Computational Methods for Some High-Dimensional Latent Variable Models

Murali Haran

Department of Statistics, Pennsylvania State University

Operations Research Colloquium, STAT 590
Penn State University, February 2015.

What This Talk is About

- Latent variable models are very useful and popular. I will provide examples.
- Computational challenges are often daunting
  - High-dimensional latent variables resulting in high-dimensional posterior distributions (or complex integrated likelihoods).
  - Constructing efficient MCMC algorithms can be a challenge.
- I will outline some approaches, skipping details:
  1. Forward simulation-based approximations to the likelihood. E.g. Approximate Bayesian computing (ABC) or Gaussian process emulation
  2. Dimension reduction of the latent variables.
  3. Fast likelihood approximations, e.g. composite likelihood.
- I will be mostly concerned with providing a broad overview of methods. Time permitting, I may discuss a few details with examples.

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, learn about hidden latent structures, subpopulations
  - E.g. learn about spread of infections from location to location (unobservable) from the 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

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), $p(\theta)$
    (Bold notation implies vectors.)
- Maximum likelihood: maximize likelihood w.r.t. θ

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

- Bayesian approach: inference based on posterior

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

- 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 of our models and the size of our data sets!**

Computational Strategy 1: Forward Simulation

Basic idea: avoid working with the likelihood, which may be very expensive to evaluate. Instead:

- Simulate $\mathbf{Y}^*$ from the "forward model" $f(Y_1, \ldots, Y_n \mid u_1, \ldots, u_k) f(u_1, \ldots, u_k \mid \theta)$ for various θ values. Obtain: $\mathbf{Y}^*(\theta_1), \ldots, \mathbf{Y}^*(\theta_k)$
- Compare the simulations to the observed data. Intuition: θ values that produced simulations that are similar to observations are "more likely" (higher posterior probability).
- In effect: replace likelihood function with an approximation based on forward simulations.

$$\pi(\theta \mid Y_1, \ldots, Y_n) \propto \hat{\mathcal{L}}(\theta; Y_1, \ldots, Y_n) p(\theta)$$

Multiple ways to do this in a systematic fashion:

- Approximate Bayesian Computation (ABC). (Beaumont et al. 2002; Marjoram et al., 2002)
- Gaussian process-based calibration (Kennedy and O'Hagan, 2001)

Strategy 1: Comments

- ABC: most useful when forward simulation is fast.
- Gaussian process approach: interpolates the behavior of the model based on relatively few forward simulations.
  - Useful when forward simulation is not fast.
  - There is smoothness in the process (nearby θ values result in similar simulated values).
  - The process is not too highly multivariate/complicated (hard to interpolate).
- Note that in both cases dimensionality of latent variable does not play a role in computational complexity except for simulation expense.

Computational Strategy 2: Dimension Reduction

- Basic idea: may be redundancy in latent variables so reduce their dimensions without information loss.
- Summary: replace $\pi(\theta, \text{latent vars} \mid Y_1, \ldots, Y_n)$

$$\propto f(Y_1, \ldots, Y_n \mid u_1, \ldots, u_k) f(u_1, \ldots, u_k \mid \theta) p(\theta)$$

with:

$$\propto f(Y_1, \ldots, Y_n \mid v_1, \ldots, v_r) f(v_1, \ldots, v_r \mid \theta) p(\theta),$$

where $v_1, \ldots, v_r$ are reduced-dimension latent variables and $r << k$.

- Can provide dramatic computational advantages.
- Reduced-dimensional approach may even have inferential/interpretability advantages over original model.
- This strategy is not very general. Most applicable when latent variables are modeling dependence. Can now consider ideas from literature on dimension reduction/sparsity.

Computational Strategy 3: Composite Likelihood

- Basic idea: approximate the likelihood function as the product of component log likelihoods.
- Each component likelihood is a likelihood function for a subset of data.
- This approximation, called a *composite likelihood function* (Lindsay, 1988) may be much faster to compute than the original log-likelihood.
- Replace $\pi(\theta, u_1, \ldots, u_k \mid Y_1, \ldots, Y_n)$

$$\propto f(Y_1, \ldots, Y_n \mid u_1, \ldots, u_k) f(u_1, \ldots, u_k \mid \theta) p(\theta)$$

with:

$$\propto \prod_{b=1}^{B} f(Z_b \mid U_b) f(U_b \mid \theta) p(\theta)$$

where Z_b is a subset (or "block") of the $Y_1, \ldots Y_n$, and U_b is a corresponding subset of $u_1, \ldots, u_k$.

Composite Likelihood with Latent Variables

- If each function $f(Z_b \mid U_b)f(U_b \mid \theta)$ can be evaluated quickly, for instance by avoiding large matrix operations for dependent data, composite likelihood offer dramatic speed-ups over likelihood function evaluations.
- These component (block) evaluations may be easily parallelized. Useful for scalability of inference.
- When dealing with latent variables, there are opportunities for analytically or numerically integrating out the latent variables within each piece. That is, find (exactly or approximately):

$$f(Z_b \mid \theta) = \int f(Z_b \mid U_b)f(U_b \mid \theta)dU_b.$$

This integrates out the latent variables, reducing dimensions for maximum likelihood and Bayesian inference.

Strategy 1 Details
Approximate Bayesian computation (ABC)

(Tavare et al., 1997; Beaumont, Zhang, Balding, 2002)

Rejection sampler:

- Simulate θ^* from prior on θ.
- Accept θ^* with probability $h(\theta^*; \mathbf{Y}) = f(\mathbf{Y} \mid \theta^*)$.
- Repeat above: accepted θ^*s have distribution $\pi(\theta^* \mid \mathbf{Y})$.

Since $h(\theta^*; \mathbf{Y})$ is intractable or too expensive (need to integrate out the latent variables), this approach is not practical.

ABC rejection sampler:

- Generate $\theta \sim p(\cdot)$.
- Simulate $\mathbf{Y}^* \sim f(\cdot \mid \theta^*)$.
- Accept θ^* if $\mathbf{Y}^* = \mathbf{Y}$.
- Repeat above: accepted θ have distribution $\pi(\theta \mid \mathbf{Y})$.

Usually not practical since probability that $\mathbf{Y}^* = \mathbf{Y}$ is generally very small (for a continuous state space, it is 0).

ABC rejection sampling [cont'd]

- Approximate rejection sampler:
  - Generate $\theta \sim p(\cdot)$.
  - Simulate $\mathbf{Y}^* \sim f(\cdot \mid \theta^*)$.
  - Accept θ^* if $\rho(\mathbf{Y}^*, \mathbf{Y}) < \epsilon$, where $\rho(\mathbf{Y}^*, \mathbf{Y}), \epsilon > 0$ are a distance and threshold defined by the user.
- As $\epsilon \to \infty$, this algorithm generates observations from the prior. As $\epsilon \to 0$, this algorithm generates from $\pi(\theta \mid \mathbf{Y})$.
- Often, the distance is defined on same summary statistics on $\mathbf{Y}$, say $S(\mathbf{Y})$, rather than on $\mathbf{Y}$ itself. That is, $\rho(\mathbf{Y}^*, \mathbf{Y}) = \rho(S(\mathbf{Y}^*), S(\mathbf{Y}))$. This is particularly useful when $\mathbf{Y}$ is high dimensional.

