

Introduction to Monte Carlo and MCMC Methods



Antonietta Mira

University of Insubria, Varese, Italy

antonietta.mira@uninsubria.it

Bayesian methods in Inverse Problems
Summer course, June, 21-24, 2004

Kuopio, Finland

INDEX

(Literature)

- Monte Carlo
- Markov chains
- Markov chain Monte Carlo
- Metropolis algorithm (1953)
- Hastings algorithm (1970)
- Gibbs Sampler (Geman & Geman, 1984)
“heat bath” physics 1979, 1976
- Green algorithm (1995)
- Examples

INDEX

(Research)

Ordering MCMC:

- Peskun ordering based on **absolute** efficiency of estimates
- Covariance ordering based on **relative** efficiency of estimates

Ways of improving MCMC algorithms:

- Slice Sampler **VS** independence M-H
- Delayed rejection algorithm **VS** M-H-G

SIMULATION:

“step by step the probabilities of separate events are merged into a composite picture which gives an **approximate** but **workable** answer to the problem”

MONTE CARLO:

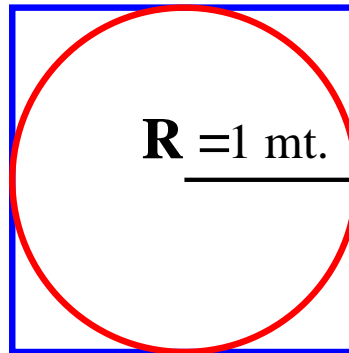
cripted name of a secret project of John von Neumann and Stanislas Ulam at the Los Alamos Scientific Laboratory. The project used random numbers to simulate complicated sequences of connected events

(**roulette**: natural random number generator)

The Monte Carlo Method,

D.D. McCracken, Scientific American, 1955

Approximation of the value π



$$\frac{\text{N. TOSSES INSIDE CIRCLE}}{\text{TOTAL N. TOSSES}} = \frac{\pi R^2}{(2R)^2} = \frac{\pi}{4}$$

6 successes in 10 tosses $\Rightarrow \hat{\pi} = 2.4$

89 successes in 100 tosses $\Rightarrow \hat{\pi} = 3.57$

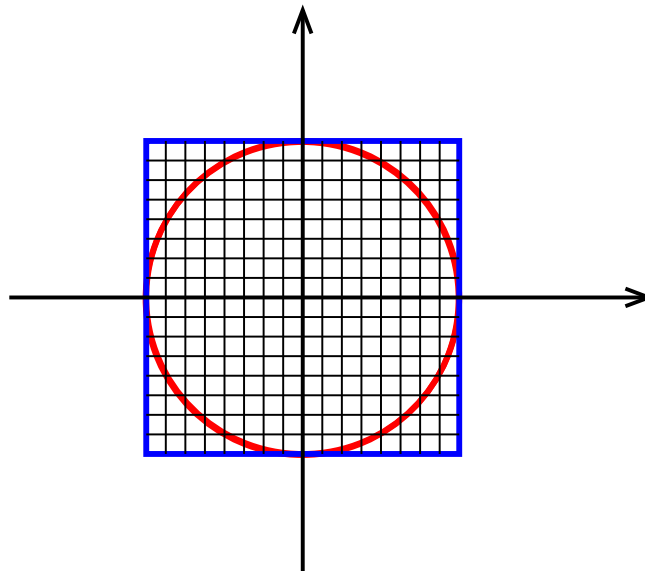
750 successes in 1000 tosses $\Rightarrow \hat{\pi} = 3$

accuracy increases with the square of the number of tosses:

to **duplicate** accuracy we have to

quadruplicate the number of experiments

Use of random numbers



generate two random numbers (i.i.d.) between $[0, 1, \dots, 36]$ the pair defines a point on the Cartesian plane

Solution of complicated integrals

As a matter of fact we are estimating/approximating an integral

$$\int_{(x,y) \in \square} f(x,y) \pi(x,y) dx dy$$

where

$$f(x,y) = \begin{cases} 1 & \text{if } x^2 + y^2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

$\pi(x,y)$ = uniform distribution on the square

Often π is so complicated that we are not able to generate i.i.d. samples from p and therefore we cannot perform Monte Carlo integration

We can construct a Markov chain that “converges” to π

We then simulate a path of the chain

X_0, X_1, X_2, \dots

And we use the X_i “as if” there were i.i.d. from π

This is known as **Markov chain Monte Carlo** (MCMC) simulation

Examples of applications

- finance: option pricing
- genetics and microarrays
- state space models :
epidemiology and meteorology
- time series analysis
- biology - physics - chemistry
- mixture models for cluster analysis:
astronomy, population studies
- operational research
traffic control, quality control,
production optimization

REFERENCES

<http://eco.uninsubria.it/dipeco>



⇒ **MCMC SERVER**

<http://www.statslab.cam.ac.uk/> ~ mcmc

⇒ Ripley

Stochastic simulation

⇒ Gilks + Richardson + Spiegelhalter

MCMC in Practice

⇒ C. Robert + G. Casella

Monte Carlo Statistical Methods

⇒ S.P. Meyn + R. Tweedie

Markov Chains and Stochastic Stability

⇒ Jun S. Liu

M.C. Strategies in Scientific Computing

NOTATION

$$Pr(d|x) = \text{likelihood} = L$$

$$Pr(x) = \text{prior} = p$$

$$Pr(x|d) = \text{posterior} = \pi$$

MOTIVATION

integration plays a fundamental role both in **classical** and **Bayesian** statistics:

$$\pi \propto L \times p$$

normalizing constant for the posterior dist.:

$$\int L \times p$$

marginalization of a joint distribution:

$$\int \pi(x_1, x_2) dx_1$$

synthesis of a complicated distribution:

$$\int f(x)\pi(x)dx$$

- **DETERMINISTIC APPROXIMATIONS**

- Laplace approximation
- Riemann approximation

- **STOCHASTIC APPROXIMATIONS**

- Monte Carlo
- Markov chain Monte Carlo

MONTE CARLO SIMULATION

In general suppose we want to evaluate

$$\mu = \int f(x)\pi(dx) = E_{\pi}f$$

- $f(x) = x \rightsquigarrow$ mean of π
- $f(x) = x^2 \rightsquigarrow$ second moment of π
- $f(x) = \mathbf{1}_{[A]} \rightsquigarrow$ probability of A under π

If we cannot compute the integral analytically but we have X_1, \dots, X_n **i.i.d.** observations from π we can estimate μ by

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

We can use:

- the **Strong Law of Large Numbers** and
- the **Central Limit Theorem**
(if $E_{\pi}f^2 < \infty$)

We thus have (w.p.1):

$$\hat{\mu}_n \rightarrow \mu$$

i.e. the estimator is asymptotically unbiased and we can estimate the variance by:

$$\hat{\sigma}^2 = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n [f(X_i) - \hat{\mu}_n]^2$$

so that

$$\frac{\hat{\mu}_n - \mu}{\hat{\sigma}} \sim \mathcal{N}(0, 1)$$

How can we generalize this idea when we do not have i.i.d. observations from π ?

