ROSLIN

Bayesian Inference using MCMC: An introduction

Osvaldo Anacleto
Genetics and Genomics, Roslin Institute
osvaldo.anacleto@roslin.ed.ac.uk

THE UNIVERSITY *of* EDINBURGH

BBSRC
20 Years of Pioneering
Great British Bioscience

Dealing with intractable posteriors

- it can be very difficult to calculate point and interval estimates depending on the density of the posterior distribution
- in this case, **stochastic (random) simulation methods** are required
- stochastic simulation provides approximate solutions to problems considered far too difficult to solve directly

stochastic simulation

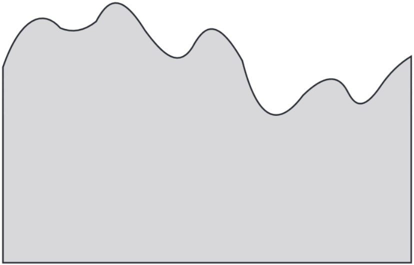
- To develop and study a **random** experiment that mimics a complex system too difficult to deal with
- Example: Monte Carlo simulation

The Monte Carlo method
A simple example

How to calculate the grey area under the curve?

Calculus can be applied to calculate the area (integration)

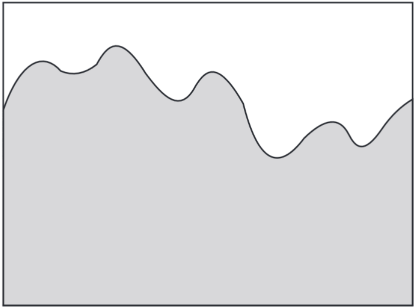
A Monte carlo alternative:



The Monte Carlo method
A simple example

How to calculate the grey area under the curve?

Calculus can be applied to calculate the area (integration)



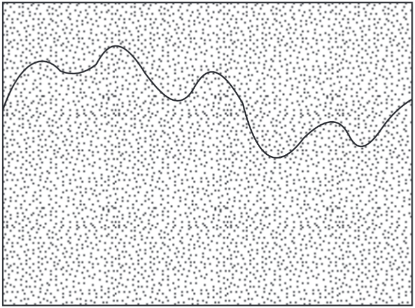
A Monte carlo alternative:

1. surround the area under the curve with a rectangle (with area A)

The Monte Carlo method
A simple example

How to calculate the grey area under the curve?

Calculus can be applied to calculate the area (integration)



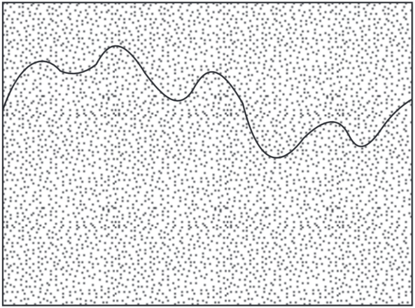
A Monte carlo alternative:

1. surround the area under the curve with a rectangle (with area *A*)
2. simulate a **large** number of points at **random positions** within the rectangle (number of points=N)

The Monte Carlo method
A simple example

How to calculate the grey area under the curve?

Calculus can be applied to calculate the area (integration)



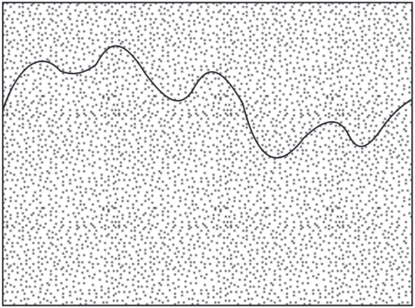
A Monte carlo alternative:

1. surround the area under the curve with a rectangle (with area *A*)
2. simulate a **large** number of points at **random positions** within the rectangle (number of points=N)
3. calculate the proportion (p) of points lying under the curve

The Monte Carlo method
A simple example

How to calculate the grey area under the curve?

Calculus can be applied to calculate the area (integration)



A Monte carlo alternative:

1. surround the area under the curve with a rectangle (with area *A*)
2. simulate a **large** number of points at **random positions** within the rectangle (number of points=N)
3. calculate the proportion (p) of points lying under the curve

- **Monte Carlo estimate for the area under the curve =** $p * A$

How to apply the Monte Carlo method in Bayesian Statistics?

