Bayesian inference for intractable distributions

Nial Friel

University College Dublin
nial.friel@ucd.ie

October, 2014

Main themes:

- The Bayesian inferential approach has had a profound impact on statistical learning.
- Monte Carlo methods which were popularised in the early 1990s provide a simulation-based approach to overcoming the intractability inherent in almost all posterior distributions. Suppose we have data y, and a likelihood function f with parameters θ:

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

Main themes:

- The Bayesian inferential approach has had a profound impact on statistical learning.
- Monte Carlo methods which were popularised in the early 1990s provide a simulation-based approach to overcoming the intractability inherent in almost all posterior distributions. Suppose we have data y, and a likelihood function f with parameters θ:

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

- However it turns out that there are many statistical models for which the likelihood function is intractable

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta).$$

Main themes:

- The Bayesian inferential approach has had a profound impact on statistical learning.
- Monte Carlo methods which were popularised in the early 1990s provide a simulation-based approach to overcoming the intractability inherent in almost all posterior distributions. Suppose we have data y, and a likelihood function f with parameters θ:

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

- However it turns out that there are many statistical models for which the likelihood function is intractable

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta).$$

- This extra level of intractability is sometimes due to the complicated structure in the data, or even due to the sheer volume of the data.

Intractable likelihoods

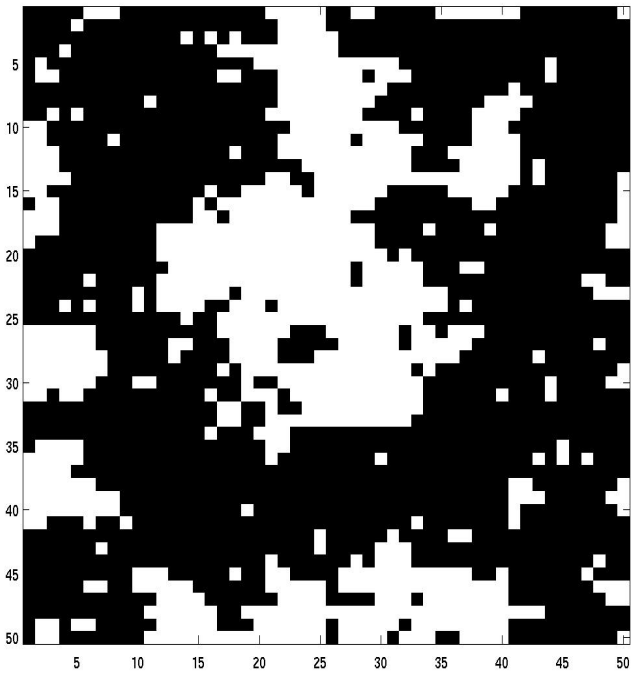
- Genetics: phylogenetics
- Ecology: spatial models for biodiversity
- Finance: volatility models
- Machine learning: deep architectures
- Epidemiology: stochastic models for disease transmissions
- Sociology: social network models

Ising model

- Defined on a lattice $y = \{y_1, \ldots, y_n\}$.
- Lattice points y_i take values $\{-1, 1\}$.
- Full conditional $f(y_i|y_{-i}, \theta) = f(y_i|\text{neighbours of } i, \theta)$.

$$f(y|\theta) \propto q_\theta(y) = \exp\left\{\frac{1}{2}\theta_1 \sum_{i\sim j} y_i y_j\right\}.$$

Here $\sim$ means "is a neighbour of".



$$f(y|\theta) = \frac{\exp(\theta^T s(y))}{z(\theta)}$$

$s(y)$ is a sufficient statistics and counts the number of 'like' neighbours.

$$z(\theta) = \sum_{y_1} \cdots \sum_{y_n} q_\theta(y).$$

Bayesian statistics

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

For many years, Bayesian statistics was considered a good idea, but impractical because there were not the tools to infer the posterior distribution, $\pi(\theta|y)$.

Bayesian statistics

$$\begin{aligned} \pi(\theta|y) &\propto f(y|\theta)\pi(\theta) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

For many years, Bayesian statistics was considered a good idea, but impractical because there were not the tools to infer the posterior distribution, $\pi(\theta|y)$.

