Approximate Bayesian inference for Gibbs random fields: noisy MCMC

Nial Friel

March, 2016

Main theme (1): Intractable likelihoods

- 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.

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

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

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

- A central theme in this talk is the notion that, although intractable, efficient statistical inference can result through simulating from the likelihood.

Main theme (1): Intractable likelihoods

- 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.

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

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

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

- A central theme in this talk is the notion that, although intractable, efficient statistical inference can result through simulating from the likelihood.

Main theme (1): Intractable likelihoods

- 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.

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

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

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

- A central theme in this talk is the notion that, although intractable, efficient statistical inference can result through simulating from the likelihood.

Main theme (2): Approximate MCMC

- It is often the case that one has an 'exact' MCMC available which target the posterior of interest.
- However, it is simply not possible to make a transition of this exact chain. Eg, a posterior with an intractable likelihood; a likelihood with too many observations etc.
- This was the topic of Daniel's talk yesterday.
- Big data: the volume of data prohibits calculation of the likelihood.
  - See, eg, "Austerity in MCMC land: cutting the MCMC budget" by Korattikara et al. (arXiv)
    "Can we make Bayesian posterior MCMC sampling more efficient when faced with very large datasets? We argue that computing the likelihood for N datapoints twice in order to reach a single binary decision is computationally inefficient."
  - "Bayesian posterior sampling via stochastic gradient Fisher scoring" by Ahn et al. (ICML 2012)
    "Can we approximately sample from a Bayesian posterior if we are only allowed to touch a small mini-batch of data-items for every sample we generate?"

Main theme (2): Approximate MCMC

- It is often the case that one has an 'exact' MCMC available which target the posterior of interest.
- However, it is simply not possible to make a transition of this exact chain. Eg, a posterior with an intractable likelihood; a likelihood with too many observations etc.
- This was the topic of Daniel's talk yesterday.
- Big data: the volume of data prohibits calculation of the likelihood.
  - See, eg, "Austerity in MCMC land: cutting the MCMC budget" by Korattikara et al. (arXiv)
    "Can we make Bayesian posterior MCMC sampling more efficient when faced with very large datasets? We argue that computing the likelihood for N datapoints twice in order to reach a single binary decision is computationally inefficient."
  - "Bayesian posterior sampling via stochastic gradient Fisher scoring" by Ahn et al. (ICML 2012)
    "Can we approximately sample from a Bayesian posterior if we are only allowed to touch a small mini-batch of data-items for every sample we generate?"

Main theme (2): Approximate MCMC

- It is often the case that one has an 'exact' MCMC available which target the posterior of interest.
- However, it is simply not possible to make a transition of this exact chain. Eg, a posterior with an intractable likelihood; a likelihood with too many observations etc.
- This was the topic of Daniel's talk yesterday.
- Big data: the volume of data prohibits calculation of the likelihood.
  - See, eg, "Austerity in MCMC land: cutting the MCMC budget" by Korattikara et al. (arXiv)
    "Can we make Bayesian posterior MCMC sampling more efficient when faced with very large datasets? We argue that computing the likelihood for N datapoints twice in order to reach a single binary decision is computationally inefficient."
  - "Bayesian posterior sampling via stochastic gradient Fisher scoring" by Ahn et al. (ICML 2012)
    "Can we approximately sample from a Bayesian posterior if we are only allowed to touch a small mini-batch of data-items for every sample we generate?"

Main theme (2): Approximate MCMC

- Doubly intractable posterior distributions.
  - Gibbs random fields which are widely used in spatial statistics and network analysis, eg, autologistic distribution, exponential random graph model.
  - For this class of models the likelihood can rarely be calculated explicitly.
  - See eg "Playing Russian roulette with intractable likelihoods" by Girolami et al. (arXiv).
    "A fundamental open problem of growing importance in the widespread application of MCMC methods for Bayesian computation is the definition of transition kernels for target distributions with data densities that are analytically or computationally intractable"
- See also, Johndrow et al. (2015, arXiv); Rudolf and Schweizer (2015, ArXiv) and many more...

The exponential random graph (or p^*) model

First proposed by Frank and Strauss (JASA, 1986).

Consider a graph with an adjacency matrix $\{y_{ij}\}$, where $y_{ij} = 1$ denote an edge connecting nodes i and j; otherwise $y_{ij} = 0$.

1. Edges y_{ij} and y_{kl} are neighbours, if they share a common node.

2. If y_{ij} and y_{kl} are not neighbours, then y_{ij} and y_{kl} are conditionally independent, given the rest of the graph.

The exponential random graph (or p^*) model

First proposed by Frank and Strauss (JASA, 1986).

Consider a graph with an adjacency matrix $\{y_{ij}\}$, where $y_{ij} = 1$ denote an edge connecting nodes i and j; otherwise $y_{ij} = 0$.

1. Edges y_{ij} and y_{kl} are neighbours, if they share a common node.

2. If y_{ij} and y_{kl} are not neighbours, then y_{ij} and y_{kl} are conditionally independent, given the rest of the graph.

The exponential random graph (or p^*) model

First proposed by Frank and Strauss (JASA, 1986).

Consider a graph with an adjacency matrix $\{y_{ij}\}$, where $y_{ij} = 1$ denote an edge connecting nodes i and j; otherwise $y_{ij} = 0$.

1. Edges y_{ij} and y_{kl} are neighbours, if they share a common node.

2. If y_{ij} and y_{kl} are not neighbours, then y_{ij} and y_{kl} are conditionally independent, given the rest of the graph.