Likelihood-free MCMC

The previous algorithm is not very general. $p(\theta)$ will generally work poorly as a proposal for $\pi(\theta \mid \mathbf{Y})$ and it may also be very difficult to find another reasonable proposal, especially if θ has more than a few dimensions.

- Recall that the Metropolis-Hastings algorithm constructs a Markov chain with stationary distribution $\pi(\theta \mid \mathbf{Y})$ by generating the next state of the Markov chain as follows:
  - If current state is θ, propose a move to θ^* according to a transition kernel $q(\cdot \mid \theta)$.
  - Calculate acceptance probability, $\alpha(\theta, \theta^*) = \min\left(1, \frac{h(\theta^*;\mathbf{Y})}{h(\theta;\mathbf{Y})} \frac{p(\theta^*)}{p(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$.
  - Accept θ^* as the next state with probability $\alpha(\theta, \theta^*)$.

Likelihood-free MCMC [cont'd]

- Again, $h(\theta^*; \mathbf{Y})$ is either intractable or too expensive.
- Likelihood-free MCMC:
  - If current state is θ, propose a move to θ^* according to a transition kernel $q(\cdot \mid \theta)$.
  - Generate $\mathbf{Y}^* \sim f(\cdot \mid \theta^*$.
  - If ($\mathbf{Y}^* \neq \mathbf{Y}$), reject θ^* (stay at θ). If ($\mathbf{Y}^* = \mathbf{Y}$) calculate acceptance probability, $\alpha(\theta, \theta^*) = \min\left(1, \frac{p(\theta^*)}{p(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$. Accept θ^* as the next state with probability $\alpha(\theta, \theta^*)$.
- Avoided evaluating h but this Markov chain has stationary distribution $\pi(\theta \mid \mathbf{Y})$. Proof is a simple reversibility argument (see Marjoram, Molitor, Plagnol, Tavare, 2003).

Likelihood-free MCMC [cont'd]

- As before, $\mathbf{Y}^* = \mathbf{Y}$ is very unlikely (or has zero probability) in most cases.
- Approximate likelihood-free MCMC:
  - If current state is θ, propose a move to θ^* according to a transition kernel $q(\cdot \mid \theta)$.
  - Generate $\mathbf{Y}^* \sim f(\cdot \mid \theta^*)$.
  - If $(\rho(S(\mathbf{Y}^*), S(\mathbf{Y})) > \epsilon)$, reject θ^* (stay at θ). If $(\rho(S(\mathbf{Y}^*), S(\mathbf{Y})) < \epsilon$ calculate acceptance probability, $\alpha(\theta, \theta^*) = \min\left(1, \frac{p(\theta^*)}{p(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$.
    Accept θ^* as the next state with probability $\alpha(\theta, \theta^*)$.
- Avoided evaluating h. The idea: if ϵ is small this Markov chain has *approximately* the right stationary distribution $\pi(\theta \mid \mathbf{Y})$. No really sound theoretical basis for convergence of estimates based on this algorithm.

Example of Strategy 1: Gravity Time Series SIR Model

SIR = Susceptible-Infected-Recovered

- Models the number of incidences of measles in K different communities (cities).
- The model has components of a discrete time-series TSIR model for local dynamics (Bjørnstad et al., 2002; Grenfell et al. 2002).
- Similar to gravity models from transportation theory, it has an explicit formulation for the spatial transmission between different host communities.
- It allows for stochasticity inherent in the disease transmission and random immigration.
- It includes seasonality in the transmission rates.

(*Jandarov, Haran, Bjornstad, Grenfell, 2013*)

Gravity TSIR Model: Notation

- I_{kt} : number of infected individuals in city k at time t
- S_{kt} : number of susceptible individuals in city k at time t
- L_{kt} : number of infected people moved to city k at time t
- d_{kj} : distance between cities k and j
- N_{kt}, B_{kt} : size and birth rate of city k at time t

Gravity TSIR Model

- Number of incidences of a disease at time $t+1$ for city k,

  $I_{k(t+1)} \sim \text{Poisson}(\lambda_{k(t+1)})$, where $\lambda_{k(t+1)} = \beta_t S_{kt}(I_{kt} + L_{kt})^{\alpha}$.

- $I_{k(t+1)}$ increases with I_{kt}, S_{kt}, and number of infected immigrants coming to city k at time t (L_{kt}).
- $\{\beta_t\}$ are 26 different parameters that are repeated every year to allow differences in seasonal transmission (26 = number of biweeks in a year).

(Xia, Bjørnstad and Grenfell, 2004)

Gravity TSIR Model

- Number of susceptible individuals at time $t+1$ for city k, $S_{k(t+1)} = S_{kt} + B_{kt} - I_{k(t+1)}$.
- Number of infected immigrants (latent) at time t for city k

$$L_{kt} \sim \text{Gamma}(m_{kt}, 1), \text{ where } m_{kt} = \theta N_{kt}^{\tau_1} \sum_{j=1, j\neq k}^{K} \frac{(Ijt)^{\tau_2}}{d_{kj}^{\rho}}.$$

- L_{kt} increases with size of city k, number of infected people in all other cities, taking into account distances.

Inference for Measles Dynamics

- Parameters of the model:
  - Reliable estimates of local transition parameters α and β are known (Bjørnstad et al. 2001).
  - Gravity parameters θ, τ_1, τ_2 and ρ are unknown.
- Sources of information:
  - The UK Registrar General's data for 952 cities in England and Wales for years 1944-1966 of biweekly incidences of measles.
  - Number of susceptibles from standard reconstruction algorithms (cf. Fine and Clarkson 1982a, Finkenstadt and Grenfell 2000).
- **Goal**: Infer gravity parameters $\boldsymbol{\Theta} = (\theta, \tau_1, \tau_2, \rho)$ from data.

Computational Challenges

- Dimensions of the data (TK): 546*952 = 519,792.
- Number of infected immigrants $\{L_{k,t}\}$ are unobserved.
- The likelihood function is complicated:
  - Involves integrating over 519,792 latent variables.
  - Expensive calculations per iteration.

An Emulation Based Solution

- Let vector of summary statistics from observations be $\mathbf{Z}$.
- Simulate realizations of the gravity TSIR model at various parameter settings $\Theta_1, \Theta_2, \ldots, \Theta_p$.
- Let $\mathbf{Y}(\Theta)$ be the vector of summary statistics obtained at parameter setting Θ.
- Consider: $(\Theta_1, \mathbf{Y}(\Theta_1)), \ldots, (\Theta_p, \mathbf{Y}(\Theta_p))$.
- Stochastic emulation: Fit a Gaussian Process (GP) to above simulations.
  - Thus for any new parameter setting Θ^*, we have a predictive distribution for the process $\mathbf{Y}(\Theta^*)$.

