Inference in the Presence of Intractable Normalizing Functions

(Joint work with Jaewoo Park)

Department of Statistics, North Carolina State University

March 2018

Murali Haran

Department of Statistics, Penn State University

Outline

Outline

Research Areas

- I. Statistical computing
- II. Climate science
- III. Infectious disease modeling
- IV. Spatial models

Lots of overlap

Example: Talk today is on a computing problem motivated by a spatial model for infectious disease

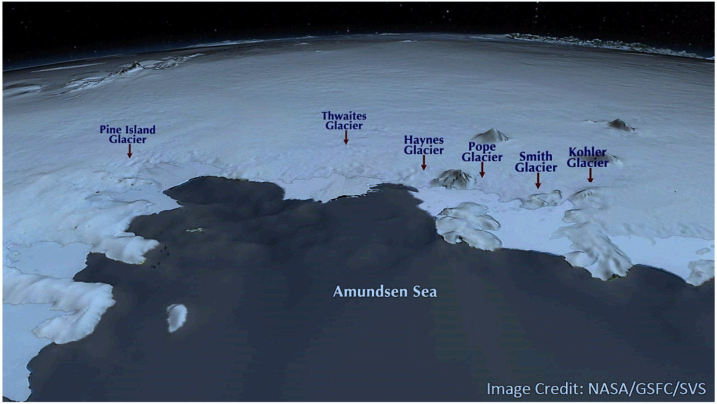
I. Statistical Computing

- Efficient Markov chain Monte Carlo (MCMC) and perfect sampling
- Stopping rules for MCMC algorithms: practical approach with theoretical justification
- Parallelizing MCMC
- Inference for climate and infectious disease models
  - Intractable likelihood functions
  - High-dimensional random effects
- Doubly intractable distributions (today's talk)

II. Climate Science

What is the future of the West Antarctic ice sheet?

(Subject of NC State talk in October 2017)



II. Climate Science: Combining Physics and Data

- Projecting future climate involves sophisticated physical models of climate systems
- Combining physics with observations
- Hierarchical modeling, dimension-reduction approaches, spatial models
- Inference for complex computer models
  - Emulation (approximation) and calibration (parameter inference) for these models using Gaussian processes
  - Challenges
    - high-dimensional spatial observations and model output
    - non-Gaussian spatial data
    - modeling data-model discrepancy

III. Infectious Disease Modeling

UNICEF India (2016)

III. Infectious Disease Research Questions

- Impact of vaccination strategies on infectious diseases
- How do the seasons affect meningitis transmission, and what does this suggest regarding vaccination strategies?
- (New NIH grant) What leads to vaccine refusal? Potential impacts?
- Methods
  - Hierarchical models for dynamics in space and time
  - Susceptible-Infected-Recovered (SIR)-type compartmental models for disease transmission.

Interdisciplinary Setting

- Involves grads, postdoc, faculty from both disciplines
- Cross-disciplinary grant support. Examples:
  - NSF Computational Data-Enabled Science & Engr
  - NSF Sustainable Climate Risk Management
  - Department of Energy (DOE)
  - NIH MIDAS (Models of Infectious Disease Agent Study)
  - Gates Foundation
- Papers published in
  - scientific journals, e.g. *Vaccine, Geoscientific Model Development, Nature Climate Change, Geophysical Research Letters, J of Climate*
  - statistics journals, e.g. *Annals of Applied Stats, JASA, Environmetrics, J of Computational and Graphical Statistics*

IV. Spatial Models

- Efficient methods for high-dimensional non-Gaussian spatial data
  (Subject of NC State talk in November 2017)
  - Latent Gaussian Markov random fields (areal data) and Gaussian process models (continuous-domain spatial data)
- Zero-inflated latent Gaussian hurdle model for modeling spatial data on beetle counts
- Markov attraction-repulsion point process spatial model for respiratory syncitial virus (RSV) infections
- Spatial point process models with dynamics for ant movement data

Outline

Summary

- Doubly intractable distributions pose serious computational challenges
- I will describe a framework for understanding algorithms for such problems
- I will discuss issues and challenges
- I will describe a new algorithm that is computationally expedient for some doubly intractable distributions

