Recent developments in Monte Carlo methods

Richard Everitt

University of Reading

February 27th, 2013

Relevance to data assimilation

- I'm going to take for granted that we want to quantify uncertainty about unknowns, which we represent using random variables.
- We have some parameters θ and a probabilistic model for $l(.|\theta)$ data y given the parameters.
- We want to infer something about θ - this is the starting point for a Bayesian statistician working on any application.
- What is special about data assimilation?
  - $l(.|\theta)$ is usually a computer model?
- Monte Carlo methods are central to solving this type of problem in the presence of non-linearities and/or non-standard distributions, i.e. real situations!

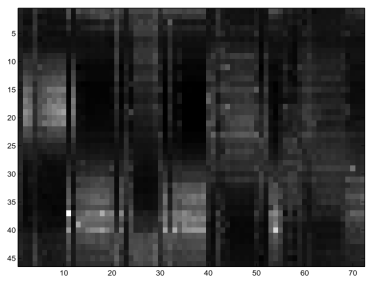
Framework

- We want to estimate parameters θ through observing data y.
- The distribution $l(y|\theta)$ is not directly available, but it is easy to see how the data arises via considering latent variables x:
  - $g(y|x,\theta)$ is available and easy to evaluate.
- Encountered in many different situations, e.g.
  - in data assimilation when the data y is an observed time series, x is a latent time series of which the data are noisy observations and θ are some parameters of either the dynamic model or the measurement model.

Framework

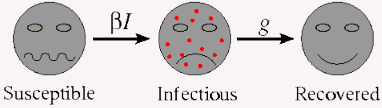
- We want to estimate parameters θ through observing data y.
- The distribution $l(y|\theta)$ is not directly available, but it is easy to see how the data arises via considering latent variables x:
    - $g(y|x, \theta)$ is available and easy to evaluate.
- Encountered in many different situations, e.g.
    - in data assimilation when the data y is an observed time series, x is a latent time series of which the data are noisy observations and θ are some parameters of either the dynamic model or the measurement model.

Noisy images



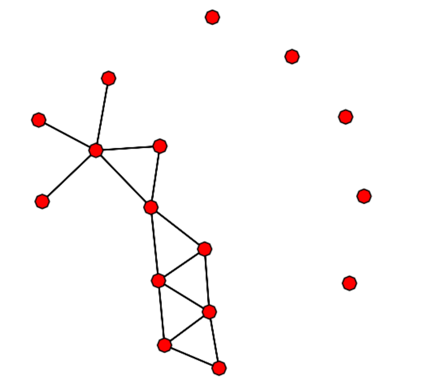
- y is the observed image (log expression of 72 genes on a particular chromosome over 46 hours).
- θ relates to the interactions between genes.
- x is a binary variable for each gene at each time, whose states represent up or down regulation.

Epidemiology



- y is information about the number of individuals infected at a number of discrete time points.
- θ is the infection and recovery rates.
- x are the unobserved times at which individual are actually infected.

Social network



- y are observed connections between actors.
- θ is the degree of transitivity, clustering, etc.
- x are unobserved connections between actors.

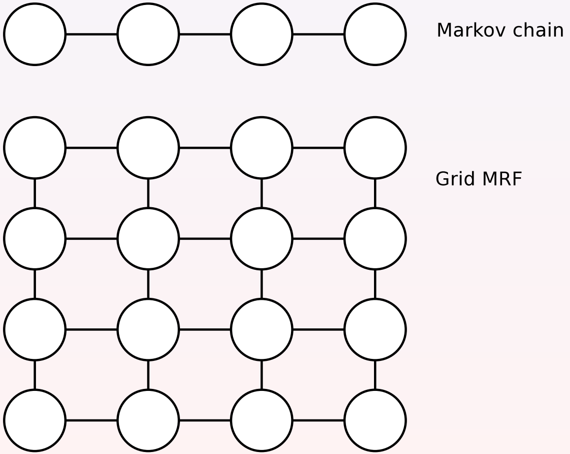
Bayesian framework

- Use a joint distribution on:
  - θ (parameters of the model);
  - x (the unobserved variables);
  - y (the observed variables).
- With factorisation:

$$p(\theta, x, y) = p(\theta) f(x|\theta) g(y|\theta, x). \tag{1}$$

- Use a simple prior for $p(\theta)$ that we can both evaluate and simulate from.