New Inferential Approach

1. Predictive distribution provides a probability model (the Gaussian process emulator) that connects the parameters to the *observed* summary statistics $\mathbf{Z}$. This gives us a likelihood function. ("emulator likelihood"), $\mathcal{L}(\boldsymbol{\Theta})$. Avoids latent variables $\{L_{k,t}\}$ in calculation, i.e., do not have to deal with $\int \mathcal{L}(\boldsymbol{\Theta}, L) dL$ or high-dimensional posterior $\pi(\boldsymbol{\Theta}, L \mid \{I_{k,t}\})$
2. ML or Bayesian inference to obtain estimates of $\boldsymbol{\Theta}$.

Skipping lots of important details: dimension reduction, computational issues, worrying about discrepancy between model and data etc. ...

Example of Strategy 2: Non-Gaussian Spatial Data Models

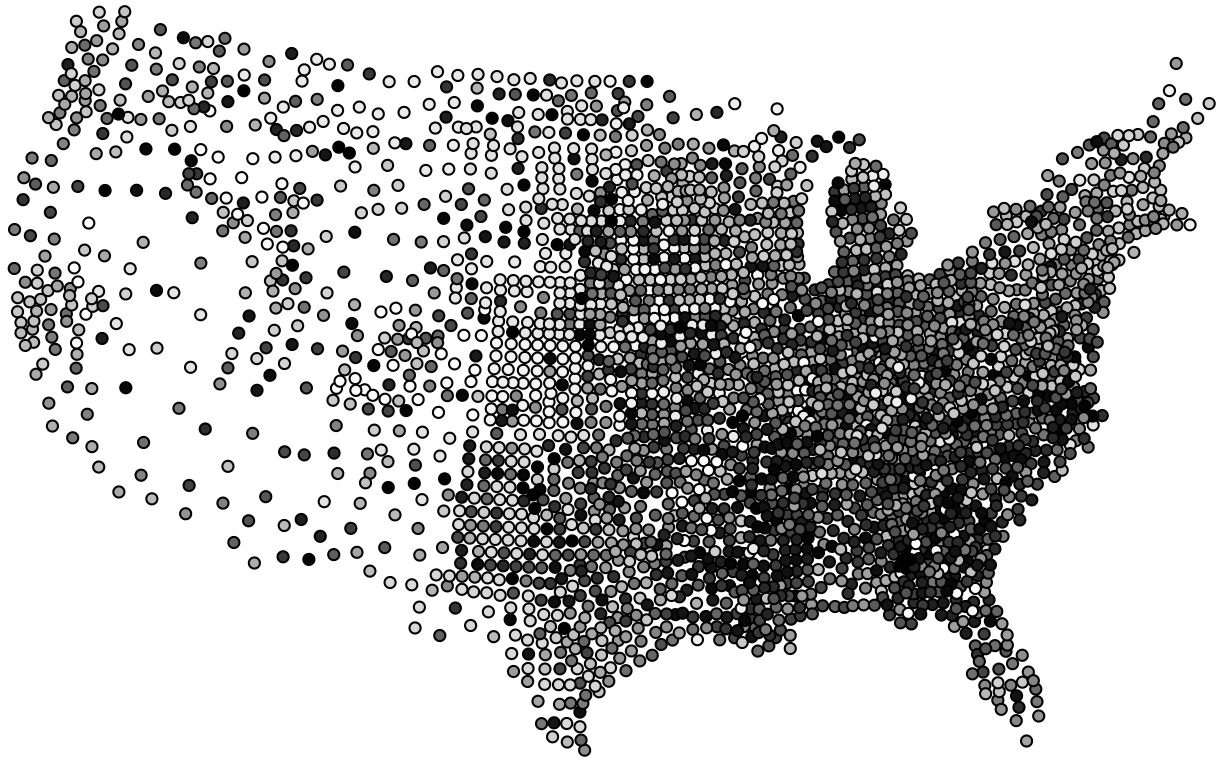
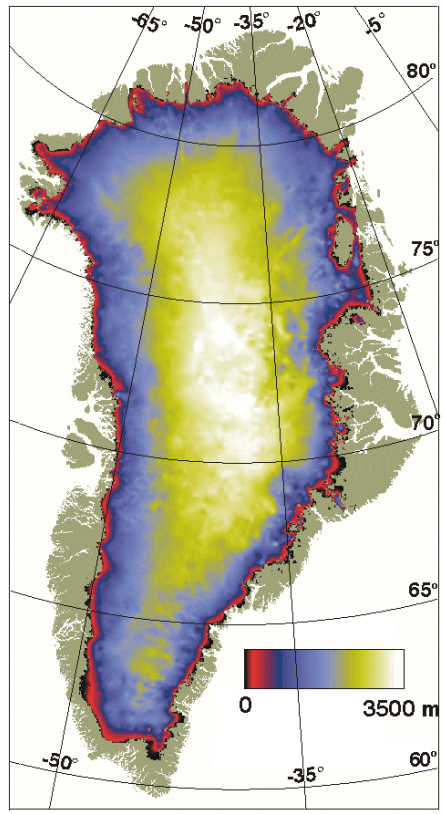


Figure: U.S. infant mortality data by county. $n = 3071$
Ratio of deaths to births, each averaged over 2002-2004.
Darker indicates higher rate.

Non-Gaussian Spatial Data Example #2



Greenland ice sheet thickness (Bamber et al., 2001)

Spatial Data on a Lattice

- Gaussian and non-Gaussian spatial data are very common and appear in a large number of disciplines.
- Common lattice data: 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.

Spatial Linear Models

- Spatial process at location $\mathbf{s}$ is $Z(\mathbf{s}) = X(\mathbf{s})\beta + W(\mathbf{s})$.
  - $X(\mathbf{s})$ are covariates at $\mathbf{s}$ and β is a vector of coefficients.
  - Model dependence among spatial random variables by imposing it on the errors (the $W(\mathbf{s})$'s).
- Gaussian Markov Random field (GMRF): Let Θ be the parameters for precision matrix $Q(\Theta)$. Then:

$$\mathbf{Z}_{n\times 1}|\Theta, \beta \sim N(\mathbf{X}_{n\times p}\beta_{p\times 1}, Q^{-1}(\Theta))$$

Spatial Linear Models: Dependence

- $Q = \text{diag}(A\mathbf{1}) - A$ where adjacency matrix A is such that $A_{ij} = 1$ if locations i and j are neighbors, 0 else
- Implications:
  - $W(\mathbf{s})$ is conditionally independent of all other Ws given its neighbors
  - uncertainty about $W(\mathbf{s})$ is inversely proportional to its number of neighbors.

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. $\log(\mu_i) = X(\mathbf{s}_i)\beta + W(\mathbf{s}_i)$
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:

$$g(E(Z_i \mid \beta, W_i)) = \mathbf{X}_i\beta + W_i = \mathbf{X}_i\beta + \mathbf{K}_i\gamma + \mathbf{L}_i\delta.$$

  $\mathbf{K}$ is collinear with $\mathbf{X}$.

This is the source of confounding. 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)

Observations

- Spatial random effects **W** are the cause of confounding issues as well as computational challenges.
- **W** are just a device to induce dependence. Not intrinsically important.
- Idea: reparameterize and reduce dimensions of **W**.

