Towards automating MCMC algorithms for spatial generalized linear models

Murali Haran

Department of Statistics
Penn State University

(collaborators: J.Flegal, G.Jones and L.Tierney)

The University of Texas M.D.Anderson Cancer Center

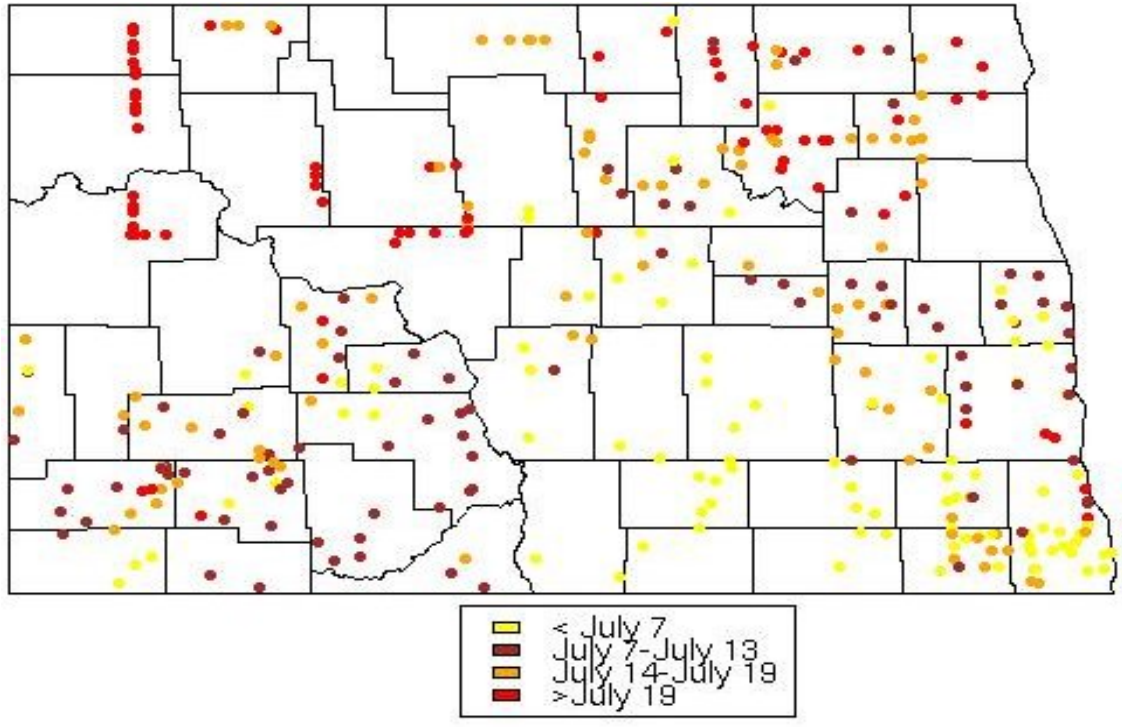
March 2009

What are spatial models?

- Models for data that are geographically referenced: Each random variable Z has a location $\mathbf{s}$ associated with it.
- Let $\mathbf{s}$ vary over index set $D \subset \mathbb{R}^d$ so as to generate the multivariate random process: $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$.
- Here I will be concerned with:
  - Geostatistical models: D is a fixed subset of $\mathbb{R}^d$. Process is infinite-dimensional (locations vary continuously in space) but observed at a finite set of locations. e.g. pollutant levels across Pennsylvania only observed at monitoring stations.
  - Areal/lattice models: D is a finite set of locations in $\mathbb{R}^d$, used to represent data often observed on or aggregated up to arbitrary spatial units such as census tracts, counties. e.g. cancer rates by county across Minnesota.

