§6 Markov chain Monte Carlo

Outline

1. Motivation (Monte Carlo integration; Markov chains)
2. MCMC (Gibbs sampling)
3. Convergence and Monte Carlo standard errors
4. Strengths and weaknesses of MCMC

1. Motivation

Bayesian inference involves expectations, in particular posterior expectations $E(f(\theta) \mid x)$ of functions $f(\theta)$ of unknown parameters θ.

For example,

- $f(\theta) = \theta$: $E(f(\theta) \mid x)$ is the posterior mean of θ.
- $f(\theta) = I[\theta < a]$: $E(f(\theta) \mid x)$ can be used to calculate credible interval (a, b) for θ. (since $E(I[\theta < a] \mid x) = P(\theta < a \mid x)$, we can find the values of a, b such that $E(I[\theta < a \mid x]) = 0.025$ and $E(I[\theta < b \mid x]) = 0.975$).

The posterior expectation of $f(\theta)$ is

$$\begin{aligned} E(f(\theta) \mid x) &= \int f(\theta) p(\theta \mid x) d\theta \\ &= \frac{\int f(\theta) p(x \mid \theta) p(\theta) d\theta}{\int p(x \mid \theta) p(\theta) d\theta} \end{aligned}$$

In practice, integrations for the calculation of $E(f(\theta) \mid x)$ usually are complex, high-dimensional and have no closed form solution.

General problem: How can we evaluate

$$E[f(\theta) \mid x] = \int f(\theta)\, p(\theta \mid x)\, d\theta \ ?$$

Numerical integration or analytic approximation (e.g. Laplace/saddle-point) can be used, but tends to work poorly if θ is high-dimensional.

A solution: draw samples $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ from $p(\theta \mid x)$. Then we can estimate

$$E[f(\theta) \mid x] \approx \frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)})$$

This is **Monte Carlo integration.**

Problem: Drawing independent samples from $p(\theta \mid x)$ is generally not feasible if $p(\theta \mid x)$ is non-standard.
However, the samples need not necessarily be independent.

Question: How do we draw dependent samples $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ from $p(\theta \mid x)$?

Solution: Draw dependent samples using a *Markov chain* having $p(\theta \mid x)$ as its equilibrium distribution.

Markov chains

A *Markov chain* is a sequence $X^{(0)}, X^{(1)}, \ldots$ of random variables such that, for each $i = 0, 1, \ldots$, the conditional probability distribution of $X^{(i+1)}$ given $X^{(0)}, X^{(1)}, \ldots, X^{(i)}$ depends only on $X^{(i)}$.

That is, $X^{(i+1)}$ is independent of $X^{(0)}, \ldots, X^{(i-1)}$ given $X^{(i)}$, denoted by

$$X^{(i+1)} \perp\!\!\!\perp X^{(0)}, \ldots, X^{(i-1)} \mid X^{(i)}$$

So, in a Markov chain, the future depends on the past only through the present.

Equilibrium distribution

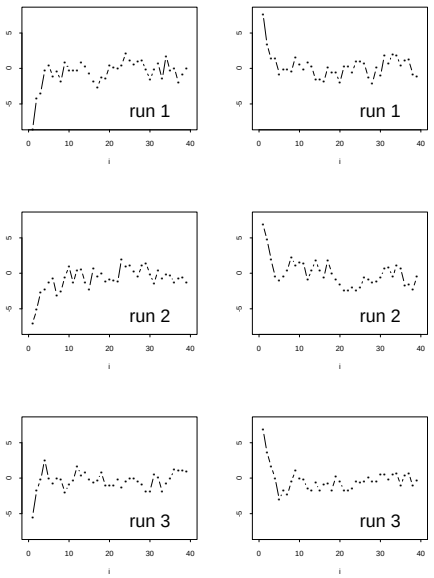
Subject to regularity conditions, as $i \to \infty$, the Markov chain *converges in distribution* to a unique *equilibrium* distribution.