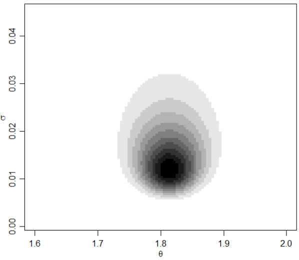
the problem: Conjugate Bayesian analysis is usually not possible for complex models

However, a **sample from the posterior distribution** can be used to make inferences about the parameters (e.g. by calculating means, modes, medians and quantiles from the Monte carlo sample)

The Monte Carlo method in Bayesian Statistics: An example

In the previous normal-model (Lecture 10), we could have used the following algorithm to sample values from the posterior distribution of mean and the variance (θ, σ^2) of a random variable:

```
do k=1, M
  sample $\sigma^{2(k)}$ from $\sigma^2|\text{data} \sim \text{inverse-gamma}(\nu_n/2, \nu_n\sigma_n^2/2)$
  sample $\theta^{(k)}$ from $\theta|\sigma^{2(k)}, \text{data} \sim \text{N}(\mu_n, \sigma^2/\kappa_n)$
end do
```



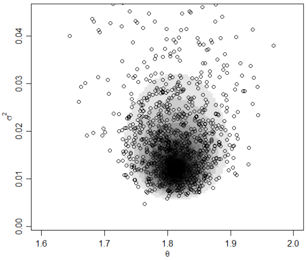
- $(\theta^{(1)}, \sigma^{2(1)}), \ldots, (\theta^{(M)}, \sigma^{2(M)})$ is a sample from the posterior distribution of (θ, σ^2) (the dots in the plot)
- $\theta^{(1)} \ldots, \theta^{(M)}$ is a sample of the marginal posterior distribution of θ given only the data

In this case, sampling from the posterior is very easy

The Monte Carlo method in Bayesian Statistics: An example

In the previous normal-model (Lecture 10), we could have used the following algorithm to sample values from the posterior distribution of mean and the variance (θ, σ^2) of a random variable:

```
do k=1, M
  • sample $\sigma^{2(k)}$ from $\sigma^2|\mathrm{data} \sim \mathrm{inverse\text{-}gamma}(\nu_n/2, \nu_n\sigma_n^2/2)$
  • sample $\theta^{(k)}$ from $\theta|\sigma^{2(k)}, \mathrm{data} \sim \mathrm{N}(\mu_n, \sigma^2/\kappa_n)$
end do
```



- $(\theta^{(1)}, \sigma^{2(1)}), \ldots, (\theta^{(M)}, \sigma^{2(M)})$ is a sample from the posterior distribution of (θ, σ^2) (the dots in the plot)
- $\theta^{(1)} \ldots, \theta^{(M)}$ is a sample of the marginal posterior distribution of θ given only the data

In this case, sampling from the posterior is very easy

What if we can’t sample from the posterior?

- One way to generate Monte Carlo samples from the posterior is to use a **Markov chain** which is related to the posterior distribution

- A Markov chain is a **sequence** of random variables (a stochastic process) in which the distribution of the next random variable depends only on the value of the current one (i.e. independent of the past)

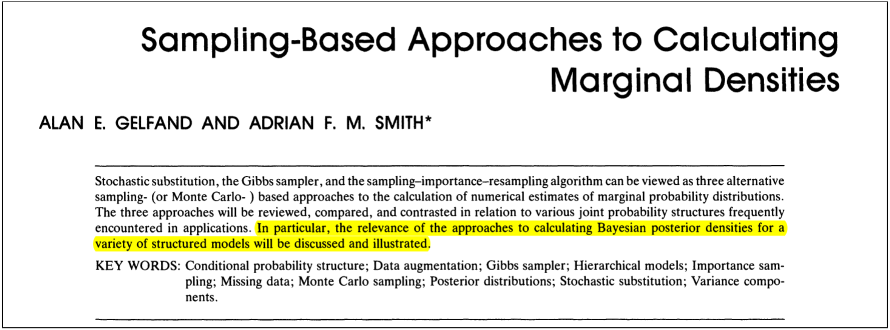
Monte Carlo method + Markov chain = **Markov chain Monte Carlo** (MCMC)

seminal MCMC methods were developed and applied well before their use in Bayesian Statistics. (see references in Gamerman and Lopes, 2006)
Early MCMC applications: Physics and Chemistry, Spatial statistics and missing data imputation.

Tutorial 11: Markov Chains

Explained Visually:
http://setosa.io/ev/markov-chains/

The paper that revolutionised Bayesian Statistics



Sampling-Based Approaches to Calculating Marginal Densities

ALAN E. GELFAND AND ADRIAN F. M. SMITH*

Stochastic substitution, the Gibbs sampler, and the sampling–importance–resampling algorithm can be viewed as three alternative sampling- (or Monte Carlo-) based approaches to the calculation of numerical estimates of marginal probability distributions. The three approaches will be reviewed, compared, and contrasted in relation to various joint probability structures frequently encountered in applications. In particular, the relevance of the approaches to calculating Bayesian posterior densities for a variety of structured models will be discussed and illustrated.

KEY WORDS: Conditional probability structure; Data augmentation; Gibbs sampler; Hierarchical models; Importance sampling; Missing data; Monte Carlo sampling; Posterior distributions; Stochastic substitution; Variance components.

