An Introduction to Markov Chain Monte Carlo Methods for Bayesian Inference

Olivier Cappé

Centre Nat. de la Recherche Scientifique
& Télécom ParisTech
46 rue Barrault, 75634 Paris cedex 13, France
http://perso.telecom-paristech.fr/~cappe/

SIMINOLE Project, Oct. 2010

Outline

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, $\mathrm{E}[x|Y]$, is not feasible except in the simplest Bayesian models.

Stationary Distribution

Definition

π is stationary for q iff

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

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 X_0 is distributed under π,

$$\mathrm{P}(X_i \in A) = \int_A \pi(x)dx$$

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

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

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 $X_1, \ldots, X_n$ of a chain with kernel q started from an arbitrary point X_1 and compute the Monte Carlo estimate

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

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 $\mathrm{E}[f(X)|Z]$ 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

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

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

The Rao-Blackwell Theorem shows that

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

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

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

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 $K_1, \ldots, K_m$ 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 $i = 1, \ldots, m$ and $\sum_{i=1}^{m} \alpha_i = 1$,

also admit π as stationary distribution. If in addition $K_1, \ldots, K_m$ 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

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

and

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

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

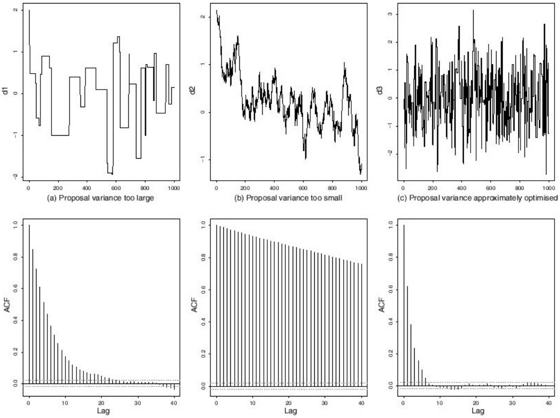
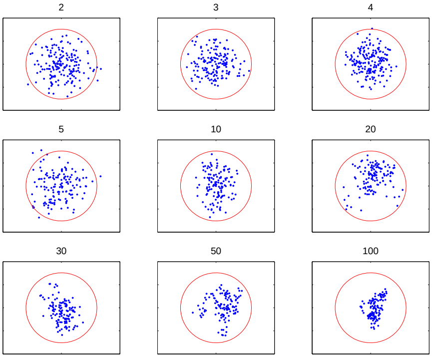


FIG. 2. *Simple Metropolis algorithm with (a) too-large variance (left plots), (b) too-small variance (middle) and (c) appropriate variance (right). Trace plots (top) and autocorrelation plots (below) are shown for each case.*

From (Roberts & Rosenthal, 2001)

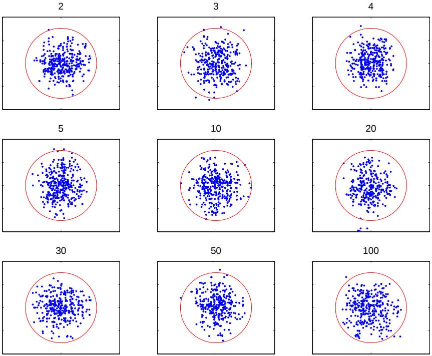
How Does the Method Scales in Large Dimensions?

(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 $2.4\,\sigma_\pi/\sqrt{d}$

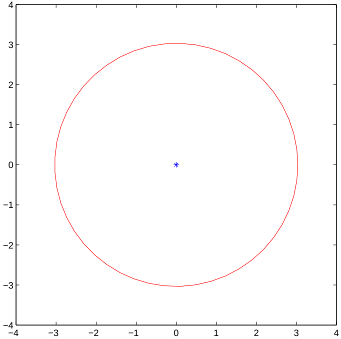
Different Proposals May Tell a Different Story

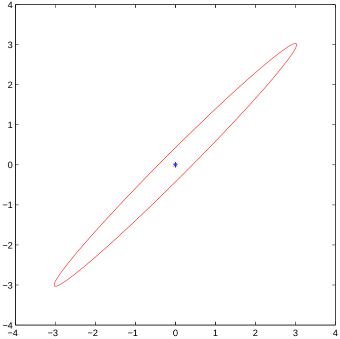


- one-at-a-time RW-MH yields d independent chains in this (very particular) case
- Numerical complexity of the alternatives must be evaluated carefully

