Some Experiments with Monte Carlo for Spatial Models

Murali Haran

Department of Statistics
Penn State University

(joint work with J.Flegal, G.Jones and L.Tierney)

Third Workshop on Monte Carlo Methods
Harvard University
May 2007

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 $D \in \mathbb{R}^d$ so as to generate the multivariate random process: $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$.
- Here we are only concerned with:
  - Geostatistical models: D is a fixed subset of $\mathbb{R}^d$. Location $\mathbf{s}$ vary continuously in space, but process is only observed at a finite set of locations. e.g. pollutant levels across Pennsylvania only observed at monitoring stations.
  - Areal/lattice models: D is a fixed collection of countably (usually finitely) many points 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.

Why focus on spatial models ?

- Spatial models are very widely used. Automated, reliable algorithms for even a few specific models will be very useful for people who want to fit these models routinely.
- Potential for exploiting structure of the model for efficient computing approaches.
- Relatively little theory on MCMC algorithms for these models.
- Strong dependence among variables can make posterior distributions challenging to simulate efficiently.
- Connections to other important models (variance component models.)

Basic spatial (linear) model

- Spatial process at location $\mathbf{s}$ is $Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s})$ 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.
- $\boldsymbol{\epsilon} = (\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n))^T$ is typically modeled as a zero mean Gaussian process (GP), for geostatistics, or Gaussian Markov random field (GMRF), for areal/lattice data.
- Gaussian Process: 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: Let Θ be the parameters for precision matrix. 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:

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

- For spatial linear model, once priors for Θ, β specified, inference is based on posterior $\pi(\Theta, \beta \mid \mathbf{Z})$.

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: Let $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$. Model $\mathbf{w}$ as either GP or GMRF.
- Stage 3: Priors for Θ, β.
- Inference based on $\pi(\Theta, \beta, \mathbf{w} \mid \mathbf{Z})$.

Monte Carlo for Posterior Inference

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

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

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

Well known problems:

- We do not know how long to run our Markov chain.
- Hard to know if CLT holds (often does not hold.)
- X_is are dependent so variance of estimator is hard to estimate $\Rightarrow$ hard to rigorously assess the accuracy of our estimates.

Question: Can we resolve these issues for at least some special cases of the spatial models discussed ?

Resolving MCMC Issues

- Theoretical approaches: convergence rates, upper bounds on distance to stationarity (cf. Rosenthal, 1995; Jones & Hobert, 2004.) Generally very difficult and often bounds can be loose.
- Usual MCMC diagnostics:
  - Easy to use, automated software exists (in `WINBUGS` for example.) Useful heuristics.
  - Not reliable, all are known to fail (cf. Cowles and Carlin, 1996).
- Standard error estimates:
  - Typically assume stationarity.
  - Not consistent.
  - Can overestimate (IMSE) or underestimate (usual batch means with fixed batch sizes).

Alternatives

1. Perfect or exact sampling:
   - Obtain *exact* draws from distribution using a Markov chain.
   - Make 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 (this is why we resort to standard Metropolis-Hastings in the first place).

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.

Case Study: A Spatial Generalized Linear Model

$Z(\mathbf{s}_i)|\mu(\mathbf{s}_i) \sim$ Poisson$(E(\mathbf{s}_i)e^{\mu(\mathbf{s}_i)}),\ i = 1, ..., N,$

$E(\mathbf{s}_i)$: estimate of expected events in region i.

$\mu(\mathbf{s}_i)$: 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 $\boldsymbol{\phi} = (\phi(\mathbf{s}_1), \ldots, \phi(\mathbf{s}_N))$.

Add priors for the precision parameters τ_h, τ_c.

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

An Approximation

- Start by finding a linear spatial model that is reasonably close to the true model.
  - Transform data (to $\mathbf{Y}$ say), use approximations (e.g. delta method) to get:

$$\mathbf{Y} \mid \boldsymbol{\theta}, \phi, \tau_h, \tau_c \sim N(\mu(\boldsymbol{\theta}, \phi), \Sigma(\tau_h, \tau_c))$$

  - Posterior for this model: $S(\boldsymbol{\theta}, \phi, \tau_h, \tau_c \mid Y)$. For convenience denote this by: $S(\boldsymbol{\theta}, \phi, \tau_h, \tau_c)$.
- Analytically integrate: $S_1(\tau_h, \tau_c) = \int S(\boldsymbol{\theta}, \phi, \tau_h, \tau_c) d\boldsymbol{\theta} d\phi$.
- From $S(\boldsymbol{\theta}, \phi, \tau_h, \tau_c)$, can obtain approximate conditional distribution of model parameters, $S_2(\boldsymbol{\theta}, \phi | \tau_h, \tau_c)$ (multivariate normal). Then, we have

$$S(\boldsymbol{\theta}, \phi, \tau_h, \tau_c) = S_1(\tau_h, \tau_c) S_2(\boldsymbol{\theta}, \phi | \tau_h, \tau_c).$$

