Monte Carlo Methods, Spatial Processes and Complex Computer Models

with Applications to Environmental Science

Murali Haran

Department of Statistics
Penn State University

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

Penn State Statistics SAC talk, April 2010

What do I work on?

- Modeling and statistical computing:
  - Monte Carlo methods/Markov chain Monte Carlo (MCMC) methods
  - Models for spatial data
  - Complex computer models
- Lots of scientific problems (“interdisciplinary research”) that use all of the above, *and* motivate new models and algorithms:
  - Climate science
  - Disease modeling
  - Geography/Ecology

Interdisciplinary research

I work with Penn State scientists on a variety of problems:

- Climate science: using computer models and data to learn about climate change and predicting future climate.
  Complex computer models; Gaussian processes; MCMC.
- Disease modeling: (i) estimating risks of crop epidemics. (ii) studying the dynamics of measles transmission.
  Space-time models; MCMC; structured covariance models.
- Geography: On studying how ecological factors can impact coastal aquatic ecosystems.
  Non-Gaussian spatial models; Markov random fields; Monte Carlo, pseudo-likelihood.

“The best thing about being a statistician is that you get to play in everyone’s backyard.” — John Tukey.

Outline

I will briefly discuss some recurring themes in my research:

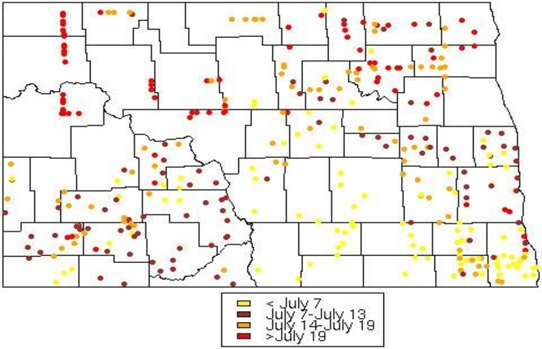
1. Gaussian random field models for Gaussian and non-Gaussian spatial processes
2. Complex computer models
3. Monte Carlo and Markov chain Monte Carlo

I. Gaussian random fields

- Gaussian random field are very popular models: useful for spatial models, models for complex computer experiments, machine learning.
- Let $\mathbf{s}$ vary over index set $D \subset \mathbb{R}^d$ so the associated spatial process is $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$.
- Useful for
  - Geostatistics: D is a fixed subset of $\mathbb{R}^d$. Process is infinite-dimensional (locations vary continuously in space).
  - Areal/lattice data: 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.

I. Geostatistics

Wheat flowering dates in North Dakota: used in estimating the risk of a crop epidemic.

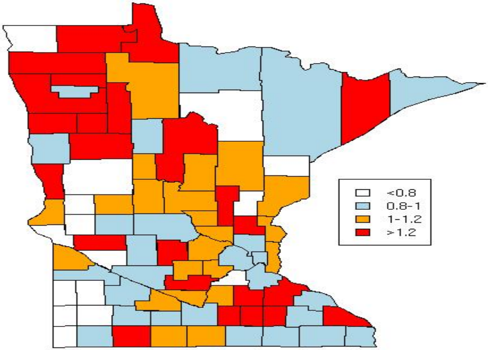


Courtesy Plant Pathology, PSU and North Dakota State.

(Haran, Bhat, Molineros, DeWolf, 2008)

I. Areal data

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



Courtesy MN Cancer Surveillance System, Dept. of Health

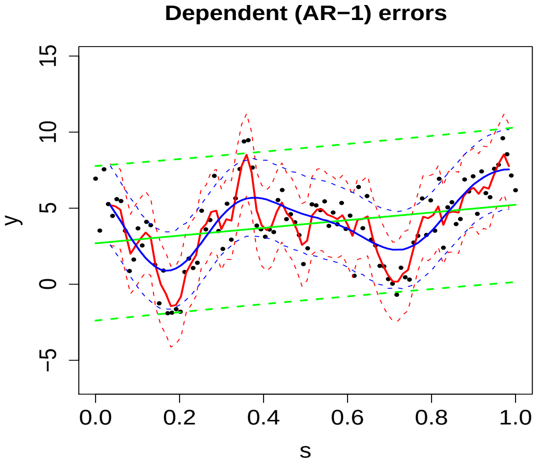
(Haran, Hodges, Carlin, 2003)

I. Basic Gaussian random field (linear) model

- Spatial process at location **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 **s** and β is a vector of coefficients.
- Model dependence via $w(\mathbf{s})$'s.
- For locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$ can be modeled via a 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 $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$.