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

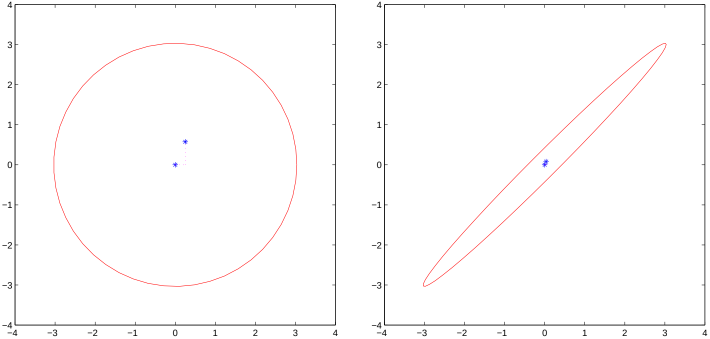
Number of Iterations 1





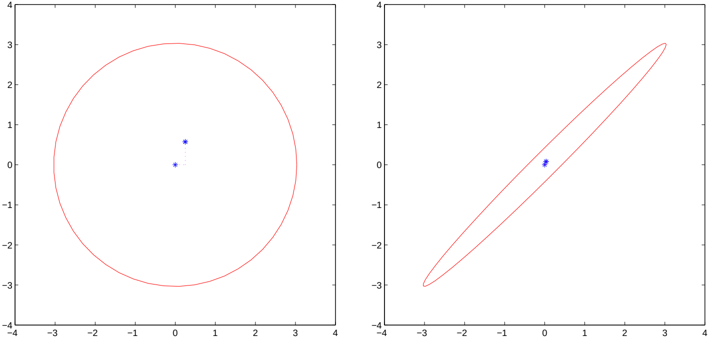
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2



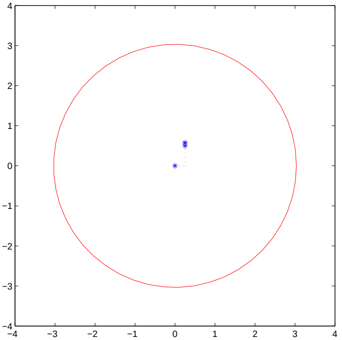
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3



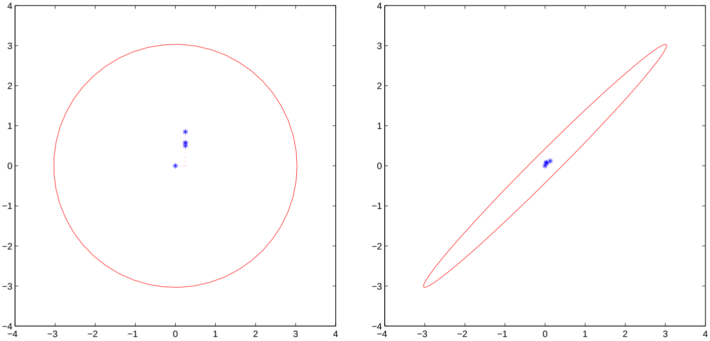
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

Number of Iterations 1, 2, 3, 4



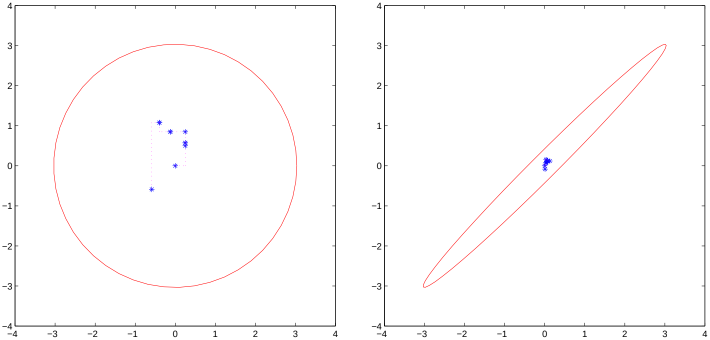
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

