Monte Carlo Methods for Bayesian Inference

Olivier Cappé

Centre Nat. de la Recherche Scientifique
& Ecole Nat. Sup. des Télécommunications
46 rue Barrault, 75634 Paris cedex 13, France
`http://www.tsi.enst.fr/~cappe/`

Ecole de Cosmologie, 28 août –2 septembre 2006
IESC, Cargèse

Outline

- Why Do We Need MCMC for Bayesian Inference?
  - Bayesian Modelling
  - Basic Monte Carlo
- MCMC Basics
  - (Minimal) Markov Chain Theory
  - MCMC Essentials
  - Metropolis-Hastings
  - Hybrid Kernels
- MCMC in Practice
  - Speed of Convergence
  - Scaling Issues
  - Convergence Diagnostics
- Variable Dimension Models and the Reversible Jump Approach
  - Variable Dimension Models
  - Reversible Jump MCMC
- Some References

The Bayesian Paradigm

Given a probabilistic model

$$Y \sim \ell(y|x), \quad x \in \mathcal{X}$$

where $\ell(y|x)$ denotes a parameterized density known as the likelihood, Bayesian inference postulates that the parameter x be embedded with a probability distribution π called the prior.

The Inference

is based on the distribution of x *conditional on the realized value of* Y

$$\pi(x|Y) = \frac{\ell(Y|x)\pi(x)}{\int_{\mathcal{X}} \ell(Y|x')\,\pi(x')\,dx'}$$

which is known as the posterior.

Feasibility of Bayesian Inference

In most of the cases, the normalizing constant (sometimes called the *evidence*)

$$\pi(x|Y) = \frac{\ell(Y|x)\pi(x)}{\int_{\mathcal{X}} \ell(Y|x')\,\pi(x')\,dx'}$$

may not be determined analytically and hence the posterior is known up to a constant only, which is usually denoted by writing

$$\pi(x|Y) \propto \ell(Y|x)\pi(x)$$

Posterior inference

Eg. determining the Minimum Mean Square Estimate of x, E$[x|Y]$, is not feasible except in the simplest Bayesian models.

Additional difficulty

For cosmological data analysis, we are only allowed to chose the prior (while in many case the statistician is also responsible for building the likelihood)
Hence, several major tools in Bayesian computational inference are useless in this context:

- latent variables
- hierarchical models
- conjugate priors
- . . .
- the Gibbs sampler
- methods that rely on likelihood slices being log-concave or with computable level sets, etc.
- . . .

Basic Monte Carlo Doesn't Solve the Problem

Standard independent Monte Carlo — with $\pi(x)$ as instrumental distribution — usually is very unreliable

Self-Normalized Importance Sampling

Simulate $\{X_i\}_{1 \le i \le n}$ from r and estimate $\mathrm{E}[f(X)|Y]$ by

$$\frac{\sum_{i=1}^{n} W_i f(X_i)}{\sum_{i=1}^{n} W_i}$$

where

$$W_i = \ell(Y|X_i)\pi(X_i)/r(X_i)$$

Works better but requires that some aspects of $\pi(x|Y)$ be known (tail behavior) and does not scale well either in large dimensions

Transition Kernel

The probability distribution of a Markov chain $\{X_i\}_{i\geq 1}$ on X is fully determined by its initial distribution $\nu(x)$ and its transition kernel $q(x, x')$, which are such that

$$\mathsf{P}(X_1 \in A) = \int_A \nu(x) dx$$

$$\mathsf{P}(X_i \in A | X_1, \dots, X_{i-1}) = \int_A q(X_{i-1}, x) dx$$

Chapman-Kolmogorov Equations

$$\mathsf{P}(X_{n+1} \in A) = \int_{x \in \mathsf{X}} \int_{x' \in A} \nu(x) q^n(x, x') dx dx'$$

where

$$\begin{aligned} q^n(x, x'') &\overset{\text{def}}{=} \int q^{n-1}(x, x') q(x', x'') dx' \\ &= \int q^{n-k}(x, x') q^k(x', x'') dx' \end{aligned}$$

for any $0 \leq k \leq n$.

- $q^n(X_1, x)$ is the conditional probability density function of X_{n+1} given X_1.

Stationary Distribution

Definition

π is stationary for q if

$$\int \pi(x)q(x,x')dx = \pi(x')$$

Hence π is a stationary point of the kernel q, viewed as an operator on probability density functions.

- It is easily checked that this implies that if $\nu = \pi$,

$$\mathsf{P}(X_i \in A) = \int \pi(x)dx$$

for all $i \geq 1$.

Detailed Balance Condition and Reversibility

Determining the stationary distribution(s) is hard in general, except in cases where the following stronger condition holds.

Detailed Balance Condition

$$\pi(x)q(x,x') = \pi(x')q(x',x) \qquad \text{for all } (x,x') \in \mathsf{X}^2$$

The chain is then said to be π-reversible and π is a stationary distribution.

