Towards Automated MCMC Algorithms for a Class of Gaussian Random Field Models

Murali Haran

Department of Statistics
Penn State University

Collaborators: J.M.Flegal (U.C.Riverside), G.L.Jones (U.Minnesota), M.M.Tibbits, J.L.Liechty (Penn State), L.Tierney (U.Iowa).

Universitat de Vàlencia

May 2010

Introduction (some virtues of sample-based inference)

- Monte Carlo-based approaches allow us to fit models we want, rather than models selected for computational convenience.
- Can obtain estimates of full joint distributions.
- In principle, can obtain estimates up to arbitrary levels of precision. That is, as the Monte Carlo sample size gets large, the standard error reduces to 0.
- Sample-based inference is an extremely useful tool for propogating uncertainty appropriately. Increasingly important to have results from one model be used as input to another model.

(analytical approximations cannot offer all of the above)

MCMC as a tool for statistical inference

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

Obvious applications: estimating expectations, probabilities with respect to an arbitrary probability distribution π.

Less obvious: estimating ratios of normalizing functions; MCMC maximum likelihood.

MCMC solution: Construct a Harris-ergodic Markov chain $X_1, X_2, \ldots$ with stationary distribution π so that if $E_\pi|g(x)| < \infty$, we have an S.L.L.N. for Markov chains:

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

Very flexible approach for fairly high-dimensional, complicated problems.

MCMC issues

Careful users of MCMC face several time consuming and challenging issues:

1. Devising/tuning the Metropolis-Hastings algorithm.
2. How are the the starting values affecting estimates? (what starting values to use.)
3. How long to run the Markov chain. That is, when is n large enough?
4. (Related to all above) When is $\bar{g}_n$ a good estimate of $E_\pi g$?
5. Theoretical basis for any of the above?

These are non-issues for i.i.d. Monte Carlo.

Options

Exact sampling (circumvent some of the problems!):

- *Perfect* draws using a Markov chain (Propp-Wilson, 1996).
- Make classical (old fashioned) Monte Carlo methods such as rejection sampling practical.

This is very hard to achieve for challenging models. And even when achievable, far less efficient than corresponding MCMC sampler. Instead, **fixed-width MCMC**:

- Construct Metropolis-Hastings so Markov chain sampler mixes well (the algorithm gives accurate estimates quickly.)
- If provably fast mixing (e.g. uniformly or geometrically ergodic), rigorous approach for estimating MCMC standard errors and stopping rules.

Fixed-width MCMC: basic idea

- Stop the sampler when standard errors of desired quantities are below a desired level. That is, simulate until confidence interval for $E_\pi g$,

$$[\bar{g}_n - c_n, \bar{g}_n + c_n]$$

  is sufficiently narrow.

Want good coverage: both theoretical justification and empirical evidence that this works well in practice.

Fixed-width MCMC theory: Markov chain CLT

Suppose at least one of the following conditions holds:

- Chain is uniformly ergodic and $E_\pi g^2 < \infty$
- Chain is geometrically ergodic and $E_\pi |g|^{2+\epsilon} < \infty$ for $\epsilon > 0$.

Then, for any initial distribution, as $n \to \infty$

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

where

$$\sigma_g^2 = \text{Var}[g(X_0)] + 2\sum_{i=1}^{\infty} \text{Cov}[g(X_0), g(X_i)]$$

Uniform, geometric ergodicity are convergence rates in total variational distance. There are many other sets of sufficient conditions (cf. Roberts and Rosenthal, 2004; Jones, 2004)

Fixed-width MCMC theory [cont'd]

Let $\hat{\sigma}_g^2$ be the *consistent batch means estimate* of MCMC standard error (with batch sizes of $\sqrt{n}$).

- Theorem: If the Markov chain is geometrically ergodic and $E_\pi g^{2+\epsilon} < \infty$ then $\hat{\sigma}_g^2 \to \sigma_g^2$ with probability 1 as $n \to \infty$.
  Note: there are other consistent estimators of MCMC standard errors, including regeneration-based estimators, consistent batch means with other batch sizes.

- Theoretical basis for fixed-width MCMC: The usual asymptotic 100(1-α)% confidence interval for $E_\pi g$ using $\hat{\sigma}_g^2$ (using appropriate critical value) is asymptotically valid.

(Jones, Haran, Caffo, Neath, 2006)

Fixed-width MCMC theory [cont'd]

- Sufficient conditions for fixed-width MCMC to be theoretically sound:
  - The Markov chain needs to mix well.
  - A consistent estimator of the asymptotic variance.
- No assumption of stationarity.

How to use fixed width MCMC

- Decide which estimates are of importance, for e.g. simple expectation or a tail probability.
- Run sampler until a desired level of accuracy (standard error) is attained for the estimate, simple, fast `R` code `http://www.stat.psu.edu/~mharan/batchmeans.R`
- Report resulting estimate (along with estimated error).