\implies **IMPORTANCE SAMPLING**

\implies **MCMC**

IMPORTANCE SAMPLING

If we cannot get iid samples from π
use an auxiliary (importance) distribution, g
that we can sample, and use this alternative
representation of μ :

$$\mu = \int f(x)\pi(x)dx = \int f(x)\frac{\pi(x)}{g(x)}g(x)dx$$

we can thus estimate μ by:

$$\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n f(X_i) \frac{\pi(X_i)}{g(X_i)}$$

For any choice of g , as long as

$\text{supp}(g) \subset \text{supp}(\pi)$

again SLLN + CLT hold

(under same regularity conditions as before)

PROS

- choose g easy to sample from
- same samples from g can be used repeatedly for different π and different f

CONTRAS

- finite variance only if

$$E_{\pi}\left[f^2 \frac{\pi(x)}{g(x)}\right] < \infty$$

- if g has tails lighter than π , i.e.
 $\sup \pi/g = \infty$ not good: weights vary widely giving too much importance to few X_i
- need $\sup \pi/g < \infty$ but if this is the case could use accept-reject to simulate from π

EXAMPLE: Student-T distribution

$$X \sim \mathcal{T}(\nu, \theta, \sigma)$$

$$\pi(x) \propto \left(1 + \frac{(x - \theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}$$

w.l.o.g. take $\theta = 0$ and $\sigma = 1$

$$f(x) = \left(\frac{\sin(x)}{x}\right)^5 \mathbf{1}_{(x>2.1)}$$

so that

$$\mu = \int_{2.1}^{\infty} \left(\frac{\sin(x)}{x}\right)^5 \pi(x) dx$$

Different possibilities

- sample directly from Student-T since:

$$T_{(\nu,0,1)} = \frac{\mathcal{N}(0, 1)}{\chi_{\nu}^2}$$

- importance sampling from $g = \mathcal{N}(0, 1)$
non-optimal
- importance sampling from $g = \text{Cauchy}(0, 1)$
OK: bounded tails

Exercise: write the code to sample from $T_{(\nu,3,2)}$

MARKOV CHAINS

A M.C. is a random process that evolves in time X_1, X_2, \dots with the property that

FUTURE indep. PAST | PRESENT

We will assume that the time is discrete and the state space is finite, $S = \{1, 2, \dots, k\}$

A M.C. is specified by giving

- **initial distribution:** λ (a vector)
 $\lambda(i) = P(X_1 = i), \quad i \in S$
- **transition probabilities:** P (a matrix)

$$P(i, j) = P(X_{t+1} = j | X_t = i), \quad i, j \in S, \quad \forall t$$

- We assume the transition probabilities do not change with time: **homogeneous**

- Notice that

$$\sum_{j \in S} P_{ij} = 1$$

since the chain must be in some state in the next step

- We have
a vector: $\lambda =$ **initial distribution**
a matrix: $P =$ **transition probabilities**
and can hardly resist the temptation of multiplying them ...

$$\begin{aligned}
\lambda P(j) &= \sum_i \lambda(i) P(i, j) \\
&= \sum_i P(X_1 = i) P(X_2 = j | X_1 = i) \\
&= \sum_i P(X_1 = i, X_2 = j) \\
&= P(X_2 = j)
\end{aligned}$$

$$\boxed{X_1 \sim \lambda} \xrightarrow{\mathcal{P}} \boxed{X_2 \sim \lambda} \xrightarrow{\mathcal{P}} \dots \xrightarrow{\mathcal{P}} \boxed{X_n \sim \lambda}$$

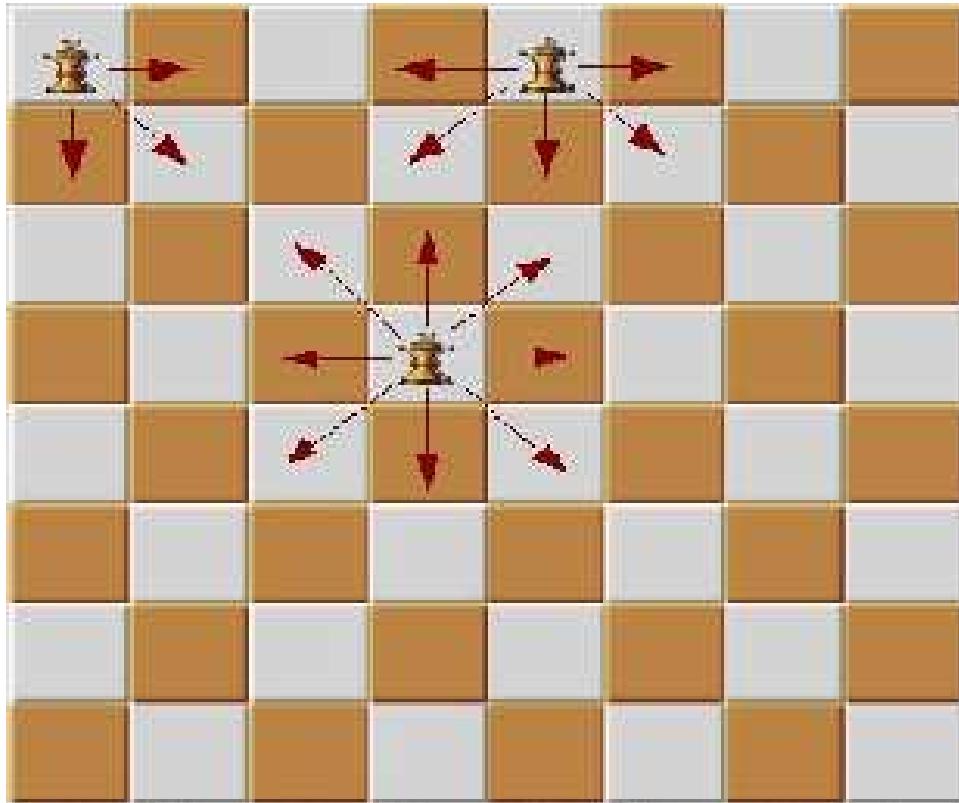
π is a **stationary distribution** for P if

$$\pi P = \pi$$

that is, if

$$\sum_i \pi(i) P(i, j) = \pi(j), \quad j \in S$$

i.e. π is the (normalized) **left eigenvector** of P with eigenvalue 1



Determine the stationary distribution of a king free to move on the chess board

What is the probability of finding the king in the top right corner?