Proof.

$$\int \pi(x)q(x,x')dx = \int \pi(x')q(x',x)dx = \pi(x')$$

□

Convergence to Stationary Distribution

If π is a stationary distribution, and under additional regularity conditions not discussed here, the following properties hold

Convergence in Distribution

$$\mathsf{P}(X_n \in A) \to \int_A \pi(x)dx \quad \text{(irrespectively of } \nu\text{)}$$

Law of Large Numbers (Ergodic theorem)

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{\text{a.s.}} \int f(x)\pi(x)dx$$

Central Limit Theorem

$$\frac{\sqrt{n}}{\sigma_{\pi,q,f}}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int f(x)\pi(x)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

. . .

Markov Chain Monte Carlo (MCMC) in a Nutshell

1. Given a target distribution π, which may be known up to a constant only, find a transition kernel which is π-reversible, i.e., such that

$$\pi(x)q(x,x') = \pi(x')q(x',x)$$

2. Simulate a (long) section $X_1, \dots, X_n$ of a chain with kernel q started from an arbitrary point X_1 and compute the Monte Carlo estimate

$$\widehat{\mathsf{E}_\pi}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

of $\int f(x)\pi(x)dx$, perhaps discarding in the sum the very first iterations (so called burn-in period).

Rao-Blackwellization

If we can find (X, Z) such that $X \sim \pi$, $Z \sim \nu$ and $\mathsf{E}\left[f(X)|Z\right]$ may be computed in closed-form,
MCMC simulation $Z_1, \ldots, Z_n$ are performed using ν as target distribution and the Rao-Blackwellized estimator

$$\widehat{\mathsf{E}_\pi^{RB}}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[f(X)|Z_i\right]$$

is used, rather than $\widehat{\mathsf{E}_\pi}(f)$.

The Rao-Blackwell Theorem shows that

$$\mathsf{Var}\left(\widehat{\mathsf{E}_\pi^{RB}}(f)\right) \leq \mathsf{Var}\left(\widehat{\mathsf{E}_\pi}(f)\right)$$

for independent simulations. This does not necessarily hold true for MCMC simulations, but empirically it does in most settings.

- Usually, Rao-Blackwellization is used with Z being a sub-component of X.

Metropolis-Hastings Algorithm

Simulate a Markov chain$\{X_i\}_{i\geq 1}$ with the following mechanism: given X_i,

1. Generate $X_\star \sim r(X_i, \cdot)$, independently of past simulations;
2. Set

$$X_{i+1} = \begin{cases} X_\star \text{ with probability } \alpha(X_i, X_\star) \overset{\text{def}}{=} \frac{\pi(X_\star)\, r(X_\star, X_i)}{\pi(X_i)\, r(X_i, X_\star)} \wedge 1 \\ X_i \text{ otherwise} \end{cases}$$

Note that the acceptance probability is computable also in cases where π is known up to a constant only

π-Reversibility of the Metropolis-Hastings Kernel

Proof.

$$\pi(x)\alpha(x,x')r(x,x') = \pi(x')r(x',x) \wedge \pi(x)r(x,x')$$

which imply that the transition kernel K associated with the Metropolis-Algorithm

$$K(x,dx') = \alpha(x,x')r(x,x')\,dx' + p_R(x)\,\delta_x(dx')$$

where $p_R(x)$ is the probability of remaining in the state x, given by

$$p_R(x) = 1 - \int \alpha(x,x')r(x,x')\,dx'$$

is $\pi(x)dx$-reversible. □

Two Simple Cases

Independent Metropolis-Hastings $r(x, \cdot)$ is a fixed — that is, independent of x — probability density function $r(\cdot)$: the proposed chain updates are i.i.d. and the acceptance probability then reduces to

$$\alpha(x, x') = \frac{\pi(x')/r(x')}{\pi(x)/r(x)} \wedge 1$$

Random Walk Metropolis-Hastings $r(x, x') = r(x' - x)$, that is, the proposals are generated as $X_\star = X_i + U$ where $U \sim r$. The acceptance probability is then

$$\alpha(x, x') = \frac{\pi(x')}{\pi(x)} \wedge 1$$

My First Sampler

Random Walk Metropolis-Hastings

```
for i = 1 ...
    x_new = x[i-1] + symmetric_perturbation(scale)
    post_new = compute_unnormalized_posterior(x_new)
    if (rand < post_new/post)
        x[i] = x_new
        post = post_new
    else(if)
        x[i] = x[i-1]
    end(if)
end(for)
```

Hybrid Kernels

Assume that $K_1, \dots, K_m$ are Markov transition kernels that all admit π as stationary distribution. Then

1. $K_{\text{syst}} = K_1 K_2 \cdots K_m$ and
2. $K_{\text{rand}} = \sum_{i=1}^m \alpha_i K_i$, with $\alpha_i > 0$ for $i = 1, \dots, m$ and $\sum_{i=1}^m \alpha_i = 1$,

also admit π as stationary distribution. If in addition $K_1, \dots, K_m$ are π reversible, K_{rand} also is π reversible but K_{syst} need not be.

Most MCMC algorithms combine several type of transitions, in particular with proposals that change only one component of X (one-at-a-time Metropolis-Hastings)

How Does This Work?

Discuss the practical use of MCMC with topics such as

1. How fast does it converges?
2. Should I use a burn-in period, parallel chains?
3. How to chose the scale of the proposal in RW-MH ?
4. How does the method scales in large dimensions?
5. What's the point of looking at the simulation path?
6. Should I trust convergence diagnostics (integrated autocorrelation time, Raftery & Lewis, Gelman & Rubin)?