Spatial Confounding: Reparameterization Solution

- Reich, Hodges and Zadnik (2006) propose solution: since **K** have no scientific meaning, delete them from the model.
- $g(E(Z_i \mid \beta, \delta)) = \mathbf{X}_i\beta + \mathbf{L}_i\delta$. Prior for random effects δ now

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

  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.
- RHZ approach does not fully account for underlying graph

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

$$g(E(Z_i \mid \beta, \delta)) = \mathbf{X}_i\beta + \mathbf{M}_i\delta.$$

  And the prior for the random effects is now

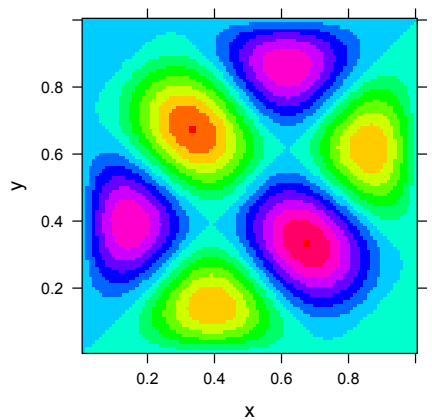
$$p(\delta \mid \tau) \propto \tau^{q/2} \exp\left(-\frac{\tau}{2}\delta'\mathbf{Q}^{**}\delta\right),$$

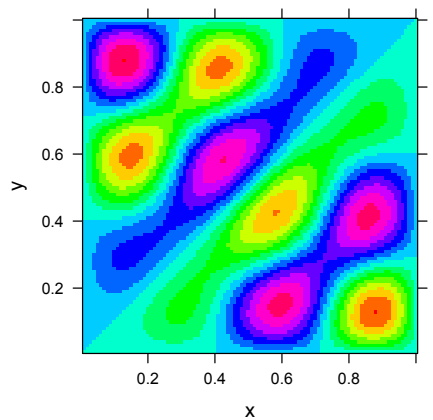
  where $\mathbf{Q}^{**} = \mathbf{M}'\mathbf{Q}\mathbf{M}$.
- Corrects issues due to confounding
- Potential for dimension reduction: if we reduce dimensions of $\mathbf{M}_i$ to q, the # parameters is reduced from $n + p + 1$ to $q + p + 1$ (q can be small)

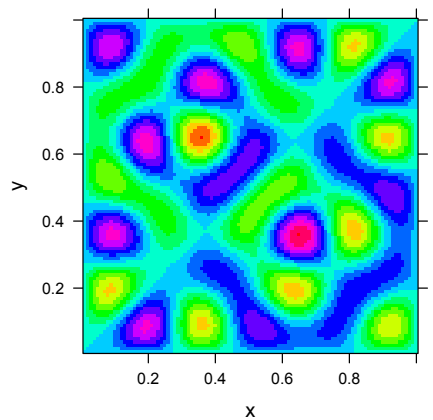
Interpreting the Resulting Reparameterization

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

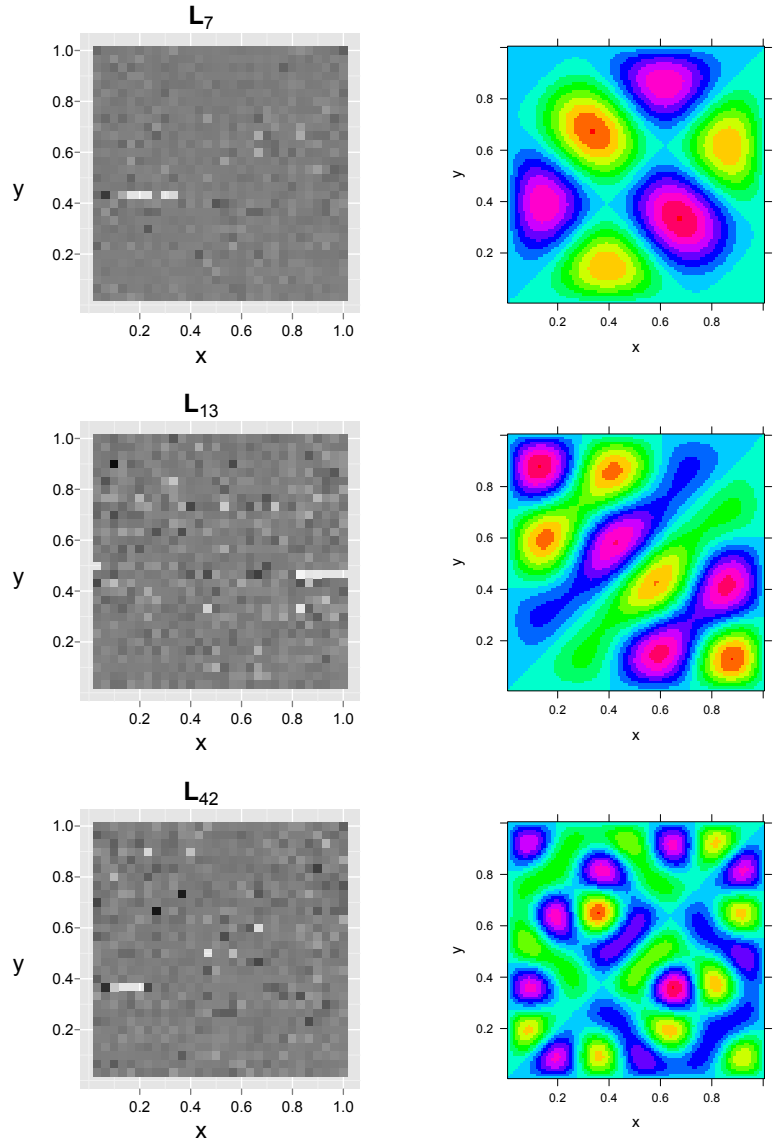
Some selected basis vectors for the 30×30 lattice.







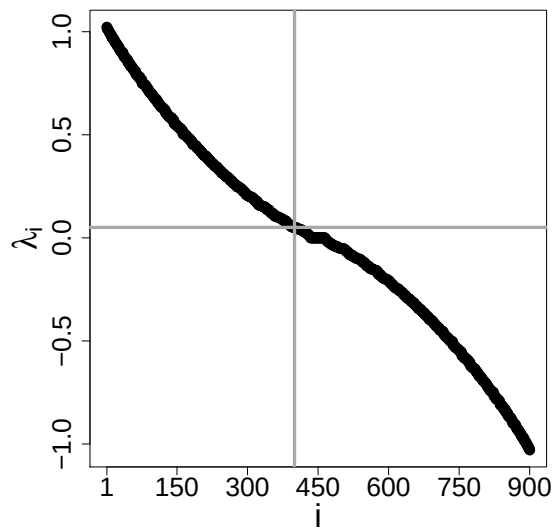
Eigenvectors 7, 13, 42 from RHZ and Moran bases