This changed, thanks largely to the seminal work by Gelfand and Smith (1990), where Markov chain Monte Carlo was introduced to the statistics community.

Bayesian statistics

$$\begin{aligned} \pi(\theta|y) &\propto f(y|\theta)\pi(\theta) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

For many years, Bayesian statistics was considered a good idea, but impractical because there were not the tools to infer the posterior distribution, $\pi(\theta|y)$.

This changed, thanks largely to the seminal work by Gelfand and Smith (1990), where Markov chain Monte Carlo was introduced to the statistics community.

It is not an understatement to say that MCMC revolutionsed Bayesian statistics.

MCMC – an historical sidenote

- In fact, Markov chain Monte Carlo had been 'invented' in the 1950s, through work by Metropolis and collaborators.
- It took almost 40 years for the statistics community to re-discover this!

MCMC – the basic idea

- Markov chain Monte Carlo provides a general (and incredibly popular) approach to simulate from the posterior distribution $\pi(\theta|y)$.
- The idea is to simulate a Markov chain $(\theta_n)_{n\in\mathbb{N}}$ with transition kernel P such that π is invariant under P: $\pi P = \pi$.

MCMC – the basic idea

- Markov chain Monte Carlo provides a general (and incredibly popular) approach to simulate from the posterior distribution $\pi(\theta|y)$.
- The idea is to simulate a Markov chain $(\theta_n)_{n\in\mathbb{N}}$ with transition kernel P such that π is invariant under P: $\pi P = \pi$.
- We can then use the approximation,

$$\frac{1}{N}\sum_{n=1}^{N} t(\theta_n) \simeq \int_{\Theta} t(\theta)\pi(\mathrm{d}\theta|y).$$

The Metropolis algorithm

Algorithm 1: The Metropolis algorithm

1. Initialise $\theta^{(0)}$;
2. At iteration i, propose $\theta' \sim h(\cdot|\theta)$;
3. With probability

$$\min\left(1, \frac{\pi(\theta'|y)}{\pi(\theta^{(i)}|y)}\right)$$

set $\theta^{(i+1)} = \theta'$, otherwise set $\theta^{(i+1)} = \theta^{(i)}$;

Remarkably: the generated Markov chain $\{\theta^{(i)}\}$ leaves the posterior distribution, $\pi(\theta|y)$ invariant.

from *SIAM News*, Volume 33, Number 4

The Best of the 20th Century: Editors Name Top 10 Algorithms

By Barry A. Cipra

Algos is the Greek word for pain. *Algor* is Latin, to be cold. Neither is the root for *algorithm*, which stems instead from al-Khwarizmi, the name of the ninth-century Arab scholar whose book *al-jabr wa'l muqabalah* devolved into today's high school algebra textbooks. Al-Khwarizmi stressed the importance of methodical procedures for solving problems. Were he around today, he'd no doubt be impressed by the advances in his eponymous approach.

Some of the very best algorithms of the computer age are highlighted in the January/February 2000 issue of *Computing in Science & Engineering*, a joint publication of the American Institute of Physics and the IEEE Computer Society. Guest editors Jack Don-garra of the University of Tennessee and Oak Ridge National Laboratory and Fran-cis Sullivan of the Center for Comput-ing Sciences at the Institute for Defense Analyses put togeth-er a list they call the "Top Ten Algorithms of the Century."

"We tried to assemble the 10 al-gorithms with the greatest influence on the development and practice of science and engineering in the 20th century," Dongarra and Sullivan write. As with any top-10 list, their selections—and non-selections—are bound to be controversial, they acknowledge. When it comes to picking the algorithmic best, there seems to be no best algorithm.

Without further ado, here's the CiSE top-10 list, in chronological order. (Dates and names associated with the algorithms should be read as first-order approximations. Most algorithms take shape over time, with many contributors.)

1946: John von Neumann, Stan Ulam, and Nick Metropolis, all at the Los Alamos Scientific Laboratory, cook up the Metropolis algorithm, also known as the **Monte Carlo method**.