How Fast Does it Converge?

Asymptotically, the error is controlled by the scaling term in the CLT: $\sigma_{\pi,q,f}/\sqrt{n}$ where

$$\sigma^2_{\pi,q,f} = \mathsf{Var}_\pi(f) \times \tau_{\pi,q,f}$$

and

$$\tau_{\pi,q,f} = 1 + 2\sum_{i=1}^{\infty} \mathsf{Corr}_{\pi,q}(f(X_0), f(X_i))$$

is the *integrated autocorrelation time*

In Contrast With Independent Monte Carlo

- Only an asymptotic result (not finite n variance)
- Estimating $\tau_{\pi,q,f}$ reliably is a hard task

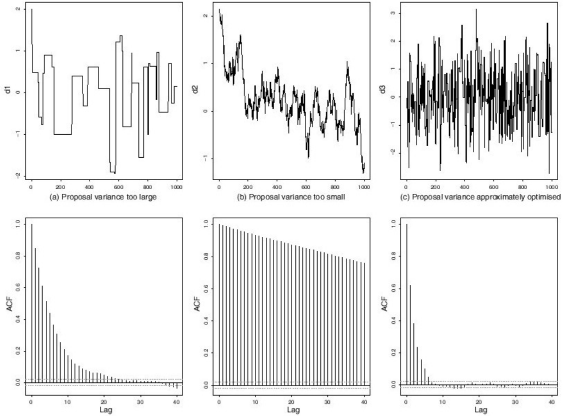
Burn-In Period and Parallel Chains

Not very popular among MCMC pundits as letting n be as large as possible is the only way to ensure convergence

- The burn-in period is mostly and issue for those who know that they are not using enough simulations
- Parallel chains are often used to assess convergence (more on this latter) and estimating $\sigma_{\pi,q,f}$
- Parallel chains are mostly of interest when parallel computing is an option (otherwise use a single chain as long as possible)

How to Chose the Scale of the Proposal in RW-MH?

Try yourself at http://www.lbreyer.com/classic.html



(a) Proposal variance too large

(b) Proposal variance too small

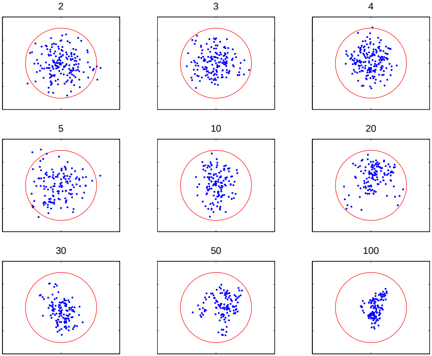
(c) Proposal variance approximately optimised

FIG. 2. *Simple Metropolis algorithm with* (a) *too-large variance (left plots),* (b) *too-small variance (middle) and* (c) *appropriate variance (right). Trace plots (top) and autocorrelation plots (below) are shown for each case.*

From (Roberts & Rosenthal, 2001)

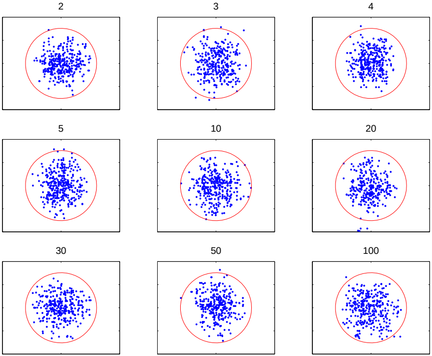
How Does the Method Scales in Large Dimensions?

(Gelman, Gilks & Roberts, 1997), (Roberts *et al.*, 1997-2001) have studied scaling properties of RW-MH in large dimensions



Optimal scaling when acceptance rate is about 23% and proposal standard deviation about $2.4\,\sigma_\pi/\sqrt{d}$

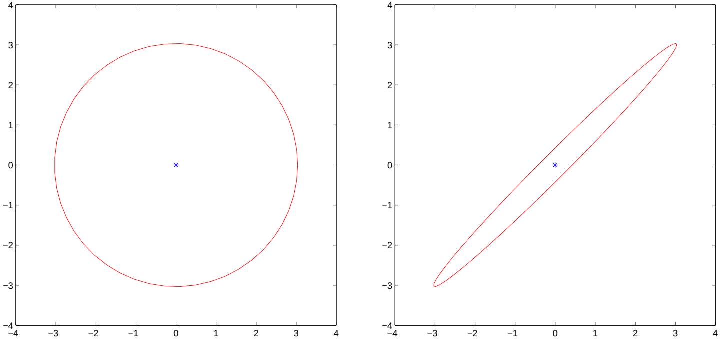
Different Proposals May Tell a Different Story



- one-at-a-time RW-MH yields d independent chains in this (very particular) case
- Numerical complexity of the alternatives must be evaluated carefully

