WRASS tutorial and workshop
Warwick, July 1998

Tutorial lectures on Markov chain Monte Carlo

by Peter Green (University of Bristol,
P.J.Green@bristol.ac.uk).

- introduction to MCMC, especially for computation in Bayesian statistics
- exploiting sparsity via graphical modelling
- basic recipes, and a sample of some techniques for improving performance
- why we (statisticians) should be interested in the topics at this meeting!
- not for experts!

Bayesian inference

Data: Y
Parameters, latent variables: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$
Likelihood: $f(Y|\boldsymbol{\theta})$
Prior: $\pi_0(\boldsymbol{\theta})$
Inference is based on the *joint posterior*

$$\begin{aligned} \pi(\boldsymbol{\theta}|Y) &= \frac{\pi_0(\boldsymbol{\theta}) f(Y|\boldsymbol{\theta})}{\int \pi_0(\boldsymbol{\theta}) f(Y|\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &\propto \pi_0(\boldsymbol{\theta}) f(Y|\boldsymbol{\theta}) \\ \text{i.e.} \quad \textit{Posterior} &\propto \textit{Prior} \times \textit{Likelihood} \end{aligned}$$

We will write $\pi(\boldsymbol{\theta})$ for $\pi(\boldsymbol{\theta}|Y)$.

Trivial example: Normal random sample

Data $Y_1, Y_2, \ldots, Y_n$ are a random sample from $N(\mu, \sigma^2)$.

Independent priors on μ and σ:

$$\begin{aligned} \mu &\sim N(\xi, \kappa^{-1}) \\ \sigma^{-2} &\sim \Gamma(\alpha, \beta) \end{aligned}$$

(these are only conditionally conjugate).

Joint posterior:

$$\begin{aligned} \pi(\mu, \sigma^2|Y) &\propto (\sigma^2)^{-\alpha-n/2-1} \\ &\times \exp\left\{-\frac{\beta}{\sigma^2} - \frac{\kappa(\mu-\xi)^2}{2} - \frac{\sum(Y_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

which is not of standard form.

Example: Weibull/Gamma experiment

Data are a random sample, possibly censored, from Weibull(ρ, κ):

$$f(Y|\rho,\kappa) = \kappa^m \rho^{m\kappa} \textstyle\prod_U Y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum Y_i^\kappa\right)$$

where m and $\prod_U$ are number of and product over uncensored observations.
Independent Gamma priors on ρ and κ:

$$\pi_0(\rho,\kappa) \propto \rho^{\alpha-1} e^{-\beta\rho} \kappa^{\gamma-1} e^{-\delta\kappa}$$

Posterior:

$$\begin{aligned} \pi(\rho,\kappa) \propto\; &\kappa^m \rho^{m\kappa} \textstyle\prod_U Y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum Y_i^\kappa\right) \\ &\rho^{\alpha-1} e^{-\beta\rho} \kappa^{\gamma-1} e^{-\delta\kappa} \end{aligned}$$

Computation for Bayesian inference

Under the posterior distribution, the parameters $\boldsymbol{\theta}$ are generally *dependent*, so we have to compute with a multivariate distribution, often in a high number of dimensions, with arbitrarily complex patterns of dependence.

Here, "compute with" could mean almost anything; an example would be to calculate a marginal (posterior) density.

Some of the possible approaches:

- Exact analytic integration is usually not available (and we don't want to be restricted in our model construction to use conjugate priors, etc., to make it possible).

- Asymptotic analytic approximations (e.g. Laplace) are awkward to set up, and can be unreliable.
- Conventional numerical methods require expertise and careful design to set up, and are only efficient in a low number of dimensions.
- Ordinary ("static") simulation is always available in principle, since the posterior distribution can be factorised as

$$\pi(\boldsymbol{\theta}|Y) = \pi(\theta_1, \theta_2, \ldots, \theta_p)$$

$$= \pi(\theta_1)\pi(\theta_2|\theta_1)\ldots\pi(\theta_p|\theta_1, \ldots, \theta_{p-1})$$

but the univariate distributions on the right hand side are rarely all available for simulation purposes (even after re-ordering).

Markov chain Monte Carlo

MCMC (a.k.a. Dynamic simulation) works even where static simulation doesn't, because

- All simulation methods rely on the Law of Large Numbers, and this remains true (the Ergodic theorem) when you have a Markov chain instead of an i.i.d. sequence.
- If you don't mind Markov dependence, then you can update the parameters $\theta_1, \theta_2, \ldots, \theta_p$ one-by-one (or in small groups).

The **objective** is to construct a Markov chain whose state space is the parameter space $\{\boldsymbol{\theta}\}$, and whose limiting distribution is the required posterior distribution $\pi(\boldsymbol{\theta})$.