Possible issues (things fixed-width cannot protect you from):

- Multiple modes. But *no approach* can protect you from this! Try different starting values.
- Very slow mixing sampler. Work on improving sampler. Again, all other approaches for determining stopping can do no better. Fixed-width still works well in many such cases (Flegal, Haran, Jones, 2008).

Fixed-width MCMC in practice

Even when we have not established mixing rates, automation, fixed width approach can be very useful:

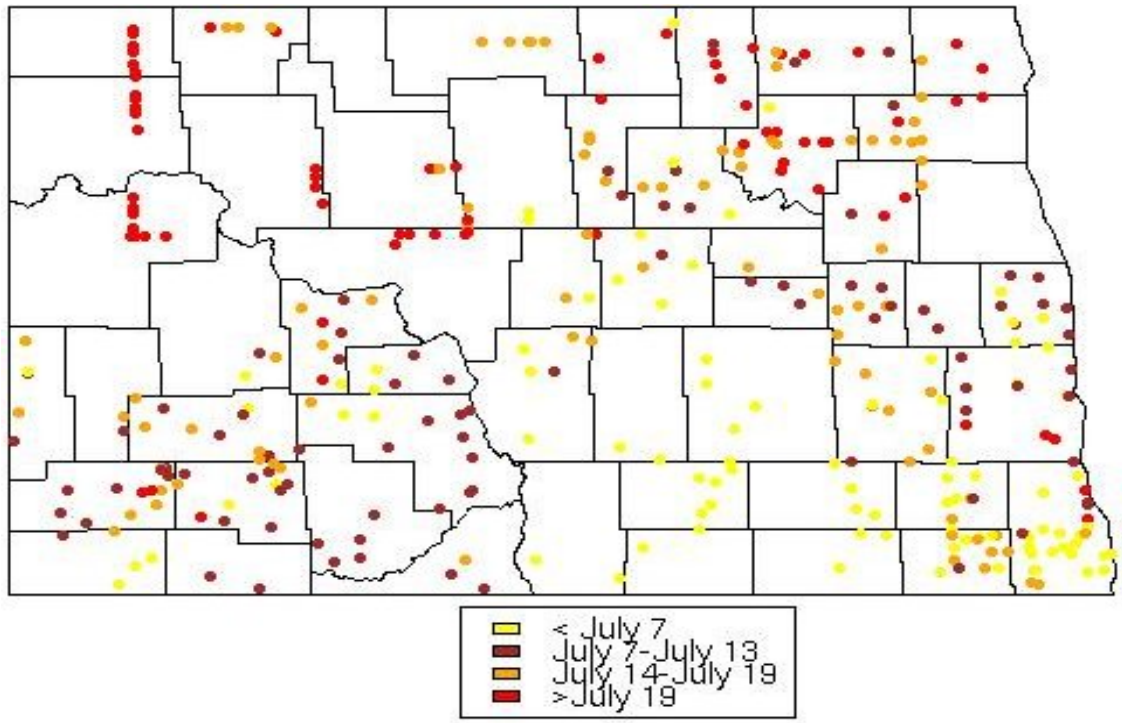
- Can use it to automate stopping for simulation studies — e.g. comparing posterior quantities for a model over randomly generated data sets.
  - MCMC algorithm mixes differently for each new data set. Bad idea to report estimates with different accuracy for each simulation!
- Automatically runs longer when Markov chain mixes poorly, or expectation is challenging (e.g. tail probabilities). Shorter if less accuracy desired, or if chain is fast mixing.
- More reliable than some other convergence diagnostics that seem mainly concerned with starting value issues/bias (Flegal, Haran, Jones, 2008).

Half-time summary

- So far: have discussed a general approach for determining when to stop an MCMC algorithm. Some theoretical background along with ideas for how **fixed width MCMC** may be used in practice.
- Now: how to construct algorithms that mix well in the context of Gaussian random field models.
- Why focus on Gaussian random field models?
  - Very useful class of models for spatial data, both lattice and continuous-domain.
  - Very useful class of models for modeling and inference for complex computer experiments (emulation + calibration.)
  - Useful for modeling non-Gaussian spatial data (binary, count data etc.) and for classification (machine learning.)

Continuous-domain spatial data (“geostatistics”)

Wheat flowering dates in North Dakota (for studying blight epidemics).