The exponential random graph model

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

- y observed graph
- $s(y)$ known vector of sufficient statistics
- θ vector of parameters
- $z(\theta)$ normalizing constant

$$z(\theta) = \sum_{\text{all possible graphs}} \exp\{\theta^t s(y)\}$$

- $2^{\binom{n}{2}}$ possible undirected graphs of n nodes
- Calculation of $z(\theta)$ is infeasible for non-trivially small graphs

The exponential random graph model

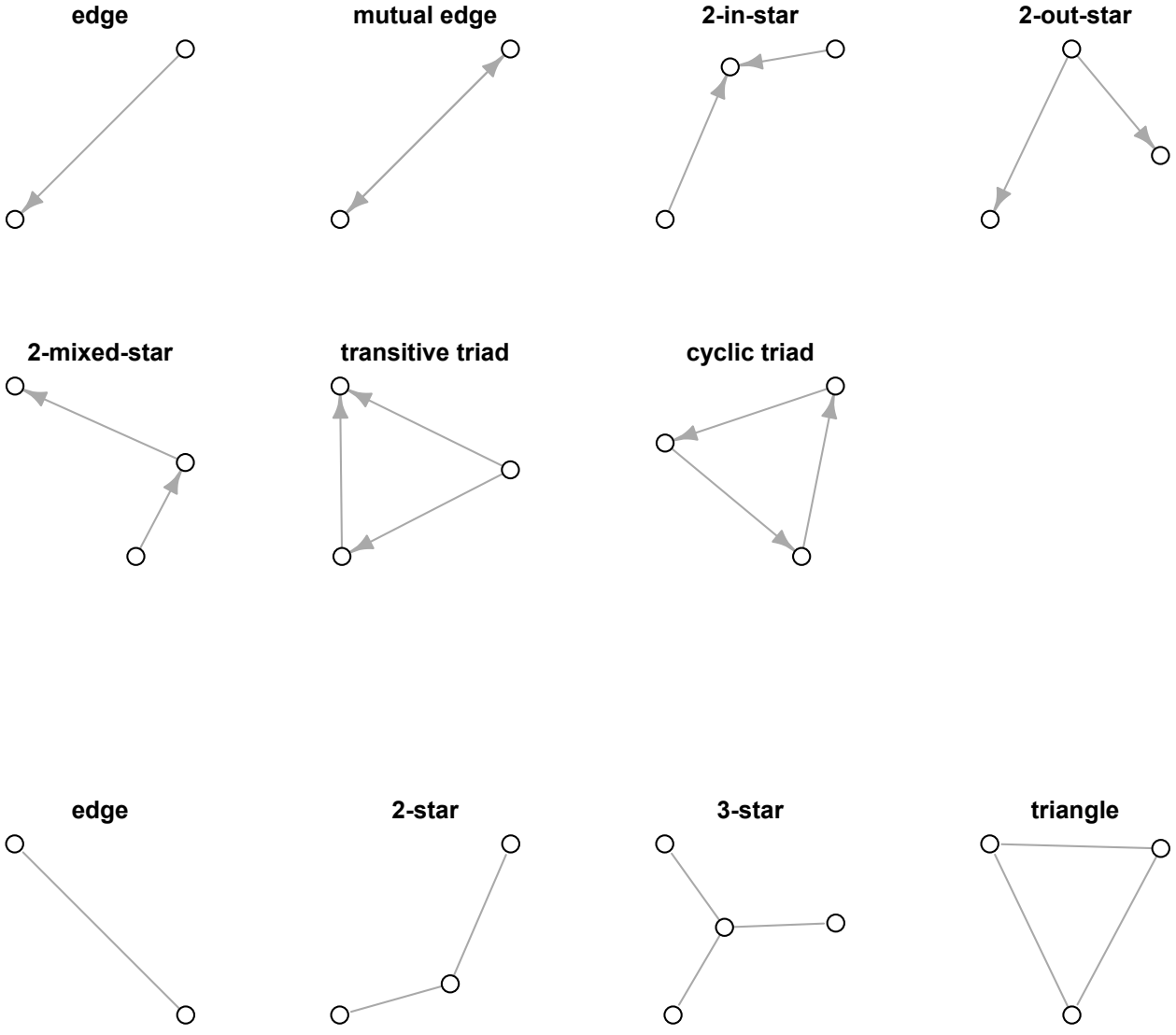
$$f(y|\theta) = \frac{\exp\{\theta^t s(y)\}}{z(\theta)} = \frac{q_\theta(y)}{z(\theta)}$$

- y observed graph
- $s(y)$ known vector of sufficient statistics
- θ vector of parameters
- $z(\theta)$ normalizing constant