The basic limit theorems

If $\{\boldsymbol{\theta}^{(t)}\}$ is an *irreducible* Markov chain with transition kernel P and *invariant distribution* π, and g is a real valued function with $\int |g(\boldsymbol{\theta})|\pi(d\boldsymbol{\theta}) < \infty$, then

$$\frac{1}{N}\sum_{t=1}^{N} g(\boldsymbol{\theta}^{(t)}) \to \int g(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}) < \infty$$

for π-almost all $\boldsymbol{\theta}^{(0)}$.

If the chain is also *aperiodic*, then there is convergence in total variation:

$$||P^t(\boldsymbol{\theta}^{(0)}, \cdot) - \pi(\cdot)|| \to 0$$

as $t \to \infty$, for π-almost all $\boldsymbol{\theta}^{(0)}$.

Richard Tweedie will give the details!

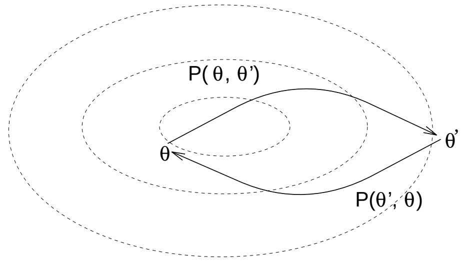
The **key idea** in most practical MCMC methods is **reversibility**. The distribution π is invariant for P if we have detailed balance (reversibility):

$$\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}', \boldsymbol{\theta})$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$.

Reversibility is sufficient but not necessary; however it is far easier to work with. We'll ignore the irreducibility and aperiodicity for the moment.

Think of reversibility as requiring a balance in the flow of probability.



The transitions described by P are neutral with respect to the contours of probability of π.

The basic MCMC sampling methods

The Gibbs sampler

Discard the current value of θ_i, and replace it by a value drawn from the *full conditional* distribution

$$\pi(\theta_i|\theta_{-i})$$

(where "$-i$" stands for $\{j : j \neq i\}$). Then we are using the kernel

$$P(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\theta_i'|\theta_{-i})I[\theta_{-i} = \theta_{-i}'],$$

and reversibility holds because given θ_{-i}, θ_i and θ_i' are i.i.d. $\sim \pi(\theta_i|\theta_{-i})$.

The Metropolis method

Draw a candidate new value (or "proposal") θ_i' from an arbitrary distribution $q(\theta_i'; \boldsymbol{\theta})$ satisfying the symmetry condition $q(\theta_i'; \boldsymbol{\theta}) = q(\theta_i; \boldsymbol{\theta}')$ (where $\theta_{-i} = \theta_{-i}'$).

Accept this with probability

$$\alpha = \min\{1, \pi(\theta_i'|\theta_{-i})/\pi(\theta_i|\theta_{-i})\};$$

otherwise leave $\boldsymbol{\theta}$ unchanged.

The Hastings sampler

As Metropolis, except that symmetry of q is not needed; the acceptance probability becomes:

$$\alpha = \min\left\{1, \frac{\pi(\theta_i'|\theta_{-i})q(\theta_i; \boldsymbol{\theta}')}{\pi(\theta_i|\theta_{-i})q(\theta_i'; \boldsymbol{\theta})}\right\};$$

the proof of correctness of each is the same: the choice of acceptance probability simply ensures that detailed balance is satisfied.

Proof of reversibility

For $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$,

$$\begin{aligned} \pi(\boldsymbol{\theta})P(\boldsymbol{\theta},\boldsymbol{\theta}') &= \pi(\theta_{-i})\pi(\theta_i|\theta_{-i})q(\theta_i';\boldsymbol{\theta})\alpha(\boldsymbol{\theta},\boldsymbol{\theta}') \\ &= \pi(\theta_{-i})\min\{R(\boldsymbol{\theta},\boldsymbol{\theta}'), R(\boldsymbol{\theta}',\boldsymbol{\theta})\} \end{aligned}$$

where $R(\boldsymbol{\theta},\boldsymbol{\theta}') = \pi(\theta_i|\theta_{-i})q(\theta_i';\boldsymbol{\theta})$. The whole expression above is therefore symmetric in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ (recall that $\theta_{-i} = \theta'_{-i}$). So reversibility holds.

This argument for the Hastings method covers Gibbs and Metropolis *a fortiori*.

Updating several variables at once

Each of the Gibbs, Metropolis and Hastings methods is equally valid if a group of variables $\theta_A = \{\theta_j : j \in A\}$ is updated simultaneously; each uses the full conditional $\pi(\theta_A|\theta_{-A})$. You could update *all* variables at once in Metropolis or Hastings. (It is a subtle question whether it is a good idea to update many variables.)

An important special case arises where the variables in θ_A are *conditionally independent* (under the full conditional). They can then be updated in parallel.

