

A Search and Jump Algorithm for Markov Chain Monte Carlo Sampling

Christopher Jennison

Department of Mathematical Sciences,
University of Bath, UK
<http://people.bath.ac.uk/mascj>

Adriana Ibrahim

Institute of Mathematical Sciences,
University of Malaya

Seminar at University of Bath

22 November 2016

Plan of talk

1. Basics of MCMC
2. The challenge: difficulties in getting a Markov chain to mix
3. Tjelmeland & Hegstad's mode-jumping algorithm
4. A two-stage approach to mode-jumping
5. Example: Phosphate levels at an archaeological site
6. Example: Electron density in the ionosphere
7. Adapting the mode-jumping method to sample from distributions with "thin" support
8. Conclusions

1. Markov chain Monte Carlo sampling (MCMC)

Aim: To sample from a complex distribution $\pi(x)$ on the state space Ω by running a Markov chain with limiting distribution π .

Typically, X is high-dimensional and π not particularly tractable.

The minimal requirement is that $\pi(x)$ can be evaluated up to a multiplicative constant.

Method: Create a Markov chain on Ω with transition matrix P satisfying

$$\pi P = \pi.$$

Let π_n denote the distribution of the state X_n after n transitions from an initial state x_0 .

Then, if the Markov chain is irreducible and aperiodic,

$$\pi_n \rightarrow \pi \text{ as } n \rightarrow \infty.$$

A couple of comments

(i) Notation:

The distribution $\pi(x)$ may be discrete or continuous.

In the discrete case, the transition matrix $P = (P_{ij})$ where

$$P_{ij} = P(X_{n+1} = j \mid X_n = i).$$

In the continuous case, $P(x_n, x_{n+1})$ specifies the conditional density of X_{n+1} given $X_n = x_n$.

(ii) Generality:

I shall describe methods and results for the continuous case.

To obtain formulae for the discrete case, replace integrals by sums.

Detailed balance

It is convenient to create a Markov chain with limiting distribution π by defining P to have **detailed balance** with respect to π , i.e.,

$$\pi(x) P(x, y) = \pi(y) P(y, x) \quad \text{for all } x, y \text{ in } \Omega.$$

The key property $\pi P = \pi$ follows since

$$\int_{\Omega} \pi(x) P(x, y) dx = \int_{\Omega} \pi(y) P(y, x) dx = \pi(y).$$

Examples of this construction are:

Metropolis-Hastings samplers, based on the work of Metropolis et al. (1953) and Hastings (1970),

The Gibbs sampler, introduced by Geman & Geman (1984).

The Metropolis-Hastings algorithm

From $X_n = x$, generate a proposal y from the kernel $q(x, y)$,

Calculate the “acceptance probability” for this proposal

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}\right\},$$

With probability $\alpha(x, y)$, accept and move to $X_{n+1} = y$,

With probability $1 - \alpha(x, y)$, reject and stay at $X_{n+1} = x$.

Detailed balance: We need to show, for all $x \neq y$,

$$\pi(x) q(x, y) \alpha(x, y) = \pi(y) q(y, x) \alpha(y, x).$$

It is straightforward to check this holds for the two cases

$$\pi(x) q(x, y) > \pi(y) q(y, x) \text{ and } \pi(x) q(x, y) < \pi(y) q(y, x).$$

The Gibbs sampler

Suppose the sample space of the target distribution is $\Omega = \mathbb{R}^k$.

Let $X(-i)$ denote the vector of elements of X excluding $X(i)$.

Denote the conditional distribution of $X(i)$ given $X(-i) = x(-i)$ when $X \sim \pi$ by

$$\pi_{X(i)|X(-i)}(x(i) | x(-i)).$$

So $\pi_{X(i)|X(-i)}$ is a 1-dimensional PDF or discrete mass function.

Denote by P_i the transition matrix when X is modified by replacing $X(i)$ with a new value sampled from $\pi_{X(i)|X(-i)}(x(i) | x(-i))$.