Pairwise Markov random fields



Ising models

- Originally used as a model for ferromagnetism in statistical physics.
- Generalisations (including the *Potts model*) are frequently used in analysing spatially structured data, especially images.
- A pairwise factorisation on a grid, where each variable can take on either the value -1 or 1.
- Each potential is:

$$\Phi(x_i, x_j|\theta_x) = \exp(\theta_x x_i x_j), \tag{2}$$

so that the joint distribution is:

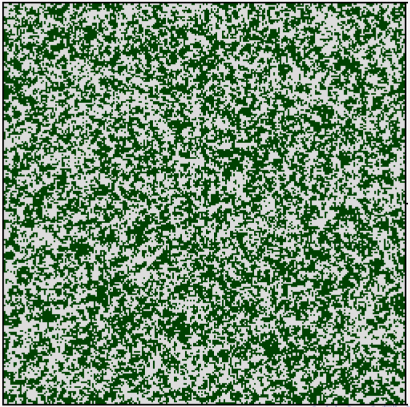
$$f(x|\theta_x) = \frac{1}{Z(\theta_x)} \exp\left(\theta_x \sum_{i,j} (x_{i,j}x_{i,j+1} + x_{i,j}x_{i+1,j})\right). \tag{3}$$

- So a larger parameter results in neighbouring variables being likely to be similar.

Ising models

- Models undergo a phase transition as θ_x increases:

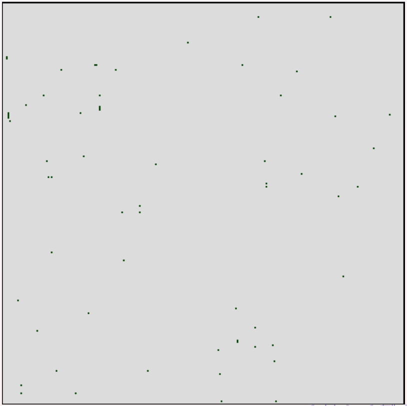
Figure : θ_x just lower than the critical value.



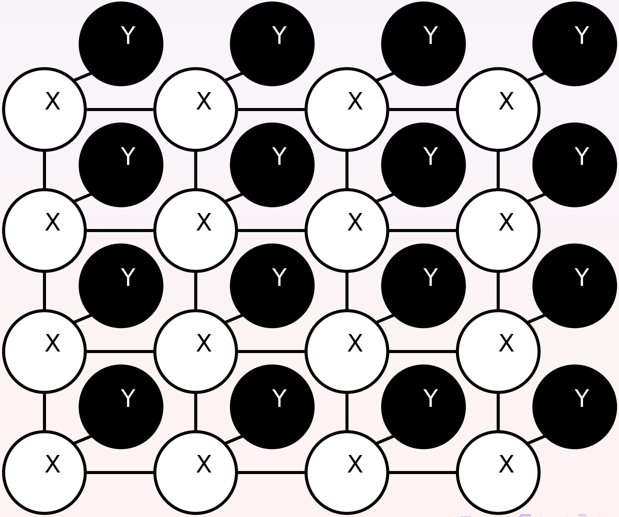
Ising models

- Models undergo a phase transition as θ_x increases:

Figure : θ_x just greater than the critical value.