Geostatistics

Wheat flowering dates in North Dakota

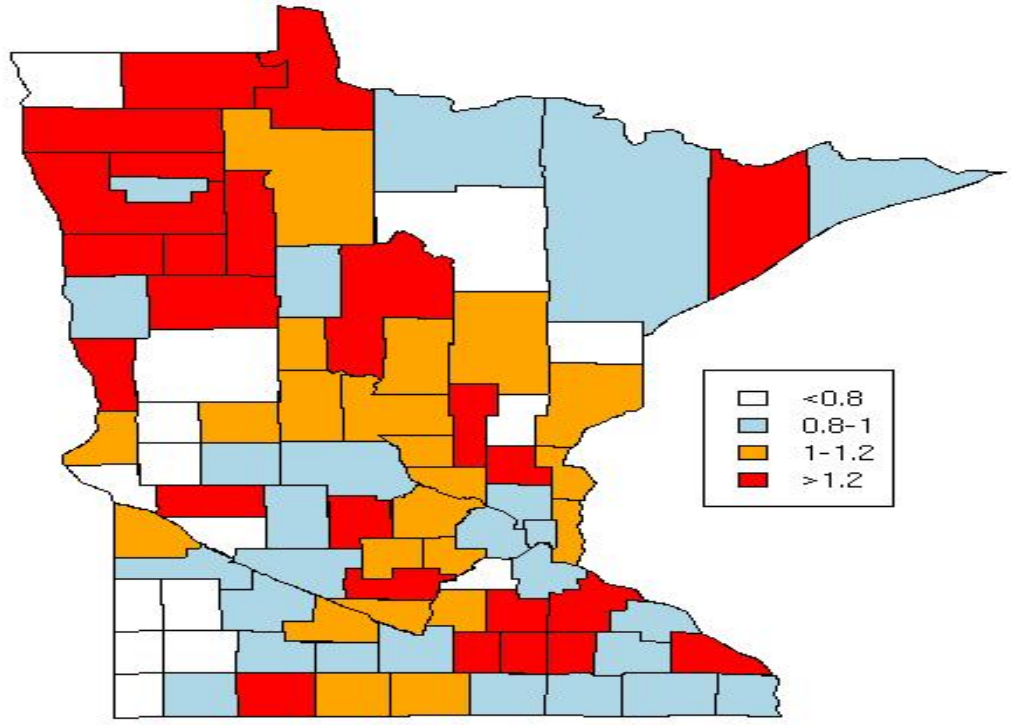


Courtesy Plant Pathology, PSU and North Dakota State.

(Haran, Bhat, Molineros, DeWolf, 2008)

Areal data

Minnesota cancer rates by county: $\frac{\text{observed}}{\text{expected}}$ counts



Courtesy MN Cancer Surveillance System, Dept. of Health

Why focus on spatial models ?

- Spatial models are very widely used by statisticians and non-statisticians. E.g. Gaussian process-based models are important for nonparametric regression, emulating complex computer experiments, classification.
- Automated, reliable algorithms for even a few specific models will be very useful for people who want to fit these models routinely.
- Relatively little theory on MCMC algorithms used.
- Strong dependence among variables can make posterior distributions challenging to simulate efficiently.
- Connections to other important models.

Unrealistic/too hard to tackle much bigger class...

Basic Gaussian random field (linear) model

- Spatial process at location $\mathbf{s}$ is $Z(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s})$ where:
  - $\mu(\mathbf{s})$ is the mean. Often $\mu(\mathbf{s}) = X(\mathbf{s})\beta$, $X(\mathbf{s})$ 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).
- For n locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n$, $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$ can be jointly modeled via a zero mean Gaussian process (GP) for geostatistics, or Gaussian Markov random field (GMRF) for areal/lattice data.
- Gaussian Process (GP): Let Θ be the parameters for covariance matrix $\Sigma(\Theta)$. Then:

$$\mathbf{Z}|\Theta, \beta \sim N(\mathbf{X}\beta, \Sigma(\Theta))$$

Spatial linear model (contd.)