In one cycle of the “Gibbs sampler”, $X(1), \dots, X(k)$ are updated in turn: the transition matrix for the full cycle is $P = P_1 \dots P_k$.

It is easy to show $\pi P_i = \pi$ for $i = 1, \dots, k$ and, hence, $\pi P = \pi$.

A variety of “move types”

We may wish to use several “types” of move, indexed by a parameter $\phi \in \Phi$, with transition matrix P_ϕ for move type ϕ .

If each P_ϕ satisfies detailed balance, we can deduce $\pi P_\phi = \pi$.

Transitions can be made using a fixed sequence of move types ϕ .

Alternatively, the move type for each transition may be selected at random (independently of the current state x).

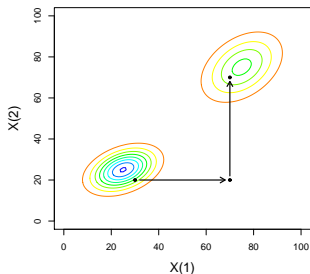
In either case, the chain has limiting distribution π , as long as the chain is irreducible and aperiodic.

We shall consider cases where one type of move produces small displacements in X , while other moves are designed to make larger jumps around the sample space Ω .

2. Mixing problems

Efficient MCMC sampling needs π_n to converge rapidly to π .

Problem 1. Multiple modes

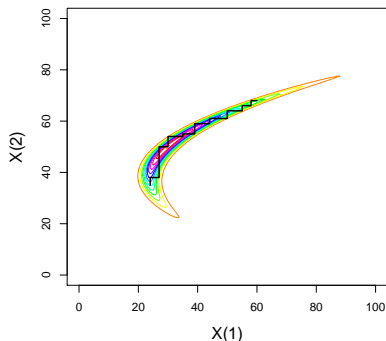


To move between modes, updating one element of x at a time, requires a visit to a state with very low $\pi(x)$ — and there is very little probability of accepting such a move.

If updating several elements of x together, the proposal must land near the middle of the other mode in order to be accepted.

Mixing problems

Problem 2. Very thin region of support for π

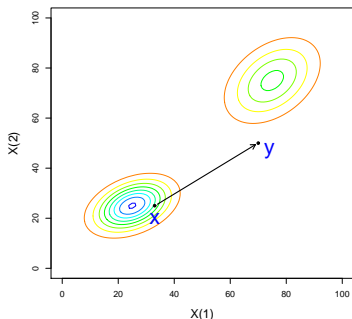


Traversing the modal region of π with updates of $X(1)$ and $X(2)$ requires a great many small steps.

3. Mode jumping method of Tjelmeland & Hegstad (2001)

Suppose the modes of π are small and in a high-dimensional space.

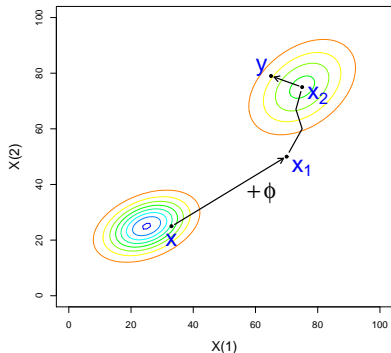
We can create a proposal kernel $q(x, y)$ that generates large jumps but the proposal y is unlikely to be at the centre of another mode.



The current state x will have fairly high $\pi(x)$ but $\pi(y)$ is small, so

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}\right\} \approx 0.$$

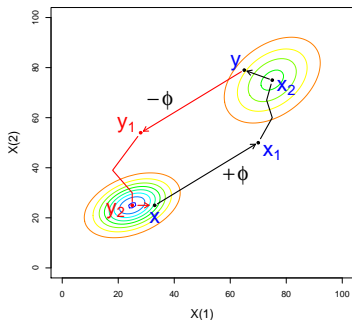
The mode jumping method of Tjelmeland & Hegstad (2001)



T & H's algorithm makes mode-jumping proposals by:

1. A large step from x to $x_1 = x + \phi$,
2. Hill climbing from x_1 to x_2 ,
3. Sample y from $h(x_2, y)$, an approximation to $\pi(y)$ at x_2 .