What is the probability of finding the king in the center of the chess board?

P is **reversible** w.r.t. π if

$$\pi(i)P(i, j) = \pi(j)P(j, i), \quad i, j \in S$$

(**detailed balance condition**)

i. e. the M. C. looks the same running forward or backward

Reversibility \rightarrow **Stationarity**

EXAMPLE

Irreducible but not reversible chain:

$$\mathbf{P} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

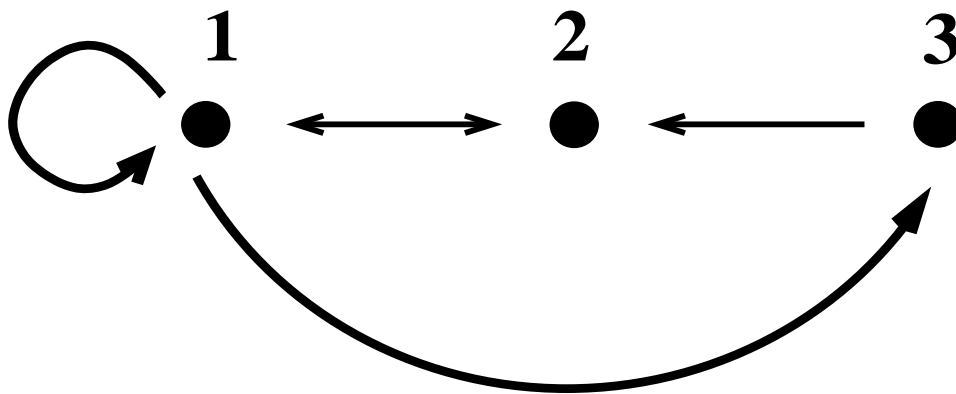
We represent the transition matrix of a MC as a **GRAPH** with

a **vertex** for each **state**

a **directed edge** from vertex i to j if there is

a **nonzero transition probability** from i to j

a **nonzero P_{ij} entry** in the transition matrix



This is a **possible** sequence

$$1 \rightarrow 3 \rightarrow 2 \rightarrow 1$$

this is an **impossible** sequence

$$1 \rightarrow 2 \rightarrow 3 \rightarrow 1$$

since $P_{23} = 0$ (an edge is missing from 2 to 3)

Thus there is a sequence of states for which it is possible to tell in which direction the simulation has occurred and the chain is **not reversible**

n -step transition matrix:

$$P^n(i, j) = P(X_{m+n} = j | X_m = i)$$

If π is stationary for P then (prove):

$$\pi P^n = \pi$$

The distribution of X_n is independent on n if and only if the initial distribution is a stationary distribution (prove).

Suppose a stationary distribution π exists and that:

$$\lim_{n \rightarrow \infty} P^n(i, j) = \pi(j), \quad j \in S,$$

then, regardless of the initial distribution, the distribution of X_n approaches π as $n \rightarrow \infty$. In such cases π is called the **limiting distribution** and the MC is said to be **ergodic**

- Not all MC have a stationary dist.
- A MC can have more than 1 stationary dist.
- Not all stat. dist. are also limiting dist.

All Markov chain used for MCMC purposes need to have a **unique stationary and limiting distribution**

A state i is said to be **accessible** from state j if, for some $n \geq 0$, $P^n(i, j) > 0$.

Two states i and j each accessible from the other are said to **communicate**

Irreducibility

if all state communicate with each other
if you can go from anywhere to everywhere

EXAMPLE

$S = \{1, 2, 3, 4\}$ = state space

$$P = \begin{bmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0.2 & 0.8 \end{bmatrix}$$

we have::

two (equivalence) classes of communicating states

two left eigenvectors of P with eigenvalue 1:

$$\sigma = (1/4, 3/4, 0, 0)$$

$$\rho = (0, 0, 1/4, 3/4)$$

depending on the initial state we get a different stationary distribution

A M.C is **PERIODIC** if there are parts of the state space that the chain can visit only at regular time intervals

A M.C. is **APERIODIC** if it is not periodic

If the diagonal elements of the transition matrix are all zero the chain may be periodic

Periodic MC are not ergodic but the difficulty can be eliminated by **sub-sampling**

If a (finite state space) M.C. is **IRREDUCIBLE**
it has a **unique stationary distribution** π
If the M.C. is also **APERIODIC**

- π is also the **limiting distribution**:

$$P(X_n \in A | X_0) \rightarrow \pi(A)$$

- **Law of Large Numbers** holds
if $E_\pi |f| < \infty$:

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \longrightarrow E_\pi f \quad a.s.$$

- **Central Limit Theorem** holds
if $E_\pi f^2 < \infty$ + other regularity conditions

EXAMPLES

$S = \{1, 2, 3, 4\} = \text{state space}$

$\pi = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}] = \text{distribution of interest}$

Possible transition matrices:

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

periodic - irreducible - non reversible

$$Q = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

periodic - **reducible** - reversible

EXAMPLES (cont.)

