Monte Carlo Methods for Cosmological Models

Olivier Cappé*

Laboratoire Traitement et Communication de l'Information,
ENST & CNRS

XXIIIrd IAP Colloquium – 6 July 2007

*Thanks to colleagues in the ECOSSTAT project: Simon Prunet, Karim Benabed, Yannick Mellier, François Bouchet, Martin Kilbinger at IAP; Olivier Le Fèvre at LAM; Christian P. Robert at CEREMADE; and Jean-François Cardoso at LTCI

Outline

The Bayesian Paradigm

Given a probabilistic model

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

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

The Inference

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

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

which is known as the posterior

Bayesian Inference = Integration w.r.t. the Posterior π

Parameter Estimation The Minimum Mean Square Estimate of $\mathcal{X}$ is $\mathrm{E}[X|Y]$ (other possible estimators include the mode of π — so-called Maximum A Posteriori estimator)

Model Comparison Based on the evidence

$$e(Y) = \int_{\mathcal{X}} \ell(Y|x)\,\rho(x)\,dx$$

or, more usefully, the Bayes factor

$$f_{1,2}(Y) = \frac{\int_{\mathcal{X}_1} \ell_1(Y|x_1)\,\rho_1(x_1)\,dx_1}{\int_{\mathcal{X}_2} \ell_2(Y|x_2)\,\rho_2(x_2)\,dx_2}$$

Predictive Inference Based on the predictive distribution

$$p(y^\star|Y) = \int_{\mathcal{X}} \ell(y^\star|x)\pi(x)$$

Feasibility of Bayesian Inference

In all but some very specific cases however,

- the normalising constant (the *evidence*)

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

  may not be determined analytically and hence the posterior is known up to a constant only
- More generally, integration with respect to π is usually not available

Monte Carlo Propaganda

Without further assumptions[†] on ℓ and ρ, Monte Carlo methods (in a general sense) are about the only approaches that guarantee an approximation error of order $1/\sqrt{n}$ in the sense that

$$\mathrm{E}\left[(\hat{\pi}_n(f) - \pi(f))^2\right] = \frac{v(f)}{n} + o\left(\frac{1}{n}\right)$$

where

- f is an (almost) arbitrary function
- $\pi(f)$ is the true expectation of h under the posterior
- $\hat{\pi}_n(f)$ is the approximation obtained after n evaluations of the likelihood

[†] Regularity, peakiness of l, etc.

There are some issues still...

In $$\mathrm{E}\left[\left(\hat{\pi}_n(f)-\pi(f)\right)^2\right]=\frac{\upsilon(f)}{n}+o\left(\frac{1}{n}\right)$$

- the error is (pseudo) random
- the $o\left(\frac{1}{n}\right)$ term could be significant for finite n (e.g. convergence to stationary distribution for Markov Chain Monte Carlo methods) and/or depend on initialisation, etc.
- $\upsilon(h)$ may vary significantly depending on the choice of f, the dimension of the problem, etc.
- $\upsilon(f)$ depends on the chosen method (and variations by factors of order 100 or more are not uncommon)

There's a Limited Range of Available Monte Carlo Methods

In general, standard independent Monte Carlo — with the prior $\rho(x)$ as instrumental distribution — is very unreliable

In Cosmological Models

- We cannot chose ℓ (and maybe not even ρ)
- There isn't only one function f of interest (we want marginals, 2D-marginals, covariances, etc.)
- ℓ has no probabilistic structure
- The cost of computing ℓ is (almost) indivisible
- We cannot use conjugate priors, control variates, data augmentation, the Gibbs sampler, Rao-Blackwellisation, blocking schemes, ...