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

I. GP model for dependence: toy 1-D example



Black: 1-D AR-1 process simulation. Green: independent error.

Red: GP with exponential, Blue: GP with gaussian covariance.

I. Spatial linear model (contd.)

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

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

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

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

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

I. Spatial generalized linear model

If data generating mechanism is non-Gaussian:

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

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

  where $g(E(Z(\mathbf{s}_i))) = \eta(\mathbf{s}_i) = X(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i), \eta$ 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 $\Theta, \boldsymbol{\beta}$.
- Inference based on $\pi(\Theta, \boldsymbol{\beta}, \mathbf{w} \mid \mathbf{Z})$.

I. Summary

So far:

- Described how Gaussian random fields are flexible models for many kinds of spatial data: Gaussian and non-Gaussian, on continuous space or on a lattice.
- Inference for these models is based on (potentially complicated) posterior distributions.

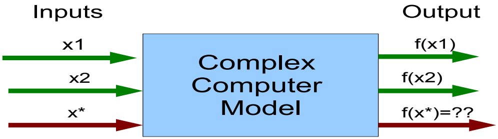
Next:

- Complex computer models. Gaussian random fields are useful here as well.

II. Complex computer models

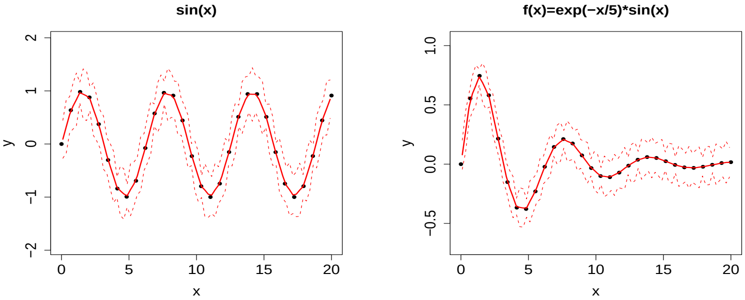
- Complex computer models are often used to mimic reality. They can be deterministic or stochastic.
- Examples:
  - Climate system models. e.g. (deterministic) General Circulation Models (GCMs).
  - Complicated disease dynamics models are used to model the spread of infectious diseases. e.g. (stochastic) gravity models.
- Given inputs (θ), the model will run and produce output.
- These models are used for (a) understanding the mechanism behind the system (model fit + estimate of 'true' θ, (b) making predictions.

II. Complex computer models



- Very complicated, no closed form.
- Computationally expensive to run at each input value.

II. Stochastic model for emulation: toy examples



Black dots: computer model output. Red: interpolation using Gaussian process.

Nice fit for both functions using *same* simple model:

$f(x) = c + w(x)$ where $\{\epsilon(x),\ x \in (0, 20)\}$ is a GP.

Also provides a probability model connecting parameter to model output.

II. ‘Computer likelihoods’

- Suppose we have observations (real data) corresponding to the physical process, say **Z**.
- We want to infer the value of ‘input’ θ based on **Z** but no explicit likelihood, i.e., $\mathcal{L}(Z \mid \theta)$ is not available.
- Instead, we *estimate* the likelihood $\hat{\mathcal{L}}(Z \mid \theta)$ by fitting a Gaussian process to the computer model output (**Y**).
- Bayesian inference is convenient here since we need to model additional sources of error and incorporate prior knowledge about θ. Inference based on:

$$\pi(\theta, \xi \mid \mathbf{Z}, \mathbf{Y}) \propto \hat{\mathcal{L}}(\mathbf{Z} \mid \theta, \xi) p(\xi) p(\theta)$$

where ξ= all other parameters (bias, dependence, variance) in the model; $p(\xi), p(\theta)$ are priors.

