed models (SGLMMs)
  - Popular for lattice or areal data

    Besag, York, Mollie (1991) $\approx$ 3,000 citations
  - and continuous-domain data

    Diggle et al. (1998) $\approx$ 2,000 citations

    Broadly: hierarchical spatial models (Banerjee et al. $\approx$ 2,500 citations)
- Shortcomings of SGLMMs:
  1. Computational challenges, especially with large data sets
  2. Regression parameter interpretation is unreliable
- I will describe projection-based methods that simultaneously resolve both these issues

Spatial Generalized Linear Mixed Models

- Spatial linear mixed models (SLMMs): for Gaussian data
- Spatial generalized linear mixed models (SGLMMs): for non-Gaussian data
- What are these models used for?
  1. interpolation (continuous-domain) or smoothing the spatial field (lattice-domain)
  2. regression while adjusting for residual spatial dependence
  3. **as a component in a multi-level hierarchical model**

Spatial Linear Mixed Models (SLMMs)

- Spatial process at location $\mathbf{s} \in D \subset \mathbb{R}^d$ 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 $W(\mathbf{s})$, the random effects
- Same framework works for both lattice data and continuous-domain data. Model for $W(\mathbf{s})$
  - Continuous domain: Gaussian process (GP)
  - Lattice data: Gaussian Markov Random field (GMRF)

Gaussian Processes

Infinite dimensional process $\{W(\mathbf{s}) : \mathbf{s} \in D\}$ such that

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

- Covariance often specified via a positive definite covariance function with parameters Θ
- E.g. (stationary) exponential covariance function
- $\Theta = (\sigma^2, \phi)$

$$\Sigma_{ij}(\Theta) = \text{Cov}(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \sigma^2 \exp(-|\mathbf{s}_i - \mathbf{s}_j|/\phi)$$

Gaussian Markov Random Fields

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

$Q(\Theta)$ is a precision matrix based on a graph that describes a neighborhood structure: adjacencies specify dependence

(skip details....)

Inference for Spatial Linear Mixed Models

- MLE involves low-dimensional optimization

$$\arg\max_{\Theta,\beta} \mathcal{L}(\Theta, \beta; \mathbf{Z})$$

- Bayesian inference:
  - Priors for Θ, β
  - Inference based on $\pi(\Theta, \beta \mid \mathbf{Z}) \propto \mathcal{L}(\Theta, \beta; \mathbf{Z}) p(\Theta) p(\beta)$
- Markov chain Monte Carlo with low-dimensional posterior

Literature on Computing for Spatial Linear Models

- Likelihood: high-dimensional matrices, $\mathcal{O}(n^3)$ operations
- Lots of excellent approaches that scale very well
  - Multiresolution methods, with parallelizations (Katzfuss, 2017; Katzfuss and Hammerling, 2014)
  - Nearest neighbor process (Datta et al., 2016)
  - Random projections (Banerjee, A., Tokdar, Dunson, 2013)
  - Stochastic PDEs (Lindgren et al., 2011)
  - Lattice kriging (Nychka et al., 2010)
  - Predictive process (Banerjee, Gelfand, Finley, Sang 2008)

Largely a "solved" problem

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. $\mathbf{W} = (W(\mathbf{s}_1), \ldots, W(\mathbf{s}_n))^T$ modeled as
   - Gaussian Markov random field model (Besag et al., 1991)
   - Gaussian processes (Diggle et al., 1998)

4. Priors for Θ, β

Commonly embedded within hierarchical models (cf. Banerjee, Carlin, Gelfand, 2014)

Problem 1. Computational Challenge

- MLE: low-dimensional optimization of *integrated* likelihood

$$\arg\max_{\Theta,\beta} \int \mathcal{L}(\Theta, \beta, \mathbf{W}; \mathbf{Z}) d\mathbf{W}$$

High-dimensional integration due to **W**

MCMC-EM or MCMC-MLE: slow, challenging to implement

(Zhang, 2002, 2003; Christensen, 2004)

- Bayesian inference based on

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

Bayes for SGLMMs

- Handle missing data easily
- Combine multiple data sets and uncertainties elegantly
- Rich inference about parameters, functions of parameters
- MCMC-based inference is easier than for MLE

Computing for SGLMMs

But... MCMC algorithms are not easy/scalable

- MCMC is slow per iteration due to high-dimensional

$$\pi(\Theta, \boldsymbol{\beta}, \mathbf{W} \mid \mathbf{Z})$$

- Markov chain is slow mixing (need longer chain) due to strong cross-correlations among $\mathbf{W}$
- Can become impractical for large N
- Impetus for very fast, popular non-MCMC approach: INLA and follow-up work (Rue, Lindgren, Simpson....*)
  Later: Our approach may be combined with INLA

