Toward Efficient Inference for High-dimensional Latent Variable Models

Murali Haran

Department of Statistics, Pennsylvania State University

MCMSKI
Lenzerheide, Switzerland. January 2016

Joint work with Yawen Guan

Talk Summary

- Latent variable models are very widely used.
- Markov chain Monte Carlo (MCMC) is a convenient approach for fitting such models.
- In practice: MCMC is often impractical when the latent variables become high-dimensional.
- I will discuss an approach for addressing these computational challenges for a class of spatial/nonparametric regression models: generalized linear mixed models with Gaussian process priors.
- The approach is based on latent variable reparameterization and dimension reduction.

Much of this is work in progress.

Latent Variable Models Review

- In sciences, latent variables are often physically meaningful.
  - E.g. unobserved immigration/carriers of disease in a disease dynamics model
- In social sciences may be a theoretical construct.
  - E.g. latent behavioral states in a psychology experiment
- Can add flexibility, help a model fit data better.
  - E.g. random intercepts or random slopes model in regression. Capture heterogeneity.
  - E.g. model dependence in non-Gaussian data

Spatial Count Data

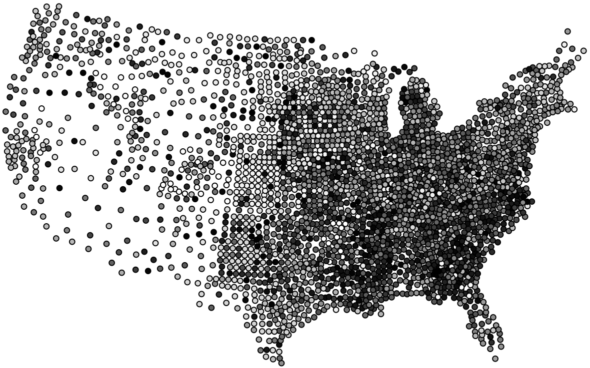
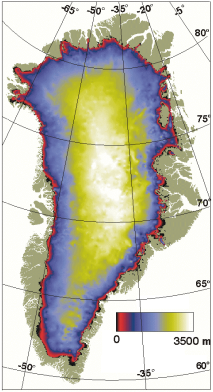


Figure: U.S. infant mortality data by county. $n = 3071$
Ratio of deaths to births, each averaged over 2002-2004.
Darker indicates higher rate.

Greenland Ice Sheet Thickness



(Bamber et al., 2001). Over 60,000 locations

Spatial Generalized Linear Mixed Models

Example model for $Z(\mathbf{s})$ for $\mathbf{s} \in D \subset \mathbb{R}^d$,

1. $Z(\mathbf{s}_i) \mid \boldsymbol{\beta}, W(\mathbf{s}_i) \sim$Poisson$(\mu(\mathbf{s}_i))$, conditionally independent for $i = 1, \ldots, n$.
2. $\log(\mu(\mathbf{s}_i)) = X(\mathbf{s}_i)\boldsymbol{\beta} + W(\mathbf{s}_i)$
3. Impose dependence: $\mathbf{W} = (W(\mathbf{s}_1), \ldots, W(\mathbf{s}_n))^T$ via
   (a) Gaussian Markov random field on a lattice,
   $$p(\mathbf{W}|\theta) \propto \theta^{(n-1)/2} \exp\left(-\frac{\theta}{2}\mathbf{W}' Q \mathbf{W}\right), \theta > 0,$$
   (b) Gaussian process for continuous-domain spatial data,
   $$p(\mathbf{W}|\theta) \sim N(0, \Sigma(\theta)).$$
4. Priors for $\theta, \boldsymbol{\beta}$

Inference based on posterior, $\pi(\theta, \boldsymbol{\beta}, \mathbf{W} \mid \mathbf{Z})$
Key references: Besag et al. (1991), Diggle et al. (1998).
Also useful for non-Gaussian nonparametric regression.

Computational Challenges with SGLMM inference

- High-dimensionality of latent variables (**W**): n. Posterior distribution is of dimension $p + k + n$ for p covariates, k covariance parameters, n data points.
- Strong cross-correlations make it hard to design efficient updating schemes. Too many low-dimensional updates may be slow, and result in poor mixing. High-dimensional updates may be computationally inefficient.
- Result (often): computationally expensive and slow mixing Markov chains.

