Toward Efficient Inference for High-dimensional Latent Variable Models

Murali Haran

Department of Statistics, Pennsylvania State University

Joint Statistical Meetings
Seattle, Washington. August 2015

Collaborators:
John Hughes (University of Minnesota Biostatistics)
Saksham Chandra and Yawen Guan (Penn State Statistics)

Talk Summary

- Latent variable models are very widely used
  - Physical sciences (physically meaningful)
  - Social sciences (of theoretical interest)
  - As a convenient device to provide a flexible model with desirable properties.
- Markov chain Monte Carlo (MCMC) is a convenient approach for fitting such models. *In principle.*
- In practice: MCMC is often impractical when the latent variables become high-dimensional
- I will discuss potential approaches for addressing these computational challenges for some classes of models

Much of this is work in progress

Why are Latent Variable Models Useful?

- Latent=hidden, unobservable
- In scientific problems, often of interest to learn about unobservable processes. Infer these processes (latent variables) via a model connecting them to the observables.
- In social science/other disciplines, specify hidden latent structures, subpopulations in the model
  - E.g. disease dynamics involve people movement (unobservable); only have data on numbers of infected at each location (observable)
- 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 via a generalized linear mixed model with dependent random effects

Spatial Count Data

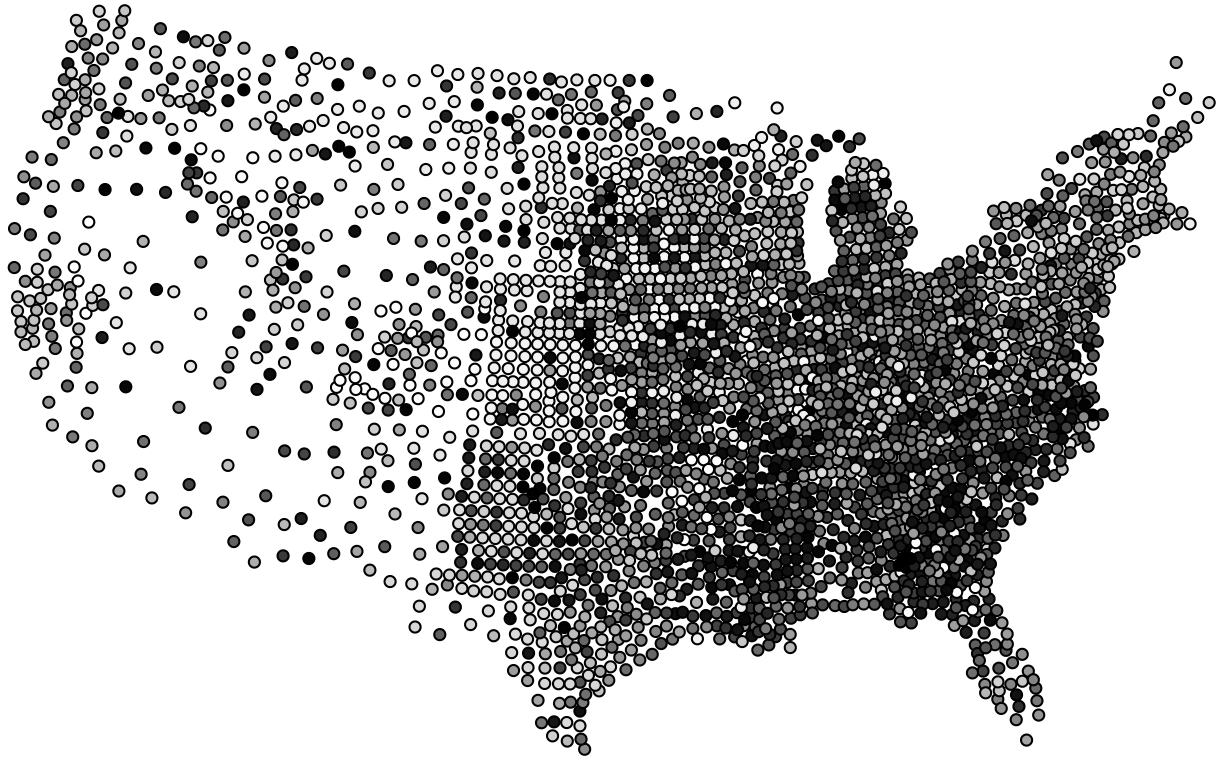
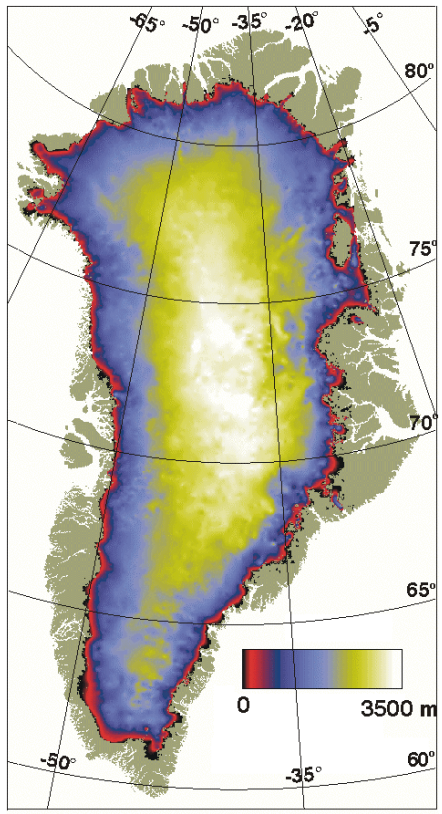