$S = \{1, 2, 3, 4\} = \text{state space}$

$\pi = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}] = \text{distribution of interest}$

Possible transition matrices:

$$R = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

aperiodic - irreducible - reversible

$$S = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}$$

aperiodic - irreducible - reversible

MARKOV CHAIN MONTE CARLO

We are interested in estimating

$$\mu = E_{\pi} f(X)$$

Construct a Markov chain that has π as its unique stationary distribution : $\pi = \pi P$

Simulate the Markov chain: $X_0, X_1, X_2 \dots \sim P$
Estimate μ with

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

How good is the estimate?

$$\begin{aligned} \mathbf{V}(\mathbf{f}, \mathbf{P}) &= \lim_{n \rightarrow \infty} n \operatorname{Var}_{\pi} [\hat{\mu}_n] \\ &= \sigma^2 \sum_{k=-\infty}^{\infty} \rho_k \end{aligned}$$

where

$$\sigma^2 = \operatorname{Var}_{\pi} f(X)$$

$$\rho_k = \frac{\operatorname{Cov}_{\pi} [f(X_0), f(X_k)]}{\sigma^2}$$

DIFFERENT PROSPECTIVE

In Markov chain theory **we are given a MC**, P and we find its equilibrium distribution

In MCMC theory **we are given distribution**, π and we construct a MC reversible wrt it

A MC can be specified by

- its **macroscopic** transition matrix
- its **microscopic** dynamics

GENERAL STRATEGY for MCMC

Current position $X_t = x$

(1) propose a candidate $y \sim q(x, \cdot)$

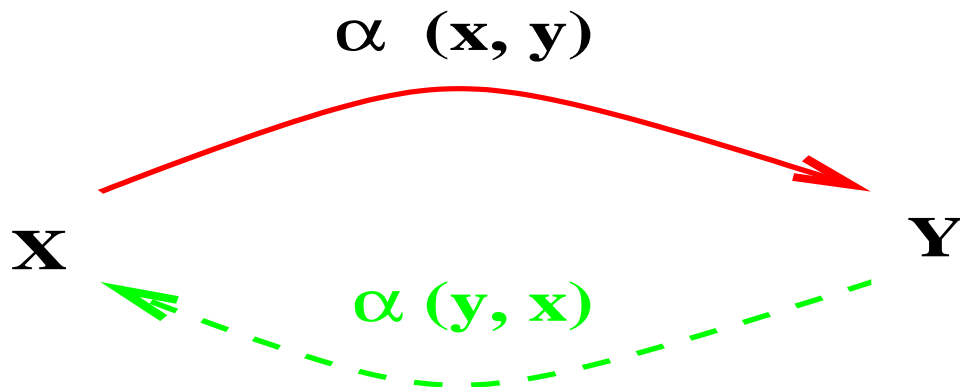
(2) with probability $\alpha(x, y)$ accept y : $X_{t+1} = y$

(3) otherwise stay where you are: $X_{t+1} = x$

the acceptance probability $\alpha(x, y)$ is computed so that **REVERSIBILITY** wrt π is preserved:

$$\int_{(x,y) \in A \times B} \pi(dx) q(x, dy) \alpha(x, y) =$$

$$\int_{(x,y) \in A \times B} \pi(dy) q(y, dx) \alpha(y, x)$$



A LITTLE BIT OF HISTORY

Metropolis et al. (1953)

if the proposal is symmetric, $q(x, y) = q(y, x)$:

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

Hastings (1970)

generic proposal in a fixed-dimension problem:

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

Since the generation and the acceptance steps are independent, the resulting **macroscopic dynamic** i.e. the transition matrix (kernel),

$$\begin{aligned} P_{xy} &= P(X_{n+1} = y | X_n = x) \\ &= q(x, y)\alpha(x, y), \quad \forall x \neq y \end{aligned}$$

and $P_{x,x}$ can be found from the requirement that $\sum_y P_{xy} = 1$

The resulting Metropolis-Hastings MC is reversible wrt π . If it is also irreducible and aperiodic we have an ergodic Markov chain with unique stationary and limiting distribution π

If $q(x, y)$ is not the transition matrix of an irreducible MC on S then $P(x, y)$ is not irreducible. However, irreducibility of q is not sufficient to guarantee irreducibility of P since it also depends on α

SPECIAL CASES

I.I.D. SAMPLING (Monte Carlo simulation)

if $q(x, y) = \pi(y)$

then $\alpha(x, y) = 1$

INDEPENDENCE M-H

the proposal distribution **does not depend** on current position of the M.C.: $q(\mathbf{x}, \cdot) = N(\mathbf{0}, \sigma^2)$

RANDOM WALK M-H

the proposal distribution **does depend** on the current position of the M.C.: $q(\mathbf{x}, \cdot) = N(\mathbf{x}, \sigma^2)$

that is: use the proposal

$$Y_t = x + \varepsilon_t,$$

where $\varepsilon_t \sim N(0, \sigma^2)$, independent of X_t .

The instrumental density is now of the form $q(y - x)$ and the Markov chain is a **random walk** if we take q to be *symmetric*

GIBBS SAMPLER

if $\mathbf{x} = (x_1, \dots, x_d) \sim \pi$ update one component at a time by using the conditional distribution of that component given everything else as the proposal

$$q(x_i, \cdot) = \pi(x_i | x_{-i}) = \text{full conditionals}$$

where “ $-i$ ” indicates $\{j : j \neq i\}$.

PRO

- the acceptance probability is always one
- no need to calibrate the proposal

CONTRAS

- full conditionals can be hard to sample from
- if high correlation on the target it takes a long time to move around the state space

RND-WALK Metropolis algorithm

Target distribution: $\pi(x) \sim N(0, 1)$

Proposal distribution: $q(x, y) \sim N(x, \sigma^2)$

(symmetric) for some σ^2 : typically hard to properly calibrate the spread of the proposal!