The Metropolis algorithm aims to obtain approximate solutions to numerical problems with unmanageably many degrees of freedom and to combinatorial problems of factorial size, by mimicking a random process. Given the digital computer's reputation for deterministic calculation, it's fitting that one of its earliest applications was the generation of random numbers.

In terms of wide-spread use, George Dantzig's simplex method is among the most successful al-gorithms of all time.

1947: George Dantzig, at the RAND Corporation, creates the **simplex method for linear programming**.

In terms of widespread application, Dantzig's algorithm is one of the most successful of all time: Linear programming dominates the world of industry, where economic survival depends on the ability to optimize within budgetary and other constraints. (Of course, the "real" problems of industry are often nonlinear; the use of linear programming is sometimes dictated by the computational budget.) The simplex method is an elegant way of arriving at optimal answers. Although theoretically susceptible to exponential delays, the algorithm in practice is highly efficient—which in itself says something interesting about the nature of computation.

1950: Magnus Hestenes, Eduard Stiefel, and Cornelius Lanczos, all from the Institute for Numerical Analysis at the National Bureau of Standards, initiate the development of **Krylov subspace iteration methods**.

These algorithms address the seemingly simple task of solving equations of the form $Ax = b$. The catch, of course, is that A is a huge $n \times n$ matrix, so that the algebraic answer $x = b/A$ is not so easy to compute. (Indeed, matrix "division" is not a particularly useful concept.) Iterative methods—such as solving equations of the form $Kx_{i+1} = Kx_i + b - Ax_i$ with a simpler matrix K that's ideally "close" to A—lead to the study of Krylov subspaces. Named

Bayesian inference

Doubly-intractable posterior

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

- Naïve Metropolis algorithm proposes the move from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ with probability:

$$\begin{aligned} \alpha &= \min\left(1, \frac{f(\mathbf{y}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right) \\ &= \min\left(1, \frac{q_{\boldsymbol{\theta}^*}(\mathbf{y})\pi(\boldsymbol{\theta}^*)}{q_{\boldsymbol{\theta}}(\mathbf{y})\pi(\boldsymbol{\theta})} \times \underbrace{\frac{z(\boldsymbol{\theta})}{z(\boldsymbol{\theta}^*)}}_{\textit{intractable}}\right) \end{aligned}$$

- Here we are interested to sample the posterior distribtion

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta).$$

- However, we can't evaluate the right hand side above, because the likelihood is unavailable.
- In turn, this implies that we may not be able to evaluate the Markov transition kernel resulting from an MCMC algorithm.

- Here we are interested to sample the posterior distribtion

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta).$$

- However, we can't evaluate the right hand side above, because the likelihood is unavailable.
- In turn, this implies that we may not be able to evaluate the Markov transition kernel resulting from an MCMC algorithm.

We focus on doubly intractable distributions, but this framework should also extend to many other situations where the actual posterior is impractical to target, eg, Big data settings etc.

- Here there is a natural kernel P, st $\pi P = \pi$. However we cannot draw $\theta_{n+1} \sim P(\theta_n, \cdot)$ for a fixed θ_n.
- A natural idea is to replace P by an approximation $\hat{P}$.

- Here there is a natural kernel P, st $\pi P = \pi$. However we cannot draw $\theta_{n+1} \sim P(\theta_n, \cdot)$ for a fixed θ_n.
- A natural idea is to replace P by an approximation $\hat{P}$.
- Ideally, $\hat{P}$ is 'close' to P, but generally $\pi\hat{P} \neq \pi$.

- Here there is a natural kernel P, st $\pi P = \pi$. However we cannot draw $\theta_{n+1} \sim P(\theta_n, \cdot)$ for a fixed θ_n.
- A natural idea is to replace P by an approximation $\hat{P}$.
- Ideally, $\hat{P}$ is 'close' to P, but generally $\pi\hat{P} \neq \pi$.
- This leads to the natural question:

  How close is the Markov chain with transition kernel $\hat{P}$ to the Markov chain with transition kernel P? Eg, is it possible to upper bound?

$$\left\| \delta_{\theta_0} \hat{P}^n - \pi \right\|.$$

- It turns out that a useful answer is given by the study of the stability of Markov chains.