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

Study: Inference for Spatial Binary

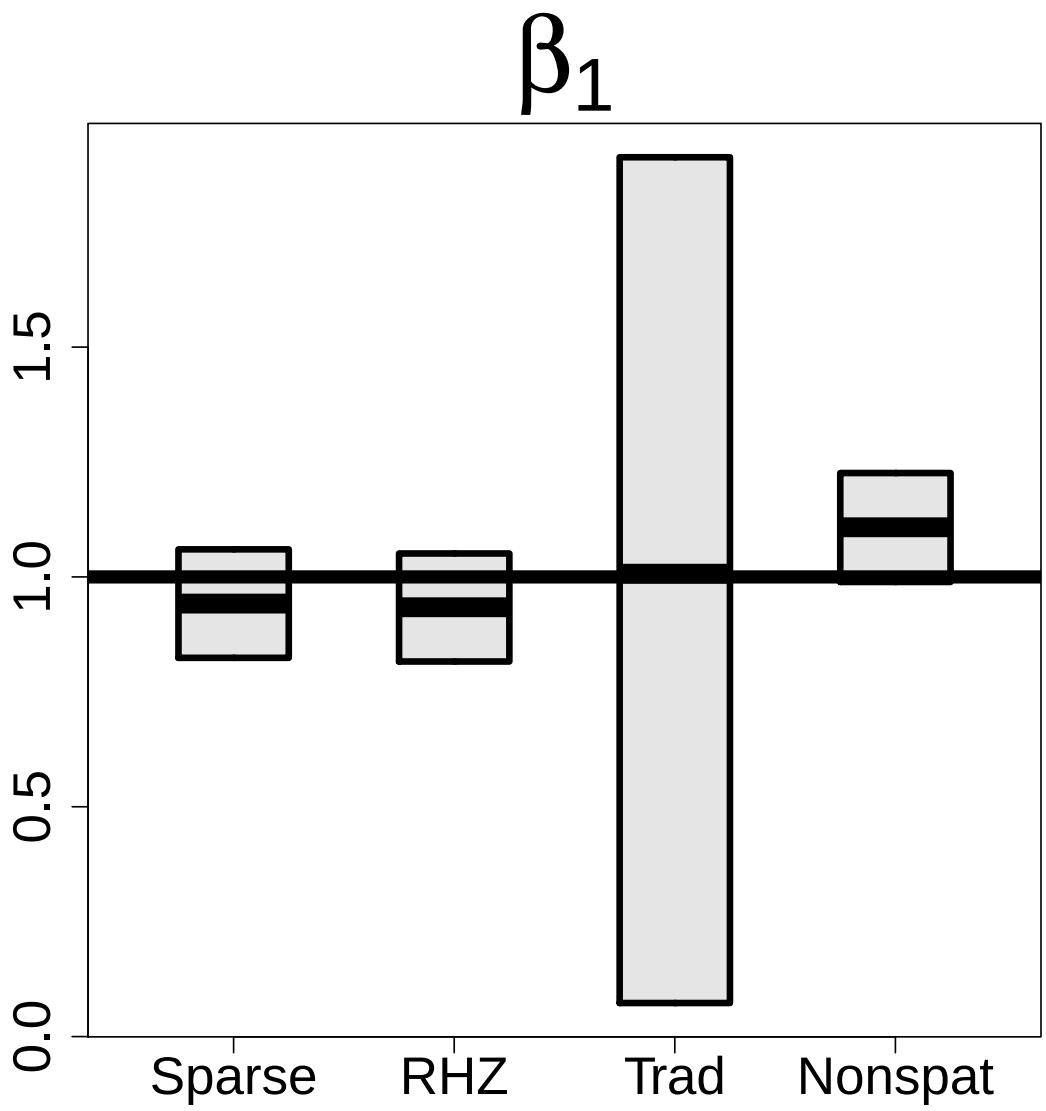
30×30 lattice simulated from RHZ model with $\beta_1 = \beta_2 = 1$.
Predictors are the coordinates of unit square.

Model	$\hat{\beta}_1$ CI(β_1)	$\hat{\beta}_2$ CI(β_2)
Sparse	1.080 (0.613, 1.556)	1.130 (0.644, 1.635)
RHZ	1.120 (0.637, 1.606)	1.192 (0.679, 1.713)
Traditional	0.500 (-2.655, 3.616)	-0.605 (-3.698, 2.577)

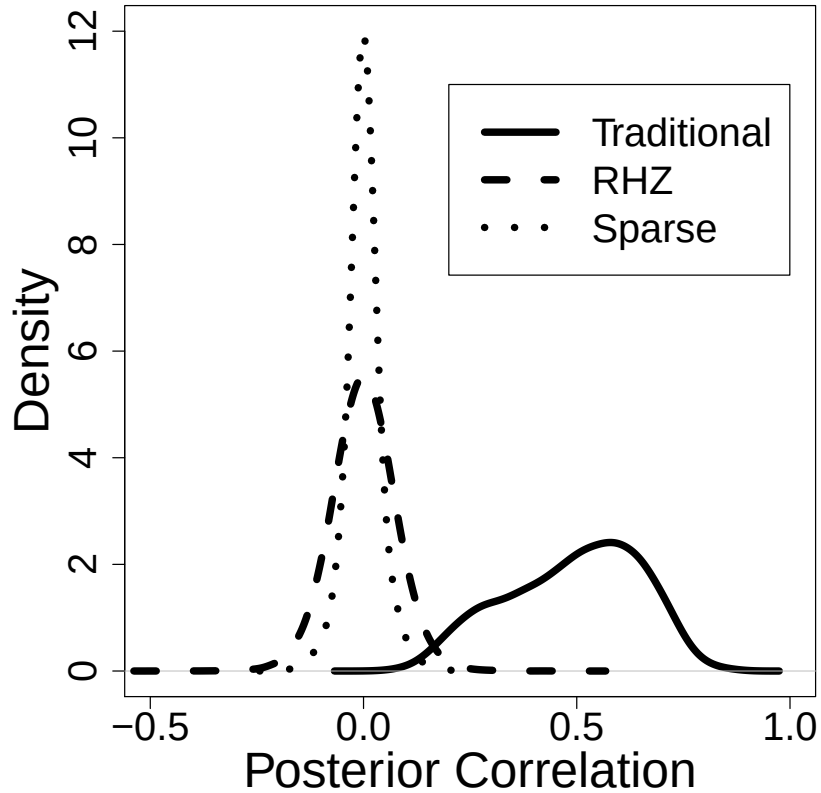
- Point and interval estimates for Traditional are very poor: 95% interval includes 0
- Sparse and RHZ produce similar (good) results