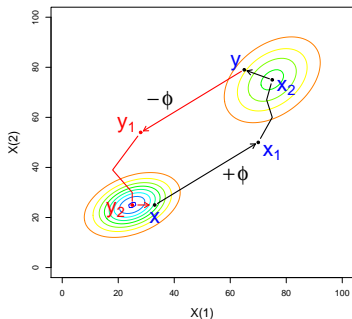
Achieving detailed balance in T & H's algorithm



4. Construct a reverse step to $y_1 = y - \phi$,
5. Hill climbing from y_1 to y_2 ,
6. Fit a local approximation $h(y_2, x)$ to $\pi(x)$ at y_2 ,
7. Accept the move from x to y with probability

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y) h(y_2, x)}{\pi(x) h(x_2, y)}\right\}.$$

T & H's proof of detailed balance



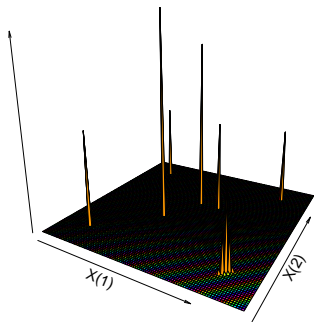
The “large step” move of type ϕ includes a random choice of $+\phi$ or $-\phi$. In the proof of detailed balance, the forward move with $+\phi$ is paired with the reverse move with $-\phi$ and vice versa.

Computationally, the return path has to be constructed in order to compute $\alpha(x, y)$ and accept or reject the proposal.

The deterministic hill climbing step may be replaced by a stochastic optimisation step (Richard Sharp, University of Bath PhD Thesis).

T & H's method: An example

$$\pi(x)^{1/5}$$



In this 2-D distribution, both $X(1)$ and $X(2)$ range from 0 to 100.

Problems are likely to grow with dimensionality.

The distribution to be sampled, π , has 7 highly compact modes, each bivariate normal or a transformed bivariate normal.

Plotting π raised to the power $1/5$ reduces the “spikiness” and makes a more meaningful plot.

T & H's method: An example

The support of each mode of π is very small — a standard MCMC sampler has very little chance of proposing a jump to a new mode.

A specialised mode-jumping method is needed.

We applied the T & H method with iterations comprising 20 local updates and 1 mode-jumping step.

Local updates:

Proposals have $N(0, 0.01^2)$ displacements in $X(1)$ and $X(2)$.

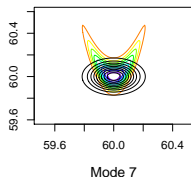
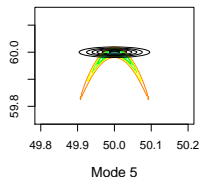
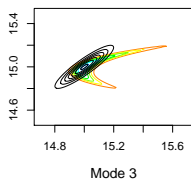
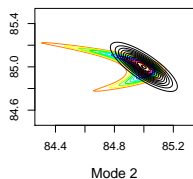
Mode jumping:

Jumps have $N(0, 50^2)$ displacements in $X(1)$ and $X(2)$,

Quasi-Newton hill climbing using numerical second derivatives,

Fitting a bivariate normal distribution at the top of the hill.

T & H's method: An example



Details of the density of π (in colour) and the approximating bivariate normal density (in black) at four of the modes.

Performance of T & H's method

The T & H method does manage to sample from this challenging distribution:

Mode jumping rate 1.1%

Evaluations of π per iteration:

Local steps 20

Mode jumping steps 125

The low success rate for mode-jumping steps is attributable to:

- Use of local approximations to make proposals at each mode,
- Reverse steps not returning to the original mode,
- Different weights for the 7 modes (0.05 to 0.4).

Note that in mode-jumping steps, the algorithm climbs the same hills over and over again.

How can this performance be improved?

4. A two-stage approach to mode jumping

Ibrahim and Jennison have proposed a two-stage approach.