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



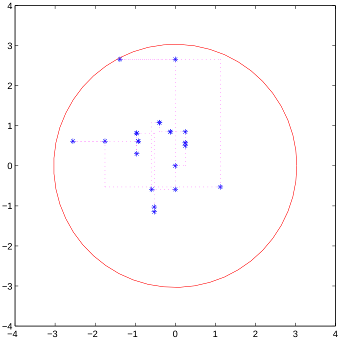
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

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



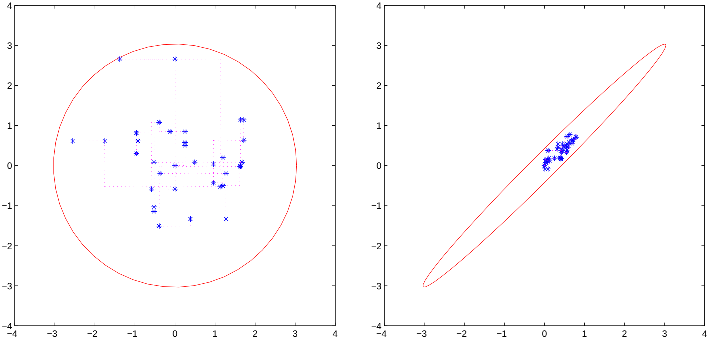
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

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



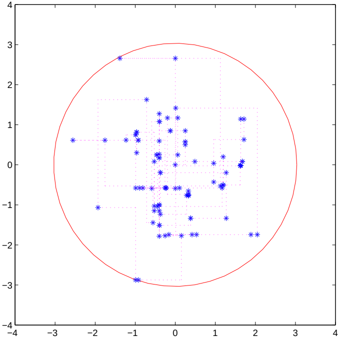
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

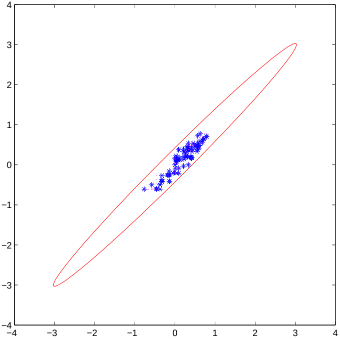
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50



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

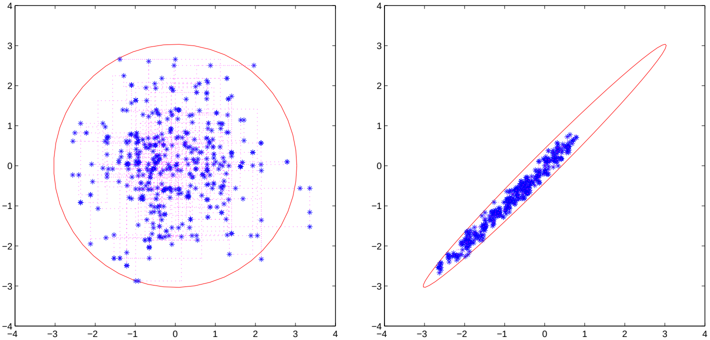
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100





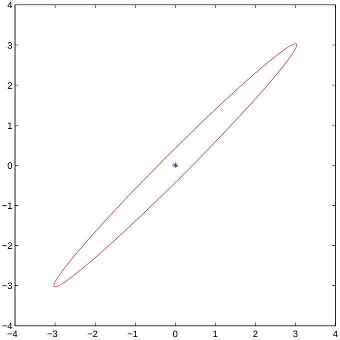
One-at-a-time Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 2$, right $\sigma_{\text{prop}} = 0.28$)

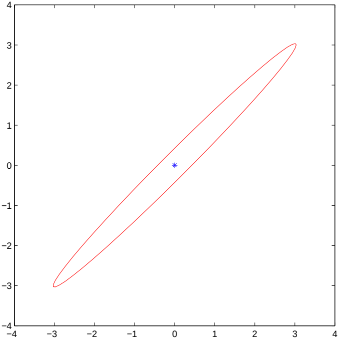
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500



