Dimension reduction and alleviation of confounding for spatial generalized linear mixed models

John Hughes¹ and Murali Haran²

¹University of Minnesota
²The Pennsylvania State University

Introduction

- SGLMM is flexible but faces two serious challenges
  1. Spatial confounding: the spatial random effects are collinear with the fixed-effects predictors
  2. Curse of dimensionality: high-dimensional spatial random effects cause computational burden

Traditional SGLMM for areal data

- \( G = (V, E) \) is underlying graph
- \( A \) is adjacency matrix for \( G \)
- **Stage 1**
  \[
g(\mu_i) = X'\beta + S_i
\]
- **Stage 2**
  \[
p(S | \tau) \propto \tau^{\text{rank}(A)/2} \exp \left( -\frac{\tau}{2} S' Q S \right)
\]
- Zero-mean GMRF with precision matrix \( Q \)
- \( Q = \text{diag}(A) - A \) is the graph Laplacian
- \( S \) and \( S_i \) are conditionally independent given their neighbors iff \( Q_{ij} = 0 \) iff \( (i, j) \notin E \)
- **Stage 3** Choose priors for \( \tau, \beta \)

Spatial confounding

- Traditional SGLMM can be rewritten to show that spatial random effects are collinear with fixed-effects predictors
- Resulting variance inflation can make significant predictors appear to be insignificant
- Traditional model also permits patterns of negative dependence among the random effects, which we do not expect to see in the phenomena to which these models are typically applied

Computational burden

- Updating \( S \) presents a two-pronged challenge
  1. MCMC is slow per iteration due to high dimensionality
  2. Resulting Markov chain is slow mixing due to dependence

Solution

- Our solution is to reparameterize the model in such a way that
  1. Confounding is alleviated,
  2. The dimension of the random effects is greatly reduced, and
  3. Only positive spatial dependence is permitted
- **Stage 1**
  \[
g(\mu_i) = X'\beta + m_i \gamma
\]
- **Stage 2**
  \[
p(\gamma | \tau) \propto \tau^{q/2} \exp \left( -\frac{\tau}{2} M' Q M \gamma \right)
\]
- The columns of \( M_{ij}, q \) are the first \( q \) eigenvectors of the Moran operator for \( X \) wrt \( G \)

The Moran operator

- Moran operator for \( X \) with respect to \( G \) is \( P \cdot A P^{-1} \), where \( P \) is the orthogonal projection onto \( C(X)^{\perp} \)
- This operator appears in the numerator of a generalized form of Moran's I statistic, a well-established measure of spatial dependence for areal data
- Eigensystem of \( P \cdot A P^{-1} \) is useful
  - Eigenvectors comprise all possible mutually distinct patterns of clustering residual to \( X \) and accounting for \( G \)
  - Positive (negative) eigenvalues correspond to varying degrees of positive (negative) spatial dependence
  - I.e., the Moran basis is a multisresolutional spatial basis "tailored" to \( X \) and \( G \)

Some example basis vectors are shown below

Performance

- Smoothing orthogonal to \( X \) alleviates confounding
  Typical 95% HPD intervals for \( \beta = 1 \)

- Avoids unrealistic dependence (repulsion)
- \( \text{dim}(\gamma) = q \ll n \)
- \( q \) equal to 50 or 100 is sufficient for most datasets
- Markov chain mixes fast because random effects are approximately uncorrelated
- Also fast per iteration because a spherical normal proposal for \( \gamma \) works well

Performance for binary data on the 30 \times 30 lattice; similar results for other types of data

<table>
<thead>
<tr>
<th>Model</th>
<th>Dimension</th>
<th>% Reduction</th>
<th>Running time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse q = 400</td>
<td>403</td>
<td>55</td>
<td>4.8 hours</td>
</tr>
<tr>
<td>Sparse q = 200</td>
<td>203</td>
<td>78</td>
<td>2.9 hours</td>
</tr>
<tr>
<td>Sparse q = 100</td>
<td>103</td>
<td>89</td>
<td>1.3 hours</td>
</tr>
<tr>
<td>Sparse q = 50</td>
<td>53</td>
<td>94</td>
<td>0.8 hours</td>
</tr>
<tr>
<td>Sparse q = 25</td>
<td>26</td>
<td>97</td>
<td>0.6 hours</td>
</tr>
</tbody>
</table>

Application to US infant mortality data

- \( n = 3,071 \)
- Our model
  - Dimension of random effects: \( q = 25, 50, 100, 200 \)
  - \( q = 50 \) gave smallest DIC
  - Running time: 8.5 hours
- Competing model
  - Dimension of random effects: \( q \approx 3,100 \)
  - Running time: 14 days

References

- Forthcoming R package ngspatial