Random Walk Metropolis-Hastings Algorithm

Simulate a trajectory of a Markov chain $(X_i)_{i\geq 1}$ with the following mechanism:

Given X_i

1. Generate $X_\star = X_i + U_{i+1}$, independently of past simulations
2. Set

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

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

Under fairly mild assumptions (when $\mathcal{X}$ is compact),

- π is the unique stationary distribution of (X_i)
- (X_i) converge in distribution to π, irrespectively of X_0,

$$\text{i.e.} \quad \mathrm{E}[f(X_i)] \xrightarrow[i\to\infty]{} \pi(f)$$

- (X_i) is ergodic:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow[n\to\infty]{} \pi(f)$$

- and satisfies $n\,\mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \pi(f)\right)^2\right] \to \upsilon(f)$ with

$$\upsilon(f) = \mathrm{Var}_\pi[f(X)]\left\{1 + 2\sum_{i=1}^{\infty} \mathrm{Corr}_{X_0\sim\pi}(f(X_0), f(X_i))\right\}$$

Summary

Recipe Leave a single chain running as long as possible and compute

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

for the functions f of interest

Options

- Burn-in period (folklore)
- Multiple parallel independent chains (good for reducing variances, diagnosing convergence or speeding up computations but biased unless each individual chain is run long enough)

Design Parameters If type of random perturbation is fixed (say Gaussian) only its scale (or covariance) has to be chosen

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

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

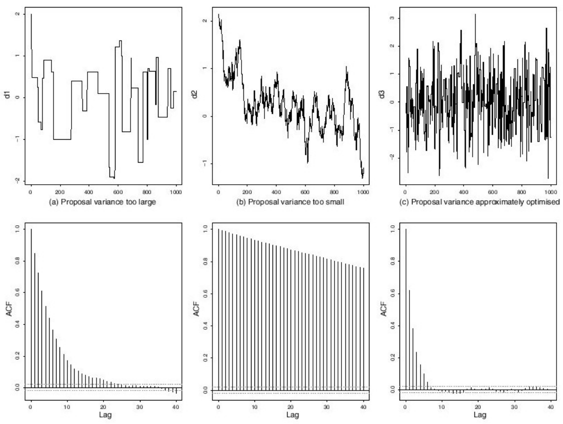


FIG. 2. *Simple Metropolis algorithm with* (a) *too-large variance (left plots),* (b) *too-small variance (middle) and* (c) *appropriate variance (right). Trace plots (top) and autocorrelation plots (below) are shown for each case.*

From (Roberts & Rosenthal, 2001)

How Does the Method Scale in Large Dimension?

The Scaling Construction (Roberts *et al.*, 1997-2001)

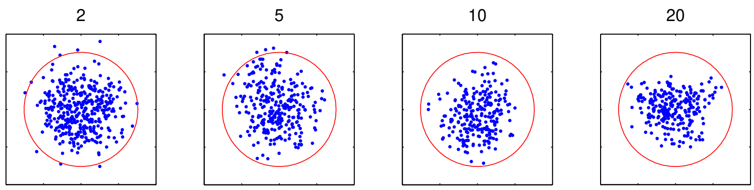
Assume a d-dimensional target of product form $\pi_{(d)}(x_1, \ldots, x_d) = \prod_{k=1}^{d} \pi(x_k)$ and an isotropic (d-dimensional) proposal Gaussian distribution with variance $\sigma^2_{(d)}$. Now let the dimension d tends to infinity

- the acceptance rate does not degenerate only if $\sigma^2_{(d)} = O(1/d)$

And for a coordinate projection

- the optimal choice of $\sigma^2_{(d)}$ is $2.38^2/d \times \mathrm{Var}_\pi(X)$, which correspond to an acceptance rate of 23%
- the asymptotic variance of the ergodic average is $\alpha d \times \mathrm{Var}_\pi(X)$ (for a Gaussian target, $\alpha \approx 3.7$)