This does not depend on $X^{(0)}$.

Example 6.1

$$\theta^{(i+1)} \sim \text{Normal}\left(\frac{\theta^{(i)}}{2}, 1\right)$$

$\theta^{(0)} = -15.0$ $\theta^{(0)} = +15.0$



The equilibrium distribution is Normal$(0, \frac{4}{3})$.

2. MCMC

If we could construct a Markov chain whose equilibrium distribution is $p(\theta \mid x)$, then, after M iterations (M is large enough), $\theta^{(M+1)}, \theta^{(M+2)}, \ldots, \theta^{(N)}$ would be dependent samples approximately from $p(\theta \mid x)$ and

$$E[f(\theta) \mid x] \approx \frac{1}{N-M} \sum_{i=M+1}^{N} f(\theta^{(i)})$$

This is *Markov chain Monte Carlo* (MCMC; ie Monte Carlo integration using Markov chains).

How do we construct a Markov chain whose equilibrium distribution is $p(\theta \mid x)$?

Using the *Metropolis-Hastings algorithm*. (Metropolis et al. 1953; Hastings, 1970)

This algorithm provides a general framework for MCMC. We shall concentrate on one of its special cases: *Gibbs Sampling*.

Gibbs sampling

Split $\boldsymbol{\theta}$ into K components $(\theta_1, \theta_2, \ldots, \theta_K)$ (components can be scalar or vector; eg $\theta_1 = \mu, \theta_2 = \tau, \ldots, \theta_K = \alpha$).

`Choose starting values` $\mu^{(0)}$, $\tau^{(0)}$, `...,` $\alpha^{(0)}$.
`set` $i = 0$.
`Repeat` {

`Sample` $\mu^{(i+1)}$ `from` $p(\mu \mid \tau^{(i)}, \ldots, \alpha^{(i)}, x)$
`Sample` $\tau^{(i+1)}$ `from` $p(\tau \mid \mu^{(i+1)}, \ldots, \alpha^{(i)}, x)$

...

`Sample` $\alpha^{(i+1)}$ `from` $p(\alpha \mid \mu^{(i+1)}, \tau^{(i+1)}, \ldots, x)$

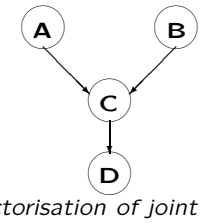
$i \leftarrow i + 1$

}

Note:
1. The most up-to-date version of $\boldsymbol{\theta}$ is used at each step.
2. Sampling is from *full-conditional distributions*.

Constructing full-conditional distributions

- Suppose we have a DAG



- By *factorisation of joint distribution*

$$p(\mathbf{V}) = \prod_{v \in \mathbf{V}} p(v \mid \text{parents}[v]), \quad \text{we have}$$

$$p(A, B, C, D) = p(A)\, p(B)\, p(C \mid A, B)\, p(D \mid C) \quad (*)$$

- *Two ways to get the full-conditional distribution for C.*
  1. Either

$$p(C \mid A, B, D) \propto \text{terms on RHS of } (*) \text{ containing } C = p(C \mid A, B)\, p(D \mid C)$$

  2. Or, based on the Markov blanket of C, we have

$$p(C \mid \mathbf{V} \backslash C) \propto p(C \mid \text{parents}[C]) \times \prod_{w \in \text{children}[C]} p(w \mid \text{parents}[w])$$

ie, $p(C \mid A, B, D) \propto p(C \mid A, B)\, p(D \mid C)$

Example 6.2: Normal, unknown mean and unknown variance

Suppose we have independent observations from a Normal(μ, τ^{-1}) distribution with unknown mean μ and unknown variance τ^{-1}:

$$X_i \sim \text{Normal}(\mu, \tau^{-1}) \qquad i = 1 \ldots n$$

Assign independent 'non-informative' priors

$$\begin{aligned} \mu &\sim \text{Normal}(0, 10^6) \\ \tau &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