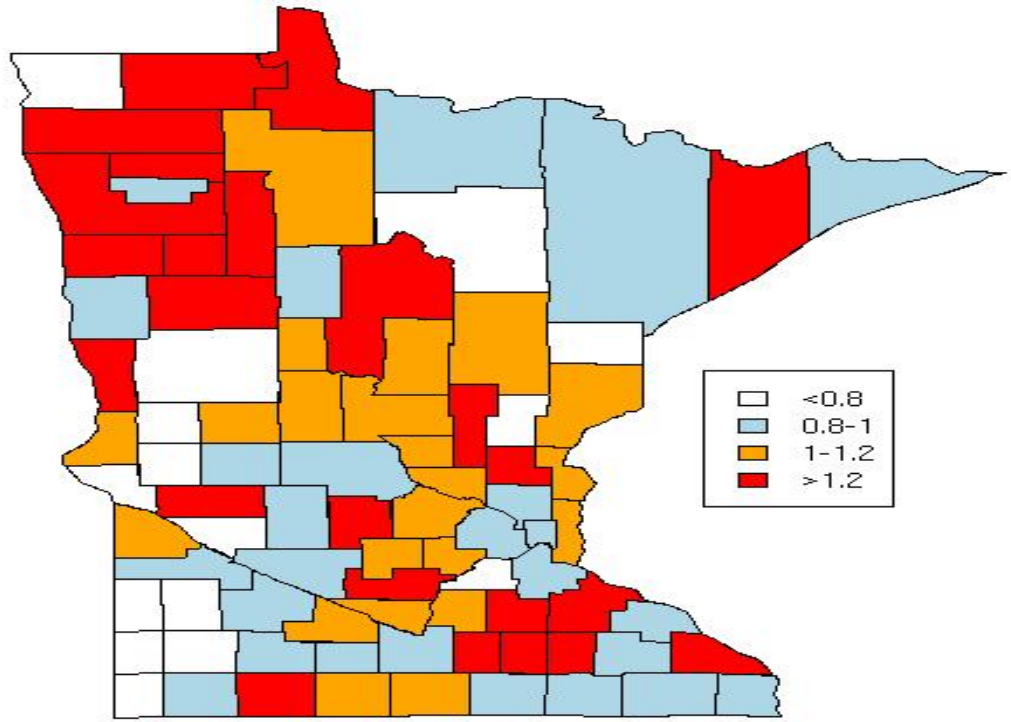


Courtesy Plant Pathology, Penn State and North Dakota State.

(Haran, Bhat, Molineros, DeWolf, 2008)

Lattice or “areal” data

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



Courtesy MN Cancer Surveillance System, Dept. of Health

(Haran, Hodges, Carlin, 2003)

Other examples: images, pixel values from remote sensing.

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 $\Sigma(\Theta)$. Let $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$. 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:

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

- 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.

Efficient MCMC for spatial linear models

- Closed form for low-dimensional (usually 2-8) marginal posterior, $\pi(\Theta, \beta \mid \mathbf{Z})$.
- Slice samplers (Agarwal, Gelfand 2005; Yan et al., 2007), multivariate slice samplers (Tibbits, Haran, Liechty, 2010) resulting in a fast mixing Markov chains.
- Fast mixing automated algorithms above can be used in conjunction with fixed-width approaches.

Spatial generalized linear model

What if data generating mechanism is non-Gaussian (Diggle et al., 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 $g(E(Z(\mathbf{s}_i))) = \eta(\mathbf{s}_i) = X(\mathbf{s}_i)\beta + w(\mathbf{s}_i)$, 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 $\pi(\Theta, \beta, \mathbf{w} \mid \mathbf{Z})$.

Efficient MCMC for spatial generalized linear models

Computing for SGLMs is more challenging:

- Higher dimensional posterior, $\pi(\Theta, \boldsymbol{\beta}, \mathbf{w} \mid \mathbf{Z})$.
- Strong dependence among components (e.g. spatial random effects) results in slow mixing chains when using univariate update MCMC approaches. Univariate updates are also more computationally expensive than block updates when matrices are dense.
- Two routes for constructing MCMC updates:
  - Approximate SGLM by linear spatial model.
  - Langevin-Hastings (Roberts and Tweedie, 1996)

Linearization of an SGLM

- Target posterior of SGLM is $\pi(\Theta, \boldsymbol{\beta}, \mathbf{w} \mid \mathbf{Z})$.
- Approximate the SGLM by a linear spatial model:
  - Transform data $\mathbf{Z}$ to $\mathbf{Y}$ and use approximation of form:

$$\mathbf{Y} \mid \Theta, \boldsymbol{\beta}, \mathbf{w} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{w}, C).$$

  - As before $\mathbf{w} \mid \Theta \sim N(0, \Sigma(\Theta))$. Denote posterior for this approximate model by $S(\Theta, \boldsymbol{\beta}, \mathbf{w})$.
- Analytically integrate: $S_1(\Theta, \boldsymbol{\beta}) = \int S(\Theta, \boldsymbol{\beta}, \mathbf{w}) d\mathbf{w}$.
- From $S(\Theta, \boldsymbol{\beta}, \mathbf{w})$, can obtain approximate conditional distribution of spatial random effects, $S_2(\mathbf{w} \mid \Theta, \boldsymbol{\beta})$ (multivariate normal). Then, we have