Outline of Strategy

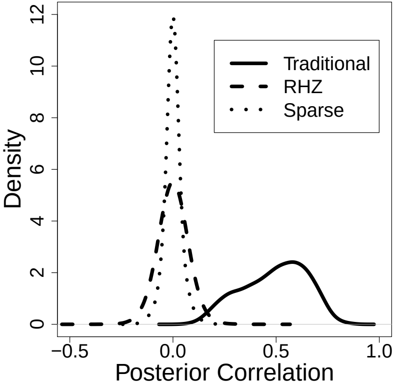
Observations:

1. Reparameterization of latent variables can help "de-correlate" them, improve mixing of MCMC algorithm.
   - Not a new idea. E.g. Christensen and Roberts (2006).
2. "Spatial confounding": dependence between latent variables and fixed effects (covariates) causes poor mixing.
3. Latent variables are merely a device for inducing dependence. May not need n variables for this.

- #2 and #3 identified in Hughes and Haran (2013) in the context of Gaussian Markov random field models.

Goal: find reparameterization to achieve above for Gaussian process (continuous-domain) model.

Distribution of Correlations



Posterior distribution of correlations (example from Hughes and Haran, 2013).

Reparameterize Random Effects

- SGLMMs:

$$g\left\{E(Z_i|\boldsymbol{\beta}, W_i)\right\} = \boldsymbol{X}_i\boldsymbol{\beta} + W_i$$
$$\boldsymbol{W}|\boldsymbol{\theta} \sim N_n(0, \Sigma(\boldsymbol{\theta}))$$

Inference based on $\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{W}|\boldsymbol{Z})$

1. Let $P^{\perp} = I - P$ denote projection on orthogonal space of $\boldsymbol{X}$, where $P = X(X'X)^{-1}X'$,
2. Approximate first m eigenvectors $\boldsymbol{H}_\theta$ of $P^{\perp}\Sigma(\theta)P^{\perp}$.
3. Replace $\boldsymbol{W}$ with $\boldsymbol{H}_{\boldsymbol{\theta}}\boldsymbol{\delta}$, where $\boldsymbol{\delta} \overset{approx}{\sim} N_m(\mathbf{0}, D_\theta)$.
   D_θ is m-dim diagonal matrix. D_{ii}=i^{th} eigenvalue.
   Reduced Model:

$$g\left\{E(Z_i|\boldsymbol{\beta}, W_i, \boldsymbol{\theta})\right\} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}_i\boldsymbol{\delta}$$
$$\boldsymbol{\delta}|\boldsymbol{\theta} \sim N_m(\mathbf{0}, D_\theta)$$

4. Inference based on $\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\delta}|\boldsymbol{Z})$

Spatial Confounding and Fast Mixing

- Step 1 is not necessary before dimension reduction: can apply algorithm to $\Sigma(\theta)$ instead of $P^{\perp}\Sigma(\theta)P^{\perp}$.
- However, 1 results in better mixing MCMC algorithm.
- (If spatial confounding issue is of interest, Step 1 addresses this as well.)

Approximate ***H*** using Random Projection

- Step #2 (computing eigenvectors) of high-dimensional $\Sigma(\Theta)$ is expensive.
- Random projections (Banerjee, Dunson, Tokdar, 2012) provides fast approximation of the leading m eigenvectors.

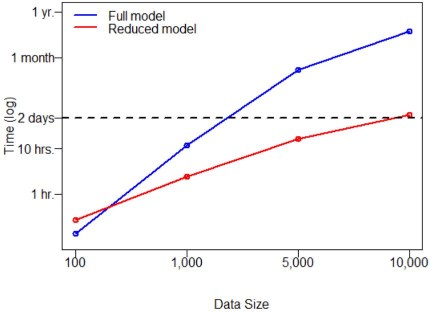
---

1. Low dimensional projection from $R^{n\times n}$ to $R^{n\times m}$:
   1.1 Simulate $\Omega_{ij} \sim N(0, \frac{1}{\sqrt{m}})$, $\Omega \in R^{n\times m}$
   1.2 Form $\Sigma\Omega$
2. Use SVD to find basis Φ^T (left vectors of $\Sigma\Omega$)
3. Nyström method to approximate eigen-decomposition:
   Approximate $\Sigma \approx HD^2H^T$

Computational Benefits

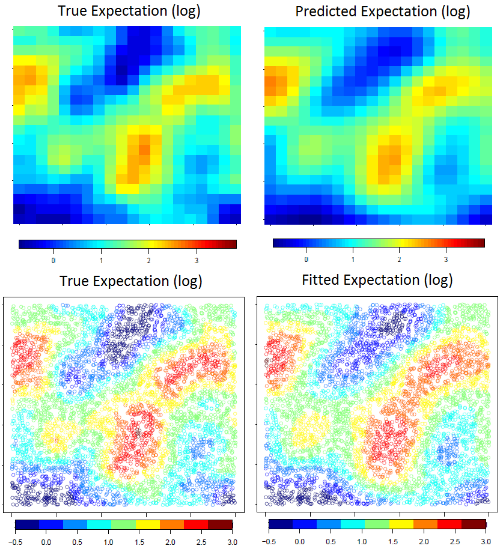
Win on both counts:

- Reduced dimensions: δ is m vs n, e.g. m=50, n=5000. Computational complexity: $O(m^2 n)$ vs $O(n^3)$.
- δ are approximately de-correlated. Improves MCMC mixing, simplifies algorithm construction. Block updates: (i) $\boldsymbol{\delta} \mid \boldsymbol{\beta}, \theta$, (ii) $\boldsymbol{\beta} \mid \boldsymbol{\delta}, \theta$, (iii) $\theta \mid \boldsymbol{\delta}, \boldsymbol{\beta}$.



Prediction Performance

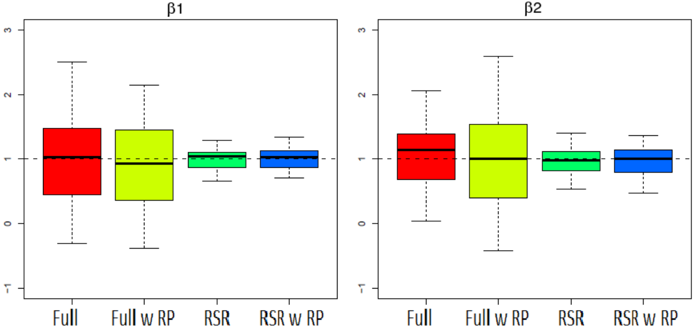
- Simulate n = 5000 spatial count data
- Prediction on 20 x 20 grid with *m*=50.



Simulation Study

(Preliminary)

- Simulate spatial count data with:
  $\boldsymbol{\beta} = (1, 1)^T$, and Matérn family$(\nu, \phi, \sigma^2) = (0.5, 0.2, 1)$
- Boxplots illustrating inference for β



Summary

- Two pronged approach: (i) reducing dimensions of posterior distribution; (ii) de-correlation of latent variables.
- Speeds up computational cost per iteration, simplifies construction of algorithm, and improves Markov chain mixing.
- Caveats:
  1. Matrix multiplication is still expensive.
  2. Approach seems impractical as n gets large, e.g. greater than around 20,000.
  3. May oversmooth when true surface is rough.

Acknowledgments and References

Support:

- NSF GEO-1240507 The Network for Sustainable Climate Risk Management (SCRiM)
- NSF-CDSE/DMS-1418090 Statistical Methods for Ice Sheet Projections

Key References:

- 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.
- Banerjee, A., Dunson, D.B., Tokdar, S.T. (2012), Efficient Gaussian process regression for large datasets, Biometrika. Projections

Reducing Dimensions/Reparameterization

- Basic idea: reparameterize the model and reduce the dimension of the random effects ($\mathbf{W}$), while preserving the desirable properties of the original model.
- Particularly worth considering when random effects are not intrinsically important, i.e., they are "nuisance parameters".
- Typical in spatial generalized linear mixed models: random effects are used to pick up residual spatial dependence, adjust for unmeasured spatially-varying covariates.

Reparameterization for Lattice-domain Data

Recall model:

- $g(\mu(\mathbf{s}_i)) = X(\mathbf{s}_i)\beta + W(\mathbf{s}_i)$
- $p(\mathbf{W}|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\mathbf{W}' Q \mathbf{W}\right)$

Let:

- $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, orthogonal projection onto $C(\mathbf{X})$
- $\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P}$
- Let $\mathbf{M} = \mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}$, where $\mathbf{A}$ is the adjacency matrix

Reparameterize as follows:

- $g(\mu(\mathbf{s}_i)) = X(\mathbf{s}_i)\beta + \mathbf{M}_i\delta$, where $\mathbf{M}_i$ is the ith row of $\mathbf{M}$
- $p(\delta \mid \tau) \propto \tau^{q/2} \exp\left(-\frac{\tau}{2}\delta' \mathbf{Q}^{**} \delta\right)$, where $\mathbf{Q}^{**} = \mathbf{M}'\mathbf{Q}\mathbf{M}$.
- If we only keep the first q columns of the matrix $\mathbf{M}$, that is, reduce dimensions of $\mathbf{M}_i$ to q for each i, the # random effects is reduced from n to q ($q \ll n$)