Theorem (Mitrophanov (2005), Corollary 3.1)

Let us assume that

- **(H1)** *the Markov chain with transition kernel P is uniformly ergodic:*

$$\sup_{\theta_0} \|\delta_{\theta_0} P^n - \pi\| \leq C\rho^n$$

for some $C < \infty$ and $\rho < 1$.

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0,

$$\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho}\right) \|P - \hat{P}\|$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

Metropolis-Hastings algorithm

Algorithm 2: Metropolis-Hastings algorithm

1 draw $\theta' \sim h(\theta'|\theta_n)$;

2 $\theta_{n+1} = \begin{cases} \theta' & \text{with proba. } 1 \wedge a(\theta_n, \theta') = \frac{\pi(\theta'|y)}{\pi(\theta|y)} \\ \theta_n & \text{otherwise.} \end{cases}$

In some applications however, it is not possible to compute exactly the ratio $a(\theta_n, \theta')$.

In this case, it is reasonable to replace this ratio by an approximation, or an estimate: we draw $x' \sim F_{\theta'}(x')$ for some suitable probability distribution $F_{\theta'}(x')$ and approximate $a(\theta_n, \theta')$ by some function $\hat{a}(\theta_n, \theta', x)$.

This leads us to consider the following *noisy Metropolis-Hastings algorithm*.

Algorithm 3: Noisy Metropolis-Hastings algorithm

1 draw $\theta' \sim h(\theta'|\theta_n)$;
2 draw $x' \sim F_{\theta'}(x')$ for some probability distribution $F_{\theta'}(x')$;
3 $\theta_{n+1} = \begin{cases} \theta' & \text{with proba. } 1 \wedge \hat{a}(\theta_n, \theta', x') \\ \theta_n & \text{otherwise.} \end{cases}$

Theoretical guarantees for the noisy MH algorithm

Corollary

Let us assume that

- **(H1)** *holds. (The Markov chain with transition kernel P is uniformly ergodic),*
- **(H2)** $\hat{a}(\theta, \theta', x')$ *satisfies:*

$$\mathbb{E}_{x' \sim F_{\theta'}} \left| \hat{a}(\theta, \theta', x') - a(\theta, \theta') \right| \leq \delta(\theta, \theta'). \tag{1}$$

Theoretical guarantees for the noisy MH algorithm

Corollary

Let us assume that

- **(H1)** *holds. (The Markov chain with transition kernel P is uniformly ergodic),*
- **(H2)** $\hat{a}(\theta, \theta', x')$ *satisfies:*

$$\mathbb{E}_{x' \sim F_{\theta'}} \left| \hat{a}(\theta, \theta', x') - a(\theta, \theta') \right| \leq \delta(\theta, \theta'). \tag{1}$$

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0,

$$\| \delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n \| \leq \left(\lambda + \frac{C \rho^\lambda}{1 - \rho} \right) 2 \sup_\theta \int \mathrm{d}\theta' \, h(\theta' | \theta) \delta(\theta, \theta')$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

Noisy MH: Gibbs random fields

Note: when the upper bound in (1) is bounded:

$$\mathbb{E}_{x' \sim F_{\theta'}} \left| \hat{a}(\theta, \theta', x') - a(\theta, \theta') \right| \leq \delta(\theta, \theta') \leq \delta < \infty,$$

then it results that

$$\| \delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n \| \leq \delta \left(\lambda + \frac{C \rho^\lambda}{1 - \rho} \right).$$

Note: when the upper bound in (1) is bounded:

$$\mathbb{E}_{x' \sim F_{\theta'}} \left| \hat{a}(\theta, \theta', x') - a(\theta, \theta') \right| \leq \delta(\theta, \theta') \leq \delta < \infty,$$

then it results that

$$\| \delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n \| \leq \delta \left(\lambda + \frac{C \rho^\lambda}{1 - \rho} \right).$$

Obviously, we expect that $\hat{a}$ is chosen in such a way that $\delta \ll 1$ and so in this case, $\| \delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n \| \ll 1$. As a consequence, letting $n \to \infty$ gives:

$$\limsup_{n \to \infty} \| \pi - \delta_{\theta_0} \hat{P}^n \| \leq \delta \left(\lambda + \frac{C \rho^\lambda}{1 - \rho} \right).$$

Exchange algorithm

(Murray, Ghahramani & MacKay 2006)

This algorithm provides a generic means to sample from a doubly intractable posterior distribution.