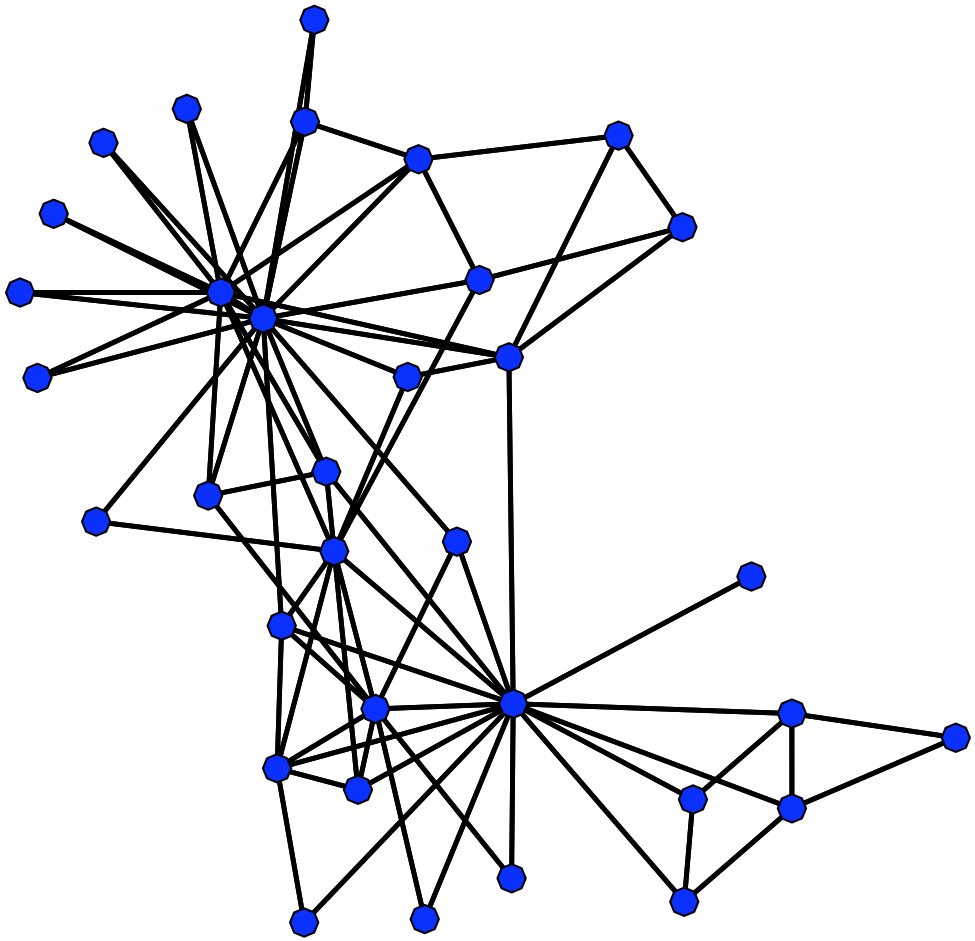
$$z(\theta) = \sum_{\textit{all possible graphs}} \exp\{\theta^t s(y)\}$$

- $2^{\binom{n}{2}}$ possible undirected graphs of n nodes
- Calculation of $z(\theta)$ is infeasible for non-trivially small graphs