Acceptance probability :

$$\begin{aligned}\alpha(x, y) &= \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right) \\ &= \exp\left(-\frac{1}{2}(y^2 - x^2)\right)\end{aligned}$$

Program in "R":

```
met.has <- function(n,x0,sigma){
  x <- array(0,n)
  x[1] <- x0
  for(t in 2:n)
  {
    y <- rnorm(1,x[t-1],sigma)
    alpha <- exp((x[t-1]^2 - y^2)/2)
    accept <- min(1,alpha)
    u <- runif(1)
    if (u <= accept) x[t] <- y
    else x[t] <- x[t-1]
  }
  plot(1:length(x),x,
       type='l',lty=1,xlab="t",ylab="x")
  x
}
```

Example: Gibbs sampler

Y_1, \dots, Y_d iid $N(\mu, \tau^{-1})$

Priors:

$$\mu \sim N(\mu_*, \sigma_*^2) \quad \tau \sim G(\alpha_*, \beta_*)$$

Posterior:

$$\begin{aligned} p(\mu, \tau | y_1, \dots, y_d) &\propto L(\mu, \tau; \mathbf{y}) p(\mu, \tau) \\ &\propto \tau^{\frac{d}{2} + \alpha_* - 1} e^{-\beta_* \tau} e^{-\frac{\tau S_d}{2} - \frac{(\mu - \mu_*)^2}{2\sigma_*^2}} \end{aligned}$$

where $S_d = \sum_{i=1}^d (y_i - \mu)^2$

Full conditionals:

$$(\mu | \tau, y_1, \dots, y_d) \sim N\left(\frac{d\bar{y}\tau + \mu_*\tau_*}{d\tau + \tau_*}, (d\tau + \tau_*)^{-1}\right)$$

where $\tau_* = (\sigma_*^2)^{-1}$

$$(\tau | \mu, y_1, \dots, y_d) \sim G\left(\alpha_* + \frac{d}{2}, \beta_* + \frac{S_d}{2}\right)$$

Start the chain at some values $\mu(0)$ and $\tau(0)$ and repeat the following algorithm

- (1) Given $\tau(t)$,
generate $\mu(t+1)$ from $\pi(\mu|\tau(t), y_1, \dots, y_d)$

- (2) Given $\mu(t+1)$
generate $\tau(t+1)$ from $\pi(\tau|\mu(t+1), y_1, \dots, y_d)$

This iterative procedure generates a M.C. on the space (μ, τ) that has the distribution of interest as its unique stationary (and limiting) distribution

```

gibbs1 <- function(nits,y,mu0=0,tau0=1){

  alphas <- 0.1
  betas <- 0.01
  mus <- 0.0
  taus <- 1.0

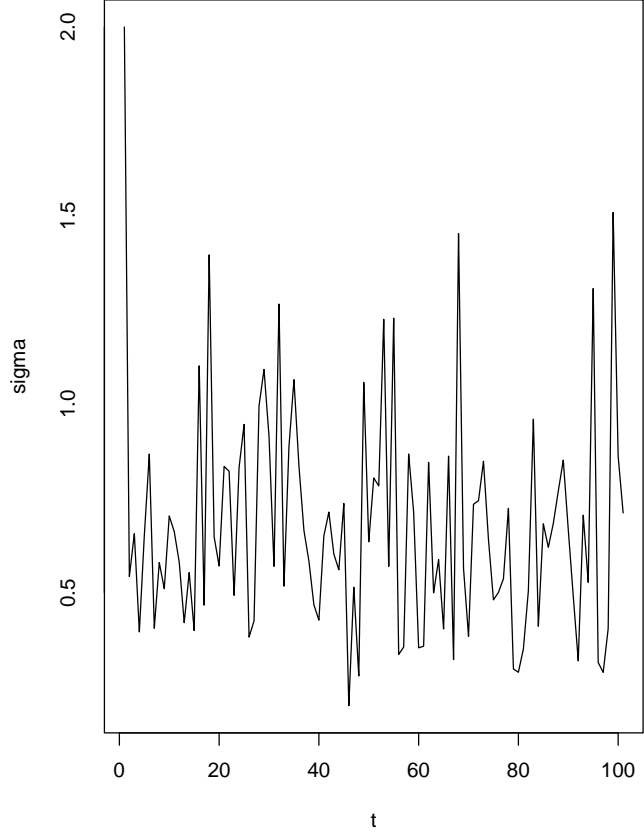
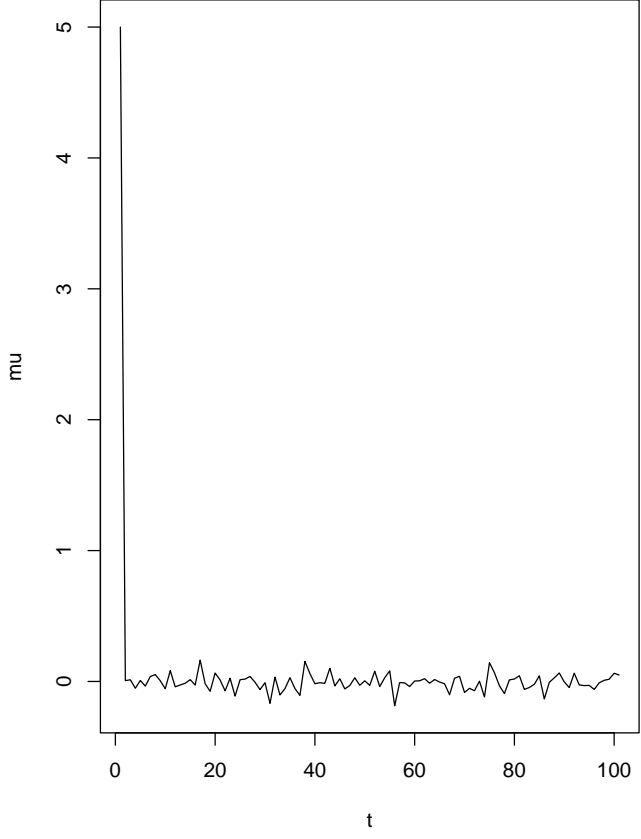
  x <- array(0,c(nits+1,2))
  x[1,1] <- mu0
  x[1,2] <- tau0
  n <- length(y)
  ybar <- mean(y)

  for(t in 2:(nits+1)){
    x[t,1]_RNORM(1,(n*ybar*x[t-1,2]+mus*taus)/
                (n*x[t-1,2]+taus),
                sqrt(1/(n*x[t-1,2]+taus)))

    sn_sum((y-x[t,1])^2)

    x[t,2]_RGAMMA(1,alphas+n/2)
                /(betas+sn/2)}
  x
}

```



Sample path:

Note: in general $\alpha(x, y)$ depends on π via the ratio $\frac{\pi(y)}{\pi(x)}$ so we can compute $\alpha(x, y)$ even if we only know π up to a normalizing constant

Calibration of the proposal

pilot runs - trial and error

aim at acceptance probability $\approx 40\%$

Sub-sampling: for memory convenience and to reduce (correlation) variance of MCMC estimators

Updating schemes:

- we can update **one variable** at a time
- we can update **groups of variables** together
- we can randomly select which variable to update: **random scan**
- we can update the variables always in the same order: **systematic scan**