Consider an augmented distribution

$$\pi(\boldsymbol{\theta}', \mathbf{y}', \boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})h(\boldsymbol{\theta}'|\boldsymbol{\theta})f(\mathbf{y}'|\boldsymbol{\theta}')$$

whose marginal distribution for $\boldsymbol{\theta}$ is the posterior of interest.

- $f(\mathbf{y}'|\boldsymbol{\theta}')$ same likelihood model for which $\mathbf{y}$ is defined
- $h(\boldsymbol{\theta}'|\boldsymbol{\theta})$ arbitrary distribution for the augmented variable $\boldsymbol{\theta}'$ which might depend on $\boldsymbol{\theta}$ (eg random walk distribution centred at $\boldsymbol{\theta}$)

Algorithm 4: Exchange algorithm

1 Draw $\boldsymbol{\theta}' \sim h(\cdot|\boldsymbol{\theta})$;
2 Draw $\mathbf{y}' \sim f(\cdot|\boldsymbol{\theta}')$;
3 With probability

$$\min\left(1, \frac{q_{\boldsymbol{\theta}}(\mathbf{y}')}{q_{\boldsymbol{\theta}}(\mathbf{y})}\frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})}\frac{q_{\boldsymbol{\theta}'}(\mathbf{y})}{q_{\boldsymbol{\theta}'}(\mathbf{y}')} \times \underbrace{\frac{z(\boldsymbol{\theta})z(\boldsymbol{\theta}')}{z(\boldsymbol{\theta})z(\boldsymbol{\theta}')}}_{1}\right)$$

set $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}'$, otherwise set $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$;

Exchange algorithm

Exchange algorithm:

$$\alpha = \min\left(1, \frac{q_{\theta'}(y)}{q_{\theta}(y)} \frac{p(\theta')}{p(\theta)} \frac{q_{\theta}(y')}{q_{\theta'}(y')}\right).$$

Exchange algorithm

Exchange algorithm:

$$\alpha = \min\left(1, \frac{q_{\theta'}(y)}{q_{\theta}(y)} \frac{p(\theta')}{p(\theta)} \frac{q_{\theta}(y')}{q_{\theta'}(y')}\right).$$

MH algorithm:

$$\alpha = \min\left(1, \frac{q_{\theta'}(y)p(\theta')z(\theta)}{q_{\theta}(y)p(\theta)z(\theta')}\right)$$

Exchange algorithm

Exchange algorithm:

$$\alpha = \min\left(1, \frac{q_{\theta'}(y)}{q_{\theta}(y)} \frac{p(\theta')}{p(\theta)} \frac{q_{\theta}(y')}{q_{\theta'}(y')}\right).$$

MH algorithm:

$$\alpha = \min\left(1, \frac{q_{\theta'}(y)p(\theta')z(\theta)}{q_{\theta}(y)p(\theta)z(\theta')}\right)$$

In fact:

$$\mathbf{E}_{y'|\theta'} \frac{q_{\theta}(y')}{q_{\theta'}(y')} = \frac{z(\theta)}{z(\theta')}.$$

Noisy MCMC

Suppose a Monte Carlo estimate of $z(\theta')/z(\theta)$ is plugged into the MH accept/reject ratio:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{q_\theta(y_i^{'})}{q_{\theta'}(y_i^{'})},$$

where $\{y_i^{'}\} \sim f(y|\theta')$.

Noisy MCMC

Suppose a Monte Carlo estimate of $z(\theta')/z(\theta)$ is plugged into the MH accept/reject ratio:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{q_\theta(y_i^{'})}{q_{\theta'}(y_i^{'})},$$

where $\{y_i^{'}\} \sim f(y|\theta')$.