Latent pairwise Markov random fields



Bayesian parameter estimation

- Observe y and use Bayesian inference:

$$p(\theta|y) \propto \int_x p(\theta)f(x|\theta)g(y|\theta,x)dx.$$

- Common approach is to use MCMC to simulate from:

$$p(\theta,x|y) \propto p(\theta)f(x|\theta)g(y|\theta,x). \tag{4}$$

Bayesian parameter estimation

- Observe y and use Bayesian inference:

$$p(\theta|y) \propto \int_x p(\theta)f(x|\theta)g(y|\theta,x)dx.$$

- Common approach is to use MCMC to simulate from:

$$p(\theta,x|y) \propto p(\theta)f(x|\theta)g(y|\theta,x). \quad (4)$$

The Metropolis-Hastings algorithm

A method for constructing an MCMC algorithm for simulating from a given target $p(\theta, x|y)$.

The Metropolis-Hastings algorithm

Returns a dependent sample $\{(\theta_i, x_i) \mid 1 \leq i \leq N\}$ from $p(\theta, x|y)$.

- For i=1:N
  - Simulate $\theta^*, x^* \sim q(.|\theta_{i-1}, x_{t-1})$
  - Simulate $u \sim \mathscr{U}[0,1]$
  - if $u < \min\left\{1, \frac{p(\theta^*,x^*|y)q(\theta_{i-1},x_{i-1}|\theta^*,x^*)}{p(\theta_{i-1},x_{i-1}|y)q(\theta^*,x^*|\theta_{i-1},x_{i-1})}\right\}$
    - $\theta_i, x_i = \theta^*, x^*$
  - else
    - $\theta_i, x_i = \theta_{i-1}, x_{i-1}$

(Example)

MCMC on multi-dimensional spaces

- When we have a posterior distribution over many variables, the algorithm is the same.
- However, choosing a proposal that moves all variables at once can be difficult.
- Most people would update θ and x separately ("Gibbs"):
  - sample from $p(x|\theta, y)$ using Metropolis-Hastings;
  - sample from $p(\theta|x, y)$ using Metropolis-Hastings.

Three problems

- Every step is a problem!
  1. **Sampling from $p(\theta|x,y)$ can be hard.** Requires the evaluation of an *intractable normalising constant*

$$Z(\theta_x) = \int_x \exp\left(\theta_x \sum_{i,j} (x_{i,j}x_{i,j+1} + x_{i,j}x_{i+1,j})\right) dx.$$

  2. **Sampling from $p(x|\theta,y)$ is hard.** A density on a large, complicated space.
  3. **Updating like this may be a bad idea anyway!** If x and θ are quite dependent in the posterior, the sampler will be poor.

- Problem 1 can be addressed by using the "exchange algorithm" (Murray et al., 2006)
  - requires exact simulation from $f(x|\theta)$.

Three problems

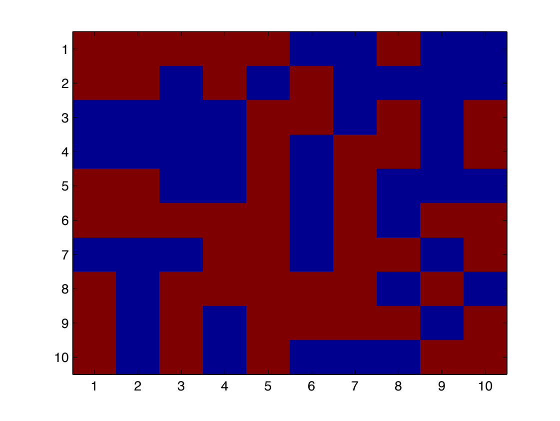
- Every step is a problem!
  1. **Sampling from $p(\theta|x,y)$ can be hard.** Requires the evaluation of an *intractable normalising constant*

$$Z(\theta_x) = \int_x \exp\left(\theta_x \sum_{i,j} (x_{i,j}x_{i,j+1} + x_{i,j}x_{i+1,j})\right) dx.$$

  2. **Sampling from $p(x|\theta,y)$ is hard.** A density on a large, complicated space.
  3. **Updating like this may be a bad idea anyway!** If x and θ are quite dependent in the posterior, the sampler will be poor.