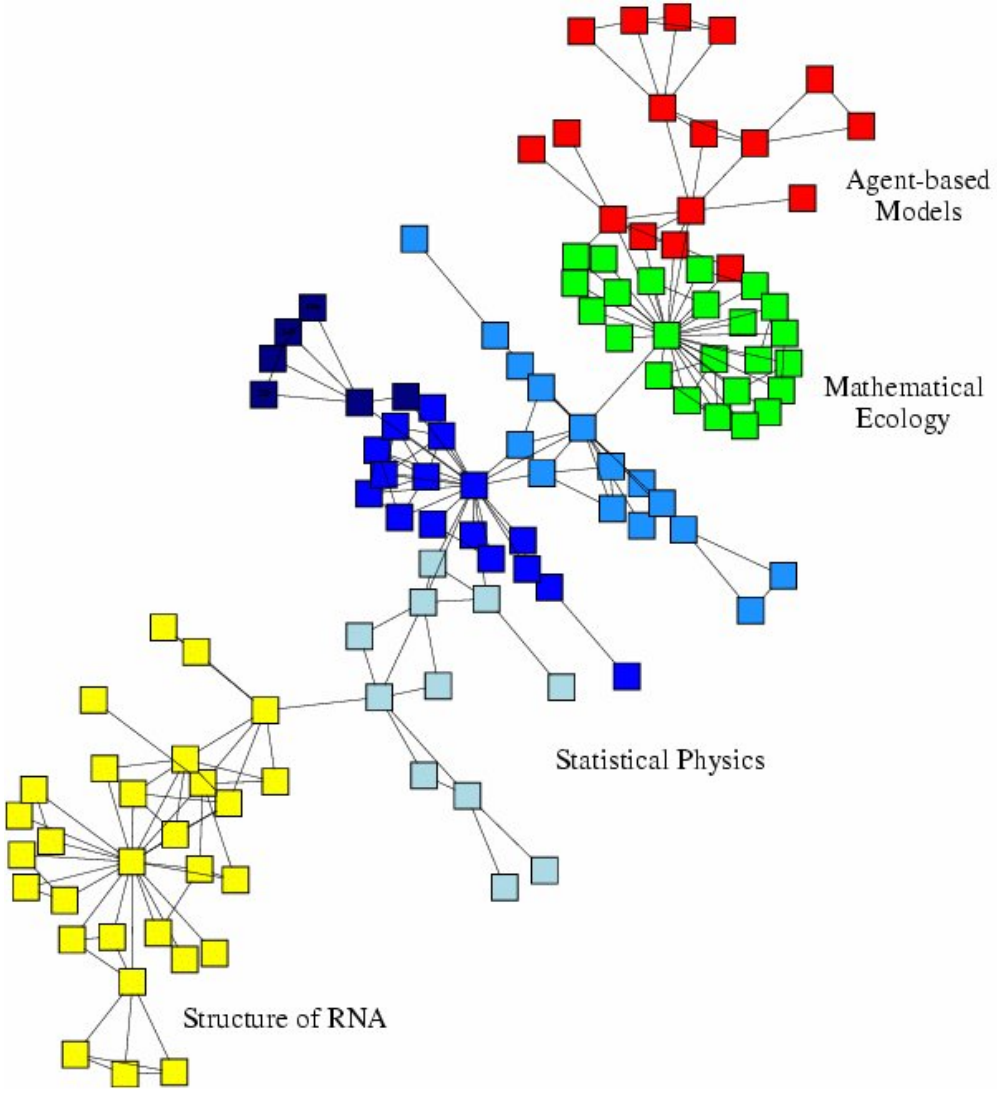
Model Specification: Network Statistics



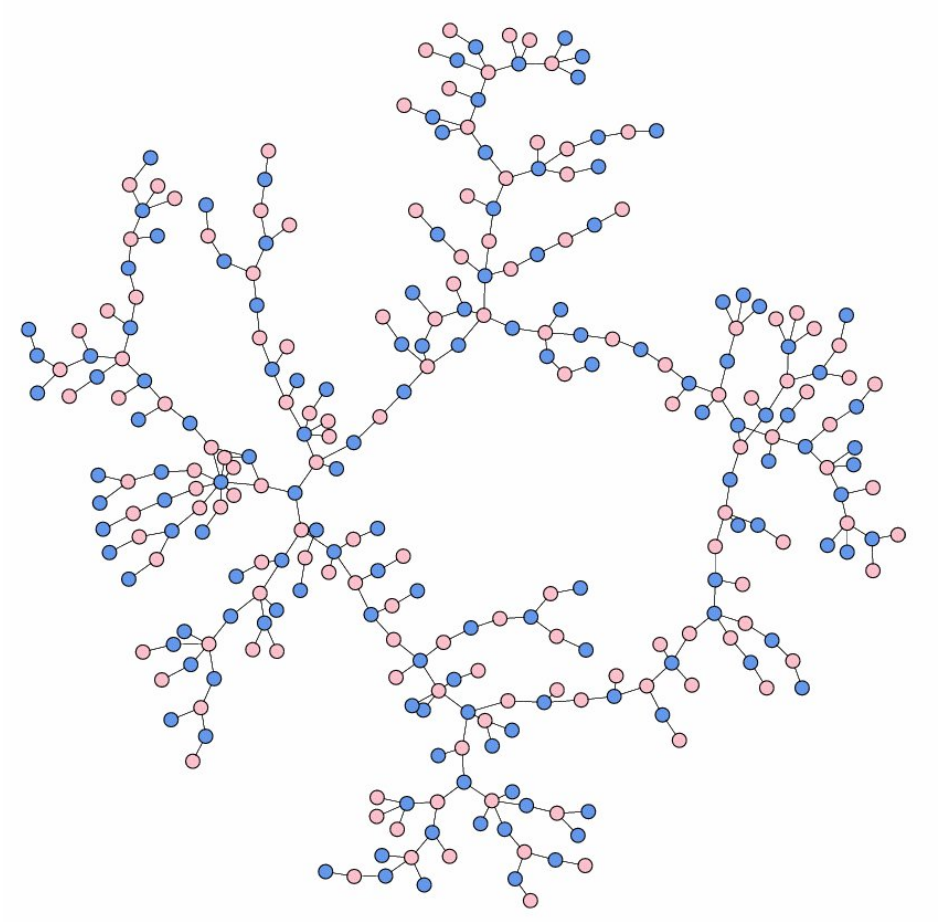
Friendships in a karate club in a US university.



Scientific collaborations



High school dating



Example: Spatial statistics – Ising model

- Defined on a lattice $y = \{y_1, \ldots, y_n\}$.
- Lattice points y_i take values $\{-1, 1\}$.
- $$f(y|\theta) \propto q_\theta(y) = \exp\left\{\frac{1}{2}\theta_1 \sum_{i \sim j} y_i y_j\right\}.$$

  Here $\sim$ means "is a neighbour of".
- The normalising constant

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

is intractable for moderately small n.

Bayesian inference

Doubly-intractable posterior

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

- Naïve Metropolis algorithm proposes the move from θ to θ^* with probability:

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

Exchange algorithm

(Murray, Ghahramani & MacKay 2006)

Sample from an augmented distribution

$$\pi(\theta', y', \theta|y) \propto f(y|\theta)\pi(\theta)h(\theta'|\theta)f(y'|\theta')$$

whose marginal distribution for θ is the posterior of interest.

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

Algorithm 1: Exchange algorithm

1. Draw $\theta' \sim h(\cdot|\theta)$;
2. Draw $y' \sim f(\cdot|\theta')$;
3. With probability

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

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

MCMC sample from the p^* model

- The main difficulty is the need to draw an exact sample $y' \sim f(\cdot|\theta')$.
- Perfect sampling is an obvious approach, if this is possible.
- A pragmatic alternative is to take a realisation from a long MCMC run with stationary distribution $f(y'|\theta')$ as an approximate draw.
- Everitt (2012) showed that, under certain regularity conditions, the corresponding stationary distribution resulting from this approximation is 'close' to the actual target distribution.

Noisy MCMC

Joint with: Pierre Alquier (Paris), Richard Everitt (Reading), Aidan Boland (UCD)

Exchange algorithm

Exchange algorithm:

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

MH algorithm:

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

In fact:

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

Exchange algorithm

Exchange algorithm:

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

MH algorithm:

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

In fact:

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

Exchange algorithm

Exchange algorithm:

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

MH algorithm:

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

In fact:

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

Noisy exchange algorithm

Suppose an 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 exchange. (Approximate)
- $N \to \infty$: MH algorithm. (Exact)

Noisy exchange algorithm

Suppose an 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 exchange. (Approximate)

$N \to \infty$: MH algorithm. (Exact)

Noisy exchange algorithm

Suppose an 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 exchange. (Approximate)
$N \to \infty$:	MH algorithm. (Exact)

Noisy exchange algorithm

Suppose an 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 exchange. (Approximate)
$N \to \infty$:	MH algorithm. (Exact)

Noisy exchange algorithm

Algorithm 2: Noisy exchange algorithm

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

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

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

Noisy Monte Carlo: MCMC with approximate transition kernels

- The noisy exchange algorithm results in a Markov chain which *does not* target $\pi(\theta|y)$.
- Essentially we have replaced an underlying transition kernel P which leaves π invariant with an approximate transition kernel $\hat{P}$.
- Can we say how 'close' this Markov chain with transition $\hat{P}$ is to the exact Markov chain with transition kernel P?
- It turns out that a useful answer is given by the study of the stability of Markov chains, particularly results from Mitrophanov (2005).