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

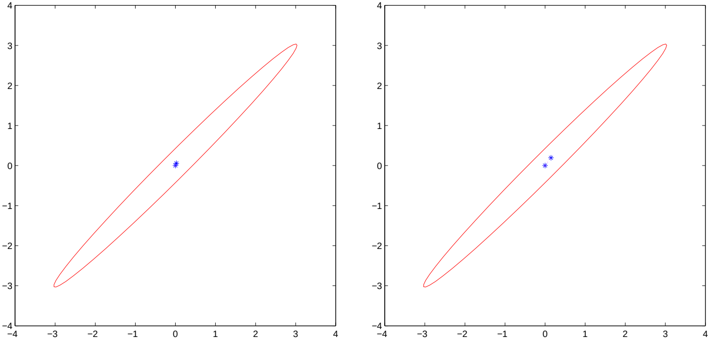
Number of Iterations 1





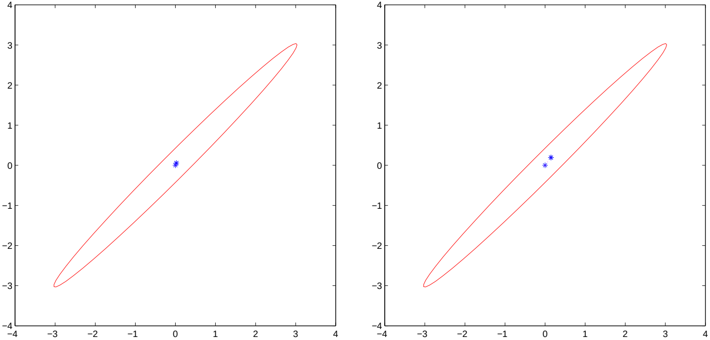
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2



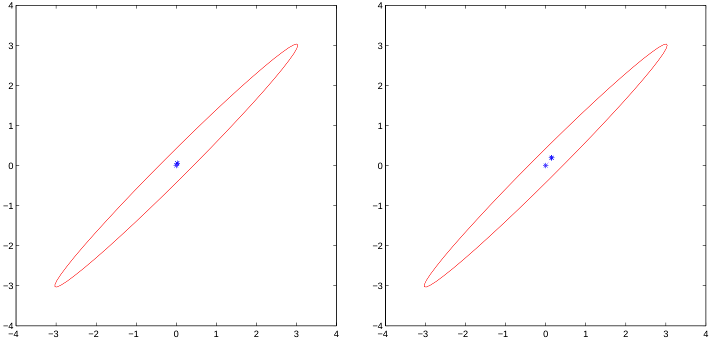
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

Number of Iterations 1, 2, 3



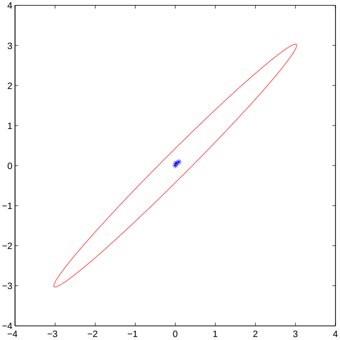
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

