Bayesian Inference in the Presence of Intractable Normalizing Functions

(Joint work with Jaewoo Park)

Conference in Honor of Charlie Geyer

University of Minnesota

April 2018

Murali Haran

Department of Statistics, Penn State University

May 2003, After a 2.5 Hr Thesis Defense

Fresh Air

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer
  - I spend the majority of my time working with climate scientists and infectious disease experts
  - My applications research drives my methods research

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer
  - I spend the majority of my time working with climate scientists and infectious disease experts
  - My applications research drives my methods research
- I run long Markov chains

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer
  - I spend the majority of my time working with climate scientists and infectious disease experts
  - My applications research drives my methods research
- I run long Markov chains
- Argue against wastage due to burn-in/thinning samples

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer
  - I spend the majority of my time working with climate scientists and infectious disease experts
  - My applications research drives my methods research
- I run long Markov chains
- Argue against wastage due to burn-in/thinning samples
- The research discussed today

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer
  - I spend the majority of my time working with climate scientists and infectious disease experts
  - My applications research drives my methods research
- I run long Markov chains
- Argue against wastage due to burn-in/thinning samples
- The research discussed today
- Even the title of this talk!
  - Normalizing *functions*

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer
  - I spend the majority of my time working with climate scientists and infectious disease experts
  - My applications research drives my methods research
- I run long Markov chains
- Argue against wastage due to burn-in/thinning samples
- The research discussed today
- Even the title of this talk!
  - Normalizing *functions*
- Give students a tough time when they write log instead of $\log$

Charlie's Influence

- You could do worse than to talk to scientists – Charlie Geyer
  - I spend the majority of my time working with climate scientists and infectious disease experts
  - My applications research drives my methods research
- I run long Markov chains
- Argue against wastage due to burn-in/thinning samples
- The research discussed today
- Even the title of this talk!
  - Normalizing *functions*
- Give students a tough time when they write *log* instead of log
- I say "woof" very often . . .

Outline

Models with Intractable Normalizing Functions

Algorithms for Bayesian Inference

An Example

A Function Emulation Approach

Models with Intractable Normalizing Functions

- Data: $\mathbf{x} \in \chi$, parameter: $\theta \in \Theta$
- Probability 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), Geyer (1999), Geyer and Møller (1994), Goldstein, Haran, Chiaromonte et al. (2015)
- Challenge: likelihood-based inference with $Z(\theta)$

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 (not compatible with a real probability model)
- Markov chain Monte Carlo Maximum Likelihood (Geyer and Thompson, 1992)
  - Elegant approach using importance sampling approximation
  - Some challenges when analytical gradients are not available E.g. Attraction-repulsion point process

Bayesian Inference

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

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

Cannot evaluate because of $Z(\cdot)$

Outline

Algorithms

Two classes of algorithms for Bayesian inference

I Auxiliary variable methods

- Generate an auxiliary random variate from model $h(\mathbf{x}|\theta)$
- Cancel $Z(\theta)$ in the acceptance ratio

II Likelihood approximation methods

- Compute Monte Carlo approximation to $Z(\theta)$, $\widehat{Z}(\theta)$
- Use $\widehat{Z}(\theta)$ in M-H acceptance ratio

Auxiliary Variable Approach

Møller et al. (2006)

- Augment distribution with auxiliary variable $\mathbf{y}$
  $\pi(\mathbf{y}, \theta | \mathbf{x}) \propto f(\mathbf{y}|\hat{\theta}) p(\theta) h(\mathbf{x}|\theta) / Z(\theta)$
  for some fixed $\hat{\theta}$
- Single iteration of Metropolis-Hastings algorithm
  1. Propose θ^* as usual, then propose $\mathbf{y}^* | \theta^*$ from $f(\cdot|\theta^*)$
  2. Accept-reject $(\mathbf{y}^*, \theta^*)$
     - Normalizing functions cancel out in Metropolis-Hastings acceptance ratio

Murray et al. (2007) suggests a related algorithm

Comments

- Asymptotically exact, that is, as $n \to \infty$, transition kernel of Markov chain converges to π
- 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 (Propp and Wilson, 1996)
  - Infeasible or very expensive for most problems
- Alternative: Double Metropolis-Hastings (Liang, 2010)
  - **Replace exact samples at each proposed θ^* with approximate draw from a Markov chain**

Double Metropolis-Hastings (DMH)

- At each iteration of a Markov chain for $\pi(\theta| \dots)$ (*outer sampler*), run a Markov chain for auxiliary $\mathbf{y}$ (*inner sampler*)
- **Asymptotically inexact** in practice
- But
  - Easy to implement
  - Computationally more efficient than other algorithms

The Adaptive Exchange Algorithm (AEX)

Liang, Jin, Song, Liu (2016)

- Idea: replace independent sampling of $\mathbf{y}$ with a re-sampling approach based on stochastic approximation Monte Carlo (Liang et al., 2007)
- Impressive result: **asymptotically exact** without perfect sampling!
- Complicated to code/tune
- Huge storage requirements unless sufficient statistics are of low dimensions

Auxiliary Variable Methods: Recap

- List
  - Møller et al. (2006) and Murray et al. (2007)
  - Adaptive exchange algorithm
  - Double Metropolis-Hastings
- Sequential algorithms, not amenable to easy parallelization
- Double M-H: asymptotically inexact but fast and easy to code

Likelihood Approximation Method

(Atchade, Lartillot and Robert (ALR), 2008)

Idea: approximate $Z(\theta)$ adaptively through weighted importance sampling (Atchade et al., 2015). Use approximation in MCMC acceptance ratio.