HYBRID ALGORITHMS

we can combine all the above recipes via MIXTURES or CYCLES:

MIXTURES:

with prob. p_i at each step select a move type or a proposal ($\sum_i p_i = 1$)

CYCLES:

each move type or proposal is used in a predetermined sequence

Example: **Metropolis within Gibbs**

Is like having different keys (proposals) to move around a building visiting all (possibly locked) rooms

Length of the Burn-in

- look if the sample path is “**sufficiently**” stable
- start more chains and wait until they get “**sufficiently**” close
- convergence diagnostic Brooks + Gelman “General methods for monitoring convergence of iterative simulations” J. Computational Graphical Stat., 1998, p. 69 - 100
- CODA and BOA: free software for convergence diagnostics

Determining the length of the burn-in is one of the major problems with MCMC. You can

never be sure that the chain has run sufficiently long to have forgotten its starting point and have thus reached stationarity. This problem is solved by the **Perfect Simulation** idea.

$X_0 = \text{STARTING POINT OF THE M.C.}?$

LLN + CLT hold for π -almost all x_0 i.e. for all starting points but a set of π measure zero

To get rid of this set of zero measure we want our chains to be ϕ -Harris recurrent

i.e. we need the existence of a distribution ϕ s.t. for all A with $\phi(A) > 0$ we have

$$P(X_t \in A \text{ i.o.} | X_0 = x) = 1 \text{ for all } x$$

If such a ϕ exists the chain is also π -Harris recurrent

$X_0 =$ STARTING POINT OF THE M.C.?

- choose something you like (!)
- sample from the prior distribution
(if you are in a Bayesian setting)
- find the modes of π and use a mixture of T-distributions centered at the modes
(the modes can be found via **pilot runs** of the M.C.)
- start from “unlikely” points to check that your algorithm is effective

n = HOW MANY SIMULATIONS?

- **single long run**

↪ you get closer to the stationary dist.

↪ reduce the **variance** of your estimator

- **many short runs**

↪ you explore the state space better

↪ reduce the **bias** of your estimator

TRADE OFF

How long is long?

Need $n \gg \tau_f$ = integrated autocorrelation time

$$V(f, P) = \sigma_f^2 \left(1 + 2 \sum_{k=1}^{\infty} \frac{\text{Cov}_{\pi}[f(X_0), f(X_k)]}{\sigma_f^2} \right) = \sigma_f^2 \tau_f$$

so τ_f is the number of correlated samples with the same variance-reduction power as one independent sample

EXAMPLE

Mixture of two normal distributions with well separated means giving two distinct modes:

$$\pi(x) = 0.5\mathcal{N}(-a, \sigma^2) + 0.5\mathcal{N}(a, \sigma^2)$$

use a Metropolis-Hastings algorithm with symmetric uniform proposal

$$q(x, y) = U[x - d, x + d]$$

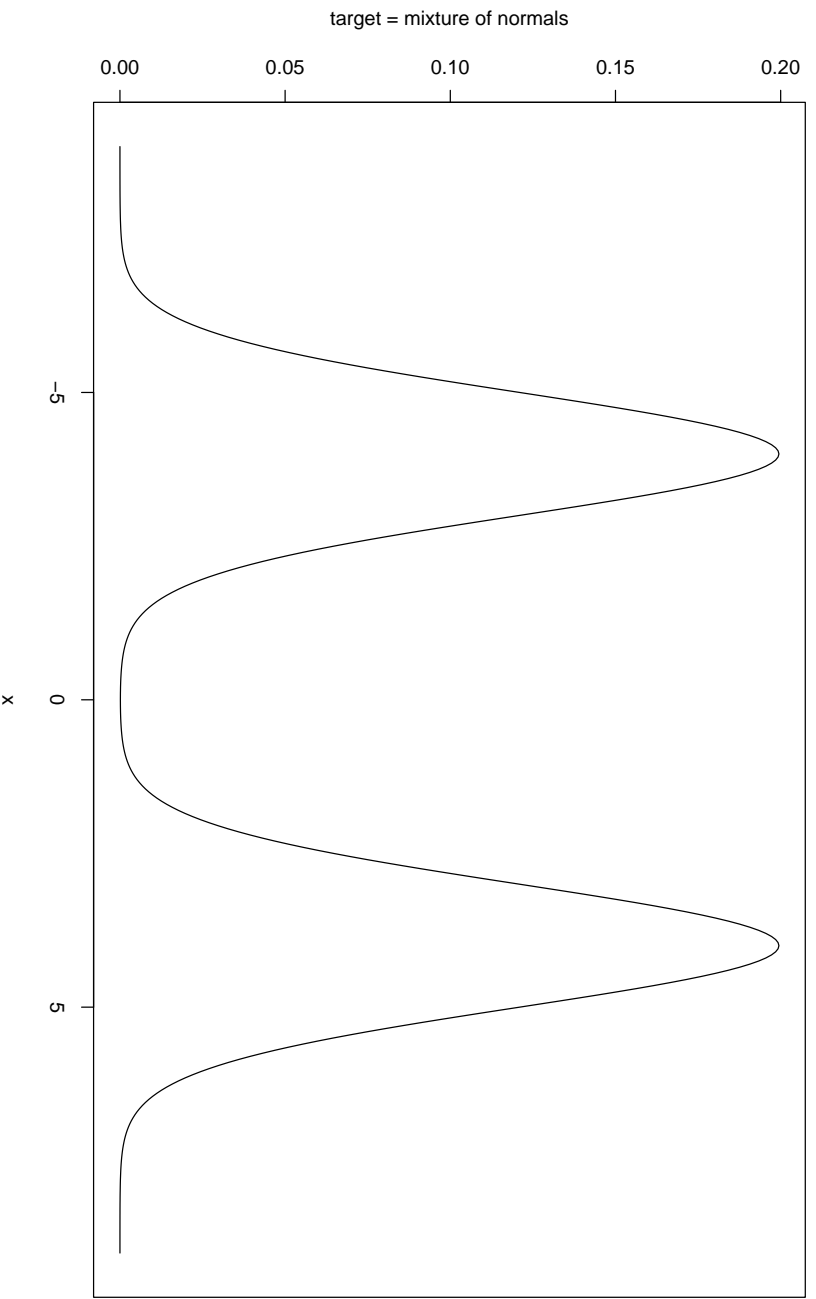
w.l.o.g. take $\sigma^2 = 1$ so that the acceptance probability is

$$\alpha(x, y) = 1 \wedge \frac{\exp(-(y + a)^2/2) + \exp(-(y - a)^2/2)}{\exp(-(x + a)^2/2) + \exp(-(x - a)^2/2)}$$