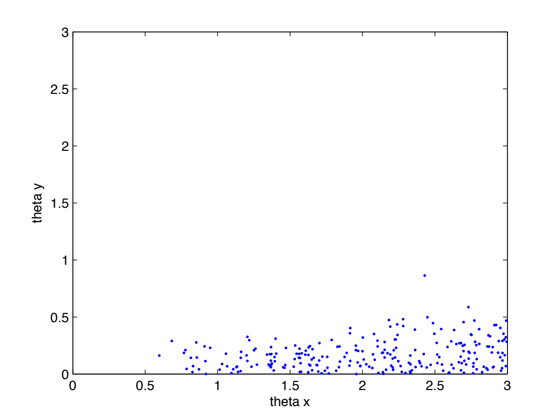
- Problem 1 can be addressed by using the "exchange algorithm" (Murray et al., 2006)
  - requires exact simulation from $f(x|\theta)$.

Example: Ising model data ($\theta_x = 0.1$, $\theta_y = 0.1$)



Example: Ising model using Gibbs

Figure : Points from the posterior using Gibbs.



Outline

Outline

Outline

Outline

What is ABC?

- Directly approximate a complicated or intractable likelihood with:

$$l_\varepsilon(y|\theta) = \int_{y'} l(y'|\theta)\pi_\varepsilon(y'|y)\mathrm{d}y' \approx \frac{1}{R}\sum_{r=1}^{R}\pi_\varepsilon(y'^{(r)}|y)$$

where $y'^{(r)} \sim l(.|\theta)$.

- In the original work $R=1$ and $\pi_\varepsilon(S_{y'^{(r)}}|S_y) \propto \delta\left(\left|S_{y'^{(r)}} - S_y\right| < \varepsilon\right)$.
- Can use rejection sampling, importance sampling, MCMC or SMC samplers to simulate from this approximate posterior.

What is ABC?

- Directly approximate a complicated or intractable likelihood with:

$$l_\varepsilon(y|\theta) = \int_{y'} l(y'|\theta)\pi_\varepsilon(y'|y)\mathrm{d}y' \approx \frac{1}{R}\sum_{r=1}^{R} \pi_\varepsilon(y'^{(r)}|y)$$

  where $y'^{(r)} \sim l(.|\theta)$.

- In the original work $R = 1$ and $\pi_\varepsilon(S_{y'^{(r)}}|S_y) \propto \delta\left(\left|S_{y'^{(r)}} - S_y\right| < \varepsilon\right)$.

- Can use rejection sampling, importance sampling, MCMC or SMC samplers to simulate from this approximate posterior.

What is ABC?

- Directly approximate a complicated or intractable likelihood with:

$$l_{\varepsilon}(y|\theta) = \int_{y'} l(y'|\theta)\pi_{\varepsilon}(y'|y)\mathrm{d}y' \approx \frac{1}{R}\sum_{r=1}^{R}\pi_{\varepsilon}(y'^{(r)}|y)$$

where $y'^{(r)} \sim l(.|\theta)$.

- In the original work $R = 1$ and $\pi_{\varepsilon}(S_{y'^{(r)}}|S_y) \propto \delta\left(\left|S_{y'^{(r)}} - S_y\right| < \varepsilon\right)$.
- Can use rejection sampling, importance sampling, MCMC or SMC samplers to simulate from this approximate posterior.

Applied to Ising models

- For our Ising model example:
  - $x^*|\theta^* \sim f(.|\theta^*)$;
  - $y^*|x^*, \theta^* \sim g(.|\theta^*, x^*)$;
  - compare S_{y^*} to S_y.
- Statistics of the data:
  - $S_y^1 = \sum_{(i,j)\in \mathbf{N}} y_i y_j$ (the number of neighbours in the same state);
  - $S_y^2 = \sum_i y_i$ (the *magnetisation*).