The posterior distribution is

$$\begin{aligned} p(\mu, \tau \mid \mathbf{x}) &\propto p(\mu)p(\tau) \prod_{i=1}^{n} p(x_i \mid \mu, \tau) \\ &\propto \exp\left(-\frac{\mu^2}{2 \times 10^6}\right) \\ &\quad \times \tau^{-0.999} \exp(-0.001\tau) \\ &\quad \times \prod_{i=1}^{n} \tau^{1/2} \exp\left(-\frac{\tau}{2}(x_i - \mu)^2\right) \end{aligned}$$

Repeat the posterior distribution:

$$\begin{aligned} p(\mu, \tau \mid \mathbf{x}) &\propto p(\mu)p(\tau) \prod_{i=1}^{n} p(x_i \mid \mu, \tau) \\ &\propto \exp\left(-\frac{\mu^2}{2 \times 10^6}\right) \\ &\quad \times \tau^{-0.999} \exp(-0.001\tau) \\ &\quad \times \prod_{i=1}^{n} \tau^{1/2} \exp\left(-\frac{\tau}{2}(x_i - \mu)^2\right) \end{aligned}$$

Then the full-conditional distributions are

$$\begin{aligned} \mu \mid \tau, \mathbf{x} &\sim \text{Normal}\left(\frac{n\tau}{10^{-6} + n\tau}\bar{x}, (n\tau + 10^{-6})^{-1}\right) \\ \tau \mid \mu, \mathbf{x} &\sim \text{Gamma}\left(0.001 + \frac{n}{2}, \right. \\ &\qquad \left. 0.001 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{aligned}$$

NB: Compare them to those by applying formulae on '§2 Bayesian Inference' p4 and p8.

Gibbs sampling involves *sampling alternately between these two full-conditional distributions*.

Sampling from full-conditional distributions

We must be able to sample from

$$p(\theta_k \mid \theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_K)$$

to do Gibbs sampling.

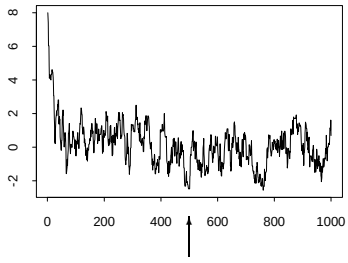
In simple problems (like Example 6.2), the full-conditional distributions have closed forms.

Otherwise, a range of algorithms is available. E.g.

- rejection sampling
- adaptive rejection sampling
- ratio-of-uniforms method

(see Chapter 5 of MCMC in Practice (Gilks et al., 1996) for more information).

3. Convergence and Monte Carlo standard errors



iteration M

Early iterations $\theta^{(1)}, \ldots, \theta^{(M)}$ reflect starting value $\theta^{(0)}$.
These iterations are called the *burn-in*.

After burn-in we say the chain has 'converged'.
$\Rightarrow \theta^{(M+1)}, \ldots, \theta^{(N)}$ are samples approximately from $p(\theta \mid x)$.

Omitting the burn-in, we estimate $E[f(\theta) \mid x]$ by using sample average,

$$\bar{f}_{MN} = \frac{1}{N - M} \sum_{i=M+1}^{N} f(\theta^{(i)})$$

Determining M

Problem: strictly speaking, convergence is only achieved for $M = \infty$.

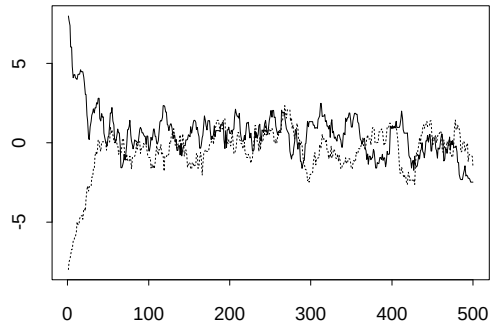
In practice: We can only make a reasonable effort to detect *lack of convergence*.
If no evidence of lack of convergence is found, we are more confident that the chain has 'converged'.