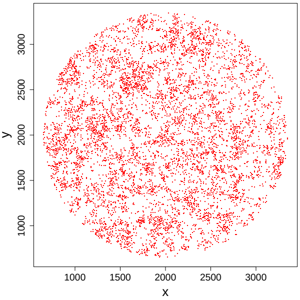
1. Framework, theory, algorithms in Park and Haran (2018) “Bayesian Inference in the Presence of Intractable Normalizing Functions” *Journal of the American Statistical Association*, to appear
2. New algorithm: manuscript in preparation

Models with Intractable Normalizing Functions

- Models with intractable normalizing functions
  - Data: $\mathbf{x} \in \chi$, parameter: $\theta \in \Theta$
  - Model: $h(\mathbf{x}|\theta)/Z(\theta)$, where $Z(\theta) = \int_{\chi} h(\mathbf{x}|\theta)d\mathbf{x}$ is intractable
- Popular examples
  - Social network models: exponential random graph models (Robins et al., 2002; Hunter et al., 2008)
  - Models for lattice data (Besag, 1972, 1974)
  - Spatial point process models: interaction models (Strauss, 1975, Goldstein, Haran et al., 2015)
- Challenge: likelihood-based inference with $Z(\theta)$

Interaction Point Process

- Biologist's interest: study progression of viral infections
- Our goal: use data from imaging of cell cultures to study the spatial structure of an infection
- An *in vitro* cell culture study identifies and locates cells infected with two strains of the human respiratory syncytial virus (RSV-A and RSV-B)



Cells infected with RSV

Question: How does the presence of an infected cell impact infections in neighboring cells?

Attraction-repulsion Model

- Previous models (e.g. Strauss process) did not allow for repulsion *and* attraction
- New point process model (Goldstein, Haran, et al., 2015) allows for both
- Allows us to easily compare interaction behavior for different strains of RSV
- This is a model with an intractable normalizing function
- Existing algorithms take days to run
- This is the motivation for studying existing algorithms and developing new algorithms

Maximum Likelihood (ML) Inference

$$\hat{\theta} = \arg\max_{\theta \in \Theta} h(\mathbf{x}|\theta)/Z(\theta)$$

- Pseudolikelihood approximation (Besag, 1975)
  - Often a poor approximation
  - Awkward in a hierarchical model (it is not compatible with a real probability model)
- Markov chain Monte Carlo Maximum Likelihood (Geyer and Thompson, 1994)
  - Sensitive to choice of importance function
  - Optimization can be unstable
  - For some models, obtaining standard errors is challenging E.g. Our point process model (Goldstein et al., 2015)

Bayesian Inference

- A Bayes approach can sidestep some of the challenges of ML inference.
- Bayesian inference for such models
  - Prior : $p(\theta)$
  - Posterior: $\pi(\theta|\mathbf{x}) \propto p(\theta)h(\mathbf{x}|\theta)/Z(\theta)$
- Inference is generally via Markov chain Monte Carlo (MCMC).
- MCMC is challenging for such models due to $Z(\theta)$.

Markov chain Monte Carlo Basics

- Construct Harris-ergodic Markov chain $\theta_1, \theta_2, \ldots$ with stationary distribution $\pi(\theta \mid \mathbf{x}))$
- Treat $\theta_1, \theta_2, \theta_3, ...$ as if they are samples from $\pi(\theta|\mathbf{x})$
- For any real-valued $g(\cdot)$, approximate $E_\pi(g(\theta))$ by

$$\hat{\mu}_n = \frac{\sum_{i=1}^{n} g(\theta_i)}{n}$$

- Under general conditions, $\hat{\mu}_n \to \mu$ as $n \to \infty$

The Metropolis-Hastings Algorithm

Recipe for constructing Markov chain. Given θ_n, obtain θ_{n+1}

$$\begin{tikzcd}
 & & \boxed{\theta_{n+1} = \theta'} \\
\boxed{\theta_n} \arrow[r] & \boxed{\text{propose } \theta' \sim q(\cdot|\theta_n)} \arrow[ur, "\text{accept}"'] \arrow[dr, "\text{reject}"] & \\
 & & \boxed{\theta_{n+1} = \theta_n}
\end{tikzcd}$$