Theorem (Mitrophanov (2005), Corollary 3.1)

Let us assume that

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

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

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

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

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

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

Metropolis-Hastings algorithm

Algorithm 3: Metropolis-Hastings algorithm

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

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

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

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 $\alpha(\theta_n, \theta')$ by some function $\hat{\alpha}(\theta_n, \theta', x)$.

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

Algorithm 4: 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 $\theta_{n+1} = \begin{cases} \theta' & \text{with proba. } 1 \wedge \hat{\alpha}(\theta_n, \theta', x') \\ \theta_n & \text{otherwise.} \end{cases}$

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)** $\hat{\alpha}(\theta, \theta', x')$ *satisfies:*

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

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

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

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

Noisy MH: Gibbs random fields

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)** $\hat{\alpha}(\theta, \theta', x')$ *satisfies:*

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

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

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

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:

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

then it results that

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

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

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

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

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

then it results that

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

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

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

Convergence of the noisy exchange algorithm

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

Lemma

Assuming that the state space Θ is bounded, then

$$\mathbb{E}_{y'}\left|\hat{\alpha}(\theta,\theta',y')-\alpha(\theta,\theta')\right| \leq \frac{1}{\sqrt{N}}\frac{h(\theta|\theta')\pi(\theta')q_{\theta'}(y)}{h(\theta'|\theta)\pi(\theta)q_{\theta}(y)}\sqrt{\mathrm{Var}_{y'\sim f(y'|\theta')}\left(\frac{q_{\theta_n}(y')}{q_{\theta'}(y')}\right)} \leq \frac{C}{\sqrt{N}}$$

Theorem

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

where C is explicitly known.

Simulation study

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

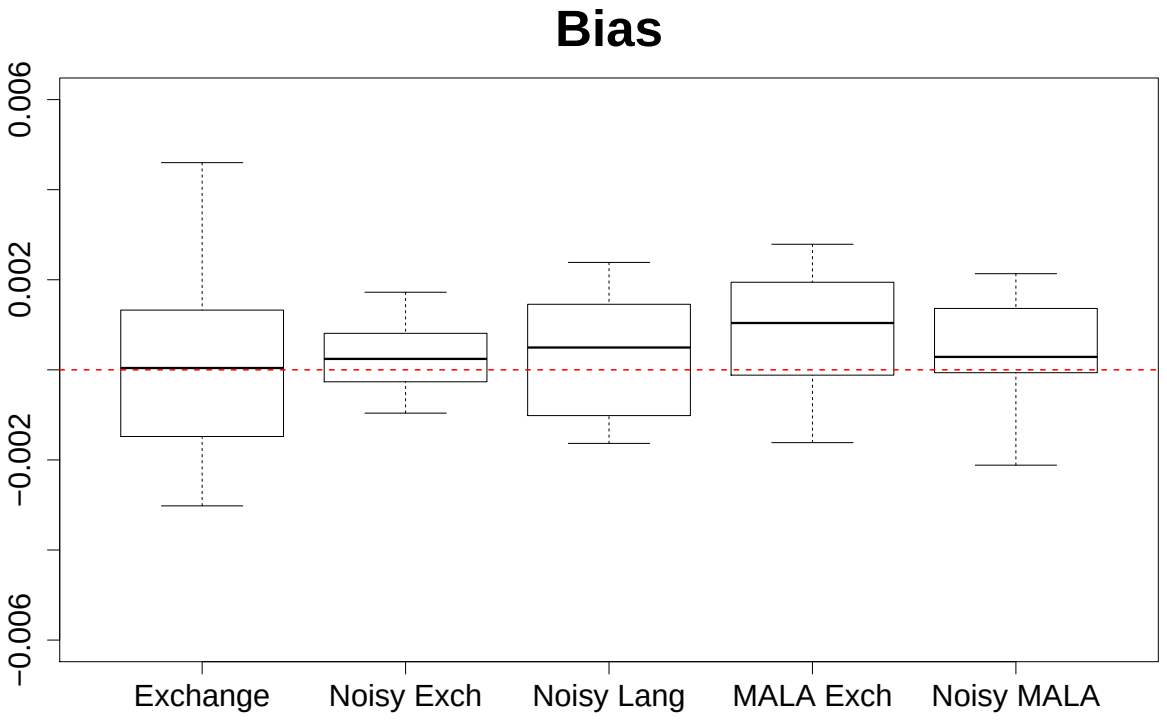
The normalising constant $z(\theta)$ can be calculated exactly for a fine grid of $\{\theta_i : i = 1, \ldots, N\}$ values (NF, Rue, 2007), which can be used to estimate

$$\hat{\pi}(y) = \sum_{i=2}^{N} \frac{(\theta_i - \theta_{i-1})}{2} \left(\frac{q_{\theta_i}(y)}{z(\theta_i)} \pi(\theta_i) + \frac{q_{\theta_{i-1}}(y)}{z(\theta_{i-1})} \pi(\theta_{i-1}) \right),$$

which in turn can be used to estimate of the posterior density at each grid point:

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

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


Bias
0.006
0.002
−0.002
−0.006
Exchange
Noisy Exch
Noisy Lang
MALA Exch
Noisy MALA

Accelerating Bayesian inference for Gibbs random fields

Joint with: Aidan Boland (UCD)

Speeding up inference by pre-computing

- The main bottle-neck of the (noisy) exchange algorithm is the requirement to sample from the likelihood at every iteration.
- It could also be considered inefficient, since the chain is likely return to a previously visited part of the start space to again draw from the likelihood.
- An alternative (and approximate) approach is to pre-compute likelihood draws over a well-chosen grid of parameter values.
- We will shortly show that this too can be places in a noisy MCMC framework.