A Heavy-Tailed Approximation

Construct heavy-tailed approximation $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\phi}, \tau_h, \tau_c)$:

- $S_1(\tau_h, \tau_c) S_2(\boldsymbol{\theta}, \boldsymbol{\phi} | \tau_h, \tau_c) \approx \pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \tau_h, \tau_c | Y)$.
- Find heavy-tailed approximation to $S_1(\tau_h, \tau_c)$, $\hat{\pi}_1(\tau_h, \tau_c)$.
- Find heavy-tailed (multi-t) approximation to $S_2(\boldsymbol{\theta}, \boldsymbol{\phi} | \tau_h, \tau_c)$, $\hat{\pi}_2(\boldsymbol{\theta}, \boldsymbol{\phi} | \tau_h, \tau_c)$. Easy: it should have same mean and variance as the multivariate normal $S_2(\boldsymbol{\theta}, \boldsymbol{\phi} | \tau_h, \tau_c)$.
- Simple sequential sampling to generate proposal from $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\phi}, \tau_h, \tau_c)$.

Heavy-Tailed Approximation (contd.)

- Sample from $\hat{\pi}$:
  1. Sample $(\tau_h, \tau_c) \sim \hat{\pi}_1(\tau_h, \tau_c)$. Easy to do for simple bivariate distribution.
  2. Sample $(\boldsymbol{\theta}, \boldsymbol{\phi}) \sim \hat{\pi}_2(\boldsymbol{\theta}, \boldsymbol{\phi} | \tau_h, \tau_c)$. Multivariate-t distribution with precision $Q(\tau_h, \tau_c)$ using Step 1.
- $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\phi}, \tau_h, \tau_c)$: proposal for Monte Carlo algorithms.
- Step 1 of proposal generation is easy (fast). Step 2: Need to be efficient since matrix operations are involved when generating proposal, evaluating Met-Hastings ratio.

Two Monte Carlo Approaches

Let $\Psi = (\boldsymbol{\theta}, \phi, \tau_h, \tau_c)$. Can show that:

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

1. Numerically maximize $\frac{\pi(\Psi)}{\hat{\pi}(\Psi)}$ to obtain $K < \infty$.
   - (a) Rejection sampling or (b) perfect tempering (Møller and Nicholls, 2006): 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, $M < \infty$, $t \in (0, 1)$, $\|\cdot\|_{TV}$ is 'total variational distance.'

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.

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 and Neath, 2006.)
  - Stopping rule (**'fixed width' approach**): When estimated standard error is below a desired level, stop the sampler (Jones et al., 2006.)
  - Not quite introductory statistics, but just as easy in practice.
- Do not need bounding constant K.

An Example

- Minnesota cancer data: 87 regions (counties), 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	2408	96 min.
independence chain	10,944	$<$4 min.

Perfect tempering performance $\approx$ rejection sampling.

An Example (contd)

- All three samplers: reasonable estimates, similar inference.
- Exact samplers are much less efficient than independence MCMC. No surprise (cf. Liu, 1996)
- In fact, for larger data set example (1046 dimensions):
  - Exact sampling is not feasible: acceptance rate ≈ 1 per 20,000 samples generated; sample generation also more expensive.
  - Independence chain still works well.
- 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.

Summary

- Started with an overall framework for some important spatial models.
- Described a method for constructing a *heavy tailed* approximation to the posterior distribution of such models.
- Two approaches are available using above approximation: exact/perfect and independence Metropolis-Hastings.
- Both approaches provide simple recipes for determining stopping rules for MCMC simulation: stop the chain as soon as a desired standard error has been achieved.
- Computational efficiency is crucial. Usual matrix operations are $O(N^3)$ so use sparse matrix algorithms (Rue, 2001).

Summary (contd.)

- Can do perfect/exact sampling for some spatial models.
  - MCMC issues are avoided. Effective when approximations are accurate and for moderate dimensional problems ($\approx$ 200 dimensions.)
  - For higher dimensions (say around 1000) and when approximation is a poor match, exact algorithms are impractical.
- Approximation can be used to construct a Markov chain with good mixing properties:
  - Estimate standard errors via consistent batch means.
  - Almost back to iid Monte Carlo scenario: understand when CLT holds, consistent standard error estimates, feel greater confidence in our results.

References

- Flegal, J., Haran, M., and Jones, G.L. "Markov chain Monte Carlo: Can we trust the third significant figure?"
- Haran, M. and Tierney, L.“Exact and approximate samplers for a Markov random field model.”
- 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*
- Galin Jones’s talk on Fixed Width MCMC.