- Gaussian Markov Random field (GMRF): Let Θ be the parameters for precision matrix $Q(\Theta)$. Then:

$$\mathbf{Z}|\Theta, \beta \sim N(\mathbf{X}\beta, Q^{-1}(\Theta))$$

- For some popular forms of the Gaussian Markov random field the precision matrix is singular so:

$$f(\mathbf{Z}|\Theta, \beta) \propto c(\Theta) \exp\left(-\frac{1}{2}(\mathbf{Z} - \mu(\mathbf{s}))^T Q(\Theta)(\mathbf{Z} - \mu(\mathbf{s}))\right).$$

- For spatial linear model, once priors for Θ, β specified, inference is based on posterior $\pi(\Theta, \beta \mid \mathbf{Z})$.
- Key observation: Θ typically has low dimensions (2-5) for both GP and GMRF models, while dimensions of $\mathbf{Z}$ can be large.

Spatial generalized linear model

What if data are non-Gaussian? (Diggle, Tawn, Moyeed, 1998)

- Stage 1: Model $Z(\mathbf{s}_i)$ conditionally independent with distribution f given parameters β, Θ, spatial errors $w(\mathbf{s}_i)$

$$f(Z(\mathbf{s}_i)|\beta, \Theta, w(\mathbf{s}_i)),$$

  where $g(E(Z(\mathbf{s}_i))) = \eta(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)\beta + w(\mathbf{s}_i)$, η is a canonical link function (for example the logit link).
- Stage 2: Again $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$. Model $\mathbf{w}$ as spatially dependent either via a GP or GMRF.
- Stage 3: Priors for Θ, β.
- Inference based on $\pi(\Theta, \beta, \mathbf{w} \mid \mathbf{Z})$.

Spatial generalized linear model for ecology

- Spatial data with excess zeros: very common in ecology.
- Of interest: (a) determining spatial field corresponding to incidence — binary outcomes, (b) determining spatial field corresponding to prevalence — counts.
- A two-stage model (Recta, Haran, Rosenberger, 2009):
  - SGLM provides elegant approach to build model based on presumed data generating mechanism.
  - $U(\mathbf{s})$ spatial binary process with logit link and latent GP.
  - $V(s)$ is defined only (conditional on) when $U(\mathbf{s}) = 1$. $V(\mathbf{s})$ has spatial truncated Poisson distribution with log link and latent GP. Zero-inflated data arise from $U, V \mid U$.
  - Separate regressions for incidence and prevalence, separate posterior predictive distribution for each.
  - Relies heavily on sample-based inference: MCMC.

MCMC for posterior inference

Goal: estimate $E_\pi g$ for real valued functions g.

MCMC: Construct a Harris-ergodic Markov chain $X_1, X_2, \ldots$ with stationary distribution π so that if $E_\pi |g(x)| < \infty$:

$$\bar{g}_n = \sum_{i=1}^{n} g(X_i)/n \to E_\pi g$$

When simulating from π for spatial models (careful) practitioners face several issues:

- Which Metropolis-Hastings algorithm should we use?
- We do not know how long to run our Markov chain.
- Hard to know if CLT holds.
- X_is are dependent so accuracy (variance) of estimator is hard to estimate.

Resolving MCMC issues

- Convergence rates, upper bounds on distance to stationarity (cf. Rosenthal, 1995; Jones & Hobert, 2004.) Generally very difficult, requires new analytical work for each new problem.
- Usual MCMC diagnostics:
  - Easy to use, automated software exists (in `WINBUGS` for example.) Useful heuristics but not reliable, all are known to fail (cf. Cowles and Carlin, 1996).
- Standard error estimates:
  - Typically assume stationarity, often not consistent, can overestimate (e.g. IMSE) or underestimate (usual batch means with fixed batch sizes).

Automation of MCMC

Ideally:

- Automated approach for constructing algorithm (proposal distributions). No ‘tuning’ necessary.
- Generate starting values automatically. No need to experiment with different starting values.
- Have a rigorous criteria for determining when to stop the chain that is *related to inferential goals*.
- Some theoretical guarantees regarding all of the above.