- Acceptance probability:

$$\alpha = \frac{\pi(\theta'|\mathbf{x})q(\theta_n|\theta')}{\pi(\theta_n|\mathbf{x})q(\theta'|\theta_n)}$$

MCMC with Intractable Normalizing Functions

- Recall:
  - Prior : $p(\theta)$
  - Posterior: $\pi(\theta|\mathbf{x}) \propto p(\theta)h(\mathbf{x}|\theta)/Z(\theta)$
- Acceptance ratio for Metropolis-Hastings algorithm

$$\alpha = \frac{\pi(\theta'|\mathbf{x})q(\theta_n|\theta')}{\pi(\theta_n|\mathbf{x})q(\theta'|\theta_n)} = \frac{p(\theta')Z(\theta_n)h(\mathbf{x}|\theta')q(\theta_n|\theta')}{p(\theta_n)Z(\theta')h(\mathbf{x}|\theta_n)q(\theta'|\theta_n)}$$

Cannot evaluate because of $Z(\theta)$

Outline

Algorithms

Two classes of algorithms for Bayesian inference

I Auxiliary variable methods

- Generate an auxiliary random variate from $f(\theta)$
- Cancel $Z(\theta)$ in the acceptance ratio

II Likelihood approximation methods

- Approximate $Z(\theta)$ using Monte Carlo
- Use approximation, $\hat{Z}(\theta)$ in acceptance ratio

Exchange Algorithm

(Moller et al. (2006); Murray et al. (2007))

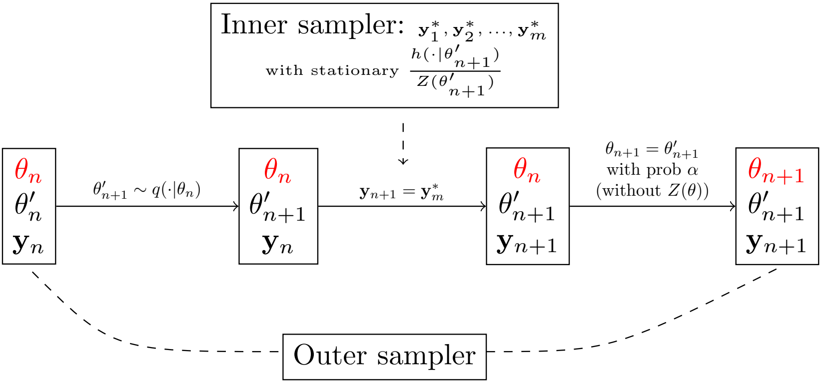
$$\boxed{\begin{matrix}\color{red}{\theta_n}\\ \theta'_n \\ \mathbf{y}_n\end{matrix}} \xrightarrow{\theta'_{n+1} \sim q(\cdot|\theta_n)} \boxed{\begin{matrix}\color{red}{\theta_n}\\ \theta'_{n+1} \\ \mathbf{y}_n\end{matrix}} \xrightarrow{\mathbf{y}_{n+1} \sim \frac{h(\cdot|\theta'_{n+1})}{Z(\theta'_{n+1})} \text{ i.i.d.}} \boxed{\begin{matrix}\color{red}{\theta_n}\\ \theta'_{n+1} \\ \mathbf{y}_{n+1}\end{matrix}} \xrightarrow[\text{with prob } \alpha \text{ (without } Z(\theta))]{\theta_{n+1} = \theta'_{n+1}} \boxed{\begin{matrix}\color{red}{\theta_{n+1}}\\ \theta'_{n+1} \\ \mathbf{y}_{n+1}\end{matrix}}$$

- Goal: obtain next state of Markov chain (θ_{n+1}) from current state (θ_n) by using auxiliary variable **y**
  - Update augmented state ($\theta_n, \theta'_n, \mathbf{y}_n$) instead of updating ($\theta_n$).
  - This cancels out $Z(\theta)$ in acceptance ratio.
  - Then take the marginal samples of (θ_n).