Some special cases:

$N = 1$: Exchange algorithm. (Exact)

Noisy MCMC

Suppose a Monte Carlo estimate of $z(\theta')/z(\theta)$ is plugged into the MH accept/reject ratio:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{q_\theta(y_i^{'})}{q_{\theta'}(y_i^{'})},$$

where $\{y_i^{'}\} \sim f(y|\theta')$.

Some special cases:

$N = 1$: Exchange algorithm. (Exact)

$1 < N < \infty$: Noisy MCMC. (Approximate)

Noisy MCMC

Suppose a Monte Carlo estimate of $z(\theta')/z(\theta)$ is plugged into the MH accept/reject ratio:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{q_\theta(y_i^{'})}{q_{\theta'}(y_i^{'})},$$

where $\{y_i^{'}\} \sim f(y|\theta')$.

Some special cases:

$N = 1$:	Exchange algorithm. (Exact)
$1 < N < \infty$:	Noisy MCMC. (Approximate)
$N \to \infty$:	MH algorithm. (Exact)

Noisy MCMC

Suppose a Monte Carlo estimate of $z(\theta')/z(\theta)$ is plugged into the MH accept/reject ratio:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{q_\theta(y_i^{'})}{q_{\theta'}(y_i^{'})},$$

where $\{y_i^{'}\} \sim f(y|\theta')$.

Some special cases:

$N = 1$:	Exchange algorithm. (Exact)
$1 < N < \infty$:	Noisy MCMC. (Approximate)
$N \to \infty$:	MH algorithm. (Exact)

We call this noisy MCMC algorithm, Noisy exchange.

Noisy exchange algorithm

1 Draw $\theta' \sim h(\theta'|\theta)$;
2 Draw $y_1', \ldots, y_N' \sim \prod_{i=1}^{N} f(y_i'|\theta')$;
3 With probability

$$\min\left(1, \frac{q_{\theta'}(y)}{q_\theta(y)} \frac{p(\theta')}{p(\theta)} \frac{1}{N} \sum_{i=1}^{N} \frac{q_\theta(y_i')}{q_{\theta'}(y_i')}\right)$$

set $\theta^{(i+1)} = \theta'$, otherwise set $\theta^{(i+1)} = \theta^{(i)}$;

Convergence of the noisy exchange algorithm

Here we show that the noisy exchange algorithm falls into our theoretical framework.

Lemma

â satisfies **(H2)** *in Corollary 3.2 with*

$$\mathbb{E}_{x' \sim F_{\theta'}} \left| \hat{a}(\theta, \theta', x') - a(\theta, \theta') \right| \leq \delta(\theta, \theta')$$

$$= \frac{1}{\sqrt{N}} \frac{h(\theta|\theta')\pi(\theta')q_{\theta'}(y)}{h(\theta'|\theta)\pi(\theta)q_{\theta}(y)} \sqrt{\mathrm{Var}_{y' \sim f(y'|\theta')} \left(\frac{q_{\theta_n}(y')}{q_{\theta'}(y')} \right)}.$$

Moreover, we can show that:

Theorem

Assuming that the space Θ is bounded, then $\hat{a}$ with

$$\delta(\theta, \theta') \leq \frac{c_h^2 c_\pi^2 \mathcal{K}^4}{\sqrt{N}},$$

and

$$\sup_{\theta_0 \in \Theta} \|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \frac{\mathcal{C}}{\sqrt{N}}$$

where $\mathcal{C} = \mathcal{C}(c_\pi, c_h, \mathcal{K})$ is explicitly known. Noisy MH result