Computing for SGLMMs

But... MCMC algorithms are not easy/scalable

- MCMC is slow per iteration due to high-dimensional

$$\pi(\Theta, \boldsymbol{\beta}, \mathbf{W} \mid \mathbf{Z})$$

- Markov chain is slow mixing (need longer chain) due to strong cross-correlations among $\mathbf{W}$
- Can become impractical for large N
- Impetus for very fast, popular non-MCMC approach: INLA and follow-up work (Rue, Lindgren, Simpson....*)
  Later: Our approach may be combined with INLA

MCMC for SGLMMs

- Markov chain is slow mixing (need longer Markov chain) due to strong cross-correlations among $\mathbf{W}$
- Block updating schemes may help. E.g. blocks:

$$\boxed{\pi(\mathbf{W} \mid \Theta, \beta, \mathbf{Z})} \quad \boxed{\pi(\Theta \mid \beta, \mathbf{W}, \mathbf{Z})} \quad \boxed{\pi(\beta \mid \Theta, \mathbf{W}, \mathbf{Z})}$$

  - Challenging to obtain good proposals for $\mathbf{W}$, especially for high-dimensions
  - Computationally expensive per update

Attempts to address these issues: Rue and Held (2005), Christensen et al. (2006), Haran and Tierney (2003/2010)

They do not scale well (problem for $N > 1000$)

Problem 2. Spatial Confounding

- Let $P = X(X^T X)^{-1} X^T$, and $P^{\perp} = I - P$

$$g\{E(\mathbf{Z} \mid \beta, \mathbf{W}, \Theta\} = X\beta + \mathbf{W} = X\beta + \boxed{P\mathbf{W}} + P^{\perp}\mathbf{W}$$

- $P\mathbf{W}$ is in span of X
- Basic regression issue: multicollinearity

Leads to variance inflation, unstable estimates of β

(Hodges and Reich 2010; Paciorek, 2010)

Hints of the symptom, without diagnosis, by others (e.g. Diggle, 1994)

Sketch of Our General Solution

- Culprit: **W** is cause of confounding as well as computational challenges
- **W**: just a device to induce dependence
- Idea: project **W** on random effects δ such that
  - Preserve spatial dependence implied by original **W**
  - δ is low-dimensional
  - δ is less dependent ("cross-correlated")
  - 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 (Hughes and Haran, 2013)
  - GPs and GMRFs: general approach using eigendecomposition (Guan and Haran, 2017)

Outline of Projection-based Approach

1. Fast approximation to the principal components of Σ_ϕ
   - Approximate first m eigenvectors $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and eigenvalues $D_m = \text{diag}(\lambda_1, ..., \lambda_m)$
2. Replace n-dimensional **W** with $UD_m^{1/2}\boldsymbol{\delta}$

   $\boldsymbol{\delta}$: lower dimensional and $\approx$ independent

   faster and better mixing MCMC algorithm
3. Project $UD_m^{1/2}\boldsymbol{\delta}$ to $C^{\perp}(X)$