- Based on Wang Landau algorithm (2001)
- **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

Summary of Likelihood Approximation Algorithms

- List with lots of overlap
  - ALR (Atchade, Lartillot, Robert, 2012) algorithm
  - Pseudo-marginal MCMC (Andrieu and Roberts, 2009), e.g. Russian roullette algorithm (Lyne et al., 2015)
  - Noisy MCMC (Alqueir et al., 2016) and hybrids
- Many clever ideas

Summary of Likelihood Approximation Algorithms

- List with lots of overlap
  - ALR (Atchade, Lartillot, Robert, 2012) algorithm
  - Pseudo-marginal MCMC (Andrieu and Roberts, 2009), e.g. Russian roullette algorithm (Lyne et al., 2015)
  - Noisy MCMC (Alqueir et al., 2016) and hybrids
- Many clever ideas
  Just because it sounds like a good idea, doesn't mean it is a good idea. – Charlie Geyer
  - The algorithms tend to be slow
  - Huge memory requirements unless there are low-dimensional sufficient statistics

Outline

Interaction Point Process Model

Inspired by Geyer (1999) and Geyer and Møller (1994)

Goldstein, Haran, Chiaromonte et al. (2015)

- Simulated example: point process with $n = 200$
  - Data $\mathbf{x} \in R^{200\times 2}$ are coordinates of point process
  - Evaluating $h(\mathbf{x}|\theta)$ requires calculating distance matrix of $\mathbf{x}$.
  - AEX, ALR are impractical (storing 200×200-dimensional distance matrices per particle per iteration)
- Practical approach: Double Metropolis-Hastings (DMH)
- DMH results are accurate if inner sampler is long enough
- For $n \approx 3,000$
  - Very efficiently coded DMH takes $\approx$ 19 hours
  - All other algorithms are infeasible
- Larger problems: even DMH is infeasible

Most of the time when people think they are screwed, they aren't really screwed. – Glen Meeden

Most of the time when people think they are screwed, they aren't really screwed. Though sometimes they *are screwed*. – Glen Meeden

Outline

A Function Emulation Approach

- Existing algorithms are computationally very expensive
- Our approach:
  1. Approximate $Z(\theta)$ using importance sampling on some design points
  2. Use Gaussian process emulation approach to interpolate this function at other values of θ
- Some theoretical justification as number of design points and number of importance sampling draws increases
- See Jaewoo Park's poster

A Function Emulation Approach

Is this just a TTD? – Charlie Geyer

A Function Emulation Approach

Is this just a TTD? – Charlie Geyer

A TTD = A Thing To Do = A method that you tout/publish, that may or may not be of any real interest or use

A Function Emulation Approach

Is this just a TTD? – Charlie Geyer

A TTD = A Thing To Do = A method that you tout/publish, that may or may not be of any real interest or use

Simulated social network (ERGM): 1400 nodes

θ_2	Mean	95%HPD	Time(hour)
Double M-H	1.77	(1.44, 2.12)	23.83
Emul_1	1.79	(1.45, 2.13)	0.45
Emul_{10}	**1.96**	(1.87, 2.05)	**1.39**

True $\theta_2 = 2$: Emul_{10} is accurate, others are not

Computational efficiency allows us to use longer chain (Emul_{10}). Corresponding DMH algorithm $\approx$ 10 days

See Jaewoo Park's poster

Details of Other Simulated Examples

Details of Other Simulated Examples

People don't care too much about the details. – Glen Meeden

The Last Slide

- Charlie taught me a lot

The Last Slide

- Charlie taught me a lot
  - statistics
  - computing
  - mathematics
  - jazz, politics, ...
- Charlie made research more fun

The Last Slide

- Charlie taught me a lot
  - statistics
  - computing
  - mathematics
  - jazz, politics, ...
- Charlie made research more fun
- Charlie set a great example

The Last Slide

- Charlie taught me a lot
  - statistics
  - computing
  - mathematics
  - jazz, politics, ...
- Charlie made research more fun
- Charlie set a great example
  - A true scholar and intellectual
  - Always willing to talk to students/colleagues

The Last Slide

- Charlie taught me a lot
  - statistics
  - computing
  - mathematics
  - jazz, politics, ...
- Charlie made research more fun
- Charlie set a great example
  - A true scholar and intellectual
  - Always willing to talk to students/colleagues
- The world (especially academia) needs more Charlies

References

- Park and Haran (2018a) Bayesian Inference in the Presence of Intractable Normalizing Functions (on arxiv.org) to appear in the *Journal of the American Statistical Association*
- Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants, *Biometrika*
- Murray, I., Z. Ghahramani, and D. MacKay (2006) MCMC for doubly-intractable distributions. *Proc of 22nd Annual Conf on Uncertainty in Artificial Intelligence* UAI06
- Liang, F. (2010) A Double Metropolis-Hastings sampler for spatial models with intractable normalizing constants.

References

- Atchade, Y., Lartillot, N. and Robert, C. (2013) Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*
- Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2015) An adaptive exchange algorithm for sampling from distributions with intractable normalising constants. *Journal of the American Statistical Association*
- Goldstein, J., Haran, M., Simeonov, I., Fricks, J., and Chiaromonte, F. (2015) An attraction-repulsion point process model for respiratory syncytial virus infections. *Biometrics*