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

[Haran (2003), Haran and Tierney (2010)]

Approximation

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

- We have: $S_1(\Theta, \beta) S_2(\mathbf{w} \mid \Theta, \beta) \approx \pi(\Theta, \beta, \mathbf{w} | Y)$.
- 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: multivariate-t with same mean and variance as the multivariate normal $S_2(\mathbf{w}|\Theta, \beta)$.
- $\hat{\pi}(\Theta, \beta, \mathbf{w})$: proposal for MCMC.
- The resulting independence-Metropolis-Hastings algorithm is uniformly ergodic (Haran and Tierney, 2010).
  - Generate starting values from proposal (heavy-tailed, over-dispersed!)
  - Stop automatically using fixed-width.
  - Fast mixing, easily parallelized.

[Haran and Tierney (2010); Haran (2010)]

Example: disease mapping model

Besag, York, Mollie (1991) model for count data on a lattice.

- Derive an approximation $\hat{\pi}$ to target π using 'linearization'.
- Generate starting values from $\hat{\pi}$: genuinely overdispersed with respect to π (cf. Gelman and Rubin, 1992).
- Construct a Metropolis-Hastings independence sampler: propose every M-H update from $\hat{\pi}$.
  - The resulting sampler is provably fast mixing: it is *uniformly ergodic* (Haran and Tierney, 2010).
  - The sampler is easily parallelized ('embarrassingly parallel').
  - Since GMRF, can use sparse matrix algorithms for fast computing (Rue, 2001).

Data examples

- Minnesota cancer data sets: 176 parameters. Southeast U.S. infant mortality: 910 parameters.
- Note: can also use rejection sampler/E-sup rejection sampler (Caffo et al., 2001) with same proposal.
- Stop algorithms when Monte Carlo standard errors are below same threshold for parameters.

data set	sample size		time taken	
	rejection	I-MH	rejection	I-MH
breast cancer	4,118	29,241	2,663s	183s
colo-rectal cancer	4,735	27,225	543s	170s
infant mortality	—	97,721	—	10,066s

Disease mapping: summary of results

- Both samplers: good estimates, similar inference. (surprising that rejection sampler works in some cases.)
- I-MH is vastly superior to rejection sampler. For large data set (910-dimensional posterior): I-MH is still practical while rejection sampler is not. Timing for I-MH can be reduced linearly according to number of processors available.
- Almost like iid Monte Carlo: know that CLT holds (generally not true for MCMC), have consistent standard error estimates, easy to determine starting values, stopping rule.

Criticism: Approximation tries to match entire posterior distribution using a multivariate normal approximation for one stage in the hierarchical model. May work poorly in some examples. (E.g. on next slide....)

Efficient MCMC using local approximations

- A two-stage SGLM model for zero-inflated data on insect populations (Recta, Haran, Rosenberger, 2009).
- Previous approximation ideas do not work well here.
- Better off exploring *local* normal approximations instead.
- Langevin-Hastings MCMC: construct local approximations using gradients at current state of Markov chain (Roberts and Tweedie, 1996; Christensen, Roberts, Sköld, 2006).
- The resulting Markov chain is geometrically ergodic under certain conditions (Christensen, Møller, Waagepetersen, 2001). Obtain rigorous MCMC errors + fixed-width stopping rule.
- This works well in practice for our SGLM.

Summary

- Fixed-width MCMC is a very useful general approach for automatically determining MCMC run lengths based on sound inferential principles (desired level of accuracy.)
- Gaussian random field models are very useful and important, but MCMC for these models is far from routine.
- For some Gaussian random field models, it is possible to construct provably fast mixing samplers and use fixed-width approaches in a theoretically justified manner.
- MCMC is still far from automatic or easy for all these models, especially as data sets get large. Use sparsity + parallel computing. Work in progress...

Select references

- Christensen, Møller, Waagepetersen (2001), “Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalized linear mixed models,” *Methodology and Computing in Applied Prob.*
- Haran, M. and Tierney, L. (2010) “On automating Markov chain Monte Carlo for a class of spatial models.”
- Haran, M. (2009) “Gaussian random field models for spatial data.” *Handbook of Markov chain Monte Carlo* (to appear).
- Tibbits, M.M., Haran, M., Liechty, J.L. (2009) “Parallel multivariate slice sampling,” (*Stat and Computing*, in press).
- Flegal, J., Haran, M., and Jones, G.L. (2008) "Markov chain Monte Carlo: Can we trust the third significant figure?" *Stat. Sci.*, `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,” *JASA*.
- Recta, V.L., Haran, M., Rosenberger, J.L. (2010) “A two-stage model for incidence and prevalence in point-level spatial count data.”