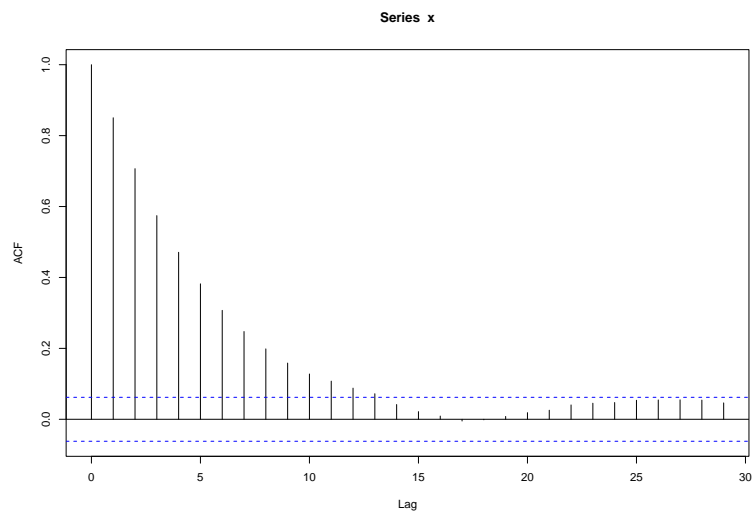
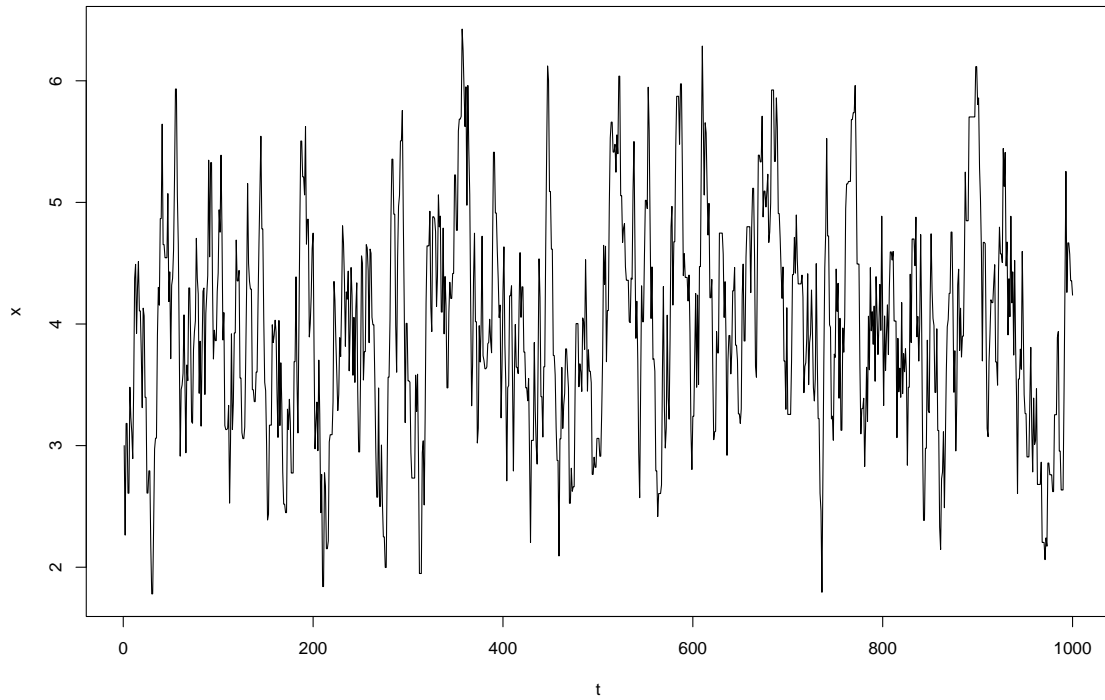
Program in "R" :

```
mixture <- function(n,x0,a,d){
  x <- array(0,n)
  x[1] <- x0
  for(t in 2:n)
  {
    y <- runif(1,x[t-1]-d,x[t-1]+d)
    alpha <- (exp(-1*(y+a)^2/2) +
              exp(-1*(y-a)^2/2))/
              (exp(-1*(x[t-1]+a)^2/2) +
              exp(-1*(x[t-1]-a)^2/2))
    accept <- min(1,alpha)
    u <- runif(1)
    if (u <= accept) x[t] <- y
    else x[t] <- x[t-1]
  }
  plot(1:length(x),x,
       type='l',lty=1,xlab="t",ylab="x")
  x
}
```