Role of full conditionals

All of the basic methods use the full conditionals $\pi(\theta_A|\theta_{-A})$, where A indexes the variables being updated. In Gibbs, you have to *draw* from this distribution; in Metropolis and Hastings, you only have to *evaluate* it (up to a multiplicative constant) at the old and new values.

Combining kernels to make an ergodic sampler

All of the methods above satisfy detailed balance, and hence preserve the equilibrium distribution: if

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$$

before the transition, then so it will afterwards.

To ensure that this is also the limiting distribution of the chain (ergodicity), we must combine such kernels to make a Markov chain transition mechanism that is irreducible (and aperiodic).

To do that, scan over the available kernels (indexed by i or A) either systematically or randomly, or in various other ways that are valid, provided you visit each variable often enough. You can use different recipes (Gibbs, Metropolis,...) for different A.

Trivial example: Normal random sample

Data are a random sample from $N(\mu, \sigma^2)$.

Independent priors on μ and σ:

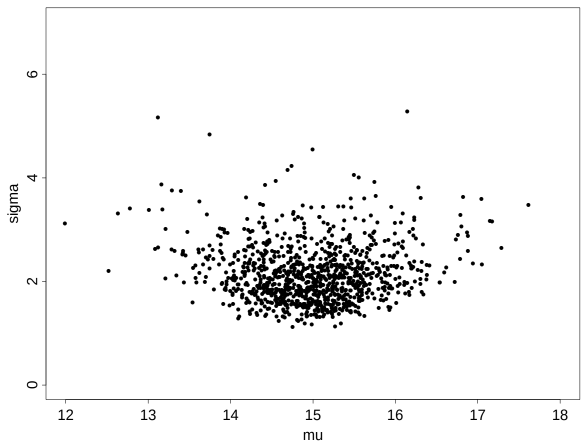
$$\begin{aligned} \mu &\sim N(\xi, \kappa^{-1}) \\ \sigma^{-2} &\sim \Gamma(\alpha, \beta) \end{aligned}$$

Full conditionals are easily found:

$$\begin{aligned} \mu|\sigma, Y &\sim N\left(\frac{\sigma^{-2}\sum Y_i + \kappa\xi}{\sigma^{-2}n + \kappa}, \frac{1}{\sigma^{-2}n + \kappa}\right) \\ \sigma^{-2}|\mu, Y &\sim \Gamma(\alpha + n/2, \beta + \textstyle\sum(Y_i - \mu)^2/2) \end{aligned}$$

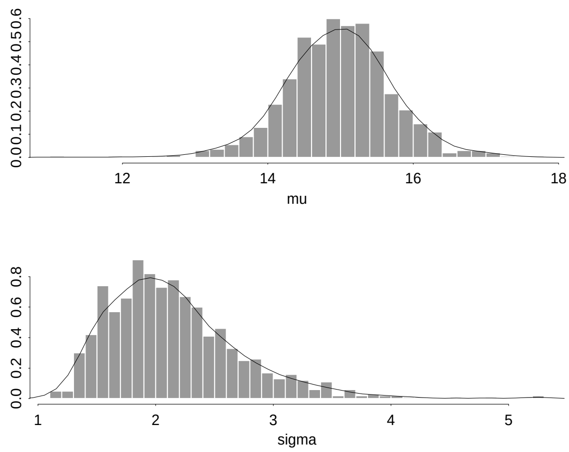
and we can implement a Gibbs sampler by alternately drawing μ and σ^{-2} from these distributions.

Gibbs sample of size 1000



Posterior sample of (μ, σ) from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

Marginal distributions



Posterior distributions of μ and σ from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

Full conditionals for Weibull/Gamma example

$$\begin{aligned} \pi(\rho|\kappa) &\propto \rho^{m\kappa} \exp\left(-\rho^\kappa \textstyle\sum Y_i^\kappa\right) \rho^{\alpha-1} e^{-\beta\rho} \\ \pi(\kappa|\rho) &\propto \kappa^m \rho^{m\kappa} \textstyle\prod_U Y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum Y_i^\kappa\right) \kappa^{\gamma-1} e^{-\delta\kappa} \end{aligned}$$

... hardly of standard form, so Gibbs is problematical, but easily evaluated for Metropolis or Hastings.

An easily implemented MCMC method would be

- alternate between ρ and κ
- propose new value from distribution symmetric about present value
- reject if out of range
- accept with probability (e.g.) $\min\{1, \pi(\rho'|\kappa)/\pi(\rho|\kappa)\}$

Using a MCMC sample

Having generated a MCMC sample $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, we are free to extract from it information about the target distribution in many different ways.