Applied to Ising models

- For our Ising model example:
  - $x^*|\theta^* \sim f(.|\theta^*)$;
  - $y^*|x^*, \theta^* \sim g(.|\theta^*, x^*)$;
  - compare S_{y^*} to S_y.
- Statistics of the data:
  - $S_y^1 = \sum_{(i,j)\in\mathbf{N}} y_i y_j$ (the number of neighbours in the same state);
  - $S_y^2 = \sum_i y_i$ (the *magnetisation*).

Are our problems solved?

1. **Intractable normalising constant when sampling from** $p(\theta|x,y)$: yes! (Grelaud et al., 2009)
2. **Sampling from** $p(x|\theta,y)$ **is hard:** yes!
3. **Posterior dependance between** x **and** θ**:** yes!

However:

- Several approximations are introduced.
- Inefficient when $l(.|\theta)$ is "vague".
- Sampling from $f(x|\theta)$ is difficult for MRFs, so problem 2 is not really avoided.

Are our problems solved?

1. **Intractable normalising constant when sampling from** $p(\theta|x,y)$**:** yes! (Grelaud et al., 2009)
2. **Sampling from** $p(x|\theta,y)$ **is hard:** yes!
3. **Posterior dependance between** x **and** θ**:** yes!

However:

- Several approximations are introduced.
- Inefficient when $l(.|\theta)$ is "vague".
- Sampling from $f(x|\theta)$ is difficult for MRFs, so problem 2 is not really avoided.

Are our problems solved?

1. **Intractable normalising constant when sampling from** $p(\theta|x, y)$**:** yes! (Grelaud et al., 2009)
2. **Sampling from** $p(x|\theta, y)$ **is hard:** yes!
3. **Posterior dependance between** x **and** θ**:** yes!

However:

- Several approximations are introduced.
- Inefficient when $l(.|\theta)$ is "vague".
- Sampling from $f(x|\theta)$ is difficult for MRFs, so problem 2 is not really avoided.

Are our problems solved?

1. **Intractable normalising constant when sampling from** $p(\theta|x,y)$: yes! (Grelaud et al., 2009)
2. **Sampling from** $p(x|\theta,y)$ **is hard:** yes!
3. **Posterior dependance between** x **and** θ: yes!

However:

- Several approximations are introduced.
- Inefficient when $l(.|\theta)$ is "vague".
- Sampling from $f(x|\theta)$ is difficult for MRFs, so problem 2 is not really avoided.

Are our problems solved?

1. **Intractable normalising constant when sampling from** $p(\theta|x, y)$: yes! (Grelaud et al., 2009)
2. **Sampling from** $p(x|\theta, y)$ **is hard:** yes!
3. **Posterior dependance between** x **and** θ: yes!

However:

- Several approximations are introduced.
- Inefficient when $l(.|\theta)$ is "vague".
- Sampling from $f(x|\theta)$ is difficult for MRFs, so problem 2 is not really avoided.

"Approximate ABC"

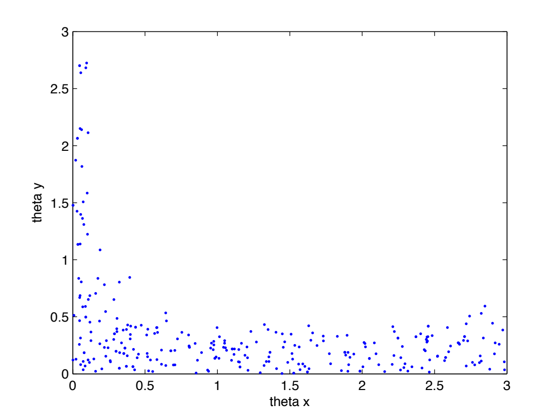
- Grelaud et al. (2009) use MCMC to sample from $f(x|\theta)$ for MRFs - introduces a further approximation.
- Let K be the MCMC kernel targeting the ABC posterior (if $f(x|\theta)$ could be simulated from exactly), L be the MCMC kernel actually used to sample from $f(x|\theta)$. If:
  - K is uniformly ergodic;
  - L is geometrically ergodic.
- Then:
  - the approximate ABC posterior gets closer to the true ABC posterior the more iterations of L are run;
  - the MCMC kernel K_L targeting the approximate ABC posterior is uniformly ergodic.
- The same result can be used to justify the "approximate exchange algorithm".