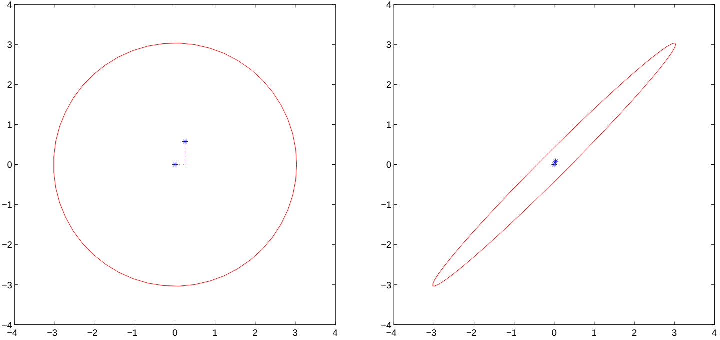
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1



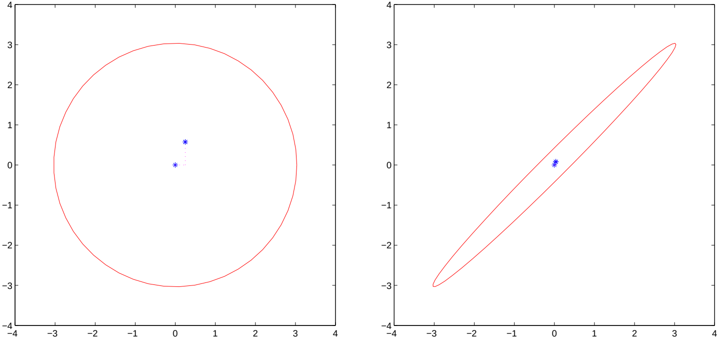
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2



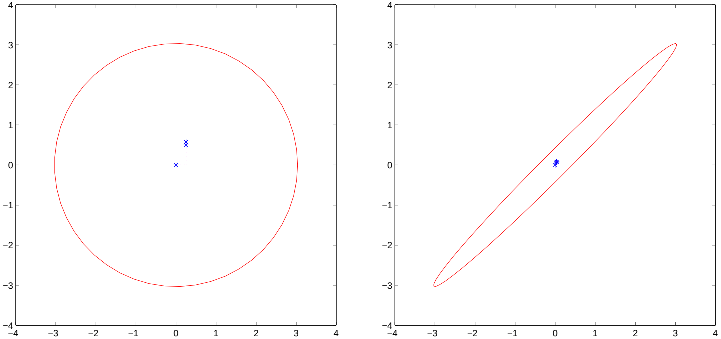
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3



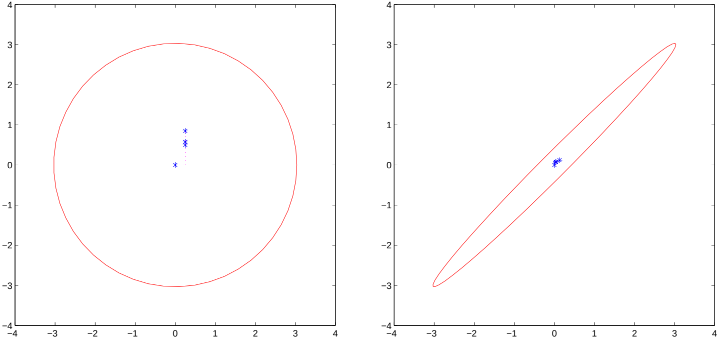
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4



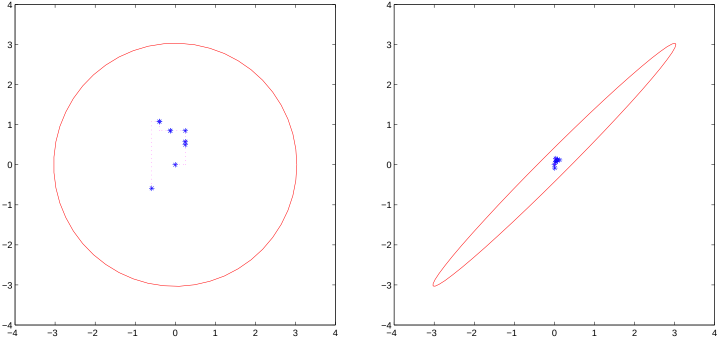
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4, 5



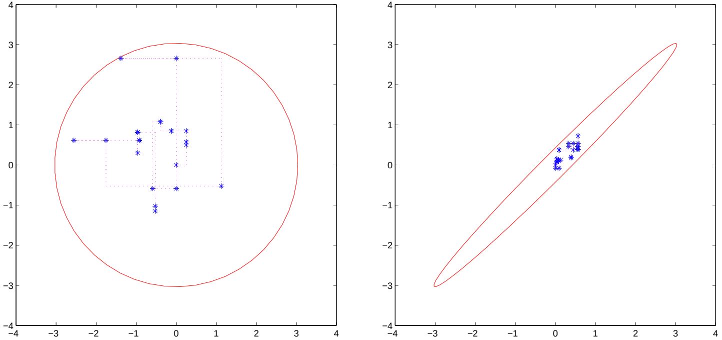
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4, 5, 10



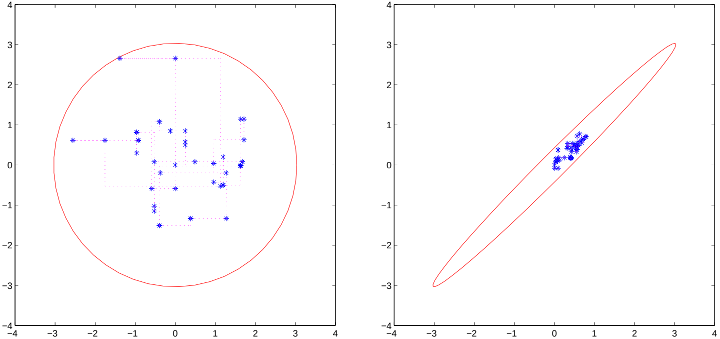
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25