Probabilities can be estimated by computing empirical frequencies:

$$\pi(A) \approx \frac{1}{N} \sum_{t=1}^{N} I[\boldsymbol{\theta}^{(t)} \in A]$$

Expectations using empirical averages:

$$E_\pi(g) = \int g(\boldsymbol{\theta}) \pi(d\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{t=1}^{N} g(\boldsymbol{\theta}^{(t)})$$

Marginal distributions arise from simply ignoring some components:

$$E_\pi(g(\theta_i)) \approx \frac{1}{N} \sum_{t=1}^{N} g(\theta_i^{(t)})$$

Conditional distributions can be obtained either by *holding components fixed* or *selecting from the sample*.

For example, if $\boldsymbol{\theta}$ is partitioned as $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and we want to estimate $E_\pi(g(\boldsymbol{\theta}_1)|\boldsymbol{\theta}_2 = \boldsymbol{c})$, we could either

(a) use a MCMC sampler formed from kernels that only update $\boldsymbol{\theta}_1$, (note that detailed balance for π is the same as detailed balance for $\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ for such kernels), initialising $\boldsymbol{\theta}_2 = \boldsymbol{c}$, or

(b) (assuming $\pi(\boldsymbol{\theta}_2 = \boldsymbol{c}) > 0$) use an unconstrained sampler and the estimate

$$E_\pi(g(\boldsymbol{\theta}_1)|\boldsymbol{\theta}_2 = \boldsymbol{c}) \approx \frac{\sum_{t=1}^N g(\boldsymbol{\theta}_1^{(t)}) I[\boldsymbol{\theta}_2^{(t)} = \boldsymbol{c}]}{\#\{t \leq N : \boldsymbol{\theta}_2^{(t)} = \boldsymbol{c}\}}$$

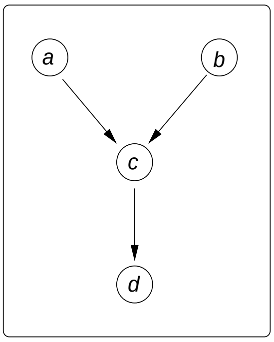
The role of graphical models

Graphical modelling provides a powerful language for specifying and understanding statistical models.

Graphs consist of vertices representing variables, and edges (directed or otherwise) that express conditional dependence properties.

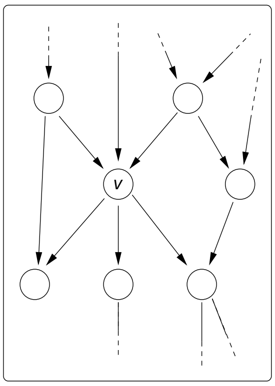
Directed acyclic graphs

For example, the DAG (directed acyclic graph) – a graph in which all edges are directed, and there are no directed loops –



expresses the natural factorisation of a joint distribution into factors each giving the joint distribution of a variable given its *parents*

$$\pi(a, b, c, d) = \pi(a)\pi(b)\pi(c|a, b)\pi(d|c)$$



In general

$$\pi(\boldsymbol{\theta}) = \prod_{v \in V} \pi(\theta_v | \theta_{\mathrm{pa}(v)})$$

which in turn implies a *Markov property*, that variables are conditionally independent of their non-descendants, given their parents.

From the perspective of setting up MCMC methods, the graphical structure assists in identifying which terms need be included in a full conditional.

$$\pi(\boldsymbol{\theta}) = \prod_{v \in V} \pi(\theta_v | \theta_{\mathrm{pa}(v)})$$

implies

$$\pi(\theta_v | \theta_{-v}) = \pi(\theta_v | \theta_{\mathrm{pa}(v)}) \prod_{w : v \in \mathrm{pa}(w)} \pi(\theta_w | \theta_{\mathrm{pa}(w)})$$

that is, one term for the variable itself, and one for each of its children.

Graphical modelling, the construction of MCMC methods through full conditional distributions, and good practice in statistical model building all exploit the same modular structure.

Performance of MCMC methods

There are three main issues to consider

- Convergence (how quickly does the distribution of $\boldsymbol{\theta}^{(t)}$ approach $\pi(\boldsymbol{\theta}|Y)$?) (*or, can we find exact/perfect MCMC sampling methods that give guaranteed convergence?*)
- Efficiency (how well are functionals of $\pi(\boldsymbol{\theta}|Y)$ estimated from $\{\boldsymbol{\theta}^{(t)}\}$?)
- Simplicity (how convenient is the method to use?)

Note that here computer effort should be measured in seconds, not iterations!

Gibbs is not necessarily superior to other methods on *any* of these three criteria.

Monte Carlo standard errors

(Not the standard error of the posterior!)

We should be concerned about the precision of simulation-based estimates. Because of Markov dependence, this is not quite straightforward, even though we (mostly) just use empirical averages as estimates.