Pre-processing ABC for image analysis

Moores, Mengersen and Robert (2015)

- Also relies on a pre-computed grid.
- Gives impressive speed-ups.
- However, it doesn't come with any convergence guarantees.

Essential idea:

1. For each grid point θ_i, generate M pseudo-datasets, $y_1, \ldots, y_M$, from the likelihood, giving rise to M summary statistics $s(y_1), \ldots, s(y_M)$.
2. Fit an auxiliary model to this collection of summary statistics, eg, a Gaussian, $\phi(\theta_i)$.
3. Non-parametric regression of $\phi(\theta)$ on θ to smooth the effect of finite M and grid..
4. Apply an on-line SMC-ABC algorithm where instead of drawing from the likelihood, draw from $\phi(\theta)$.

Two useful facts

1. Gradient:

$$\nabla_\theta \log \pi(\theta|y) = s(y) - \mathbf{E}_{y'|\theta} s(y') + \nabla \log \pi(\theta).$$

2. Hessian:

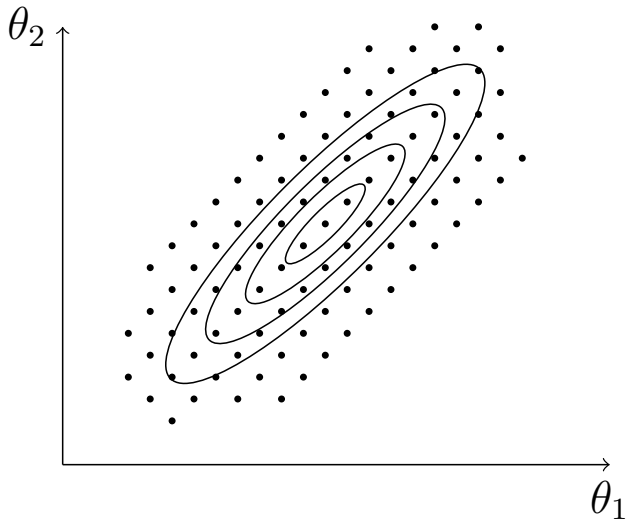
$$H \log \pi(\theta|y) = -Cov_{y'|\theta} s(y') + H \log \pi(\theta).$$

1. and 2. can be estimated via Monte Carlo.

Defining a grid over the posterior

1. Estimate the MAP, θ^*, using a stochastic approximation algorithm, eg, Robbins-Monro (R-M) algorithm.
2. Estimate the Hessian matrix $\mathbf{H}$ at the estimated mode. Let $\Sigma = \mathbf{H}^{-1}$, and the eigendecomposition of $\mathbf{H}^{-1}$ be $\Sigma = V\Lambda V^T$. The standardised variable $\mathbf{z}$ is used to explore the parameter space using,

$$\theta(\mathbf{z}) = \theta^* + V\Lambda^{1/2}\mathbf{z}.$$



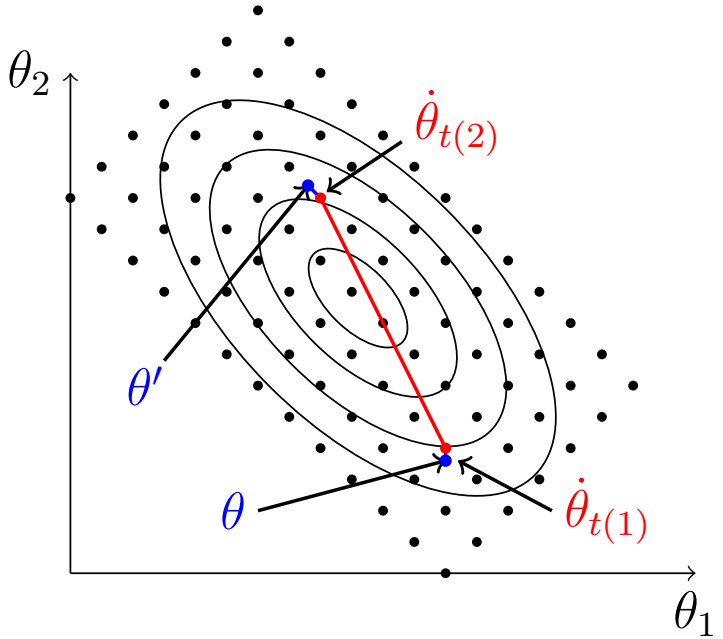
Pre-computing strategy

1. Assume now that a grid of parameter values is selected: $\{\dot{\theta}_1, \ldots, \dot{\theta}_M\}$.

2. At each grid point θ_i we simulate draws $y_i^1, \ldots, y_i^N \sim f(y|\theta_i)$ and calculate the corresponding sufficient statistics $s(y_i^1), \ldots, s(y_i^N)$.