Gelfand and Smith (1990) showed how the MCMC can be used to obtain samples from posterior distributions

MCMC: the basics

- In the previous examples (website) the Markov Chains were simulated by setting a transition matrix
- If a Markov chain satisfies some properties, its values follows a stationary (or equilibrium) distribution after sampling many times from it
- A Markov chain can be built such that its stationary distribution is the **posterior distribution** of interest

How to approximate posterior distributions using MCMC?

1. Set up a Markov chain that has the posterior distribution as its stationary distribution
2. Sample from the Markov chain
3. Use the sampled values to make inferences about the unknown quantities of interest (the parameters)

there are **several** MC methods for approximating posteriors: importance sampling, Gibbs sampling, Metropolis-Hastings, reversible jump MCMC, slice sampling, hybrid MC, sequential MC, INLA, ...

Gibbs Sampling: an example

- in the conjugate normal-normal model (Lecture 10), the prior for the mean θ depended on the variance σ^2
- this dependence may not always hold
  - a non-informative prior for θ may not be possible when σ^2 is very small
- to specify the prior uncertainty about θ **independently** of σ^2, we need

$$f(\theta, \sigma^2) = f(\theta) f(\sigma^2)$$

- for example $\theta \sim N(\mu_0, \tau_0)$ and $\sigma^2 \sim$ inverse-gamma$(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0)$
- **problem:** when θ and σ^2 are independent a priori, the posterior of σ^2 doesn't follow any known distribution (not the inverse-gamma as in the normal-normal model of Lecture 10)

$$\sigma_2 | \text{data} \sim ??????$$

- However, if we assume we know the value of θ, it can be shown that the posterior distribution of σ^2 is

$$\sigma^2|\text{data}, \theta \sim \text{inverse-gamma}(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2(\theta))$$

  (details omitted)

- from the conjugate normal-normal model, we also know that

$$\theta|\sigma^2, \text{data} \sim \text{N}(\mu_n, \sigma^2/\kappa_n)$$

$(\theta|\sigma^2, \text{data})$ and $(\sigma^2|\text{data}, \theta)$ are called **full conditional distributions** (conditional distribution of a parameter given everything else)

- So in this example is easy to sample from the full conditional distributions

can we use the full conditional distributions to sample the joint posterior distribution?

- given initial conditions, we can sample from the full conditional distributions $(\theta|\sigma^2, \text{data})$ and $(\sigma^2|\text{data}, \theta)$ to **approximate** the joint posterior of (θ, σ^2) **(Hammersley-Clifford theorem)**

- Then the marginal posterior distributions of θ and σ^2 are obtained from the joint posterior (θ, σ^2)

- This is the core idea of the Gibbs Sampling algorithm

Gibbs sampling: the algorithm

Suppose a vector of parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_p\}$ whose information is measured by a probability distribution $f(\boldsymbol{\theta}) = f(\theta_1, \ldots, \theta_p)$ (the **target distribution**). Given a starting point $\boldsymbol{\theta}^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_p^{(0)}\}$, the Gibbs sampler generates $\boldsymbol{\theta}^{(s)}$ from $\boldsymbol{\theta}^{(s-1)}$ as follows:

1. sample $\theta_1^{(s)}$ from $f(\theta_1|\theta_2^{(s-1)}, \theta_3^{(s-1)} \ldots, \theta_p^{(s-1)})$
2. sample $\theta_2^{(s)}$ from $f(\theta_2|\theta_1^{(s)}, \theta_3^{(s-1)} \ldots, \theta_p^{(s-1)})$

$\vdots$

p. sample $\theta_p^{(s)}$ from $f(\theta_p|\theta_1^{(s)}, \theta_2^{(s)} \ldots, \theta_{p-1}^{(s)})$

- the algorithm generates a **dependent** sequence of vectors $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}$
- this sequence is a **Markov chain** because $\boldsymbol{\theta}^{(s)}$ depends only on $\boldsymbol{\theta}^{(s-1)}$
- under some technical conditions and for a large number of samples (large S), the distribution of $\boldsymbol{\theta}^{(S)}$ follows the target distribution

The Metropolis Hastings Algorithm

Recap: Bayesian Inference

- **likelihood:** $f(\mathbf{data}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$
- **prior distribution:** initial beliefs about θ: $g(\theta)$
- **posterior distribution:** combination of initial beliefs with observed data using **Bayes theorem**

The density of the posterior distribution is

$$g(\theta|\mathbf{x}) = kg(\theta)f(\mathbf{x}|\theta)$$

where k is a constant which doesn't depend on θ: **the normalising factor**

- calculating the normalising factor is often very difficult in practice

The Metropolis Hastings Algorithm

Suppose we want to approximate the posterior distribution of a **single** parameter