RW-MH Scaling Example

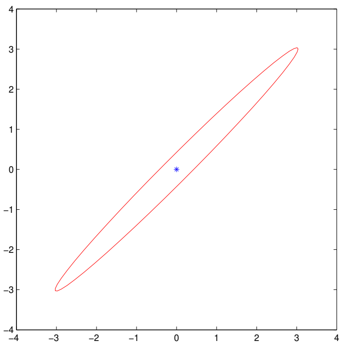


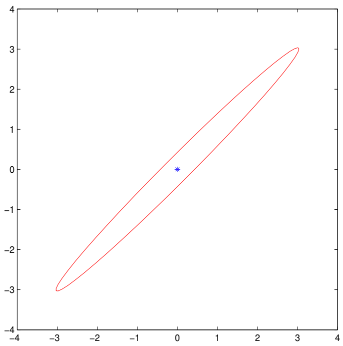
First 1 000 samples of the RW-MH algorithm started from mode for a multivariate Gaussian distribution with optimal proposal scaling (arbitrary 2D marginal for increasing values of d)

The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

Number of Iterations 1

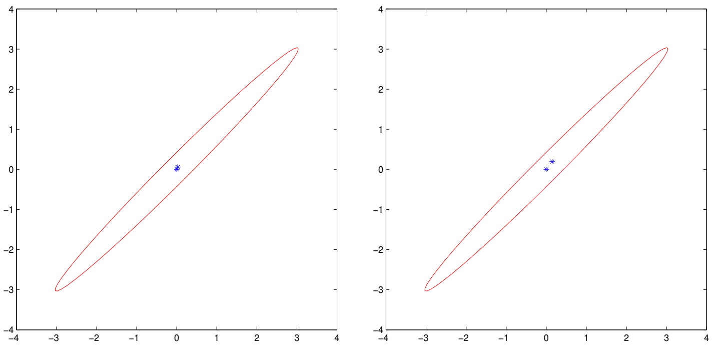




The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

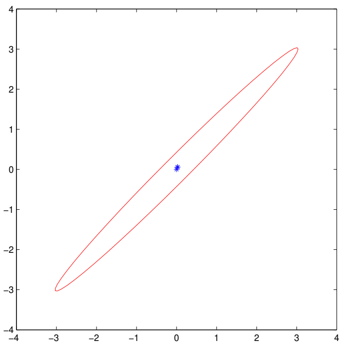
Number of Iterations 1, 2

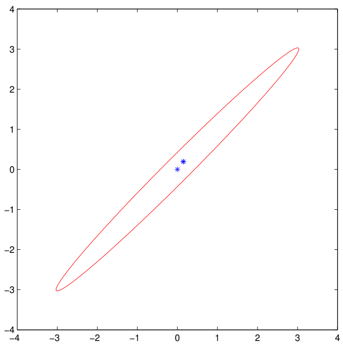


The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

Number of Iterations 1, 2, 3

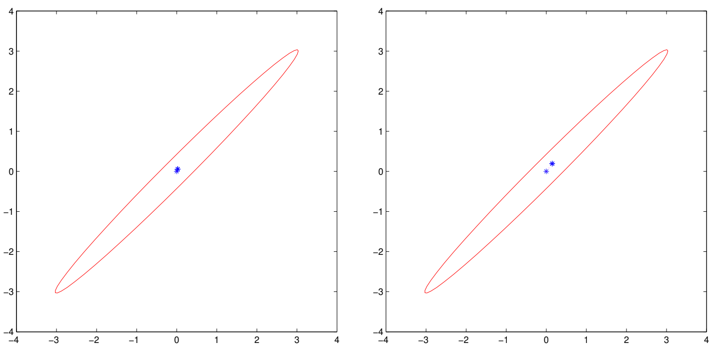




The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

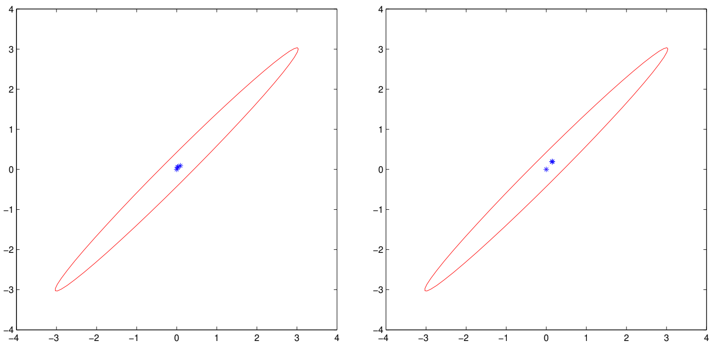