Stage 1

- Search for modes using multiple runs of local optimisation from random (or systematically spaced) starting points,
- Apply cluster analysis to reduce the results to a set of distinct modes and select a representative member of each cluster,
- Fit a local approximation to π at each mode.

Stage 2

- Run an MCMC sampler with a mixture of local updates and mode jumping steps between the modes found in Stage 1.

Details of the two-stage approach to mode jumping

Our aim is to present a generic, widely applicable method.

Stage 1

Searching for modes

Local optimisation can be by simulated annealing with a fast cooling schedule — a small modification of an MCMC sampler.

Since we do not require detailed balance, the search process is simpler than that built into the T & H algorithm.

Many runs should be conducted to ensure all modes are found.

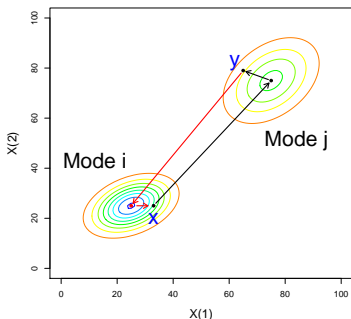
Modelling modes

A multivariate normal approximation to the target distribution at a mode provides an estimate of the overall probability of the mode.

These weights can be used in defining proposal probabilities.

Details of the two-stage approach to mode jumping

Stage 2



Current state: x

Nearest mode to x : Mode i

Propose to jump from Mode i
to Mode j (probability P_{ij})

Sample y from $h_j(y)$, the local
approximation to π at Mode j

Check y is nearest to Mode j
— if not, reject y .

For detailed balance within jump moves between Modes i and j ,
accept the move from x to y with probability

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y) P_{ji} h_i(x)}{\pi(x) P_{ij} h_j(y)}\right\}.$$

Applying our two-stage approach to the previous example

We applied our approach as follows:

Initial search for modes:

We carried out 500 runs of simulated annealing with a logarithmic cooling schedule over 50 iterations, followed by hill climbing.

We applied cluster analysis to the 500 mode locations and chose the mode in each cluster with highest $\pi(x)$.

We fitted a bivariate normal distribution at each mode and found associated weights w_i .

Sampling

Iterations comprised 20 local updates and 1 mode-jumping step — as in our application of the T & H method.

In mode jumping, we took $P_{ij} \propto w_j$ for $j \neq i$.

Performance of the two-stage method

	2-stage method	T & H
Mode jumping rate	47%	1.1%
Computation		
<i>Initial search</i>		
Function calls to π	500×120	0
<i>Sampling: calls to π per iteration</i>		
Local steps	20	20
Mode jumping steps	1	125
<i>Total function calls in 10^5 iterations</i>		
Initial exploration	6×10^4	0
Sampling	2×10^6	14×10^6

Other performance measures also show an efficiency gain of ~ 300 .

What about asymptotic theory?

When a standard MCMC algorithm is run for N steps, we have theory for the convergence of π_n , the distribution of X_N , and estimates of $E_\pi(g(X))$ based on X_1, \dots, X_N as $N \rightarrow \infty$.

Consider the two-stage method with N_1 runs of the initial search and a sampling phase of length N_2 , and let both N_1 and $N_2 \rightarrow \infty$.

Standard results hold for the sampling phase as $N_2 \rightarrow \infty$, conditional on having found all the modes.

Overall properties as N_1 and $N_2 \rightarrow \infty$ follow from the fact that

$$Pr\{\text{Fail to find all modes}\} \sim e^{-\lambda N_1} \quad \text{for some } \lambda > 0.$$

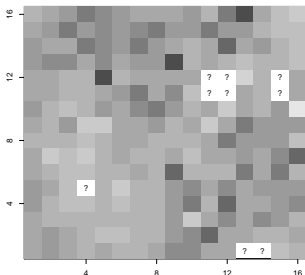
Note: T & H suggest a hybrid approach with an initial search, then more jump steps in the sampling stage.

However, for a fixed amount of computing, performing all the mode searching up-front has clear advantages.

5. Example: Phosphate levels at an archaeological site

Besag, York & Mollie (1991) analyse data from an archaeological site in Laconia, Greece (see Cavanagh et al, 1988).