- As with any MCMC method, the Metropolis-Hastings algorithm generates a sequence (Markov chain) of sample values, such that the **distribution** of values closely approximates the posterior distribution as sequence gets longer.
- at each iteration *s*, the algorithm generates a candidate value of the sequence based on a **proposal** distribution, which is easy to sample from.
- then, with some (acceptance) probability, the candidate is either
  - accepted: candidate value is the next the value of the sequence
  - rejected: the candidate is discarded and the current value becomes the next one in the sequence
- the acceptance probability depends on ratio of the posterior density evaluated at the candidate value and the posterior density evaluated at the current value (the normalization factor cancels out in the ratio)

The Metropolis Hastings Algorithm

- The Gibbs sampling is a special case of the Metropolis Hastings algorithm (acceptance probability is always 1)

- The Metropolis-Hastings algorithm can be applied to generate vectors of parameters

- A crucial step of the algorithm is to define a suitable proposal distribution to minimize the length of the chain (therefore the computational effort) and the “quality” of the chain values

Dealing with samples from MCMC

MCMC methods generate a sequence of values which are samples from the posterior distribution, **if the sequence is long enough**

- How to evaluate whether the Markov chain has **converged** to the posterior distribution?
  - there are several strategies for assessing convergence (see Geyer, 1992)
  - but usually, only lack of convergence can be assessed
  - samples generated before reaching convergence must be discarded (the burn-in period)

Dealing with samples from MCMC

- If the Markov chain has converged, does it produce a representative sample of the posterior?
  - an efficient MCMC algorithm must produce a representative sample of the parameter space (good mixing)
  - careful choice of the initial values of the chain makes the algorithm more efficient
  - correlation between parameters can severely affect efficiency (eg. hierarchical models, stochastic epidemic models)
  - MCMC generates **dependent** samples from the posterior: affects MC error (thining is required)

Some software for applying MCMC to “standard” statistical models

- in R: CRAN Task View: Bayesian Inference:
  `cloud.r-project.org/web/views/Bayesian.html`
  (lists many packages to implement MCMC for Bayesian inference)

- Winbugs:
  `mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbug`

- Stan: `http://mc-stan.org/`

MCMC for stochastic epidemic models

It was assumed so far that epidemic data was complete:

- infection times: $\boldsymbol{i} = (i_2, i_3, \ldots i_n)$
- removal times: $\boldsymbol{r} = (r_1, r_2, \ldots r_n)$

However, epidemic data are usually **partially** observed.

Examples of incomplete epidemic data

- **only removals observed**
- Infection times observed at fixed time periods (see talk tomorrow) (e.g an individual infected between week 1 and 2)

Also, inference might be needed before epidemic finishes.

MCMC for stochastic SIR models: example

Assumptions:

- removal times are known
- epidemic is observed until its end at time T
- livestock data, $I_1 = 0$ (artificial infection)
  (it is straightforward to adapt the MCMC otherwise)

- As in the complete observation case, gamma priors are considered for the transmission rate β and recovery rate γ
- then, full conditional distributions of the transmission rate and recovery rate are gamma distributions
- Since the infection times $\boldsymbol{i} = (i_2, i_3, \dots i_n)$ are unknown, they are treated as **parameters** (renamed as $(\phi_1, \phi_2, \phi_3, \dots \phi_n)$) and are also estimated from the available data
- However, due to the complex structure of the likelihood, the full conditional distributions of the infection times do not follow any standard distribution

MCMC for stochastic SIR models (O’Neill & Roberts, 1999)

In this case, the Metropolis-hastings and Gibbs Sampling algorithms can be combined to generate samples from the joint posterior distribution of the parameters

Therefore, given initial values for all parameters, the following MCMC algorithm can be applied:

```
do k=1, M
  1 sample $\beta^{(k)}$ from a gamma distribution
  2 sample $\gamma^{(k)}$ from a gamma distribution
  3 chose a infection time $\phi_j$ and sample it using metropolis hastings
end do
```

- see details in O’Neill & Roberts, 1999
- the algorithm can be extended to the case when the epidemic is not observed until its end

Seminal papers on MCMC for stochastic epidemic models

- Gibson, Gavin J. ”Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology.” Journal of the Royal Statistical Society: Series C (Applied Statistics) 46, no. 2 (1997): 215-233.
- Gibson, G. J., & Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. Mathematical Medicine and Biology, 15(1), 19-40.
- ONeill, P. D., & Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society), 162(1), 121-129.

References

- Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. The American Statistician, 46(3), 167-174.
- Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. The American Statistician, 49(4), 327-335.
- Geyer, Charles J. Practical Markov Chain Monte Carlo. Statistical Science (1992): 473-483.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. Chapman & Hall
- Gamerman, D., & Lopes, H. F. (2006). Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press.
- Sorensen, D., & Gianola, D. (2007). Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer Science & Business Media.

Introduction to MCMC for epidemic models:

- ONeill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical biosciences, 180(1), 103-114.