Number of Iterations 1, 2, 3, 4



The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

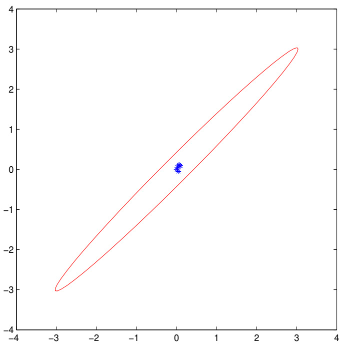
Number of Iterations 1, 2, 3, 4, 5

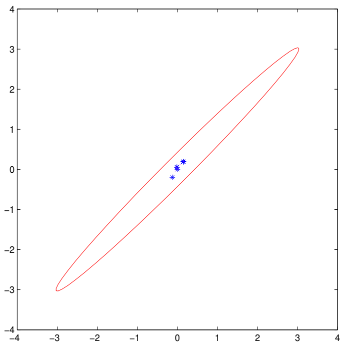


The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\rm prop} = 0.04\, I_2$; right: $\sigma^2_{\rm prop} = 1.4\, \Sigma_\pi$)

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

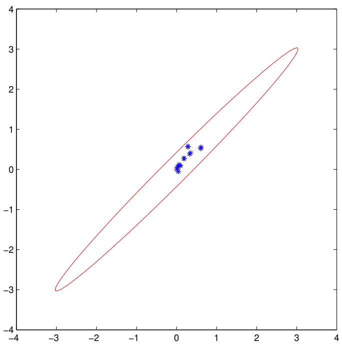


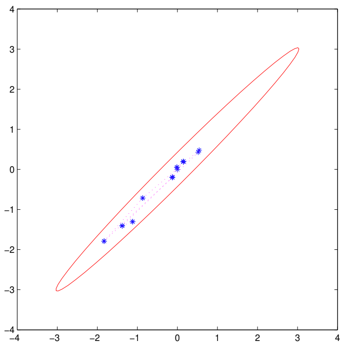


The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

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

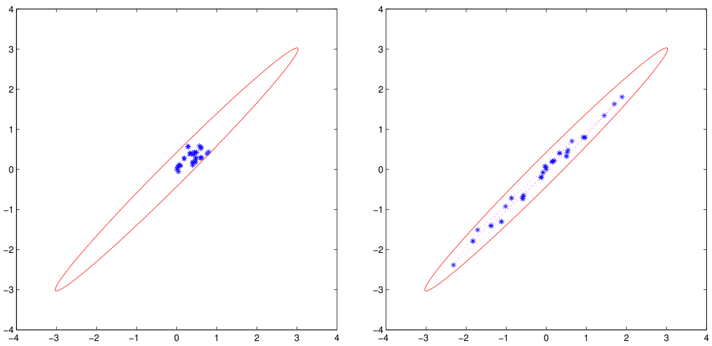




The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

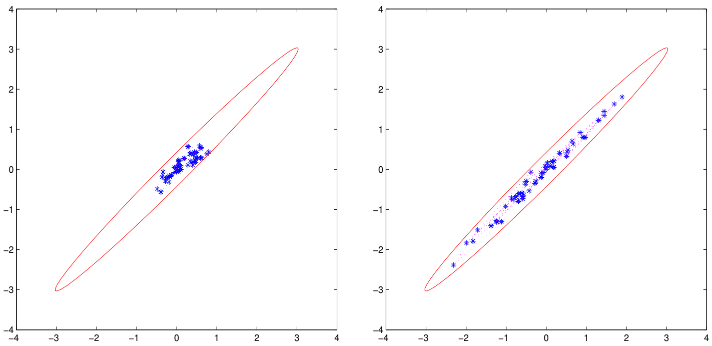
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50