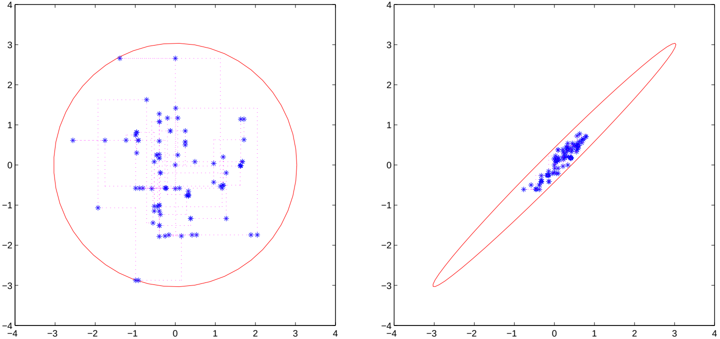
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50



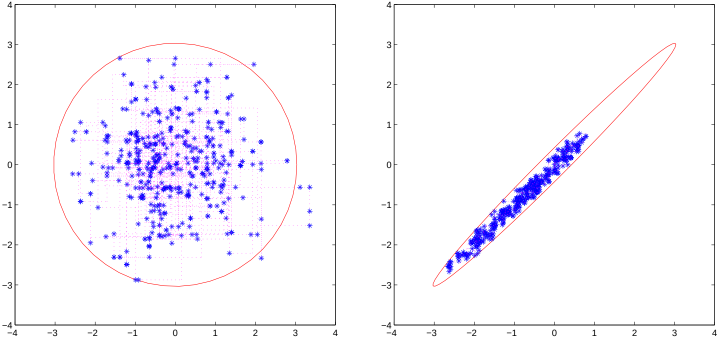
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100



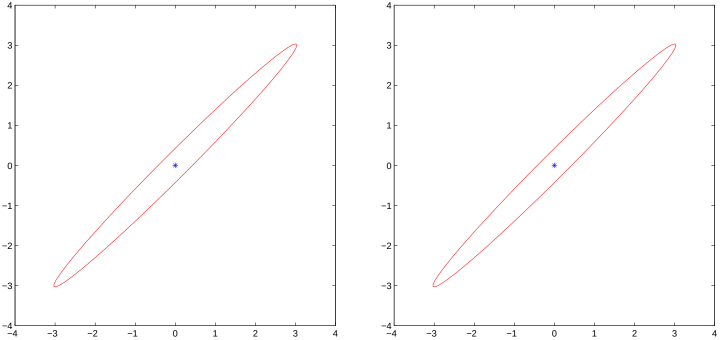
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500



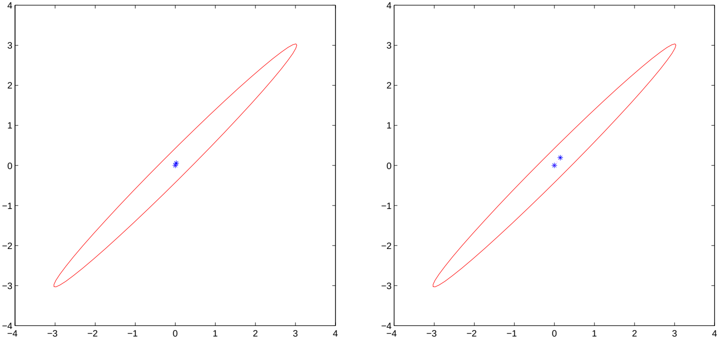
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1



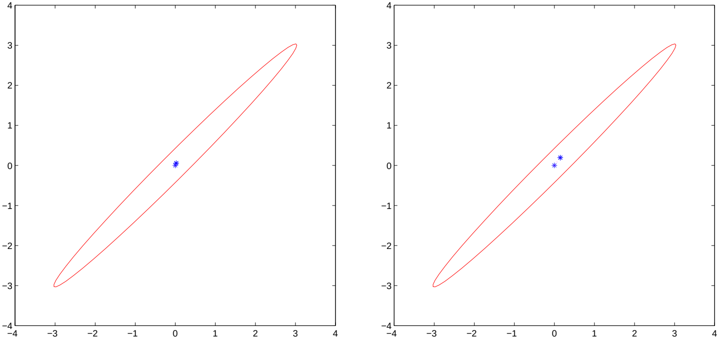
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2



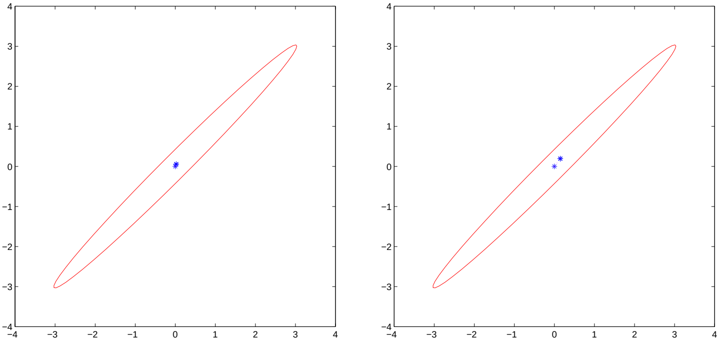
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2, 3