   Makes random effects orthogonal to fixed effects

   handles confounding issues
4. Fit the reduced model under Bayesian framework

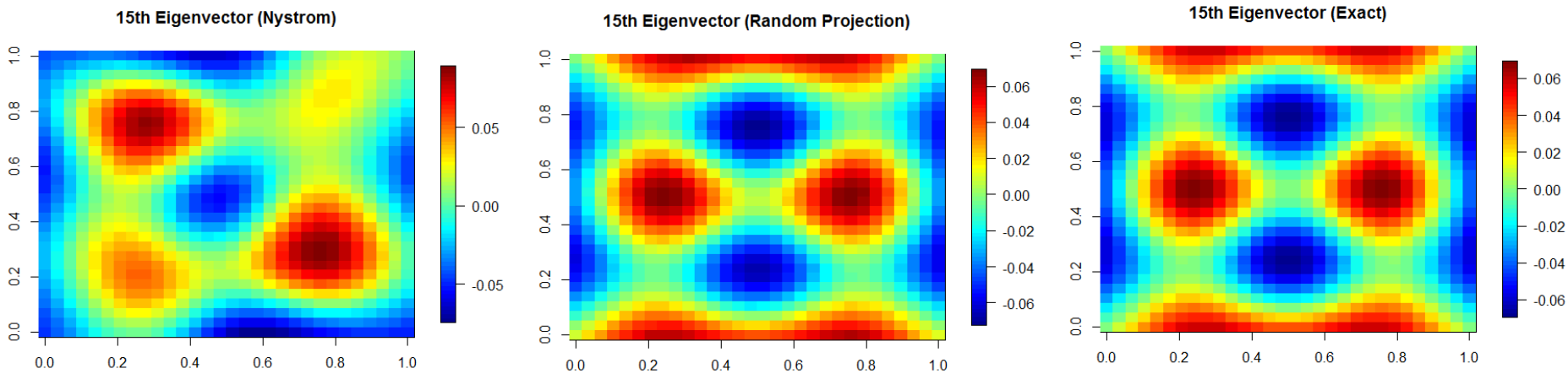
Step 1: Eigendecomposition

For speed we use a fast *approximate* eigendecomposition

Left: deterministic approximation

Center: **random approximation**

Right: exact eigendecomposition



- **Random projections** used in Banerjee, Tokdar, Dunson (2013); also Sarlos (2006), Halko et al. (2009)

Step 2: Reducing Dimensions via Projection

- Approximates the leading m eigencomponents of the covariance matrix Σ_ϕ
- **Replace W with** $UD_m^{1/2}\delta$

Step 3: Projection to Handle Confounding

- Let $P = X(X^T X)^{-1} X^T$, and $P^{\perp} = I - P$
- Recall: $P\mathbf{W}$ is in span of X, causes confounding
- Solution: Remove it

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

[cf. Reich et al., 2006; Hughes and Haran, 2013]

Step 4: Inference Based on Reparameterization

- Spatial generalized linear mixed models

  Usual: inference based on $\pi(\boldsymbol{\beta}, \sigma^2, \phi, \mathbf{W} \mid \mathbf{Z})$
- Obtain U, D_m of Σ_ϕ
- D_m is m-dim diagonal matrix with $D_{ii} = i^{th}$ eigenvalue
- FRP: replace $\mathbf{W}$ with $UD_m^{1/2}\boldsymbol{\delta}$ to approximate SGLMM or

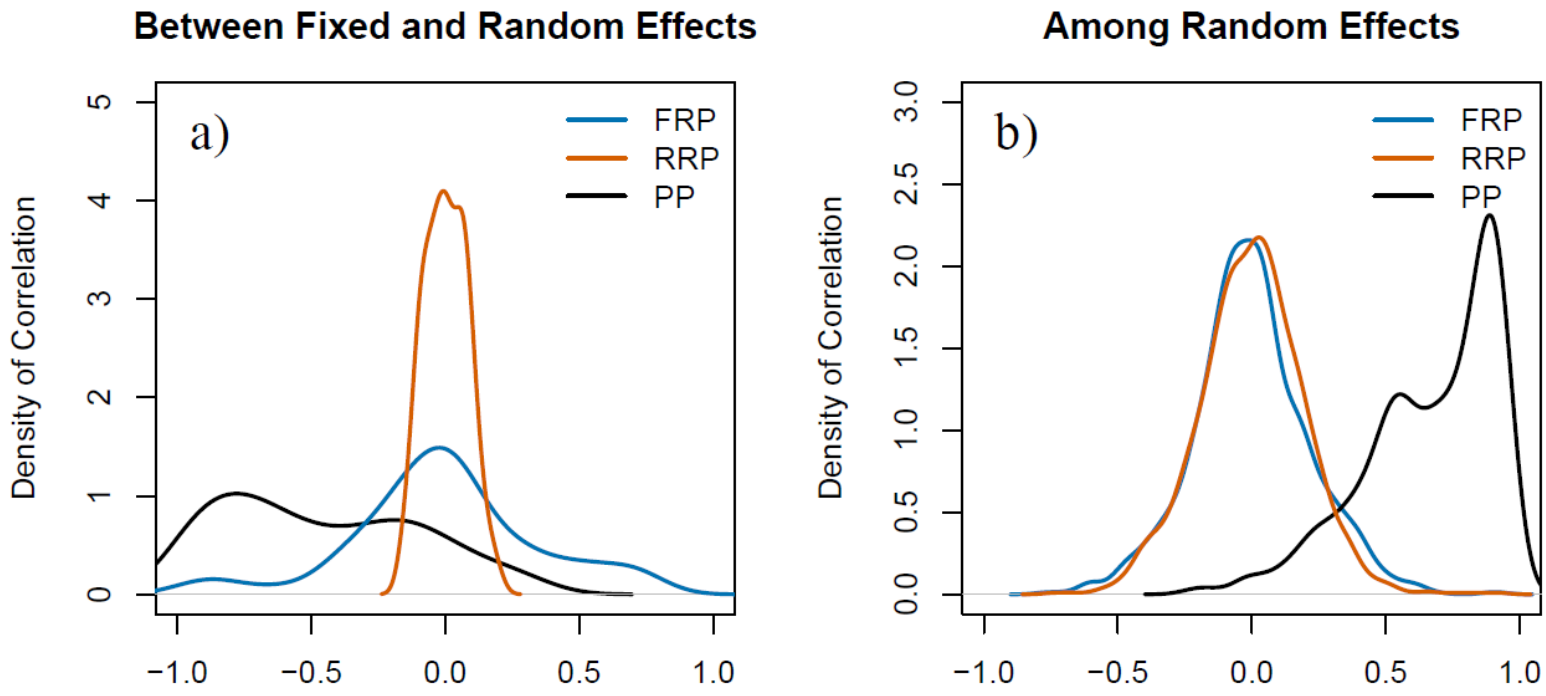
  RRP: replace $\mathbf{W}$ with $P^{\perp}UD_m^{1/2}\boldsymbol{\delta}$ to approximate restricted spatial model
- Reduced Model:

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

$$\boldsymbol{\delta} \mid \ldots \overset{approx}{\sim} N_m(\mathbf{0}, \sigma^2 I)$$

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

Reduced Correlations



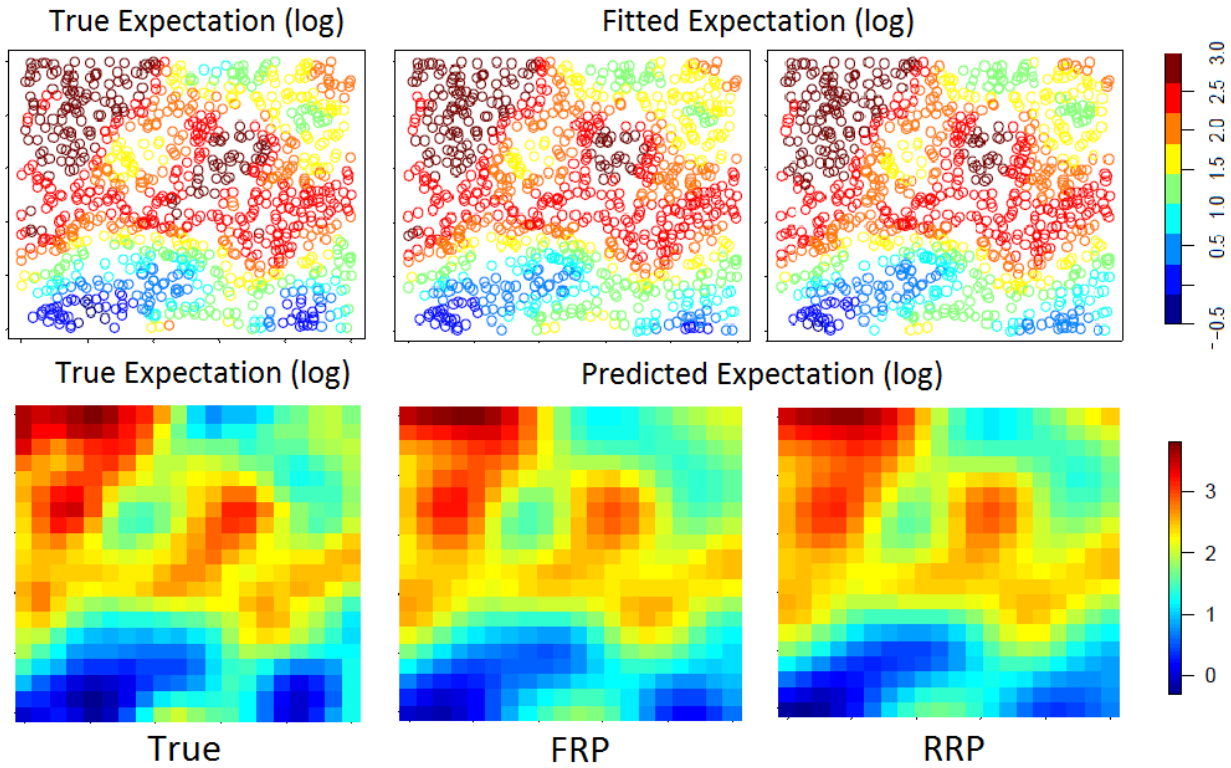
- Reparameterized random effects are approximately independent of each other *and* fixed effects