Log phosphate levels



Question marks denote missing values

High phosphate levels (brighter pixels) suggest human activity.

The aim is to identify regions of human activity, with the expectation that this occurs in localised areas.

A Bayesian analysis can give a probability of activity in each pixel.

Phosphate levels at an archaeological site

Let $X = \{X(i)\}$ be an array of binary variables, with $X(i) = 1$ indicating human activity at pixel i .

As a prior distribution for X , Besag et al. assume a binary random field model

$$P(x) \propto \exp(-\beta\nu),$$

where ν is the number of pairs (i, j) of neighbouring pixels (horizontally, vertically or diagonally) with $x(i) \neq x(j)$.

Given X , log phosphate levels $Y(i)$ are assumed to be independent with

$$Y(i) \sim \begin{cases} N(4.0, \sigma^2) & \text{if } X(i) = 0, \\ N(4.5, \sigma^2) & \text{if } X(i) = 1. \end{cases}$$

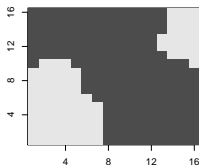
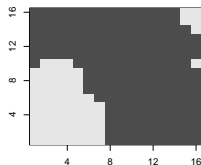
We wish to sample π , the posterior distribution of X given Y .

The case $\beta = 0.78$ and $\sigma^2 = 0.125$ gives a challenging problem.

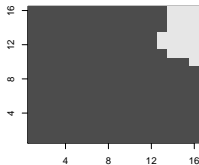
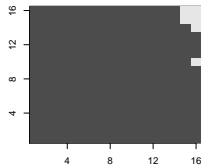
Sampling the posterior distribution, $\pi(X)$

We can run a Gibbs sampler, updating each pixel in turn.

Trial runs show that π has four main modes:



In the bottom left, an area of activity is present or absent,



In the top right, an area of activity is large or small.

Demo: The Gibbs sampler deals poorly with the bottom left area.

We would like to apply our two-stage approach to this problem.

Applying the two-stage approach to sample $\pi(X)$

We can find modes of π by multiple runs of fast simulated annealing, then use cluster analysis to identify the major modes.

Since each element of X is 0 or 1, we cannot fit a multivariate normal approximation to π at each mode.

Instead, we create a proposal Y by performing one cycle of Gibbs sampling from the new mode (Richard Sharp, Bath PhD Thesis).

Multiplying the conditional probabilities for each pixel update gives the probability $h_j(y)$ of reaching state y after starting at Mode j .

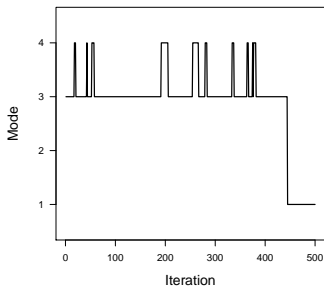
As before, we maintain detailed by accepting a jump from state x near Mode i to state y near Mode j with probability

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y) P_{ji} h_i(x)}{\pi(x) P_{ij} h_j(y)}\right\}.$$

Demo: Applying the two-stage approach.

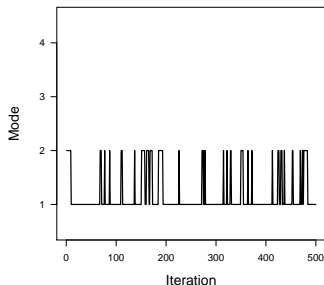
Gibbs sampling the posterior image distribution, $\pi(X)$

Starting at Mode 3: Mode index sequence



One iteration is 5 cycles of Gibbs sampling

Starting at Mode 2: Mode index sequence

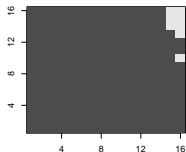
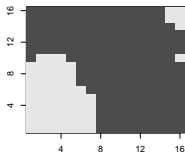


One iteration is 5 cycles of Gibbs sampling

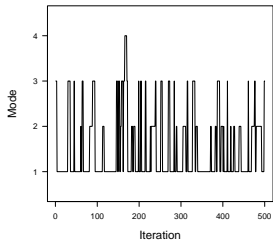
Jumps are rare between Modes 1 and 2 (activity in bottom left area) and Modes 3 and 4 (no activity in bottom left).