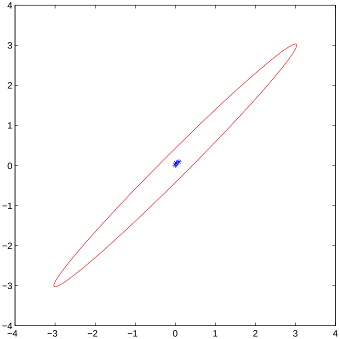
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

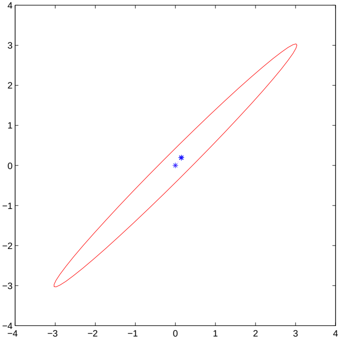
Number of Iterations 1, 2, 3, 4



Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

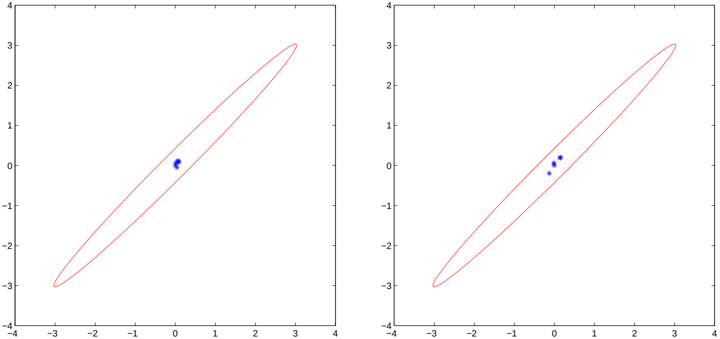
Number of Iterations 1, 2, 3, 4, 5





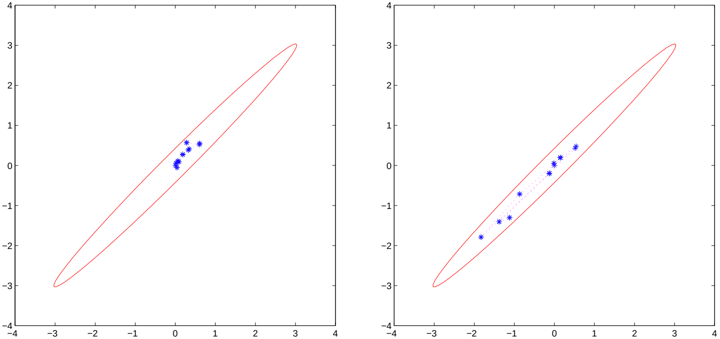
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2, 3, 4, 5, 10



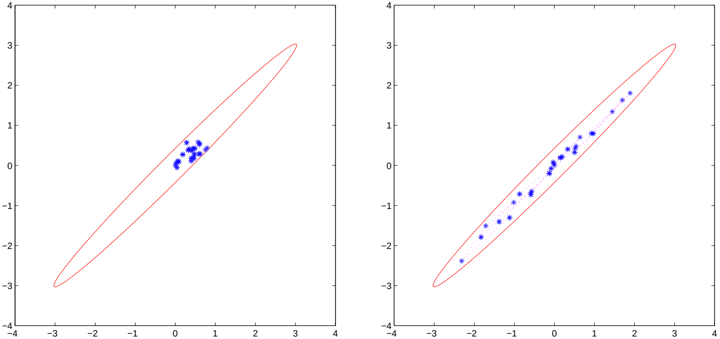
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25



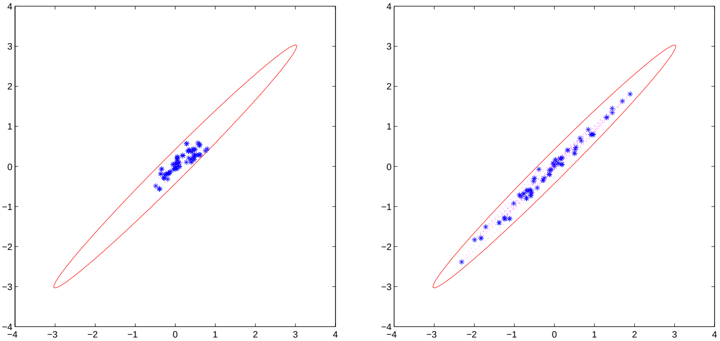
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50



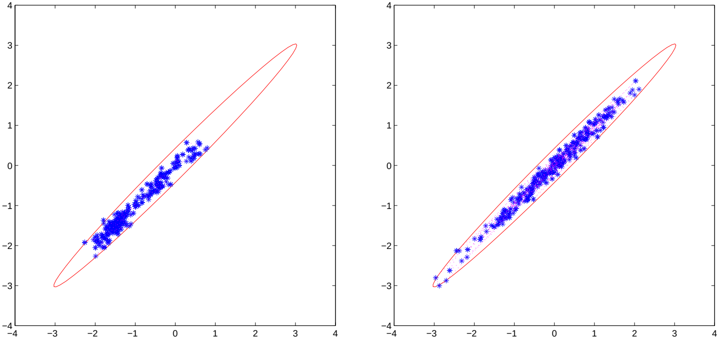
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100



Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with $\underline{\text{knowledge of } \Sigma_\pi}$ and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500



When Should the Chain be Stopped?