- Using *trace plots*. Once convergence has been reached, samples should look like a random scatter about a stable value.
- Using *convergence diagnostics* to determine M for the 'burn-in'.
  Many convergence diagnostics have been proposed.

The Gelman-Rubin diagnostic (1992)

Intuition

A single chain can be misleading. So, run several chains, with widely differing starting values. After burn-in, the behavior of all chains should be approximately the same.



Specifically, for a certain parameter θ_k, the variance within the chains should be the same as the variance across the chains.

Methodology

- Run q chains, each of length $2n$: keep last n samples from each chain
- For each parameter θ_k, calculate $R = \sqrt{\frac{V}{W}}$, where

  W = under-estimate of true posterior variance, $\sigma_k^2 = \text{Var}(\theta_k \mid x)$

  V = over-estimate of σ_k^2

  and $V, W \to \sigma_k^2$ as $n \to \infty$
- $R \to 1$ as $n \to \infty$
- Rule-of-thumb: $R < 1.05 \Rightarrow$ 'practical' convergence
- Calculate R for all parameters (or at least several if there are many parameters)

The Brooks-Gelman diagnostic (1998) is a variant of Gelman-Rubin.
Again, require $R < 1.05$

Determining N

Q: After burn-in, how long should we run the chain?

A: It is reasonable to run the chain until the **Monte Carlo standard error** (MCSE), SE($\bar{f}_{MN}$), is sufficiently small.

Q: How small should MCSE be?

A: We want MCSE small in relation to posterior standard deviation of $f(\theta)$.
Rule of thumb: run the chain until the MCSE of each parameter is less than 5% of the parameter's posterior standard deviation.

Q: For a given run length N, how can we estimate SE($\bar{f}_{MN}$), taking account of autocorrelations in

$$f(\theta^{(M+1)}), \ldots, f(\theta^{(N)}) \qquad ?$$

A: One method is *batching*.

Batching

- Divide the sequence
$$\theta^{(M+1)}, \theta^{(M+2)}, \ldots, \theta^{(N)}$$
into Q equal-length batches of size L.

- Calculate
$$b_q = \frac{1}{L} \sum_{i \in \text{batch } q} f(\theta^{(i)})$$
- Check that $b_1, \ldots, b_Q$ are approximately uncorrelated.
E.g., *estimated lag-1 autocorrelation* gives an indication of whether batches are approximately uncorrelated. If autocorrelation is high, larger batches are needed.
- Estimate
$$\widehat{\text{SE}}(\bar{f}_{MN}) = \sqrt{\frac{1}{Q(Q-1)} \sum_{i=1}^{Q} (b_i - \bar{b})^2}$$

4. Strengths and weaknesses of MCMC

Strengths

- Can offer freedom in modelling
  - in principle, no limits
- Can offer freedom in inference
  - in principle, no limits
  - can estimate arbitrary functions of model parameters (e.g. ranks, probabilities of threshold exceedence, etc)
- Can coherently integrate uncertainty
- Is the only available method for complex problems

Weaknesses and dangers

- Can be slow: may need to generate very long chains to
  - achieve convergence
  - reduce MCSE to acceptable level
- May fail to diagnose lack of convergence

  My MCMC has converged because
  - *I ran it for 10,000 iterations;*
  - *my wife called out 'coffee's ready';*
  - *WinBUGS crashed;*
  - *the plots were still going down.....*

  — T. O'Hagan
- May be difficult to validate the computer code written for the implementation of the MCMC

Outline revisited

1. Motivation (Monte Carlo integration; Markov chains)
2. MCMC (Gibbs sampling)
3. Convergence and Monte Carlo standard errors
4. Strengths and weaknesses of MCMC

Next week: WinBUGS