Some possibilities:

- Blocking (Hastings)
- Time-series methods (= estimating spectral density at 0) (e.g. Sokal)
- Initial series estimates (Geyer)
- Regeneration (Tierney, Mykland and Yu)

For reversible samplers, there is also a Central Limit theorem for Markov chain averages (Kipnis and Varadhan).

Estimating $\int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ by $N^{-1}\sum_{t=1}^{N} g(\boldsymbol{\theta}^{(t)}) = \bar{g}_N$.

$$\mathrm{var}(\bar{g}_N) \sim N^{-1} \sum_{-\infty}^{\infty} \gamma_t$$

where $\gamma_t = \mathrm{cov}\{g(\boldsymbol{\theta}^{(s)}), g(\boldsymbol{\theta}^{(s+t)})\}$.

Blocking (or batching)

Divide run of length N into b consecutive blocks of length k.

$$\mathrm{var}(\bar{g}_N) \approx \{b(b-1)\}^{-1} \sum_{i=1}^{b} \{\bar{g}_{k,i} - \bar{g}_{N,1}\}^2$$

where

$$\bar{g}_{k,i} = k^{-1} \sum_{j=(i-1)k+1}^{ik} g(\boldsymbol{\theta}^{(j)}).$$

This extends to nonlinear functions of empirical averages.

Using empirical covariances

Cannot estimate $\sum_{-\infty}^{\infty} \gamma_t$ consistently by $\sum_{-\infty}^{\infty} \widehat{\gamma}_t$: use windowed estimate $\sum_{-\infty}^{\infty} w(t)\widehat{\gamma}_t$ instead.

Initial series estimators

Geyer (1993, Stat. Sci.) observes that, for a reversible ergodic chain, $\widehat{\gamma}_{2t} + \widehat{\gamma}_{2t+1}$ is non-negative, decreasing and convex in t: truncate $\sum \widehat{\gamma}_t$ when one or other of these properties is violated.

Regeneration

Look for regeneration points in the Markov chain path, in practice aided by Nummelin's splitting technique; tours between regenerations are i.i.d., so renewal theory and ratio estimation give estimates of posterior expectations, and simulation standard errors that are valid without quantifying Markov dependence.

Regeneration using Nummelin's splitting

Suppose the transition kernel $P(\boldsymbol{\theta}, A)$ satisfies

$$P(\boldsymbol{\theta}, A) \geq s(\boldsymbol{\theta})\nu(A)$$

where ν is a probability measure, and s is a non-negative function such that $\int s(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}) > 0$.

Let $r(\boldsymbol{\theta}, \boldsymbol{\theta}')$ denote the Radon-Nikodym derivative

$$r(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{s(\boldsymbol{\theta})\nu(d\boldsymbol{\theta}')}{P(\boldsymbol{\theta}, d\boldsymbol{\theta}')} \leq 1.$$

Now, given a realisation $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots$ from P, construct conditionally independent 0/1 random variables $S^{(0)}, S^{(1)}, \ldots$ with

$$P(S^{(t)} = 1 | \ldots) = r(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)})$$

Then by simple probability calculus we find

$$P(S^{(t)} = 1 | \boldsymbol{\theta}^{(\leq t)}, S^{(<t)}) = s(\boldsymbol{\theta}^{(t)})$$

and

$$P(\boldsymbol{\theta}^{(t+1)} \in A | \boldsymbol{\theta}^{(\leq t)}, S^{(<t)}, S^{(t)} = 1) = \nu(A)$$

that is, we can post-process the chain stochastically to generate binary 'splitting variables'. Whenever $S^{(t)} = 1$, the next state $\boldsymbol{\theta}^{(t+1)}$ is drawn from ν, independent of the past! The chain regenerates.

The problem with using the technique in practice is that in the Markov chains we tend to create for Bayesian computation, $P(\boldsymbol{\theta}, A)$ is difficult to handle algebraically, and/or impossible to bound below by $s(\boldsymbol{\theta})\nu(A)$ as required.

Prediction

Formally, in Bayesian inference, we can group all unobserved variables (parameters, latent variables, missing data, future data) as $\boldsymbol{\theta}$. But temptation to do so with MCMC should be restrained.

For example, suppose we have actual data Y and future data Y^+ that are conditionally independent given $\boldsymbol{\theta}$. Then

$$\pi(\boldsymbol{\theta}, Y^+ | Y) = \pi(\boldsymbol{\theta} | Y)\pi(Y^+ | \boldsymbol{\theta})$$

and MCMC should be used for the first factor, but direct forwards simulation for the second. There are other much more subtle cases where combinations of MCMC and direct simulation will be more efficient: these are examples of "partial conditioning". You need to look at the conditional independences in the model rather closely to see what is valid.