3. Using importance sampling, we can estimate the ratio $z(\dot{\theta}_i)/z(\dot{\theta}_j)$ for any two grid points:

$$\frac{z(\dot{\theta}_i)}{z(\dot{\theta}_j)} \approx \sum_{n=1}^{N} \frac{q_{\dot{\theta}_i}(y_i^n)}{q_{\dot{\theta}_j}(y_j^n)} = \frac{1}{N} \sum_{n=1}^{N} \exp\left\{(\dot{\theta}_i - \dot{\theta}_j)^T s(y_j^n)\right\}$$



In an on-line phase of the algorithm, for *any* two parameters θ, θ':

$$\begin{aligned}\frac{Z(\theta)}{Z(\theta')} &= \frac{Z(\theta)}{Z(\dot{\theta}_{t(1)})} \times \frac{Z(\dot{\theta}_{t(1)})}{Z(\dot{\theta}_{t(2)})} \times \frac{Z(\dot{\theta}_{t(2)})}{Z(\theta')} \\ &\approx \frac{\widehat{Z(\theta)}}{Z(\dot{\theta}_{t(1)})} \times \frac{\widehat{Z(\dot{\theta}_{t(1)})}}{Z(\dot{\theta}_{t(2)})} \times \left(\frac{\widehat{Z(\theta')}}{Z(\dot{\theta}_{t(2)})} \right)^{-1} . \end{aligned} \tag{2}$$

using the off-line pre-computed likelihood draws!

Algorithm 5: Griddy exchange algorithm

Select the grid. (Off-line) ;

Choose a collection of grid points $\{\dot{\theta}_j\}$. ;

For each $\dot{\theta}_j$;

draw $\mathbf{y_j} = \{y^n_{\dot{\theta}_j}\} \sim f(\cdot|\dot{\theta}_j)$;

MCMC sampling using pre-computed grid. (On-line) ;

For $i = 1$ to I;

Draw $\theta' \sim h(\cdot|\theta_i)$;

Set $\theta_{i+1} = \theta'$ with probability:

$$\hat{\alpha}(\theta_i, \theta', \mathbf{y_i}) = \min\left(1, \frac{q_{\theta'}(y)\pi(\theta')h(\theta_i|\theta')}{q_{\theta_i}(y)\pi(\theta_i)h(\theta'|\theta_i)} \times \frac{\widehat{Z(\theta_i)}}{Z(\theta')}\right)$$

Otherwise $\theta_{i+1} = \theta_i$;

Convergence guarantees for the griddy exchange algorithm

Our approach is very similar to the case of noisy exchange algorithm.

Lemma

Assuming that the state space Θ is bounded, our approximate acceptance ratio, $\hat{\alpha}(\theta, \theta', \mathbf{y})$ satisfies

$$\begin{aligned}\mathbb{E}_{\mathbf{y}} \left| \hat{\alpha}(\theta, \theta', \mathbf{y}) - \alpha(\theta, \theta') \right| &\leq \delta(\theta, \theta') \\ &= \frac{C}{\sqrt{N}} \left(\frac{1}{\sqrt{N}} + K^4 \right).\end{aligned}$$

Here K is a constant which depends on the grid-size.

Theorem

$$\sup_{\theta_0 \in \Theta} \| \delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n \|_{TV} \leq \frac{C}{\sqrt{N}} \left(\frac{1}{\sqrt{N}} + K^4 \right)$$

where C is explicitly known.

Ising study

Here 24 lattices of size 80×80 were simulated.

The posterior density was estimated for each graph using a long run (24 hours) of the exchange algorithm.

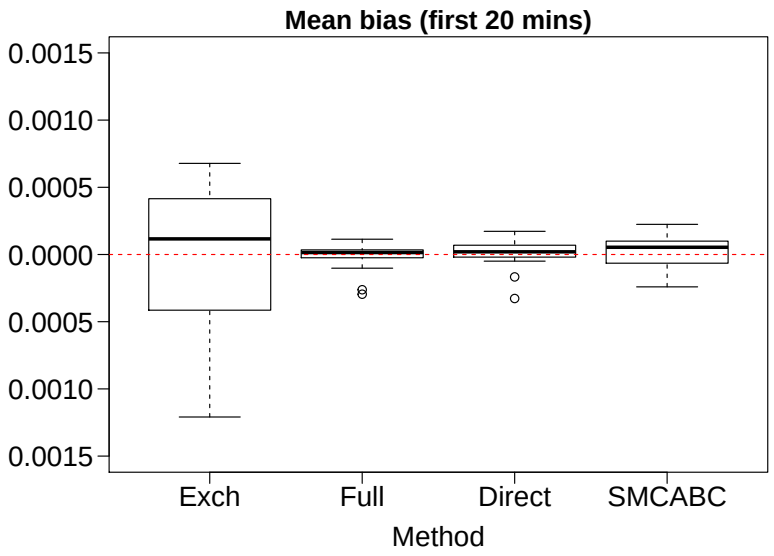
Performance assessment

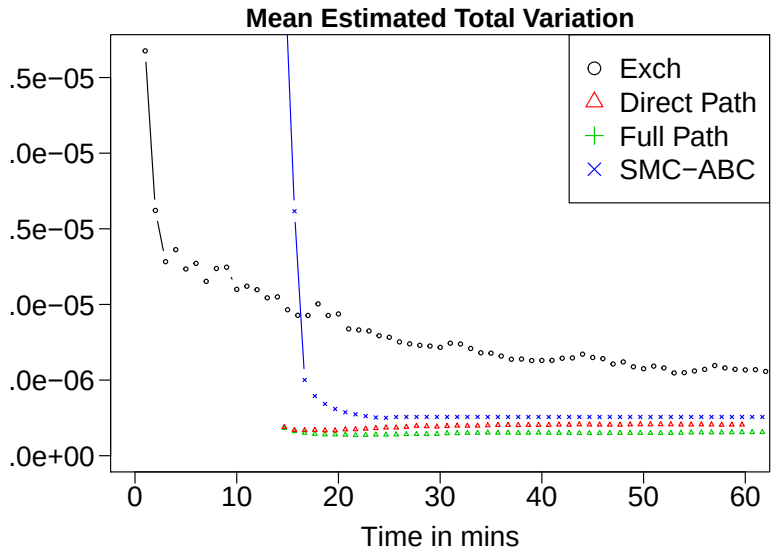
We assess performance of each algorithm in terms of the total variation distance.

$$\|\pi(\theta|y) - \tilde{\pi}(\theta|y)\|_{TV} = \frac{1}{2} \int |\pi(\theta|y) - \tilde{\pi}(\theta|y)| \, d\theta,$$

For two-dimensional targets the total variation distance was approximated by splitting the state-space into bins with a pre-defined window size.