"Approximate ABC"

- Grelaud et al. (2009) use MCMC to sample from $f(x|\theta)$ for MRFs - introduces a further approximation.
- Let K be the MCMC kernel targeting the ABC posterior (if $f(x|\theta)$ could be simulated from exactly), L be the MCMC kernel actually used to sample from $f(x|\theta)$. If:
  - K is uniformly ergodic;
  - L is geometrically ergodic.
- Then:
  - the approximate ABC posterior gets closer to the true ABC posterior the more iterations of L are run;
  - the MCMC kernel K_L targeting the approximate ABC posterior is uniformly ergodic.
- The same result can be used to justify the "approximate exchange algorithm".

Example: Ising model posterior using ABC

Figure : Points from the posterior of θ_x and θ_y.



Pseudo-marginal approach

- Ideally, we would target $p(\theta|y)$.
- Beaumont (2003) and Andrieu and Roberts (2009) describe the idea of targeting instead an importance sampling approximation to this idealised situation:

$$\widetilde{p}^{N}(\theta|y) = \frac{1}{N}\sum_{k=1}^{N}\frac{p(\theta, x^{(k)}|y)}{q(x^{(k)}|\theta)}, \tag{5}$$

  where $x^{(k)} \sim q(.|\theta)$.
- In general, an MCMC algorithm that targets an unbiased estimator of $p(\theta|y)$ will give points from $p(\theta|y)$ itself.

Example: Ising model using the pseudo-marginal approach

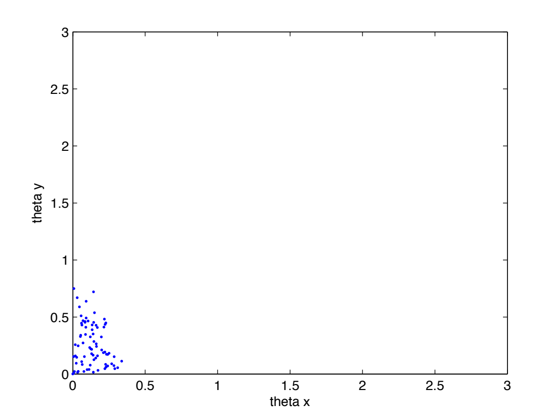


Figure : Points from the posterior using the pseudo-marginal approach.

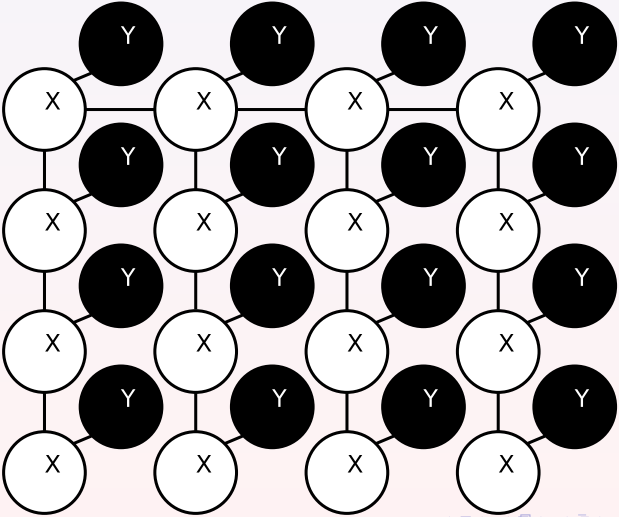
SMC samplers

- SMC sampler:
  - choose a sequence of target distributions $\pi_1, \ldots, \pi_T$, where π_1 is easy to sample from, π_T is the distribution of interest and π_{t+1} is not too different from π_t;
  - perform importance sampling sequentially on this sequence of targets, using a kernel to move the points at each step.