Applying the two-stage approach to sample $\pi(X)$

Image modes used in jump moves



Mode index sequence



One iteration: 5 cycles of Gibbs

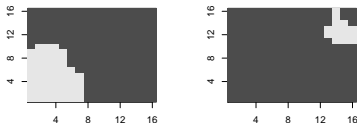
sampling + 1 jump move

There are frequent jumps between Modes 1 and 2 (activity in bottom left area) and Modes 3 and 4 (no activity in bottom left).

Comments on the two-stage approach to sample $\pi(X)$

Application is straightforward, requiring extensions of the Gibbs sampler for searching and jumping, plus use of cluster analysis.

The method can be further developed to make use of spatial properties of the distribution $\pi(x)$ by extracting “image elements” from the modal images:

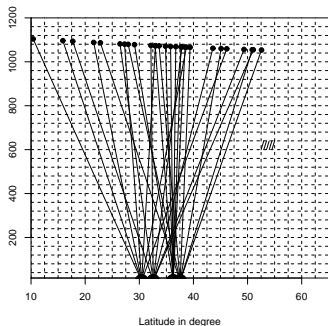


In jump steps, we “add” or “subtract” these elements, then

- It is not necessary to find all modes initially — each local feature should be present and absent in two different modes,
- In the jump steps we only need to apply Gibbs sampling around the updated “element”.

6. Example: Electron density in the ionosphere

Khorsheed, Hurn & Jennison (2011) discuss estimation of the electron density in the ionosphere, which is important for making precise GPS measurements.



Measurements between a satellite and ground receivers provide line integrals of electron content.

Khorsheed et al. use a Bayesian analysis to solve the inverse problem and, hence, create a map of electron density.

The direction of the line integrals in the data are between about 60° and vertical — making an under-determined system and a challenging inverse problem.



Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Mapping electron density in the ionosphere: A principal component MCMC algorithm

Eman Khorsheed^a, Merrilee Hurn^{b,*}, Christopher Jennison^b^a Department of Mathematics, University of Bahrain, Bahrain^b Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK

ARTICLE INFO

Article history:

Received 26 November 2008

Received in revised form 30 April 2010

Accepted 30 April 2010

Available online 9 May 2010

Keywords:

Bayesian modelling

Ionospheric mapping

Inversion

Markov chain Monte Carlo

Principal components

Tomography

ABSTRACT

The outer layers of the Earth's atmosphere are known as the ionosphere, a plasma of free electrons and positively charged atomic ions. The electron density of the ionosphere varies considerably with time of day, season, geographical location and the sun's activity. Maps of electron density are required because local changes in this density can produce inaccuracies in the Navy Navigation Satellite System (NNSS) and Global Positioning System (GPS). Satellite to ground based receiver measurements produce tomographic information about the density in the form of path integrated snapshots of the total electron content which must be inverted to generate electron density maps. A Bayesian approach is proposed for solving the inversion problem using spatial priors in a parsimonious model for the variation of electron density with height. The Bayesian approach to modelling and inference provides estimates of electron density along with a measure of uncertainty for these estimates, leading to credible intervals for all quantities of interest. The standard parameterisation does not lend itself well to standard Metropolis–Hastings algorithms. A much more efficient form of Markov chain Monte Carlo sampler is developed using a transformation of variables based on a principal components analysis of initial output.

© 2010 Elsevier B.V. All rights reserved.

Example: Electron density in the ionosphere

The physics of the ionosphere implies that vertically, electron densities follow “Chapman profiles” which have, approximately, the shape of normal densities.

Let X be the set of the 66 parameters defining 22 vertical profiles. We expect these parameters to vary smoothly in space and reflect this in the prior model of a Bayesian analysis.

For inference, we wish to sample $\pi(x)$, the posterior distribution of X given satellite-to-receiver data Y .