II. Challenges

- Specifying a flexible model when the output is high dimensional or functional, e.g. time series or spatial.
- Computing becomes intractable when data sets get large (likelihood evaluations are too expensive). One solution is to use *kernel convolutions* instead of simple Gaussian processes: (Bhat, Haran, Tonkonojenkov, Keller, 2009).
- MCMC algorithms for sampling from posterior distribution can be difficult to construct.

II. Summary

So far:

- **Modeling**: described how Gaussian random fields are useful for spatial processes and complex computer models.
- Inference based on these models involves complicated distributions.

Next:

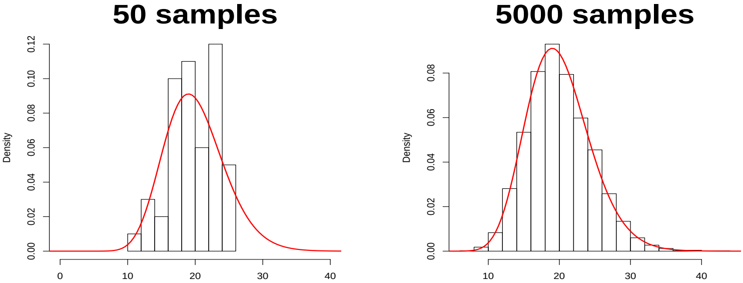
- **Statistical computing**: Monte Carlo/MCMC methods for carrying out statistical inference.

III. Monte Carlo-based inference (review!)

Idea: if we can sample from π, can perform statistical inference.

Simple e.g. X $\sim \pi$ where π is Gamma($\alpha = 20, \beta = 1$).

Simulate: $X_1, \ldots, X_N \overset{iid}{\sim}$Gamma(20,1). Red=true density.



50 samples

5000 samples

N=5000: estimates $E_\pi X$=20.05199 and Pr($X > 22$)=0.3102.

Truth: $E_\pi X = 20$ and Pr$(X > 22) = 0.306027$.

III. Markov chain Monte Carlo (MCMC) solution

In general, may be very difficult to sample from π.

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

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

Several difficult issues:

- Need to construct an efficient algorithm (not automatic!)
- Starting values?
- How long to run the Markov chain?
- Accuracy of the estimator is hard to estimate.

III. Automation of MCMC

Ideally (the dream):

1. Automated approach for constructing algorithm. No tuning necessary.
2. Generate appropriate starting values automatically.
3. Have a rigorous criteria for determining when to stop the chain that is *related to inferential goals*. E.g. How accurate do you want your estimates to be?
4. Some theoretical guarantees regarding all of the above.

This has been the subject of much of my research, past and current.

III. Efficient MCMC for spatial linear models

- Closed form for low-dimensional (usually 2-8) marginal posterior, $\pi(\Theta, \boldsymbol{\beta} \mid \mathbf{Z})$. Slice samplers (Agarwal, Gelfand 2005; Yan et al., 2007) involve univariate updates.
- To improve mixing, can use block updates: multivariate slice sampling (Tibbits, Haran, Liechty, 2010) resulting in a faster mixing algorithm. Since search for proposals at each step of the algorithm is very expensive, this is done in parallel on a graphics processing unit (GPU).
- Aside: matrix operations at each iteration are expensive for large data sets. Need to take advantage of some form of sparsity (banded matrices, tapering, reduced-rank approaches).

III. Efficient MCMC for spatial generalized linear models

Computing for SGLMs is more challenging:

- Higher dimensional posterior, $\pi(\Theta, \boldsymbol{\beta}, \mathbf{w} \mid \mathbf{Z})$.
- 'Block sampling' involves updating multiple components at the same time. This can solve multiple issues:
  - Strong dependence among components (e.g. spatial random effects) results in slow mixing chains: Markov chain mixes better with blocking.
  - Block updates can also be computationally more efficient per iteration.