Computational Speed-up

- Drastic reduction in dimension of random effects, e.g. m = 50 for n = 1,000, or m = 60 for n = 3,000,...
- Reparameterized random effects are approximately independent of each other and fixed effects
- Easy to construct fast-mixing MCMC algorithm
- Eg. 10 to 50 to 300-fold reduction in compute time
- Scale beyond $n > 10,000$?
  - computational cost is of order nm^2
  - discretization of space/pre-computing
  - new decomposition algorithms/parallelization

Prediction Study: Poisson SGLMM

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



FRP: full model

RRP: restricted model (orthogonalized random effects)

Summary of Projected SGLMM

- reduces dimensions + better MCMC mixing
- adjusts for spatial confounding
- simple to implement, mostly “automated”

Summary of Projected SGLMM

- reduces dimensions + better MCMC mixing
- adjusts for spatial confounding
- simple to implement, mostly “automated”
- Our approach does *not* result in exchangeability between observed and predicted. (Predictive process does.)
- But we use optimal (minimal truncation error) projection
- And prediction is still straightforward
- Other approaches
  - may be better for the basic linear model
  - our approach works better for SGLMMs
  - our approach and predictive process approach: easy for more complex hierarchical settings

FAQ (Stuff/Wit/Insight/Matters of Great Importance*)

(* cf. Brown, Coeurjolly, Duchesne et al., 2017)

1. Why not just use INLA?

FAQ (Stuff/Wit/Insight/Matters of Great Importance*)

(* cf. Brown, Coeurjolly, Duchesne et al., 2017)

1. Why not just use INLA?
   - Great idea!

FAQ (Stuff/Wit/Insight/Matters of Great Importance*)

(* cf. Brown, Coeurjolly, Duchesne et al., 2017)

1. Why not just use INLA?
   - Great idea!
   - Our reparameterization can be combined with INLA
   - For multi-level hierarchies or if people are interested in a fully sample-based approach, our approach still applies

FAQ (Stuff/Wit/Insight/Matters of Great Importance*)

(* cf. Brown, Coeurjolly, Duchesne et al., 2017)

1. Why not just use INLA?
   - Great idea!
   - Our reparameterization can be combined with INLA
   - For multi-level hierarchies or if people are interested in a fully sample-based approach, our approach still applies
2. How is this different from a reduced-rank approach?

FAQ (Stuff/Wit/Insight/Matters of Great Importance*)

(* cf. Brown, Coeurjolly, Duchesne et al., 2017)

1. Why not just use INLA?
   - Great idea!
   - Our reparameterization can be combined with INLA
   - For multi-level hierarchies or if people are interested in a fully sample-based approach, our approach still applies
2. How is this different from a reduced-rank approach?
   - It is not different

FAQ (Stuff/Wit/Insight/Matters of Great Importance*)

(* cf. Brown, Coeurjolly, Duchesne et al., 2017)

1. Why not just use INLA?
   - Great idea!
   - Our reparameterization can be combined with INLA
   - For multi-level hierarchies or if people are interested in a fully sample-based approach, our approach still applies
2. How is this different from a reduced-rank approach?
   - It is not different
   - Similar to predictive process approach. But... more efficient, more accurate, and automated (knot choice)
3. How about log-Gaussian Cox processes?

FAQ (Stuff/Wit/Insight/Matters of Great Importance*)

(* cf. Brown, Coeurjolly, Duchesne et al., 2017)

1. Why not just use INLA?
   - Great idea!
   - Our reparameterization can be combined with INLA
   - For multi-level hierarchies or if people are interested in a fully sample-based approach, our approach still applies
2. How is this different from a reduced-rank approach?
   - It is not different
   - Similar to predictive process approach. But... more efficient, more accurate, and automated (knot choice)
3. How about log-Gaussian Cox processes?
   - Under consideration....

Acknowledgments

- Yawen Guan, SAMSI/NC State
- John Hughes, U of Colorado-Denver
- Conversations
  - Alan Gelfand (Duke U.)
  - Peter Hoff (Duke U.)
  - Bo Li (Purdue U.)
  - Doug Nychka (NCAR)
  - Dorit Hammerling (NCAR)
- Support from **NSF-CDSE/DMS-1418090**

Key References

- Guan and Haran (2017), A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models, *arxiv.org*
- Hughes and Haran (2013), Dimension reduction and alleviation... *Journal of the Royal Statistical Society (B)*
  - R package (CRAN) `ngspatial`
- Banerjee A, Tokdar, S., Dunson, D. (2013) Efficient Gaussian process regression for large datasets, *Biometrika*
- Reich et al. (2006), Effects of residual smoothing on the posterior of the fixed effects... *Biometrics*
- Haran (2011) Gaussian random field models for spatial data, *Handbook of MCMC*