Options

1. Exact sampling:
   - Obtain *exact* draws from π using a Markov chain.
   - Make classical (old fashioned) Monte Carlo methods (such as rejection sampling) practical.
2. Construct Metropolis-Hastings so Markov chain sampler mixes well (uniformly or geometrically ergodic):
   - Can estimate Monte Carlo standard errors consistently.
   - We know when CLT holds.

Option (1) is generally very hard for spatial models.

Option (2) is hard to achieve and usually hard to prove.

However, these options may be available when an approximation $\hat{\pi}$ is available that is: (i) close to target (π), (ii) heavy-tailed (with respect to π), (iii) easy to simulate from.

An approximation

- Consider SGLM so inference is based on $\pi(\Theta, \beta, \mathbf{w} \mid \mathbf{Z})$.
- Start by finding a linear spatial model that is reasonably close to this model.
  - Transform data $\mathbf{Z}$ (to $\mathbf{Y}$ say), use approximations (e.g. Laplace) to obtain:

$$\mathbf{Y} \mid \Theta, \beta, \mathbf{w} \sim N(\mu(\beta, \mathbf{w}), \Sigma(\Theta))$$

  - Posterior for this model: $S(\Theta, \beta, \mathbf{w} \mid \mathbf{Y})$. For convenience denote this by: $S(\Theta, \beta, \mathbf{w})$.
- Analytically integrate: $S_1(\Theta, \beta) = \int S(\Theta, \beta, \mathbf{w}) d\mathbf{w}$.
- From $S(\Theta, \beta, \mathbf{w})$, can obtain approximate conditional distribution of spatial random effects, $S_2(\mathbf{w} \mid \Theta, \beta)$ (multivariate normal). Then, we have

$$S(\Theta, \beta, \mathbf{w}) = S_1(\Theta, \beta) S_2(\mathbf{w} \mid \Theta, \beta).$$

A heavy-tailed approximation

Construct heavy-tailed approximation $\hat{\pi}(\Theta, \beta)$:

- $S_1(\Theta, \beta) S_2(\mathbf{w} \mid \Theta, \beta) \approx \pi(\Theta, \beta, \mathbf{w}|Y)$ as just described.
- Find heavy-tailed approximation to $S_1(\Theta, \beta)$: $\hat{\pi}_1(\Theta, \beta)$.
- Find heavy-tailed (multi-t) approximation to $S_2(\mathbf{w}|\Theta, \beta)$, $\hat{\pi}_2(\mathbf{w}|\Theta, \beta)$. Easy: it should have same mean and variance as the multivariate normal $S_2(\mathbf{w}|\Theta, \beta)$.
- Simple sequential sampling to generate proposal from $\hat{\pi}(\Theta, \beta, \mathbf{w})$.

Heavy-tailed approximation (contd.)

- Sample from $\hat{\pi}$:
  1. Sample $(\Theta, \boldsymbol{\beta}) \sim \hat{\pi}_1(\Theta, \boldsymbol{\beta})$.
  2. Sample $(\mathbf{w}) \sim \hat{\pi}_2(\mathbf{w} \mid \Theta, \boldsymbol{\beta})$. Multivariate-t distribution with precision $Q(\Theta)$ or covariance $\Sigma(\Theta)$ using $\Theta, \boldsymbol{\beta}$ sample from Step 1.
- $\hat{\pi}(\Theta, \boldsymbol{\beta}, \mathbf{w})$: proposal for Monte Carlo algorithms.
- Note:
  - Step 1 of proposal generation is typically easy (fast) since low-dimensional.
  - Step 2: Matrix operations (of order $O(N^3)$) are involved when generating proposal, evaluating Met-Hastings ratio. Fast if GMRF, slow if GP unless sparse covariance.