III. Linearization of an SGLM

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

- We have: $S_1(\Theta, \boldsymbol{\beta}) S_2(\mathbf{w} \mid \Theta, \boldsymbol{\beta}) \approx \pi(\Theta, \boldsymbol{\beta}, \mathbf{w} | Y)$.
- Find heavy-tailed approximation to $S_1(\Theta, \boldsymbol{\beta})$: $\hat{\pi}_1(\Theta, \boldsymbol{\beta})$.
- Find heavy-tailed (multi-t) approximation to $S_2(\mathbf{w}|\Theta, \boldsymbol{\beta})$, $\hat{\pi}_2(\mathbf{w}|\Theta, \boldsymbol{\beta})$. Easy: multivariate-t with same mean and variance as the multivariate normal $S_2(\mathbf{w}|\Theta, \boldsymbol{\beta})$.

[details in Haran and Tierney (2009); Haran (2009)]

- We can now use $\hat{\pi}$ as a proposal for Monte Carlo algorithms to sample from π.

III. Example: disease mapping

Besag, York, Mollie (1991) model: spatial generalized linear mixed models for count data.

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

Recap:

- Our goal is to construct an automated sampler for Gaussian random field models.
- We have a general approach for constructing a heavy-tailed approximation to $\pi(\boldsymbol{\theta}, \phi, \tau_h, \tau_c | \mathbf{Z})$.

III. Automated MCMC for disease mapping example

- Construct an independence Metropolis-Hastings (I-MH) sampler using $\hat{\pi}$ (cf. Tierney, 1994):
  - **Definition**: Let $P^n(x, A)$ be the n-step transition kernel of the Markov chain, and π be its stationary distribution. If $\|P^n(x, \cdot) - \pi(\cdot)\| \leq C(x)t^n$ where $t \in (0, 1)$ and $C : X \to [0, \infty]$, with $C(x)$ bounded, chain is *uniformly ergodic.*

    Theorem: The I-MH Markov chain is uniformly ergodic (Haran and Tierney, 2009).
- Starting values from $\hat{\pi}$ + no tuning + easily parallelized + can use sparse matrix algorithms for fast computing.

III. Stopping rules, estimating standard errors

Since the Markov chain is fast mixing (uniformly ergodic), can obtain rigorous estimates of standard errors for expectations based on MCMC runs as well as rigorous stopping rules for MCMC.

- CLT holds quite generally for such samplers.
- Consistent batch means (Jones, Haran, Caffo, Neath, 2006) provides a *consistent* estimate of the Monte Carlo standard error. Simple stopping rule (**'fixed width' approach**): When estimated standard error is below a desired level, stop the sampler. Works well in practice (Flegal, Haran, Jones, 2008; Jones et al., 2006).

III. Data examples

- Minnesota cancer data sets: 176 parameters. Southeast U.S. infant mortality: 910 parameters.
- Stop algorithms when Monte Carlo errors $<$threshold.

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

- For these models: automatic, efficient algorithm (I-MH) with fixed-width stopping rule.
- For other models: can always use fixed-width stopping rule.

Summary

- Spatial linear and generalized linear models are a flexible, useful class of models for a wide variety of problems.
- Need to develop flexible models for new problems and scientific questions.
- Lots of computational challenges. Solutions may involve:
  - Improving/studying MCMC algorithms: new more automatic algorithms, theoretical properties of such algorithms (Markov chain theory).
  - Taking advantage of modern parallel computing.
  - Developing new models that allow for fast computing. (spatial process theory; matrix identities.)
- Scientific collaborations motivate interesting statistical research.
- Lots of open modeling and computation problems.

Ongoing projects

Some ongoing/recently completed projects:

- Inference for complex computer models used to study climate change: joint work with Sham Bhat, Roman Tonkonojenkov, Klaus Keller.
- Models for zero-inflated spatial data used for modeling insect populations: joint work with Jean Recta, Jim Rosenberger.
- Parallel automated MCMC algorithms: joint work with Matt Tibbits, John Liechty.
- Models and computing for spatial binary, non-Gaussian data: joint work with John Hughes.
- Inference for computationally expensive disease dynamics models: joint work with Roman Jandarov, Ottar Bjornstad.

Where to look for more information

`http://www.stat.psu.edu/~mharan`