Simulation study

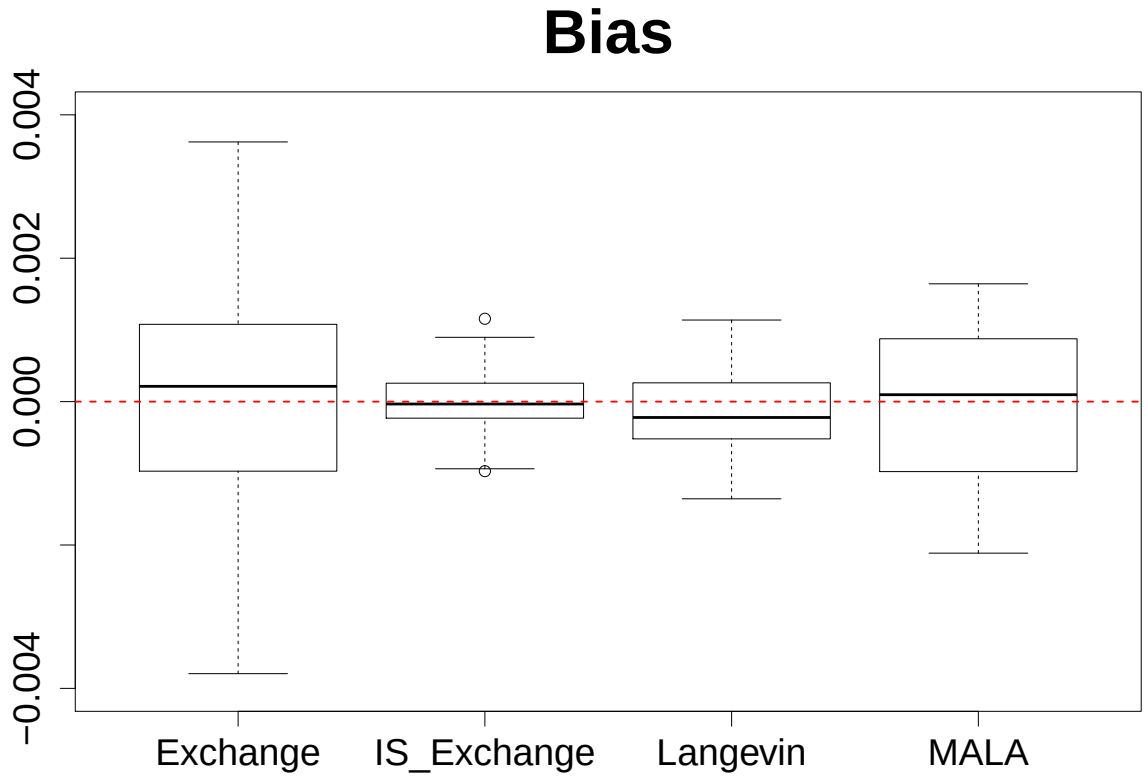
20 Datasets were simulated from a first-order Ising model defined on a 16×16 lattice, with a single interaction parameter $\theta = 0.4$.

Here the lattices are sufficently small to allow a very accurate estimate of the posterior density:

The normalising constant $z(\theta)$ can be calculated exactly for a fine grid of $\{\theta_i\}$ values (NF, Rue, Biometrika, 2007), which can then be plugged into the right hand side of:

$$\pi(\theta_i|y) \propto \frac{q_{\theta_i}(y)}{z(\theta_i)}\pi(\theta_i), \quad i = 1, \ldots, n.$$

Here we used a fine grid of $8,000$ points in the interval $[0, 0.8]$.


Bias
0.004
0.002
0.000
−0.004
Exchange
IS_Exchange
Langevin
MALA

Future directions

- Our framework give bounds on the total variation between a desired target distribution, and the invariant distribution of a noisy MC algorithm.
- An important question for future work concerns the statistical efficiency of the resultant estimators. This is a key question since the use of noisy MC will usually be motivated by the inefficiency of a standard alternative algorithm.
- This framework also applies in more general situations, apart from Gibbs random field models, and it will be important to generalise our results and findings.

References

- Alquier, Friel, Everitt, Boland (2015) Noisy MCMC. *Statistics and Computing* (to appear).
- Caimo and Friel (2011) Bayesian inference for the ERG model. *Social Networks*.
- Mitrophanov (2005) *Sensitivity and convergence of uniformly ergodic Markov chains*. Journal of applied probability.
- Murray, Ghahramani, and MacKay. (2006) *MCMC for doubly-intractable distributions*. In Proceedings of the 22nd annual conference on uncertainty in artificial intelligence