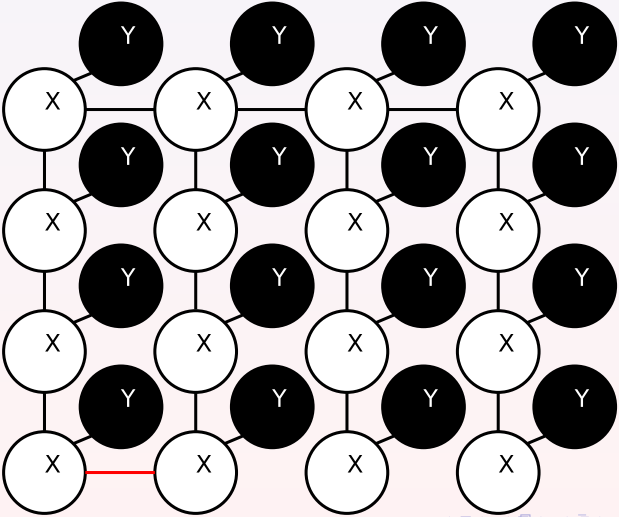
SMC samplers for Ising models

- Begin with $\pi_1 = \gamma_{\text{tree}}(x|\theta, y)$.
  - can be sampled from exactly, and the normalising constant can be calculated exactly.
- Add an arc to make each new target, with the final target being a grid (known as "hot coupling" Hamze and De Freitas, 2004).

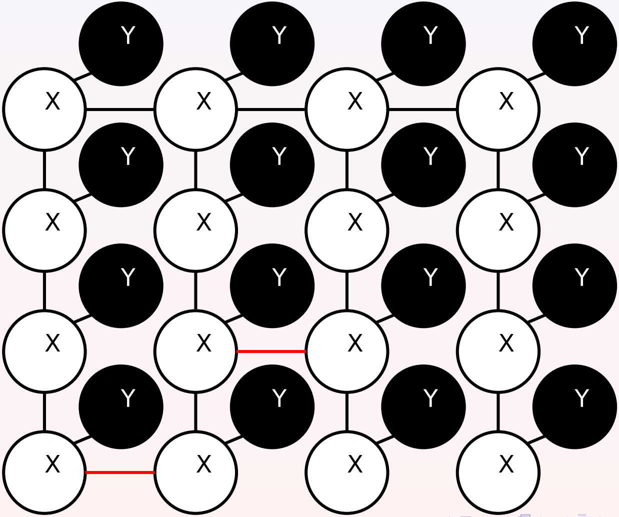
Hot coupling



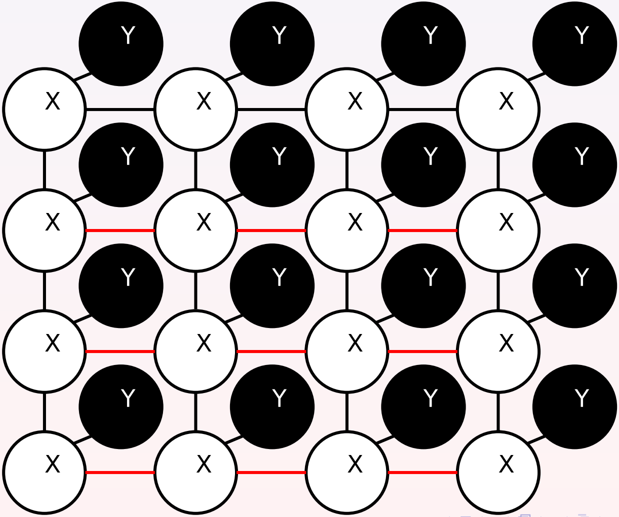
Hot coupling



Hot coupling



Hot coupling



Particle MCMC

- Sequential Monte Carlo (SMC) samplers are particularly suited to sampling from some spaces.
- Andrieu et al. (2010) formalise the idea of using an SMC sampler as a proposal within an MCMC algorithm - known as *particle MCMC:*
    - simulate $\theta^* \sim q(.|\theta)$;
    - run an SMC sampler targeting $p(x|y,\theta^*)$ to find approximations $\widehat{p}(x|y,\theta^*)$ to $p(x|y,\theta^*)$ and $\widehat{\phi}(y,\theta^*)$ to the normalising constant $\int_x p(x|y,\theta^*)dx$;
    - simulate $x^* \sim \widehat{p}(x|y,\theta^*)$ and accept (θ^*, x^*) with probability:

$$1 \wedge \frac{p(\theta^*)}{p(\theta)} \frac{\widehat{\phi}(\theta^*, y)}{\widehat{\phi}(\theta, y)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}. \tag{6}$$

Does this solve our problems?

1. **Intractable normalising constant when sampling from $p(\theta|x,y)$:** require merging PMCMC with the exchange algorithm.
2. **Sampling from $p(x|\theta,y)$ is hard:** SMC samplers can help a lot.
3. **Posterior dependance between x and θ:** no longer an issue.

However:

- PMCMC can be computationally expensive.

Does this solve our problems?

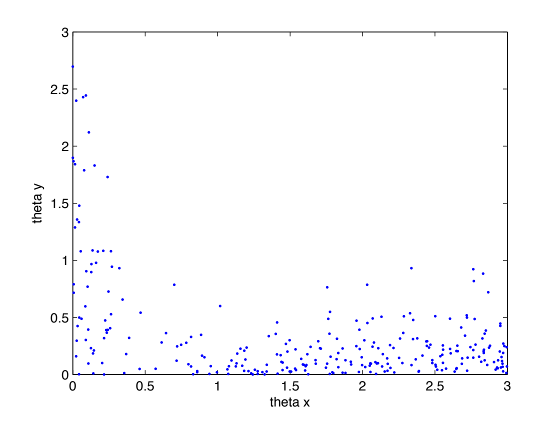
1. **Intractable normalising constant when sampling from $p(\theta|x, y)$:** require merging PMCMC with the exchange algorithm.
2. **Sampling from $p(x|\theta, y)$ is hard:** SMC samplers can help a lot.
3. **Posterior dependance between x and θ:** no longer an issue.

However:

- PMCMC can be computationally expensive.

Example: Ising model using PMCMC

Figure : Points from the posterior using the PMCMC.



Discussion

- Have considered two alternatives to the standard approach.
- ABC:
  - superficially easy to use;
  - justification of use of MCMC for simulating from $l(.|\theta)$;
  - approximations can be hard to quantify.
- PMCMC:
  - targets the correct distribution (almost!);
  - requires the design of an effective SMC sampler;
  - would benefit from parallelisation.

Discussion

- Have considered two alternatives to the standard approach.
- ABC:
  - superficially easy to use;
  - justification of use of MCMC for simulating from $l(.|\theta)$;
  - approximations can be hard to quantify.
- PMCMC:
  - targets the correct distribution (almost!);
  - requires the design of an effective SMC sampler;
  - would benefit from parallelisation.

Discussion

- Have considered two alternatives to the standard approach.
- ABC:
  - superficially easy to use;
  - justification of use of MCMC for simulating from $l(.|\theta)$;
  - approximations can be hard to quantify.
- PMCMC:
  - targets the correct distribution (almost!);
  - requires the design of an effective SMC sampler;
  - would benefit from parallelisation.

Paper and acknowledgements

- Everitt, R. G. (2012) Bayesian parameter estimation for latent Markov random fields and social networks, JCGS.
  - includes full description of exchange PMCMC algorithm;
  - additional application to exponential random graphs (social networks);
  - proof of result about approximate algorithms.
- Thanks to Christophe Andrieu, SuSTaIn at the University of Bristol, and the University of Oxford.
- Also, for more on ABC, see Didelot, X., Everitt, R. G., Johansen, A. M. and Lawson, D. J. (2011) Likelihood-free estimation of model evidence, Bayesian Analysis.