Within each bin the absolute difference of the frequencies was calculated, such that TV takes values in [0,1]





It takes the exchange algorithm 3 hours to reach the same total variation distance as the noisy grid exchange does after 30 minutes.

Autologistic model: Absence/Presence of red deer

The picture represents the presence (red) or absence (black) of deer in a 1km^2 square region in the Grampian region of Scotland.

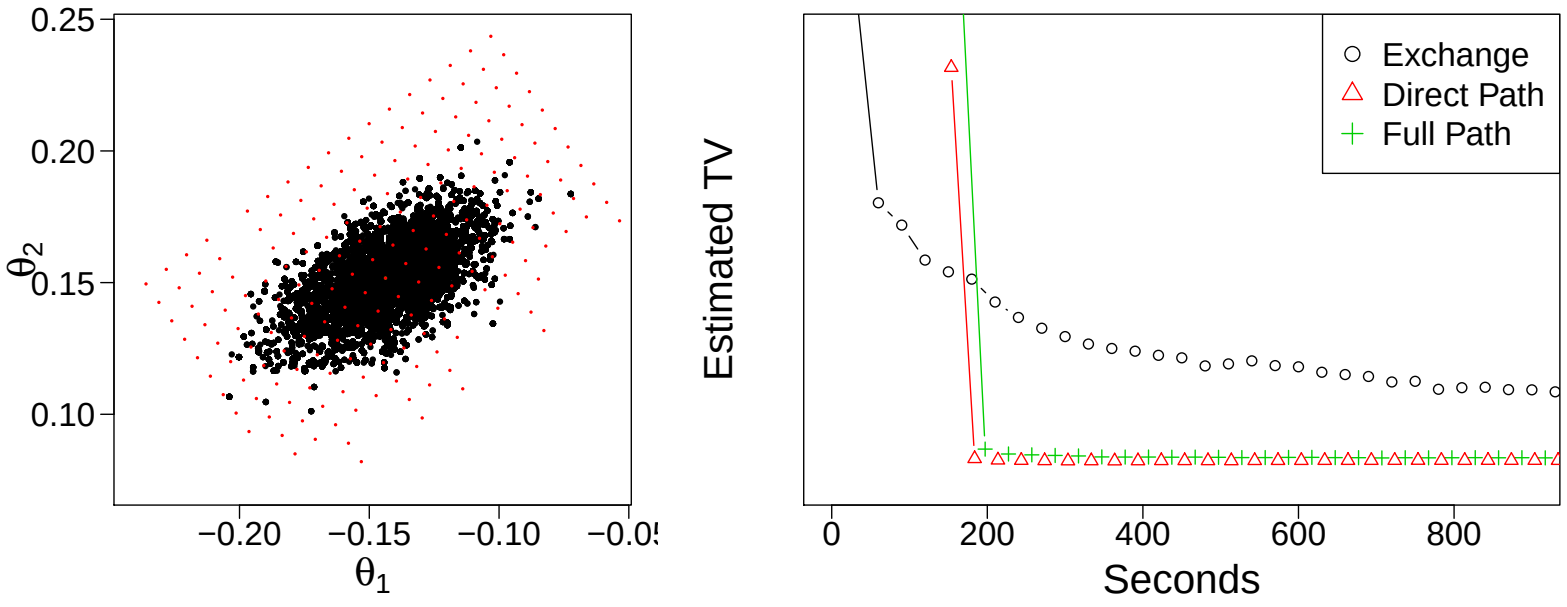
$$f(y|\theta) \propto \exp(\theta_1 s_1(y) + \theta_2 s_2(y)),$$

where $s_1(y) = \sum_{i=1}^{N} y_i$ and $s_2(y) = \sum_{i \sim j} y_i y_j$ with $i \sim j$ denoting node i and node j are neighbours.

The parameter θ_1 controls the relative abundance of -1 and $+1$ values while θ_2 controls the level of spatial aggregation.

	θ_1		θ_2	
	Mean	Var	Mean	Var
Exchange (long)	-0.1435429	0.00028611	0.1516334	0.00016096
Exchange	-0.1424322	0.00026794	0.1530567	0.00014771
Griddy exchange	-0.1436186	0.00028256	0.1515273	0.00016495

Posterior mean and variance estimates for model parameters for the exchange algorithm and griddy exchange for a fixed computational time of 4 minutes.



Note: It takes the exchange algorithm 45 minutes to reach the same total variation distance as the griddy exchange algorithm after 4 minutes.

Future directions: Noisy MCMC

- Our framework give bounds on the total variation distance 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.
- We have further noisy Monte Carlo algorithms, in particular, noisy Langevin algorithm (where we approximate the gradient of the log target).

A quick advertisement!

GiRaF

Gibbs Random Fields: An R package. Stoehr, Pudlo and Friel.

- Normalisation constant calculations.
- Exact sampling.
- And many more to come.

References

- Alquier, Friel, Everitt, Boland. (2016) Noisy MCMC. *Statistics and Computing* (to appear).
- 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