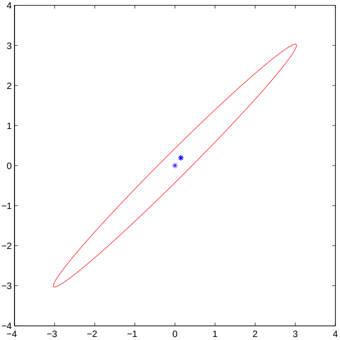
Number of Iterations 1, 2, 3, 4



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

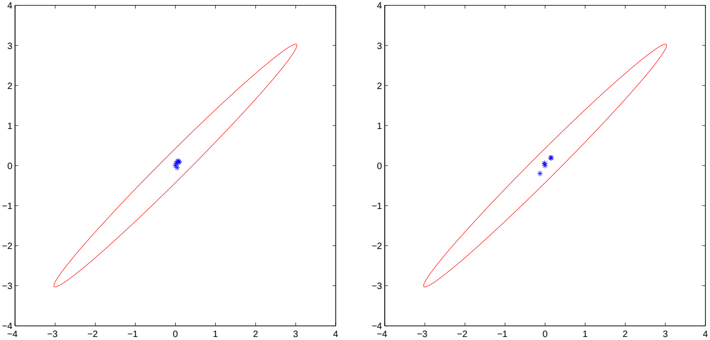
Number of Iterations 1, 2, 3, 4, 5





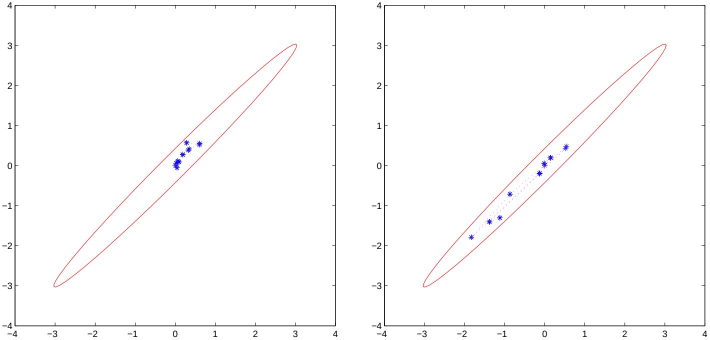
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

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



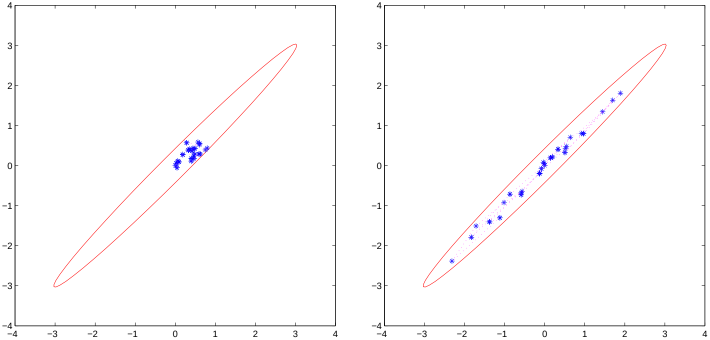
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

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



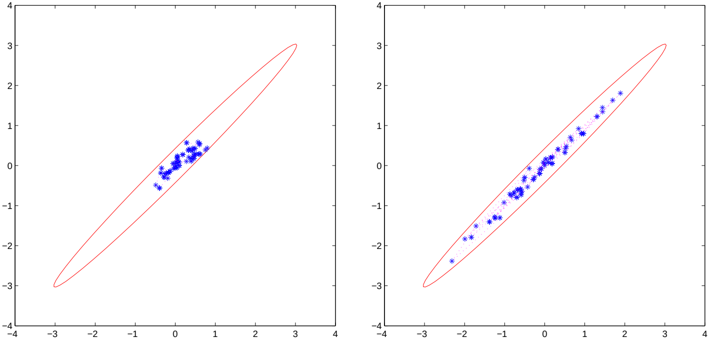
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

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



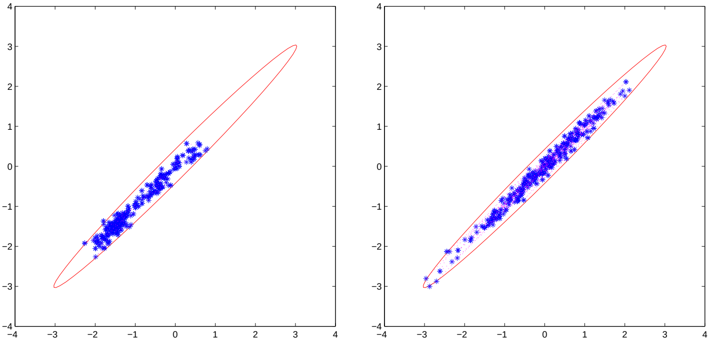
Gaussian RW-MH with accept. rate 50% (left $\sigma_{\text{prop}} = 0.2$; right, with knowledge of Σ_π and $\sigma_{\text{prop}} = 1.2$)

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 knowledge of Σ_π 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

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

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

$$B_n = \frac{1}{M} \sum_{m=1}^{M} (\bar{\xi}_m - \bar{\xi})^2 ,$$

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

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

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.