Figure: U.S. infant mortality data by county. $n = 3071$
Ratio of deaths to births, each averaged over 2002-2004.
Darker indicates higher rate.

Greenland Ice Sheet Thickness



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

Inference for a Latent Variable Model

- Generic model:
  - Data given latent variables: $f(Y_1, \ldots, Y_n \mid u_1, \ldots, u_k)$
  - Latent variable model $f(u_1, \ldots, u_k \mid \theta)$
  - Prior (if Bayesian approach), $p(\theta)$
- ML: maximize likelihood w.r.t. θ,

$$\mathcal{L}(\theta) = \int f(\mathbf{Y} \mid u_1, \ldots, u_k) f(u_1, \ldots, u_k \mid ; \theta) du_1 \ldots du_k$$

- Bayesian approach: inference based on joint posterior

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

- Want: marginal posterior, $\pi(\theta \mid \mathbf{Y})$
- In both cases: computation may be challenging if $u_1, \ldots, u_k$ is large in number and it is not easy to integrate them out analytically.
- Computing is getting faster but not fast enough to keep up with the increasing complexity/size of our models/data. (Few exceptions: some parallel methods.)

Spatial Generalized Linear Mixed Models

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. $\log(\mu(\mathbf{s}_i)) = X(\mathbf{s}_i)\beta + W(\mathbf{s}_i)$
3. Impose dependence: $\mathbf{W} = (W(\mathbf{s}_1), \ldots, W(\mathbf{s}_n))^T$

$$p(\mathbf{W}|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\mathbf{W}'Q\mathbf{W}\right), \tau > 0.$$

   Gaussian Markov random field model on a lattice.
   Q is completely specified by the $n \times n$ adjacency matrix $\mathbf{A}$ for the lattice where $\mathbf{A}_{ij}$ is 1 if i, j are neighbors, 0 else

4. Priors for Θ, β

Inference based on posterior, $\pi(\Theta, \beta, \mathbf{W} \mid \mathbf{Z})$
(Besag et al. (1991), Diggle et al. (1998))

Computational Challenges with SGLMM inference

- High-dimensional latent variables. If there are p covariates, k covariance parameters, and n data points, the posterior is $p + k + n$ dimensional.
- 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): slow mixing Markov chains with computationally expensive updates

Computational Strategies

1. Reduced-dimensional approximations/reparameterizations
2. Composite likelihood-based approaches
3. Approximate integration approaches
4. 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)
5. Some combination of the above

Focus here: (1)