Exchange Algorithm

- Asymptotically exact, that is, $\hat{\mu}_n \to \mu$ as $n \to \infty$
- Very clever and simple (in theory)
- **Requires that we draw exact samples from probability model for each proposed θ**
  - Need to do perfect sampling with Markov chains
  - Infeasible or very expensive in general
- Alternative: Double Metropolis-Hastings (Liang, 2010)

Double Metropolis-Hastings (DMH)



- Theory assumes length of inner and outer sampler go to infinity. **Asymptotically inexact** in practice
- Most practical of algorithms we considered

The Adaptive Exchange Algorithm (AEX)

Liang et al. (2016)

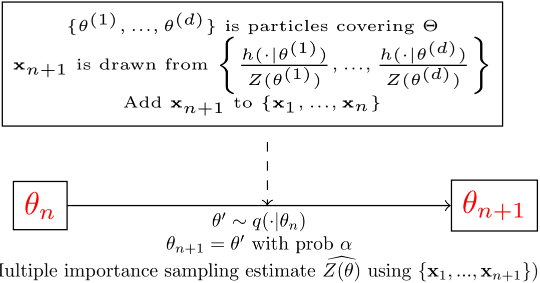
- Basic idea: AEX replaces independent sampling of $\mathbf{y}$ with a re-sampling method.
- With increasing iterations, more samples get added to $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$: re-sampling $\approx$ exact sampling of $\mathbf{y}$ from $h(\mathbf{y}|\theta')/Z(\theta')$.
- Theory: **Asymptotically exact** without perfect sampling
- Slow
- (Extremely) complicated to code/tune
- Huge storage requirements unless sufficient statistics are of low dimensions

Auxiliary Variable Methods: Summary

- Sequential algorithms, not amenable to easy parallelization
- Each iteration involves runnning a (sequential) MCMC algorithm
- Double M-H: asymptotically inexact but easy to code

Likelihood Approximation Method

Atchade, Lartillot and Robert (ALR) Algorithm



- Basic idea: approximate $Z(\theta)$ adaptively through weighted importance sampling. (Atchade et al., 2015)

ALR Algorithm

- With increasing iterations, more samples get added to $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$: approximation $\widehat{Z}_{n+1}(\theta)$ becomes more accurate.
- **Asymptotically exact** without independent sampling
- Memory issues: have to store large number of sampled data used in importance sampling
- Comparable to AEX algorithm in speed
- Outer algorithm is sequential but "inner algorithm" to update importance sampling estimates is amenable to parallelization

Outline

Interaction Point Process Models

θ_1	Mean	95%HPD	ESS	Time(minute)
DMH	1.196	(1.028,1.335)	1343.106	3.935
Gold standard	1.192	(1.031,1.339)	4397.390	

Table: 40,000 MCMC samples

- Simulated low-dimensional point process data $n = 200$. Real problem: $n = 13,000$
  - Data $\mathbf{x} \in R^{200\times 2}$ is coordinates of point process
  - Evaluating $h(\mathbf{x}|\theta)$ requires calculating distance matrix of $\mathbf{x}$.
  - AEX and ALR are impractical (need to store 200×200-dimensional distance matrix with each iteration).
- DMH (**inexact**) is only feasible approach.

Computational Costs

- Computational complexity as a function of data size n
  - Exponential family models: $\mathcal{O}(\mathbf{n})$, point process: $\mathcal{O}(\mathbf{n^2})$
  - Except for fixed costs, complexity of all algorithms are similar
- Memory
  - Without low-dimensional sufficient statistics, huge memory usage in AEX and ALR (**n^2** versus **p**).
  - No memory issues with other algorithms.
- Computational efficiency
  - Likelihood approximation shows better mixing. (high effective sample size).
  - Auxiliary variable approaches are less expensive per iteration. (high effective sample size per second).
  - Hence, roughly speaking, auxiliary variable methods are

Outline

New Emulation-Based Algorithm

- All existing algorithms are computationally very expensive
- An alternative is desirable
- Basic idea:
  - Approximate $Z(\theta)$ using importance sampling on some design points
  - Use Gaussian process emulation approach to interpolate this function at any new value
  - We have some theory to justify this work as number of design points and number of importance sampling draws increases

Computational Benefits

- Can compute in parallel; much of this is done “offline”, before running the algorithm
- Preliminary results: dramatic reduction in computing time.

References

Park and Haran (2018) Bayesian Inference in the Presence of Intractable Normalizing Functions, *Journal of the American Statistical Association*, to appear