Similar results for Gaussian (linear) and Poisson

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

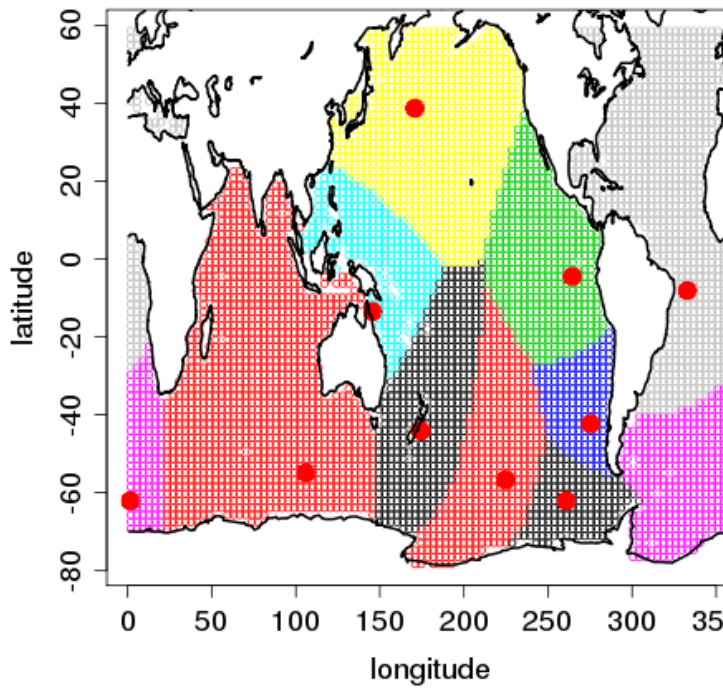
- MCMC algorithm is
  - faster per iteration (far fewer random effects)
  - mixes faster (random effects are "decorrelated")
- Far greater speed-ups with much smaller q, e.g. 25-50 is adequate for our examples (we are also being *extremely* careful by running very long chains!)
  Real data example: 14 days (traditional) versus 2-8 hours

Summary

- I have described three fairly different frameworks for thinking about fast computing for high-dimensional latent variable models.
- Different strategies are better suited to different models.
- Lots of opportunities for interesting research, including ways for combining these methods.

Approach 2: Block Composite Likelihood

- Composite likelihood for spatial data (Vecchia, 1988; Stein et al., 2004; Caragea and Smith, 2006; Eidsvik et al., 2013)
- Block composite likelihood (Caragea and Smith 2006):
  - Divide spatial field into M blocks
  - Conditional independence between blocks given their block means
  - Large scale trend captured by dependence between block means
  - Small scale variation captured by dependence within each block
  - Valid probability model: Important for rigorous Bayesian inference



Composite Likelihood Basics

- Approximating log likelihood function as sum of sub-log likelihoods
- Each sub-likelihood is likelihood based on part or summary of data
- Maximum composite likelihood estimator (MCLE): consistency and asymptotic normality under same conditions as MLE

Emulation Using Composite Likelihood

- We approximate original log likelihood $\ell(\mathbf{Y}|\boldsymbol{\xi})$ by

$$c\ell(\mathbf{Y}|\boldsymbol{\xi}) \quad \propto \quad \underbrace{\ell(\bar{\mathbf{Y}}|\boldsymbol{\xi})}_{\text{likelihood for block means}} \quad + \quad \underbrace{\sum_{i}^{M} \ell(\mathbf{Y}_{(i)}|\bar{\mathbf{Y}}_{(i)}, \boldsymbol{\xi})}_{\text{likelihood for block output}} \ ,$$

  with emulator parameter $\boldsymbol{\xi}$, collection of block means $\bar{\mathbf{Y}}$, and ith output $\mathbf{Y}_{(i)}$ and mean $\bar{\mathbf{Y}}_{(i)}$.

- MCLE $\hat{\boldsymbol{\xi}}$ is consistent under original emulation model.

  ▸ original model

Calibration Using Composite Likelihood

- Approximate original log likelihood $\ell(\mathbf{Z}|\mathbf{Y}, \theta, \xi_\delta)$ by

$$c\ell(\mathbf{Z}|\mathbf{Y}, \theta, \xi_\delta, \hat{\xi}) \propto \underbrace{\ell(\bar{\mathbf{Z}}|\bar{\mathbf{Y}}, \theta, \xi_\delta, \hat{\xi})}_{\text{likelihood for block means}} + \underbrace{\sum_i^M \ell(\mathbf{Z}_{(i)}|\mathbf{Y}_{(i)}, \bar{\mathbf{Z}}_{(i)}, \theta, \xi_\delta, \hat{\xi})}_{\text{likelihood for block observations}},$$

  with covariance parameter for discrepancy ξ_δ, block means for observations $\bar{\mathbf{Z}}$, ith block observation $\mathbf{Z}_{(i)}$ and mean $\bar{\mathbf{Z}}_{(i)}$.

- Infer θ through "composite" posterior

$$\log \pi(\theta, \xi_\delta|\mathbf{Y}, \mathbf{Z}, \hat{\xi}) \propto c\ell(\mathbf{Z}|\mathbf{Y}, \theta, \xi_\delta, \hat{\xi}) + \log p(\theta) + \log p(\xi_\delta)$$

  with prior densities $p(\theta)$ and $p(\xi_\delta)$.

Theorems on Posterior Mode

- For posterior mode $\hat{\theta}_n$,
  - Under probability model for composite posterior, asymptotic covariance for $\hat{\theta}_n$ is $\mathbf{Q}_n^{-1}$.
  - Under original model, asymptotic covariance for $\hat{\theta}_n$ is $\mathbf{G}_n^{-1} = \mathbf{Q}_n \mathbf{P}_n^{-1} \mathbf{Q}_n$, inverse of Godambe information matrix.

  $\mathbf{P}_n$: Covariance matrix for gradient of composite likelihood.
  $\mathbf{Q}_n$: Information matrix for composite likelihood.

- Proofs follow directly from Chernozhukov and Hong (2003). Details in Chang et al. (2013)

Covariance Adjustment

- For posterior mode $\hat{\theta}_n$ and true value θ_0,
  - Under composite posterior
    $\mathbf{Q}_n^{\frac{1}{2}} \left(\hat{\theta}_n - \theta_0 \right) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}),$ as $n \to \infty$
  - Under original posterior $\mathbf{G}_n^{\frac{1}{2}} \left(\hat{\theta}_n - \theta_0 \right) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}),$ as $n \to \infty$

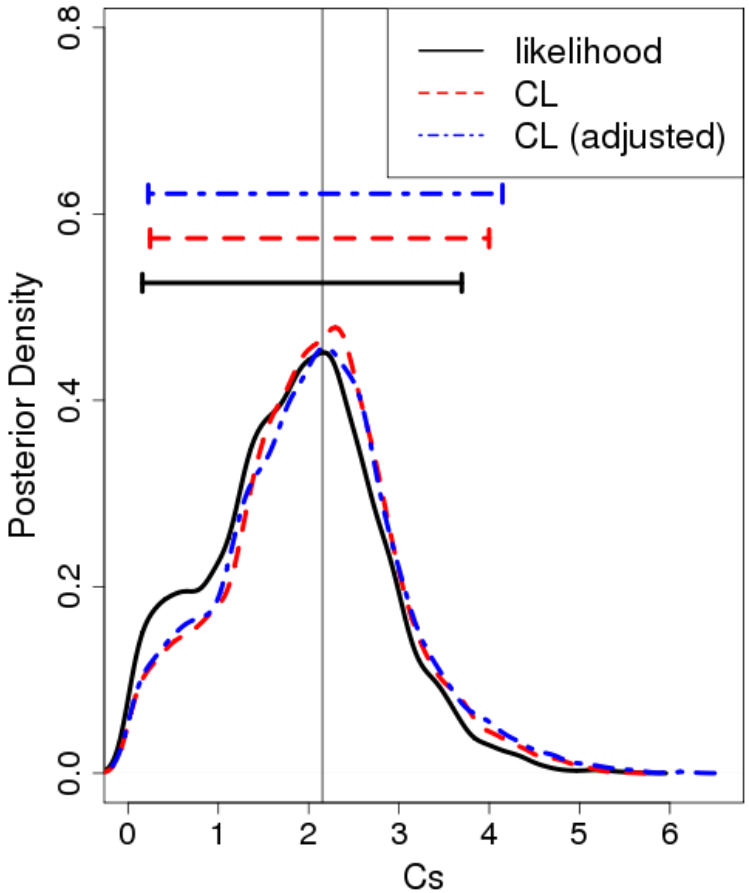
- Asymptotic covariance of $\hat{\theta}_n$ computed using MCMC sample from composite posterior converges to $\mathbf{Q}_n^{-1}$