Credibility intervals

Many computations on the posterior that would be quite complicated analytically reduce to trivial enumeration from the MCMC sample. For example, an estimated $100(1-\alpha)\%$ credibility interval for θ_i, given Y, is $[\theta_i^{[j]}, \theta_i^{[N+1-j]}]$ where $j = N\alpha/2$, and $\theta_i^{[j]}$ is the j^{th} order statistic of the MCMC sample of θ_i.

Simultaneous credibility intervals

Often, particularly in function and image estimation, we are interested in *simultaneous* credibility intervals. This is still quite straightforward. Compute the ordinary intervals as above, then choose j to be the largest integer such that

$$\theta_i^{[j]} \leq \theta_i^{(t)} \leq \theta_i^{[N+1-j]}$$

for all i, for at least $N(1-\alpha)$ values of t.

This has nice equivariance properties (exactly equivariant to strictly monotone componentwise transformations of $\boldsymbol{\theta}$), and can be computed by ordering one vector of maximum folded ranks, in addition to the MCMC samples.

Sensitivity analysis via MCMC

It is important to study effect on posterior of changes to model, especially variations in the prior.

$$\begin{aligned} \pi(\boldsymbol{\theta}) &\propto \pi_0(\boldsymbol{\theta}) f(Y|\boldsymbol{\theta}) \\ \pi^*(\boldsymbol{\theta}) &\propto \pi_0^*(\boldsymbol{\theta}) f^*(Y|\boldsymbol{\theta}) \end{aligned}$$

Can just repeat MCMC computation on the new model: note that even where the base model is rather tractable (e.g. $\pi_0(\boldsymbol{\theta})$ conjugate to $f(Y|\boldsymbol{\theta})$), responsible analysis will involve alternatives that are not.

π		π^*
explicit calculation	$\Rightarrow$	MCMC
Gibbs sampler	$\Rightarrow$	Hastings

Sensitivity analysis via importance sampling

$$E_{\pi^*}(g) = E_\pi\left(\frac{\pi^*}{\pi} g\right)$$

... estimate from MCMC run aimed at π, by

$$\frac{\sum_{t=1}^N w(\boldsymbol{\theta}^{(t)}) g(\boldsymbol{\theta}^{(t)})}{\sum_{t=1}^N w(\boldsymbol{\theta}^{(t)})}$$

where

$$w(\boldsymbol{\theta}) \propto \frac{\pi^*(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}$$

Problem: except in very low dimensional problems, $w(\boldsymbol{\theta}^{(t)})/\sum_t w(\cdot)$ is effectively concentrated on very few samples, $\Rightarrow$ very poor efficiency.

This can sometimes be mitigated by considering infinitesimal perturbations instead: $\pi_\epsilon(\boldsymbol{\theta}) \propto (\pi(\boldsymbol{\theta}))^{(1-\epsilon)}(\pi^*(\boldsymbol{\theta}))^\epsilon$ (or by running another chain).

Some tools for improving performance

- Grouping variables for simultaneous updating
- Re-parameterisation (e.g. hierarchical centering)
- Antithetic variables/over-relaxation methods
- Adaptive algorithms
- Enlarging state-space:
  i. Auxiliary variables (e.g. Swendsen-Wang)
  ii. Multigrid methods
  iii. Simulated tempering
  iv. Inventing additional models (and using reversible jump MCMC)
  v. Hybrid MCMC (momentum variables)

Improving performance by augmenting the state space

Perhaps counter-intuitively, it is sometimes possible to improve MCMC performance by augmenting the state vector to include additional components. Two successful recipes are those in which the original model appears as a *conditional* distribution in an augmented model (simulated tempering) and in which it appears as a *marginal* (auxiliary variables).

Simulated tempering

Combat slow mixing by embedding desired model in a family of models, indexed say by α, and treat α now as an additional dynamic variable. Design the family so that for some α, the chain mixes much better.

$$\pi(\boldsymbol{\theta}|Y) \Rightarrow \pi^\star(\boldsymbol{\theta}, \alpha_0|Y)$$

Run MCMC on $\pi^\star(\boldsymbol{\theta}, \alpha|Y)$, and condition on $\alpha = \alpha_0$ by selecting from the output.

Simulated tempering, by changing the temperature

This was the original Marinari/Parisi idea; we set

$$\pi^\star(\theta, \alpha|Y) \propto \{\pi(\theta|Y)\}^\alpha$$

where $\alpha = \alpha_0 = 1$ corresponds to the original model, and $\alpha \to 0$ makes the probability surface 'flatter', or in physical terms, 'warmer'.