Two Monte Carlo approaches

Let $\Psi = (\Theta, \mathcal{B})$. If we can show (as in Haran and Tierney, 2009):

$$\sup_{\Psi} \frac{\pi(\Psi)}{\hat{\pi}(\Psi)} < \infty.$$

1. Numerically maximize $\frac{\pi(\Psi)}{\hat{\pi}(\Psi)}$ to obtain $K < \infty$.
   - (a) Rejection sampling or (b) perfect tempering (Møller and Nicholls, 1999): use simulated tempering to construct a perfect sampler.
2. Metropolis-Hastings 'independence sampler' (cf. Tierney, 1994): propose every M-H update from $\hat{\pi}$.
   - Sampler is uniformly ergodic (Mengersen, Tweedie, 1996)

     $$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq Mt^n,$$

     for any x in the state space. P^n is the n-step transition kernel, $M < \infty$, $t \in (0, 1)$.

Stopping rules and estimating standard errors

Consider the first category of algorithms:

- Rejection sampler/perfect tempering: iid draws.
  - Central Limit theorem holds if $E_\pi g(x)^2 < \infty$:

$$\sqrt{n}(\bar{g}_n - E_\pi g) \to N(0, \sigma^2) \text{ in distribution}$$

  - Standard error estimation: Estimate σ^2 by s^2, sample variance.
  - Stopping rule: When estimated standard error ($s/\sqrt{n}$) is below a desired level, stop the sampler.
  - These are ideas from introductory statistics!
- Need the bounding constant K which can be difficult to obtain when $\pi, \hat{\pi}$ are complicated.
- Note: Strictly speaking these are *almost* exact samplers since K still has to be estimated.

Stopping rules, estimating standard errors (contd.)

Consider the second category, using $\hat{\pi}$ as a proposal:

- Independence chain is uniformly ergodic so:
  - Central Limit theorem holds if $E_\pi g(x)^2 < \infty$:

$$\sqrt{n}(\bar{g}_n - E_\pi g) \to N(0, \sigma^2) \text{ in distribution}$$

  - Standard error estimation: Estimate σ^2 by consistent batch means (Jones, Haran, Caffo, Neath, 2006).
  - Stopping rule (**'fixed width' approach**): When estimated standard error is below a desired level, stop the sampler (Jones et al. (2006); Flegal, Haran, Jones (2008)).
  - Not quite introductory statistics, but just as easy in practice.
- Do not need bounding constant K.

Case study: disease mapping model

Consider Besag, York, Mollie (1991) model used for disease mapping.

- Suppose $Z(\mathbf{s}_i)$'s are counts.
- Let $E(\mathbf{s}_i)$: estimate of expected number of events in region i (assume this is known.)
- $Z(\mathbf{s}_i)|\mu(\mathbf{s}_i) \sim$ Poisson$(E(\mathbf{s}_i)e^{\mu(\mathbf{s}_i)}),\ i = 1, ..., N$, where
  - $\mu(\mathbf{s}_i)$: log-relative risk of event

$$\mu(\mathbf{s}_i) = \theta(\mathbf{s}_i) + \phi(\mathbf{s}_i)$$

- $\theta(\mathbf{s}_i)$'s are non-spatial:

$$\theta(\mathbf{s}_i)|\tau_h \overset{iid}{\sim} N(0, 1/\tau_h)$$

Case study (contd)

$\phi(\mathbf{s}_i)$'s form a GMRF. $i \sim j \Rightarrow i, j$ are neighbors.

$$\phi(\mathbf{s}_i)|\phi(\mathbf{s}_{-i}), \tau_c \sim N\left(\frac{\sum_{i\sim j}\phi(\mathbf{s}_j)}{n_i}, \frac{1}{\tau_c n_i}\right)$$

n_i= number of neighbors of ith region. Alternatively,

$$f(\boldsymbol{\phi} \mid \tau_c) \propto \tau_c^{(N-1)/2} \exp\left(-\frac{1}{2}\boldsymbol{\phi}^T Q(\tau_c)\boldsymbol{\phi}\right),$$

where $\boldsymbol{\phi} = (\phi(\mathbf{s}_1), \ldots, \phi(\mathbf{s}_N))$ and $Q(\tau_c)$ is an adjacency matrix.