- **Open-faced sandwich adjustment** (Shaby 2012): For each posterior draw for θ, compute
  $\tilde{\theta}^{open} = \hat{\theta}_n + \mathbf{Q}_n^{-1} \mathbf{P}_n^{\frac{1}{2}} \mathbf{Q}_n^{\frac{1}{2}} (\theta - \hat{\theta}_n)$.
  $\Rightarrow$ This adjusts covariance from $\mathbf{Q}_n^{-1}$ to $\mathbf{G}_n^{-1}$.

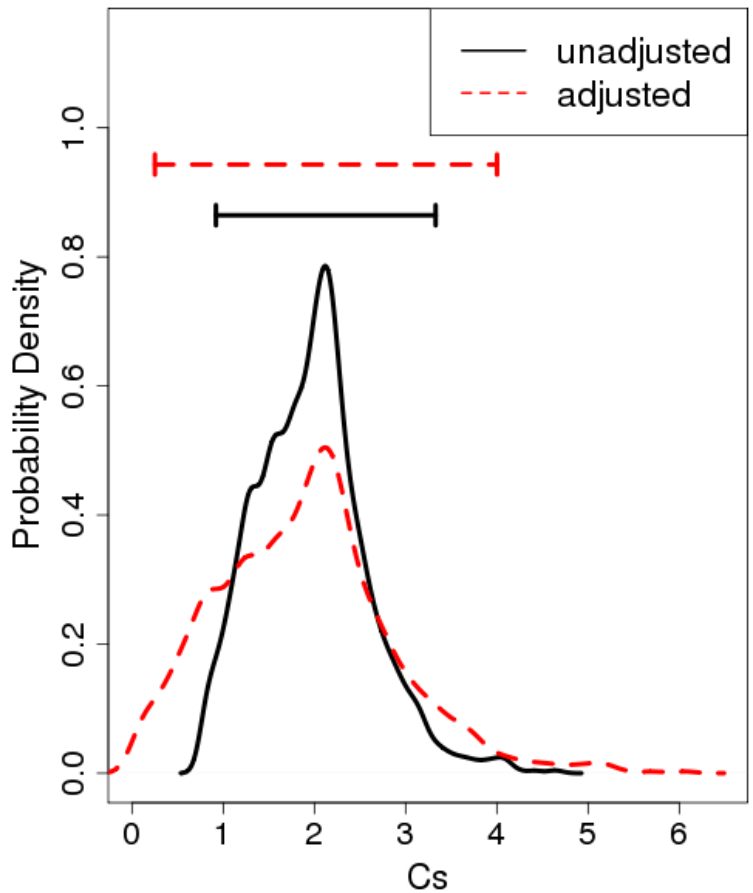
Results

Climate sensitivity estimation using sea temperature anomaly:

Simulated Example (n=1000, # of blocks=10)



Real Data (n=5903, # of blocks=200)



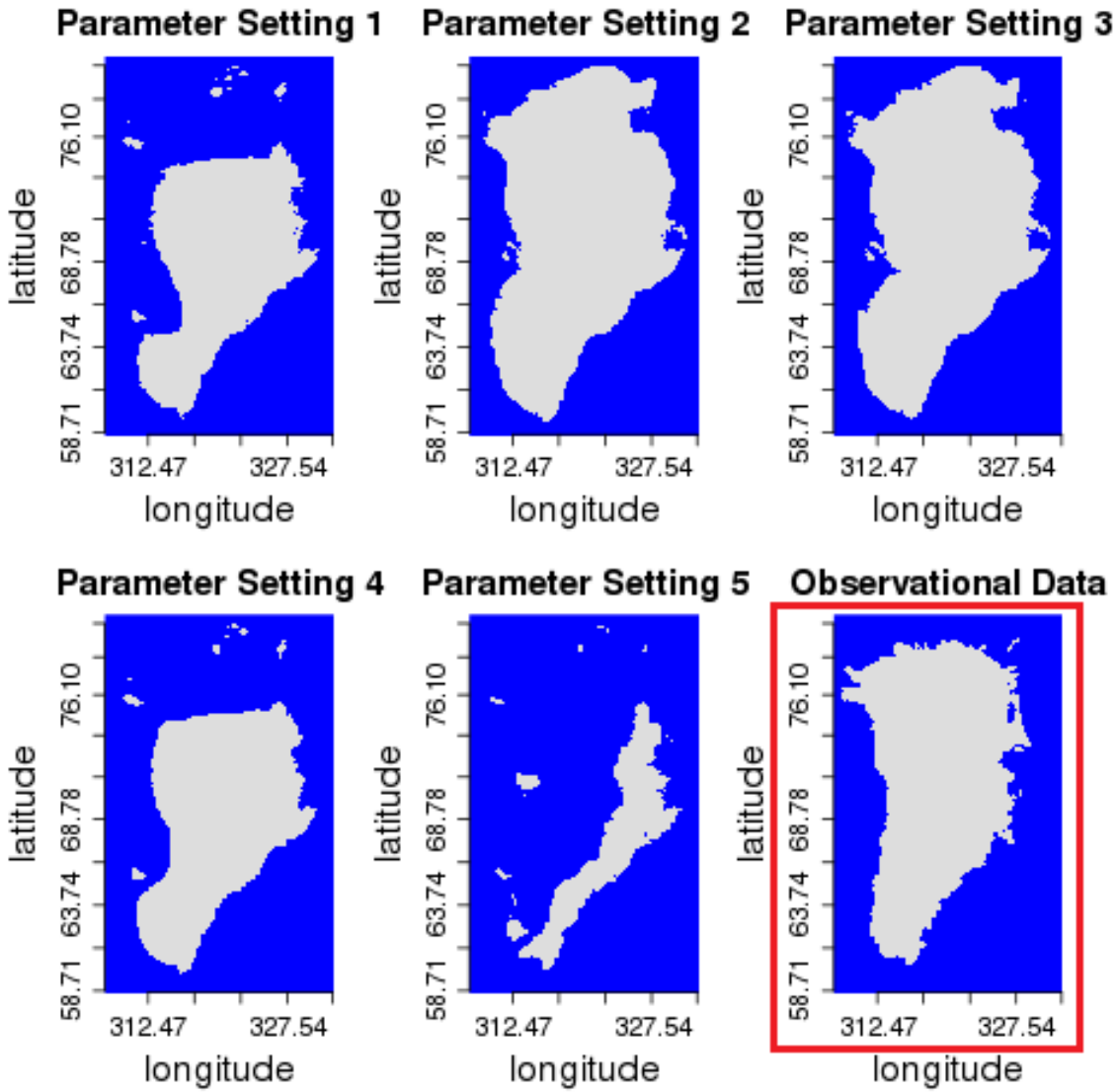
▸ other results

Discussion

- Dimension reduction-based approach:
  - Very fast, scales well with n, number of spatial locations
  - Very easy to use: Automatic emulation step
- Composite likelihood-based approach:
  - Formal way to account for information loss due to blocking

Ongoing Research: Calibration Problem for Binary Output

Again, which output best matches the observations?