The full conditionals change in the same way:

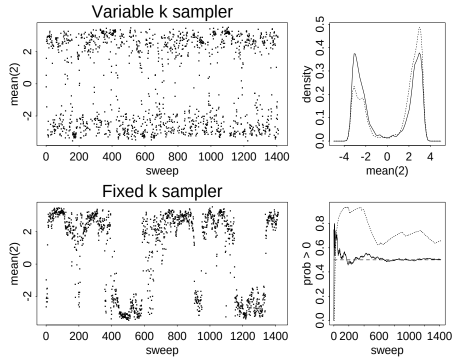
$$\pi^\star(\theta_i|\theta_{-i}, \alpha, Y) \propto \{\pi(\theta_i|\theta_{-i}, Y)\}^\alpha$$

so implementation is very easy.

We place a (discrete) artificial prior on α so that the *marginal* for α is approximately uniform.

Simulated tempering, by inventing models

An example from mixture analysis: allowing the number of components to vary gives much better mixing. (*This uses reversible jump MCMC: see later.*)



Auxiliary variables

Edwards and Sokal proposed a way to improve mixing by augmenting the state space so that the original target appears as the *marginal* equilibrium distribution.

Starting from $\pi(\boldsymbol{\theta})$, take some additional variables u, with $\pi(u|\boldsymbol{\theta})$ arbitrarily chosen. Then the joint is $\pi(\boldsymbol{\theta}, u) = \pi(\boldsymbol{\theta})\pi(u|\boldsymbol{\theta})$, for which $\pi(\boldsymbol{\theta})$ is certainly the marginal for $\boldsymbol{\theta}$.

We could now run a MCMC method for the *joint* target $\pi(\boldsymbol{\theta}, u)$ (usually a method that updates $\boldsymbol{\theta}$ and u alternately), and simple ignore the u variable in extracting information from the simulation.

When might this idea be useful? Suppose $\pi(\boldsymbol{\theta})$ factorises as:

$$\pi(\boldsymbol{\theta}) = \pi_0(\boldsymbol{\theta})b(\boldsymbol{\theta})$$

where $\pi_0(\boldsymbol{\theta})$ is a (possibly unnormalised) distribution that is easy to simulate from, and $b(\boldsymbol{\theta})$ is the awkward part, often representing the 'interactions' between variables that are slowing down the chain.

Then take a one-dimensional u with $u|\boldsymbol{\theta} \sim U[0, b(\boldsymbol{\theta})]$: we find

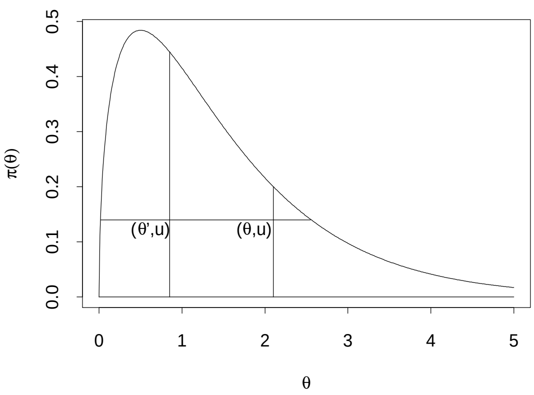
$$\pi(\boldsymbol{\theta}, u) = \pi(\boldsymbol{\theta})\pi(u|\boldsymbol{\theta}) = \pi_0(\boldsymbol{\theta})b(\boldsymbol{\theta})\frac{I[0 \le u \le b(\boldsymbol{\theta})]}{b(\boldsymbol{\theta})}$$

so that

$$\pi(\boldsymbol{\theta}|u) \propto \pi_0(\boldsymbol{\theta})$$

restricted to (conditional on) the event $\{\boldsymbol{\theta} : b(\boldsymbol{\theta}) \ge u\}$. At least when this $\pi(\boldsymbol{\theta}|u)$ can be sampled without rejection, we can easily implement a Gibbs sampler, drawing u and $\boldsymbol{\theta}$ in turn.

This method has recently been popularised under the name of the 'slice sampler', reflecting the fact that if $\pi_0(\boldsymbol{\theta}) =$ constant, both $\pi(u|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|u)$ are uniform distributions, corresponding to vertical and horizontal slices through the graph of $\pi(\boldsymbol{\theta})$.



Bayesian model determination

It is wrong to behave as if the statistical model for our data was not subject to question.

Suppose we have a (countable) collection of models that we wish to entertain: $M_1, M_2, \ldots, M_k, \ldots$.

A priori, we assign probabilities to these: $p(k)$.

For each model, there is a parameter vector $\theta = \theta^{(k)} \in \mathcal{R}^{n_k}$ say, with a prior: $p(\theta^{(k)}|k)$, and a likelihood for the observed data Y: $p(Y|k, \theta^{(k)})$.