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

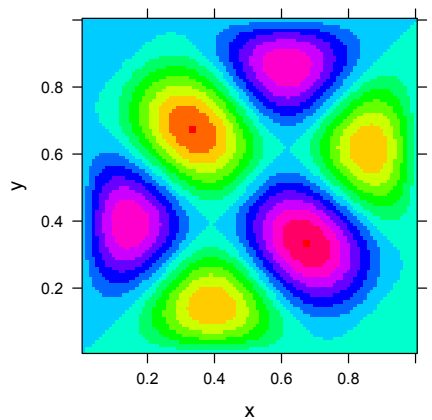
Comments

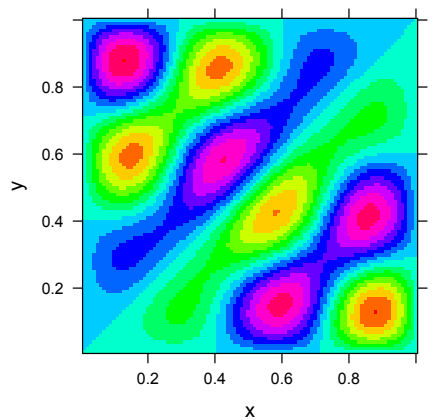
- Intuition: projected spatial random effects orthogonal to the predictors and in the direction specified by the graph.
- Inference is now based on $\pi(\Theta, \beta, \delta \mid \mathbf{Z})$
  $q + p + 1$-dimensional
- Dimension reduction works because of an ordering: highest to lowest (including negative)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

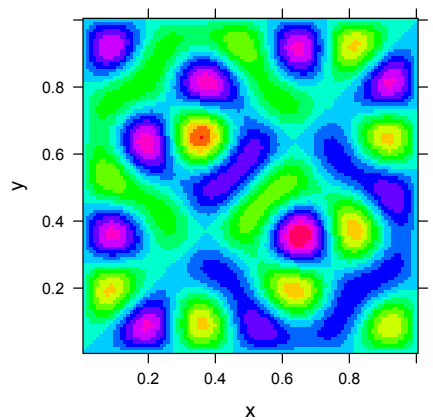
Interpreting the Resulting Reparameterization

- "Tailored" to $\mathbf{X}$ 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.







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.

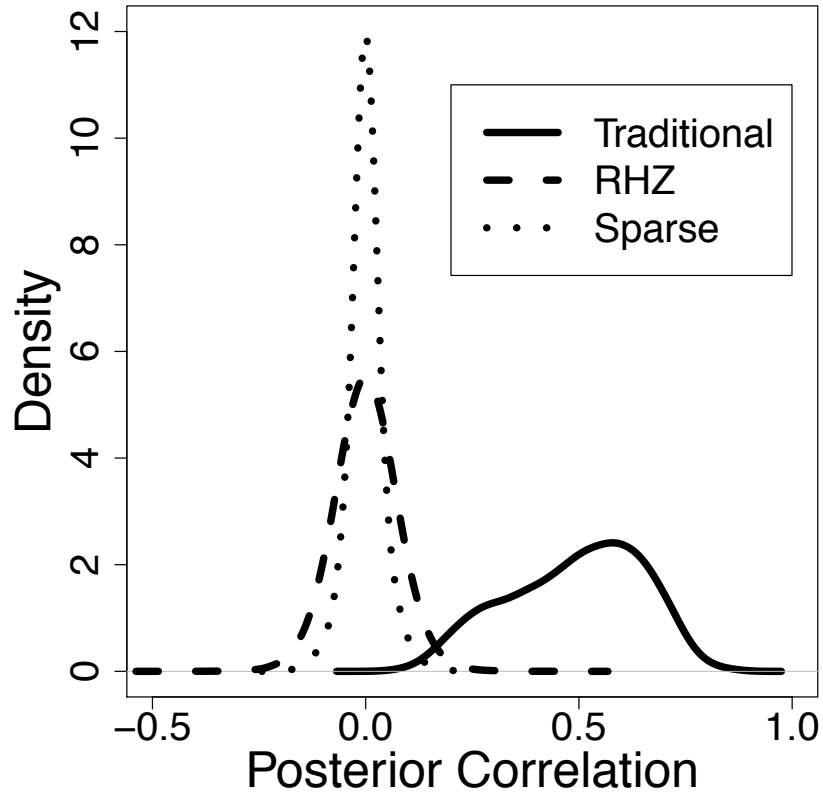
Preliminary Results

- 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

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)

Summary

- Constructing efficient MCMC algorithms for high-dimensional latent variable models is challenging
- In special cases, dimension-reduction techniques and likelihood approximations may be efficient
- Finding general methods for a large class of models is non-trivial
- Other methods are worth considering: approximate integration, simulation-based. Ideally: combine multiple methods

Acknowledgments

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

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.