Hughes and Haran (2013)

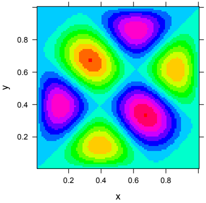
Comments

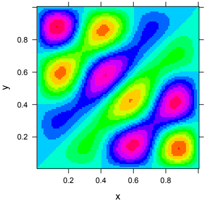
- Intuition: projected spatial random effects orthogonal to the predictors and in the direction specified by the graph.
- Inference is now based on $\pi(\Theta, \boldsymbol{\beta}, \boldsymbol{\delta} \mid \mathbf{Z})$
  $q + p + 1$-dimensional
- Dimension reduction works because of an ordering: highest to lowest (including negative) spatial dependence (Boots and Tiefelsdorf, 2000)

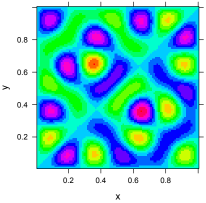
Interpreting the Resulting Reparameterization

- "Tailored" to **X** and G: eigenvectors comprise all possible patterns of clustering residual to **X** and accounting for G

Some selected basis vectors for the 30×30 lattice.







Reducing Dimensions for Continuous-Domain Processes

- Unlike in the lattice case, there is no graph/adjacency matrix to work with.
- Alternative: use an idea from Banerjee, Dunson and Tokdar (2012): "random projections" of data into a lower-dimensional subspace
- Apply a fast algorithm to obtain reduced-dimensional random effects, replacing $\mathbf{W}$ (n-dimensional) with V (m-dimensional) with $m \ll n$.
- Same idea: we project latent variables to obtain a reduced-dimensional posterior distribution. Easier to construct efficient MCMC algorithms.

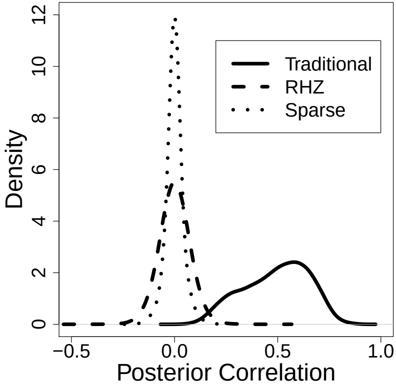
Preliminary Results

- Prediction: reduced-dimensional approach gives similar results as regular methods
- Inference: better or worse, depending on the assumed true model. If interpreting parameters is not important, this is a non-issue. But if it is, need to think harder about spatial confounding-related issues. (Hanks et al., 2015)

(JSM 2015 poster by Yawen Guan)

Pros

- Random effects are much smaller in number.
- They are approximately "de-correlated". That is (by construction) they no longer exhibit as much dependence. Easy to construct fast mixing MCMC



Cons

- Highly specialized approach
- There may be scaling issues: as dimensions and complexity of the model increases, may still need a significant fraction of the latent variables.

Can improve inference while in other cases can induce problems

Computational Strategies

1. Composite likelihood-based approaches
2. Approximate integration approaches
3. Simulation-based approaches: study how the forward (probability) model generates data for different parameter settings. Then compare the simulations to the real observations.
   - Approximate Bayesian Computing (ABC)
   - Gaussian process approximations ("emulation-calibration"). (Jandarov, Haran, Bjornstad, Grenfell, 2014)
4. Reduced-dimensional approximations/reparameterizations
5. Some combination of the above

Composite Likelihood

Has potential to address inferential and scaling issues

- Inference with latent variables $u_1, \ldots, u_k$, joint posterior distribution, $\pi(\theta, u_1, \ldots, u_k \mid \mathbf{Y})$

$$\propto f(\mathbf{Y} \mid u_1, \ldots, u_k) f(u_1, \ldots, u_k \mid \theta) p(\theta).$$

- Basic idea: replace above with $\prod_{b=1}^{B} f(\mathbf{Y}_b^C \mid u_b^C) f(u_b^C \mid \theta) p(\theta)$, where $\mathbf{Y}_b^C$ and u_b^C, for $b = 1, \ldots, B$, are each subsets (blocks) of the vectors $\mathbf{Y}$ and $u_1, \ldots, u_k$ respectively
- Evaluating this approximation can be much more computationally efficient than evaluating the joint distribution
- Separating the latent variables into blocks suggest convenient block-MCMC schemes. Many choices for composite likelihood (e.g. Caragea and Smith, 2003)

(JSM 2015 poster by Saksham Chandra)

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 $\times$ 30 lattice.