The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

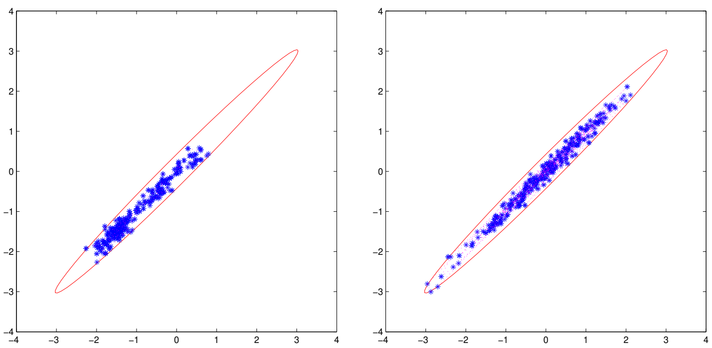
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100



The Geometry of the Target is Very Important

2D Gaussian RW-MH (Gaussian target) with accept. rate 50%
(left: $\sigma^2_{\text{prop}} = 0.04\, I_2$; right: $\sigma^2_{\text{prop}} = 1.4\, \Sigma_\pi$)

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



Another Available Algorithm

Self-Normalized Importance Sampling

Simulate $(X_i)_{1 \leq i \leq n}$ independently from q and estimate $\pi(f)$ by

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

where

$$W_i = \frac{\ell(Y|X_i)\rho(X_i)}{q(X_i)}$$

How Does the Method Scale in Large Dimension?

The asymptotic variance is given by

$$\upsilon(f) = \mathrm{E}_\pi \left[(f(X) - \pi(f))^2 \frac{\pi(X)}{q(X)} \right]$$

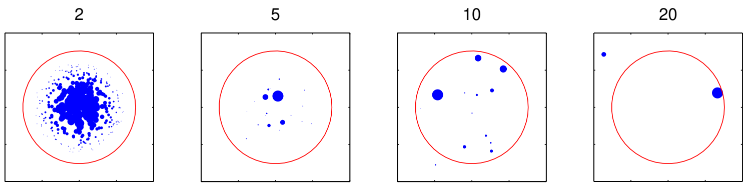
If both target and proposal are product distributions,

$$\begin{aligned} \text{ie.} \quad & \pi_{(d)}(x_1, \ldots, x_d) = \textstyle\prod_{k=1}^{d} \pi(x_k) \\ \text{and} \quad & q_{(d)}(x_1, \ldots, x_d) = \textstyle\prod_{k=1}^{d} q(x_k), \end{aligned}$$

and $f_{(d)}(x_1, \ldots, x_d) = f(x_1)$ is a function of a single component

$$\upsilon_{(d)}(f_{(d)}) = \upsilon_{(1)}(f) \times \left\{ \underbrace{\mathrm{E}_\pi \left[\frac{\pi(X)}{q(X)} \right]}_{> 1 \text{ unless } \pi = q} \right\}^{d-1}$$

Importance Sampling Scaling Example



Importance sampling (1 000 samples) for a truncated multivariate Gaussian distribution with uniform proposals (arbitrary 2D marginal for increasing values of d, radius of points increase with the importance weights)

Here $\mathrm{E}_\pi\left[\frac{\pi(X)}{q(X)}\right] = 2.26$

Random-Walk Metropolis-Hastings

- For a 5–10 parameters model, variance is at least about $30 \times \mathrm{Var}_\pi(X)$
- Some degradation with dimension but not major (in the range 5-20)
- Adjusting the geometry of the proposal is very important

Importance Sampling

- Extremely unreliable in dimension more than 10, unless q is very carefully chosen
- Can reach a variance of $\mathrm{Var}_\pi(X)$ for a proper choice of q
- Independent draws mean parallelisable code (up to weight normalisation)