References

- PCA approach:
  - Chang, W., Haran, M., Olson, R., and Keller, K. (2013) Fast dimension-reduced climate model calibration, *accepted for publication in the Annals of Applied Statistics, arXiv:1303.1382*.
  - Chang, W., Applegate, P., Haran, M. and Keller, K. (2013) Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties, *under revision*.
- Composite likelihood approach:
  - Chang, W., Haran, M., Olson, R., and Keller, K. (2013) A composite likelihood approach to computer model calibration with high-dimensional spatial data, *tentatively accepted by Statistica Sinica, arXiv:1308.0049*.

This work was supported by the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507.

General Asymptotic Properties of Maximum CL

Theorem

(Lindsay, 1988) Under the same regularity conditions as used for ordinary likelihood, the maximum composite likelihood estimator (MCLE) $\hat{\psi}_n^{CL}$ *has*

- *Consistency:* $\hat{\psi}_n^{CL} \xrightarrow{\mathcal{P}} \psi$ *as* $n \to \infty$*, where* ψ^* *is the true value of* ψ*.*
- *Asymptotic Normality:*

$$\mathbf{G}_n^{\frac{1}{2}} \left(\hat{\psi}_n^{CL} - \psi \right) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}),$$

where $\mathbf{G}_n = \mathbf{Q}_n \mathbf{P}_n^{-1} \mathbf{Q}_n$ *is the Godambe information matrix (Godambe, 1960),* $\mathbf{P}_n$ *is the covariance matrix of the gradient* $\nabla c\ell_n$*, and* $\mathbf{Q}_n$ *is the information matrix of* $c\ell_n$*.*

Asymptotic Results

Theorem: For posterior mode $\hat{\theta}_n$ and true value θ_0,

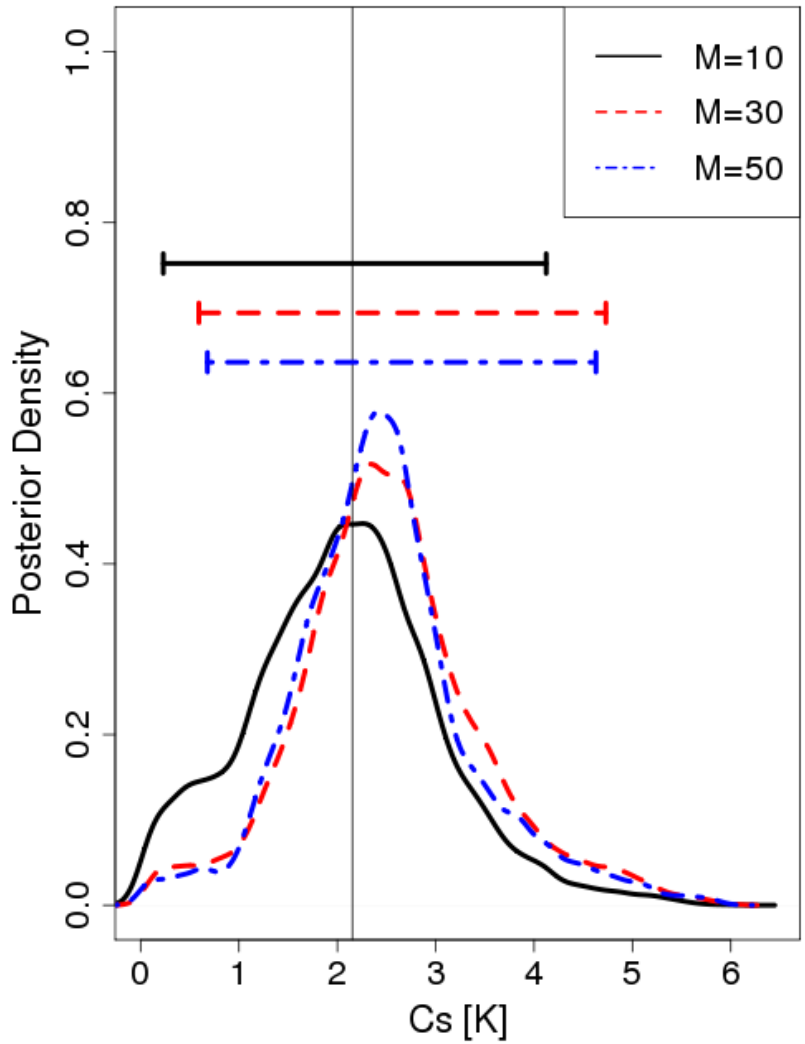
1. **Consistency**: Posterior density of θ degenerates on θ_0 in total variation.
   For finite n, covariance of θ is approximately $\mathbf{Q}_n^{-1}$.
2. **Normality**: For Godambe information matrix $\mathbf{G}_n = \mathbf{Q}_n \mathbf{P}_n^{-1} \mathbf{Q}_n$,

$$\mathbf{G}_n^{\frac{1}{2}} \left(\hat{\theta}_n - \theta_0 \right) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}), \text{ as } n \to \infty,$$

   utilizing results from Chernozhukov and Hong (2003).
   $\mathbf{P}_n$: Covariance matrix of gradient.
   $\mathbf{Q}_n$: Information matrix of composite likelihood.

Effect of No. of Blocks

- Trade-offs between block sizes (n_i) and number of blocks (M)
- More blocks: Faster computation, but slower convergence
- Heuristic guideline: Keep block size larger than effective range ($\frac{3}{\text{range parameter}}$)



back

Pseudo Likelihood

- Computing exact likelihood for autologistic model involves computing intractable constant.
- Conditional composite likelihood (Besag, 1975):

$$c\ell(\mathbf{Z}, \boldsymbol{\eta}(\boldsymbol{\theta}) | \psi, \boldsymbol{\theta}) = \sum_{j=1}^{n} \log f(Z(\mathbf{s}_j) | \mathbf{Z}_{-j}, \psi, \boldsymbol{\eta}(\boldsymbol{\theta})) + \log f(\boldsymbol{\eta}(\boldsymbol{\theta}) | \boldsymbol{\theta}).$$

  where $\mathbf{Z}_{-j} = \{Z_k | k \in \mathcal{N}_j\}$.
- Other approaches are not feasible for large n.