We found the Metropolis-Hastings MCMC algorithm to move very slowly around the distribution π , with runs from different starting points becoming stuck at different end points.

Further investigations led us to conclude that $\pi(x)$ was almost completely confined to a (non-linear) subspace of around 20 dimensions, rather than 66.

Example: Electron density in the ionosphere

For certain values of parameters in the prior model, the sub-space supporting most of $\pi(x)$ was almost linear.

In this case

- We explored this sub-space in a long initial MCMC run with very many small steps,
- We found principal components of the data points generated,
- We used these to define an efficient MCMC sampler with large updates within “the subspace” and small orthogonal steps.

A further challenge remains:

For other smoothing parameters, the sub-space is curved and the directions of principal components vary with X .

Computing the matrix of numerical second derivatives of $\log\{\pi(x)\}$ is time-consuming: we would rather not do this repeatedly to find good directions for every MCMC step.

7. Search and jump sampling for “thin” distributions

Example:

Consider the 2-D random variable $X = (X(1), X(2))$, which we write in polar co-ordinates as (R, θ) .

Suppose θ has the marginal distribution

$$f_{\theta}(\theta) = \frac{1 + 0.5 \sin(2\theta)}{2\pi} \quad \text{for } \theta \in (0, 2\pi)$$

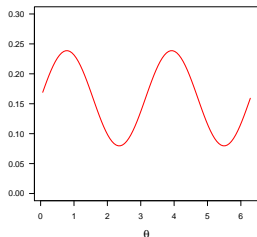
and conditional on θ , R is normally distributed as

$$R|\theta \sim N\left(\frac{1 + \theta}{2\pi}, \sigma^2\right).$$

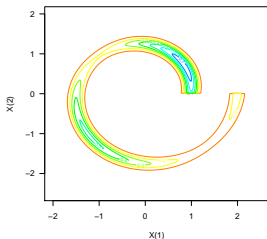
Then, for a small value of σ^2 , X lies close to the spiral curve $(r, \theta) = ((1 + \theta)/(2\pi), \theta)$, for $\theta \in (0, 2\pi)$.

Example: A distribution with “thin” support

Marginal density of θ



Contour plot of $\pi(x)^{1/500}$



Here $\sigma^2 = 0.005^2$ and the region in which π is significantly non-zero is very thin.

We have raised π to the power $1/500$ in order to make a readable contour plot.

Example: A distribution with thin support

A Metropolis-Hastings sampler updating $X(1)$ and $X(2)$ in turn makes very slow progress around the distribution π .

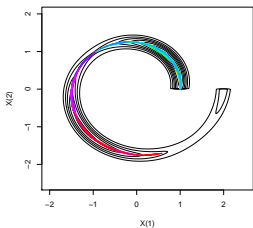
In implementing our two-stage method:

- We create a sample of points by making short runs of a Metropolis-Hastings sampler with decreasing step lengths,
- We apply cluster analysis to produce a set of “skeleton” points covering the main area of support of π ,
- At each skeleton point, we calculate numerical second derivatives of $\log\{\pi(x)\}$ and find eigen-vectors to use as proposal directions, with the variance of step lengths based on the associated eigen-values (or a local 1-D exploration if the eigen-value is negative).

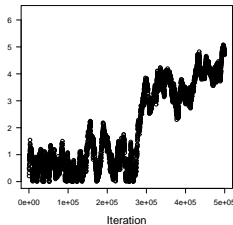
Demo:

Example: A distribution with thin support

Progress of a simple
Metropolis-Hastings sampler
over 500,000 iterations



Sequence of θ values in
500,000 M-H iterations

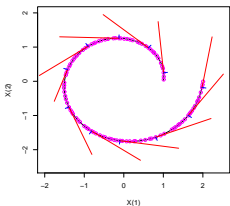


Step lengths have to be small in order to propose states with reasonably high values of $\pi(x)$ (acceptance rate = 29%).