Three types of convergence:

Convergence to the Stationary Distribution Minimal requirement for approximation of simulation from π

Convergence of Averages convergence of the empirical averages

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathsf{E}_{\pi}(f)$$

most relevant in the implementation of MCMC algorithms

Convergence to i.i.d. Sampling How close a sample $X_{i_1}, \ldots, X_{i_d}$ is to being i.i.d.?

This is Not an Easy Task!

Theoretical Answers Only in very restricted class of models and algorithms; nonetheless provide interesting insights (eg. importance of tail behavior)

Graphical Methods Looking at trajectories of X_n, at partial sums $1/n \sum_{i=1}^{n} f(X_i)^*$, estimating the cumulated autocorrelations, comparing half chain boxplots, monitoring the acceptance rate, etc.

- None of this is effective in presence of a severe mixing problem

[*] (Raftery & Lewis, 1992) corresponds to a (very) approximate criterion computed on binary functions f

Multiple Runs are Helpful

(Gelman & Rubin, 1992) suggest a numerical criterion based on the comparison of

$$\begin{aligned} B_n &= \frac{1}{M} \sum_{m=1}^{M} (\bar{\xi}_m - \bar{\xi})^2 \,, \\ W_n &= \frac{1}{M} \sum_{m=1}^{M} \frac{1}{n} \sum_{i=1}^{n} (\xi_i^{(m)} - \bar{\xi}_m)^2 \,, \end{aligned}$$

with

$$\bar{\xi}_m = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(m)}, \qquad \bar{\xi} = \frac{1}{M} \sum_{m=1}^{M} \bar{\xi}_m \qquad \text{and } \xi_i^{(m)} = f(X_i^{(m)})$$

B_n and W_n represent the between- and within-chains variances

Variable Dimension Model

A variable dimension model is defined as a collection of models (here, identified with parameter spaces),

$$\mathcal{X}_r, \quad r = 1, \ldots, R\,,$$

associated with a collection of priors on these spaces,

$$\pi_r(x_r), \quad r = 1, \ldots, R\,,$$

and a prior distribution on (the indices of) these spaces,

$$\varrho(r), \quad r = 1, \ldots, R\,.$$

The model-and-parameter space is defined as

$$\mathcal{X} = \bigcup_{r=1}^{R} \{r\} \times \mathcal{X}_r$$

Bayesian Posteriors in Variable Dimension Models

Structure of the Posterior Distribution

Given observations Y, the posterior is such that

$$\pi(x|Y) = \pi(r, x_r|Y) = \frac{\overbrace{\ell_r(Y|x_r)}^{\text{data likelihood}} \; \overbrace{\pi_r(x_r)}^{\text{parameter prior}} \; \overbrace{\varrho(r)}^{\text{model prior}}}{\underbrace{\sum_{r=1}^{R} \int_{\mathcal{X}_r} \ell_r(Y|x_r)\pi_r(x_r)\varrho(r)\, dx_r}_{\text{non-computable normalizing constant}}}$$

- How do we design MCMC moves that can connect points from a smaller dimensional space $\mathcal{X}_s$ to a larger dimensional one $\mathcal{X}_l$?

Reversible Jump Approach (Green, 1995)

1. The algorithm is of Metropolis-Hastings type (where proposed moves are, or are not, accepted).
2. Move proposals must be very simple, as we must be able to compute the probability of jumping from $x_s \in \mathcal{X}_s$ to any reachable $x_l \in \mathcal{X}_l$ as well as the converse.
3. The simplest solution is to make the move from $\mathcal{X}_l$ to $\mathcal{X}_s$ deterministic.

Note that each individual move may not be able to reach all the points in $\mathcal{X}_l$; but the combination of all possible moves (incl. fixed-dimensional moves) has to, in order to ensure irreducibility.

The Basic Case: Birth / Death Moves

Birth When in $x_s \in \mathcal{X}_s$, with probability $P_{s,l}$, draw an independent $V_\star \sim p$ and let $x_l = (x_s, V_\star)$[†].

Death When in $x_l \in \mathcal{X}_l$, with probability $P_{l,s}$ truncate x_l to its $\dim(\mathcal{X}_s)$ first components.

The acceptance probability for the birth move may be written as $A(x_s, x_l) \wedge 1$ where

$$A(x_s, x_l) = \frac{\varrho(l)\pi_l(x_l)\ell_l(Y|x_l)P_{l,s}}{\varrho(s)\pi_s(x_s)\ell_s(Y|x_s)P_{s,l}\,p(V_\star)}$$

The acceptance probability for the death move is $A^{-1}(x_s, x_l) \wedge 1$.

[†] Hence, $V_\star$ is of dimension $\dim(\mathcal{X}_l) - \dim(\mathcal{X}_s)$.