Add priors for the precision parameters τ_h, τ_c, say Inverse Gamma densities.

Posterior: $\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \tau_h, \tau_c|\mathbf{Z})$, of $2N + 2$ dims.

Examples

- Minnesota cancer data sets: 176 parameters.
- Use exactly same heavy tailed proposal, $\hat{\pi}$, for rejection sampler, perfect tempering, and Independence chain.
- Stop all algorithms when Monte Carlo standard errors are below same threshold for parameters.

algorithm	samples required	time taken
exact sampling	2,408	96 min.
independence chain	10,944	$<$4 min.

Perfect tempering performance $\approx$ rejection sampling.

Examples (contd)

- All three samplers: reasonable estimates, similar inference.
- Exact samplers are much less efficient than independence MCMC (related theoretical discussion in Liu (1996)).
- For other examples: similar results — typically the harder the problem, the greater the payoff from running the independence chain.
- Consider larger data set example. County-level infant mortalities: 910-dimensional posterior.
  - Exact sampling is not feasible: acceptance rate $\approx$ 1 per 20,000 samples generated; sample generation also more expensive. For example: $\approx$ 60 hrs.: 56 samples.
  - Independence chain still works (though slow.) Get estimates with desired accuracy in $\approx$ 60 hrs.

Observations

- Can (surprisingly) do exact sampling for some non-toy spatial models.
  - MCMC issues are avoided. Effective for moderate dimensional problems ($\approx$ 200 dimensions.)
  - Impractical for much higher dimensions.
- Approximation can be used to construct a uniformly ergodic Markov chain.
  - Almost like iid Monte Carlo: know when CLT holds, have consistent standard error estimates.
  - Practical for higher dimensions ($\approx$ 1000 dimensions.)

Observations (contd)

Heavy tailed approximation $\hat{\pi}$ is genuinely 'overdispersed' with respect to the target π (cf. Gelman and Rubin, 1992).

- Can be used for starting simulations of multiple chains on parallel machines if very worried about multimodality.

Criticisms:

- Approximation tries to match entire posterior distribution.
- Examples so far rely on a normal approximation; may work very poorly in many cases.
- However, the approximation scheme discussed is more general: other (better) versions are possible.

Currently under investigation: Algorithms that are fast mixing and automated but allow for more 'divide and conquer' style MCMC via parallel computing.

Summary

- Spatial generalized linear models are a flexible, useful class of models.
- Possible to construct a nearly automated algorithm using a combination of approximation and theoretical results:
  - Known mixing properties (uniformly ergodic).
  - Automatically generate starting values.
  - Consistent estimate of standard errors.
  - Rigorous stopping rule ('fixed width') directly connected to inferential goals. Simple idea: when desired accuracy is attained, stop the chain.
- Worth considering this approach for automation: construct fast mixing chain using approximation + use fixed width methodology.

References

- Haran, M. and Tierney, L. (2009) “Exact and approximate samplers with applications to spatial generalized linear models.”
- Flegal, J., Haran, M., and Jones, G.L. (2008) "Markov chain Monte Carlo: Can we trust the third significant figure?" *Statistical Science.* `www.stat.psu.edu/~mharan/batchmeans.R`
- Jones, G.L., Haran, M., Caffo, B.S. and Neath, R. (2006). “Fixed Width Output Analysis for Markov chain Monte Carlo,” *Journal of the American Statistical Association*
- Recta, V.L., Haran, M., Rosenberger, J.L. (2009). “A two-stage model for incidence and prevalence in point-level spatial count data.”
- Haran,M., Bhat, K.S., Molineros, J. and Dewolf E. (2009) “Estimating the risk of a crop epidemic from coincident spatiotemporal processes.” *J. of Ag. Biol. Envir.Statistics*