The joint distribution of all variables is

$$p(k, \theta^{(k)}, Y) = p(k)p(\theta^{(k)}|k)p(Y|k, \theta^{(k)})$$

Observing Y provides information about both the model indicator k and the corresponding parameter vector $\theta^{(k)}$, through their posterior distributions:

$$p(k|Y) = \frac{\int p(k, \theta^{(k)}, Y)d\theta^{(k)}}{\sum_k \int p(k, \theta^{(k)}, Y)d\theta^{(k)}}$$

and

$$p(\theta^{(k)}|Y, k) = \frac{p(k, \theta^{(k)}, Y)}{\int p(k, \theta^{(k)}, Y)d\theta^{(k)}}$$

involving integrals that as usual seem to need MCMC!

There are two main approaches: within-model and across-model simulation.

Within-model simulation

Here we treat each model M_k separately.

The posterior for the parameters $\theta^{(k)}$ is in any case a within-model notion:

$$p(\theta^{(k)}|Y, k) = \frac{p(\theta^{(k)}|k)p(Y|k, \theta^{(k)})}{\int p(\theta^{(k)}|k)p(Y|k, \theta^{(k)})d\theta^{(k)}}$$

As for the posterior model probabilities, since

$$\frac{p(k_1|Y)}{p(k_0|Y)} = \frac{p(k_1)}{p(k_0)}\frac{p(Y|k_1)}{p(Y|k_0)}$$

(the *Bayes factor* for model M_{k_1} vs. M_{k_0}), it is sufficient to estimate the *marginal likelihoods*

$$p(Y|k) = \int p(\theta^{(k)}, Y|k)d\theta^{(k)}$$

separately for each k, using individual MCMC runs.

Estimating the marginal likelihood

There are many possible estimates based on importance sampling, some of which are well-studied, for example

$$\widehat{p}_1(Y|k) = N \bigg/ \sum_{t=1}^{N} p(Y|k, \theta_t^{(k)})$$

based on a MCMC sample $\theta_1^{(k)}, \theta_2^{(k)}, \ldots$ from the posterior $p(\theta^{(k)}|Y, k)$, or

$$\widehat{p}_2(Y|k) = N^{-1} \sum_{t=1}^{N} p(Y|k, \theta_t^{(k)})$$

based on a sample from the *prior* $p(\theta^{(k)}|k)$.

Both of these has its faults, and composite estimates can perform better.

Across-model simulation

Here we conduct a *single* simulation that traverses the entire $(k, \theta^{(k)})$ space. Since the dimension n_k of $\theta^{(k)}$ typically varies with k, this requires a MCMC sampler that works in more general spaces than $\mathcal{R}^d$.

The Metropolis-Hastings recipe extends to arbitrary measure spaces: "reversible-jump MCMC". We will use a range of *move types* m, each providing a transition kernel P_m, and insist on detailed balance for each:

$$\int_{\boldsymbol{\theta} \in A} \pi(d\boldsymbol{\theta}) P_m(\boldsymbol{\theta}, B) = \int_{\boldsymbol{\theta}' \in B} \pi(d\boldsymbol{\theta}') P_m(\boldsymbol{\theta}', A)$$

for all sets of parameter values A, B.

The Metropolis-Hastings idea still works, but you need to work a bit to make the acceptance ratio make sense:

$$\alpha = \min\left\{1, \frac{\pi(d\boldsymbol{\theta}')q(\boldsymbol{\theta}', d\boldsymbol{\theta})}{\pi(d\boldsymbol{\theta})q(\boldsymbol{\theta}, d\boldsymbol{\theta}')}\right\}$$

where numerator and denominator need to have densities with respect to a common dominating measure ("dimension-balancing"). That is, we find a dominating measure for the *joint* distribution of the current state (in equilibrium) *and the next one*.

In concrete terms, think of how the program will do it:

$$(\boldsymbol{\theta}, u) \leftrightarrow (\boldsymbol{\theta}', u')$$

where u are the random numbers you will draw to combine with $\boldsymbol{\theta}$ to make the proposed new state $\boldsymbol{\theta}'$, and vice-versa. Then the acceptance probability becomes

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')q(u')}{\pi(\boldsymbol{\theta})q(u)} \left|\frac{\partial(\boldsymbol{\theta}', u')}{\partial(\boldsymbol{\theta}, u)}\right|\right\}$$

The ratio is of joint densities with the same degrees of freedom, together with the Jacobian needed to account for the change of variable.

This expression applies for straightforward moves that do not change the dimension, as well.

Strengths of MCMC

- Freedom in modelling
- Freedom in inference
- Well-adapted for models defined on sparse graphs
- Addresses questions only posed after simulation completed (e.g. ranking and selection)
- Opportunities for simultaneous inference
- Allows/encourages sensitivity analysis
- Model comparison/criticism/choice

Weaknesses of MCMC

- Order $N^{-1/2}$ precision
- Possibility of slow convergence, especially when not diagnosable (meta-stability)