The More Elaborate Case: Split / Merge

Split When in $x_s \in \mathcal{X}_s$, with probability $P_{s,l}$, draw an independent $V_\star \sim p$ and let $x_l = m(x_s, V_\star)$ where m is an invertible transform.

Merge When in $x_l \in \mathcal{X}_l$, with probability $P_{l,s}$ truncate $m^{-1}(x_l)$ to its $\dim(\mathcal{X}_s)$ first components.

The acceptance ratio is now given by

$$A(x_s, x_l) = \frac{\varrho(l)\pi_l(x_l)\ell_l(Y|x_l)P_{l,s}}{\varrho(s)\pi_s(x_s)\ell_s(Y|x_s)P_{s,l}\,p(V_\star)} J_{s,l}(x_l)$$

where

$$J_{s,l}(x_l) = \left|\frac{\partial m(x_s, v)}{\partial(x_s, v)}\right|_{(x_s,v)=m^{-1}(x_l)} = \left|\frac{\partial m^{-1}(x_l)}{\partial x_l}\right|^{-1} .$$

is the determinant of the Jacobian of m.

Typical Choices of the Split Mapping

Most often, the split mapping operates on just one component of x_s, say $x_s(i)$ and the split is done according to, e.g.,

- $m(x_s(i), V_\star) = (x_s(i) - V_\star, x_s(i) + V_\star)$ with $V_\star \sim \mathsf{N}(0, \sigma^2)$[‡], if $x_s(i)$ is a real parameter (mean, regression coefficient, etc.)
- $m(x_s(i), V_\star) = (x_s(i)\,\mathsf{e}^{-V_\star}, x_s(i)\,\mathsf{e}^{V_\star})$ with $V_\star \sim \mathsf{N}(0, \sigma^2)$, if $x_s(i)$ is a positive parameter (variance, etc.)
- ...

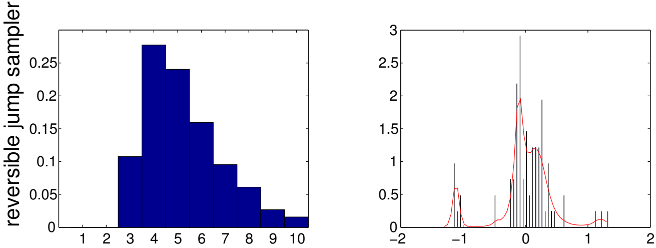
[‡] Or other symmetric distribution on $\mathbb{R}$.

A (Reasonably) Simple Example: Gaussian Mixtures

A C code example in `http://www.tsi.enst.fr/~cappe/ctrj_mix/` for the model

$$p(y|\theta_r) = \sum_{i=1}^{r} \frac{w_i}{\sqrt{2\pi\upsilon_i}} \exp\left[-\frac{(y-\mu_i)^2}{2\upsilon_i}\right]$$

assuming independent observations $Y_1, \ldots, Y_n$ (and r unknown!).



Birth or Death Moves

When in a r components configuration, we propose a new component from the prior according to

1. $w^\star_{r+1} \sim \text{Beta}(1, r)$ with $w_{1:r+1} = \left((1 - w^\star_{r+1})w_{1:r}, w^\star_{r+1}\right)$
2. $\mu^\star_{r+1} \sim \text{Normal}(0, \kappa)$
3. $\upsilon^\star_{r+1} \sim \text{Inverse-Gamma}(\alpha, \beta)$

The acceptance ratio for the birth move is

$$\frac{\ell(Y_1, \ldots, Y_n | \theta_{r+1})}{\ell(Y_1, \ldots, Y_n | \theta_r)} \times \frac{P_D(r+1)}{P_B(r)} \wedge 1$$

Note that the choice of the prior as proposal simplifies the acceptance ratio.

Split or Merge Moves

When in a r components configuration, we propose to split component i according to

1. $w_i \longrightarrow (w', w_i'') = (V_w^\star w_i, (1 - V_w^\star) w_i)$ with $V_w^\star \sim \text{Beta}(\gamma_S, \gamma_S)$
2. $\mu_i \longrightarrow (\mu_i', \mu_i'') = (\mu_i - V_\mu^\star, \mu_i + V_\mu^\star)$ with $V_\mu^\star \sim \text{Normal}(0, \rho_S)$
3. $\upsilon_i \longrightarrow (\upsilon_i', \upsilon_i'') = (\upsilon_i / V_\upsilon^\star, \upsilon_i V_\upsilon^\star)$ with $V_\upsilon^\star \sim \text{Log-Normal}(0, \nu_S)$

Some References

- C. P Robert & G Casella, *Monte Carlo statistical methods*, Springer, 1999.
- G. O. Roberts & J. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms*, Statistical Science, 2001, Vol. 16, No. 4, 351–367 (and references therein).
- A. Gelman & D. B. Rubin, *Inference from iterative simulation using multiple sequences*, Statistical Science, 1992, Vol. 7, No. 4, pp. 473–483, see also, C. J. Geyer *Practical Markov chain Monte Carlo* (pp. 473–483 in the same issue) as well as discussion of both papers (pp. 483–511).
- P. J. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika, 1995, Vol. 82, pp. 711–732.

The End

Thank you for your attention.