Even though $f_{\theta}(\theta)$ stays well away from zero, there is only one transition between modes at $\theta = \pi/4$ and $\theta = 5\pi/4$ in 500,000 iterations — and no visit at all to the mode at $\theta = 2\pi$.

Eigen-vector sampling for a distribution with thin support

Skeleton points,
eigen-vector directions,
s.d.s for M-H proposals



Results from 500 short
M-H runs yielded 105
clusters.

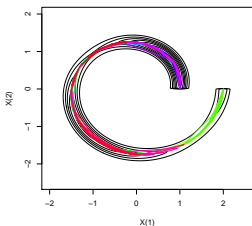
Jump directions of 11 of
the 105 skeleton points
are displayed

For each move:

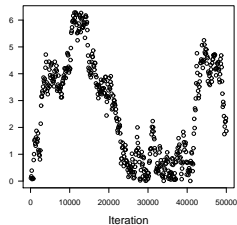
- Find the nearest skeleton point,
- Propose a move with increment based on that point's eigen-vector directions and associated s.d.s,
- Construct reverse move and find M-H acceptance probability,
- Accept or reject the proposed move.

Eigen-vector sampling for a distribution with thin support

Progress of eigen-vector
sampler over
50,000 iterations



Sequence of θ values
in 50,000 iterations



Moves are local but in optimised directions, as described above.

Step lengths used in creating proposals were tuned to give the best possible performance (acceptance rate = 17%).

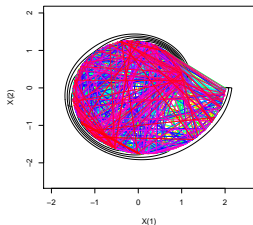
The whole state space is visited, with some repetition, in one tenth of the run length of the simple M-H sampler.

Search and jump sampling for a distribution with thin support

In a jump move:

- Current state is x ,
- Choose a skeleton point,
- Draw a proposal, y , from the local bivariate normal approximation to π ,
- Considering moves to y via the 20 nearest skeleton points, calculate the M-H acceptance probability,
- Accept or reject the proposed move.

Progress of the
jump sampler over
5,000 iterations

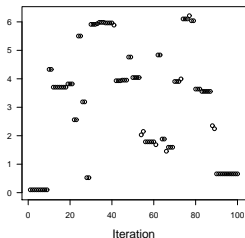


Local moves and jump steps were applied alternately.

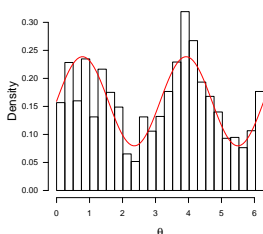
Acceptance rates were 17% for local moves, 46% for jump steps.

Search and jump sampling for a distribution with thin support

Sequence of θ values
in first 100 iterations



Histogram of θ from
5,000 iterations



Jumps between the skeleton points lead to excellent mixing of the Markov chain.

After just 5,000 iterations, the estimate of $f_{\theta}(\theta)$ is already quite accurate.

8. Discussion

We have proposed a two-stage “search and jump” algorithm for MCMC sampling of multi-modal distributions.

There are other methods designed to do this, such as simulated tempering (Marinari & Parisi, 1992, and Geyer & Thompson, 1995) or tempered transitions (Neal, 1996).

Tjelmeland & Hegstad’s (2001) method is very different in that it moves directly between modes without pausing at intermediate, low probability states.

We have taken a further, radical step in separating the exploration and sampling stage, with substantial efficiency gains as a result.

Discussion

Swendsen & Wang (1987) proposed their algorithm for sampling a binary random field, the problem faced in our image analysis example.

Our two-stage method simplifies matters by separating the search for modes from the sampling process and avoids known problems experienced by the Swendsen-Wang algorithm.

There is much current interest in MCMC sampling on (or close to) manifolds (e.g., Girolami & Calderhead, 2011). Our search and jump methods offer an innovative approach to this problem.

Another potential area of application of our methods is in sampling a distribution with a state space of variable dimension. (Brooks, Giudici & Roberts (2003) have noted the difficulties in creating acceptable “reversible jump” for dimension changing steps.)