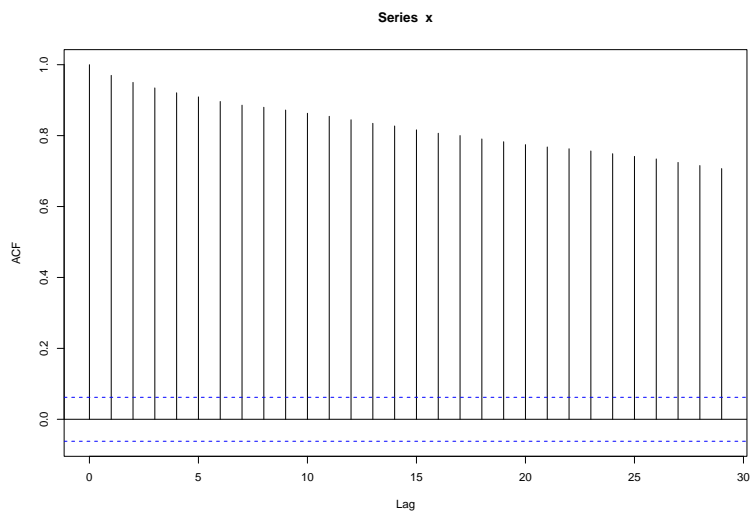
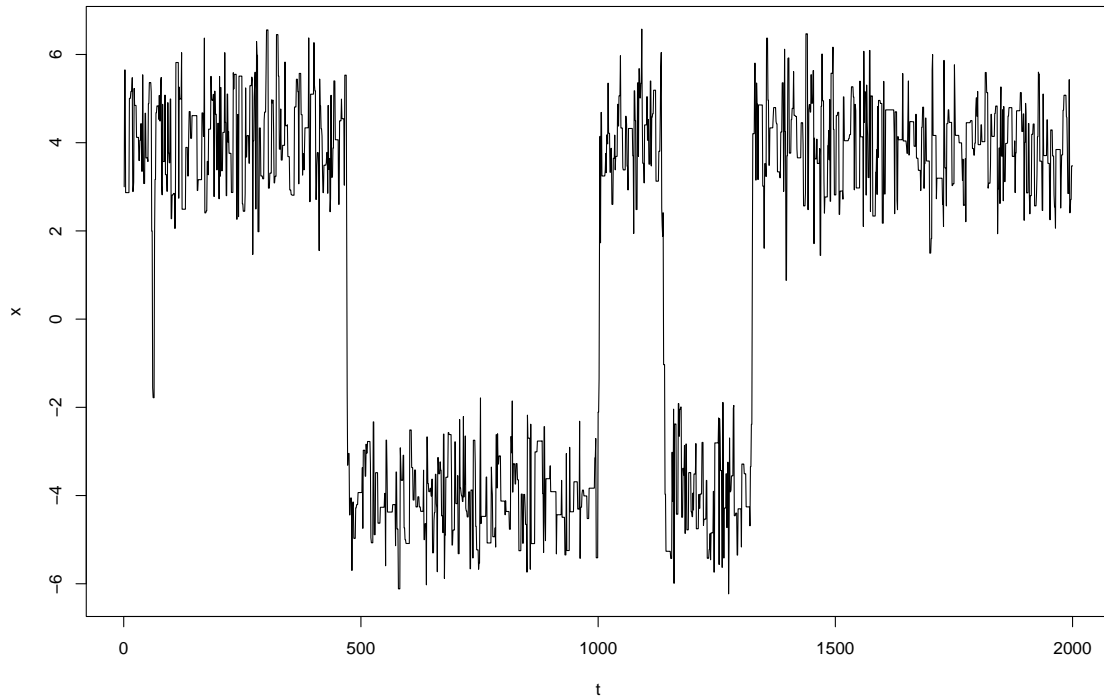
Target distribution with modes in 4 and -4
and standard-deviation =1



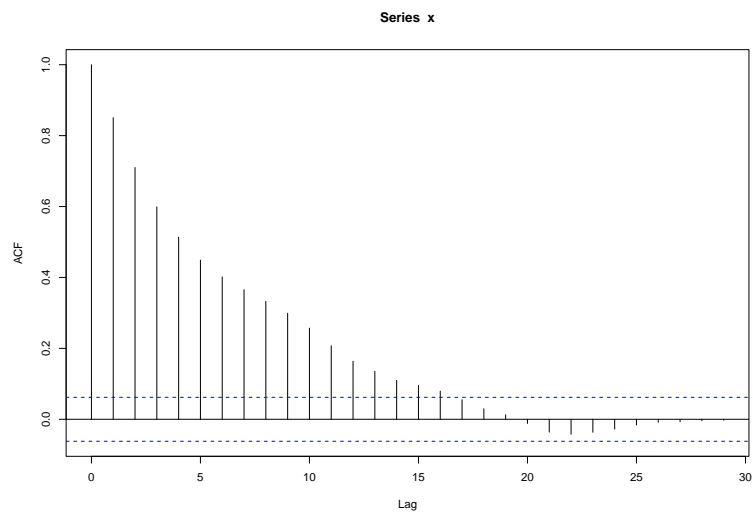
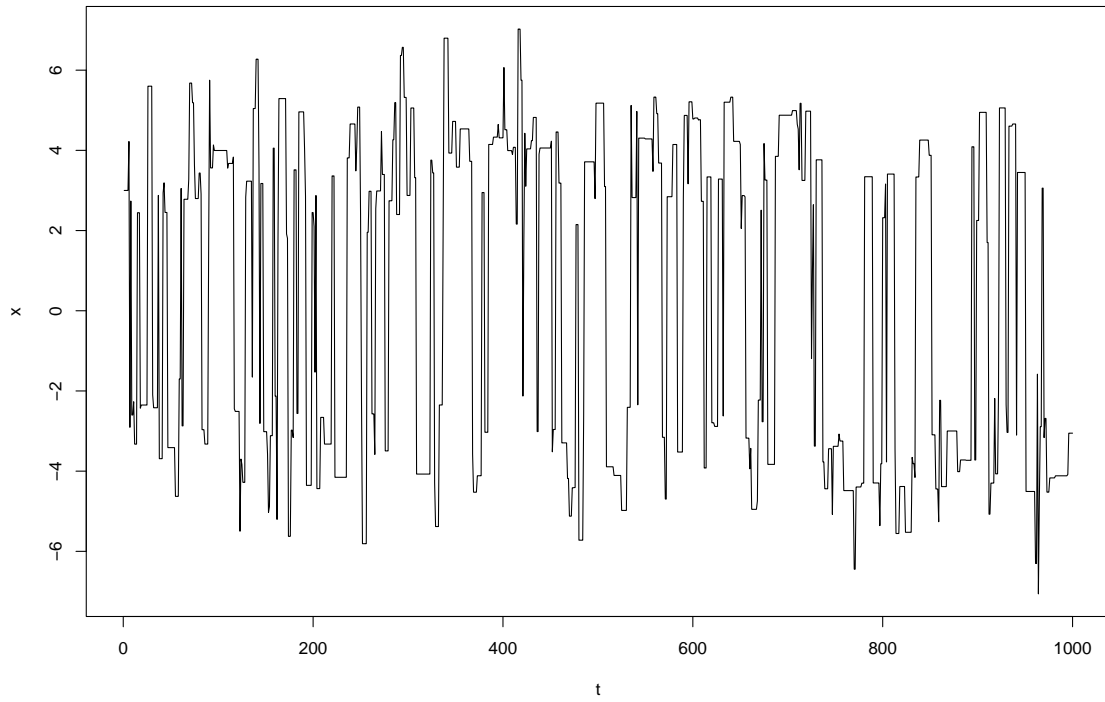
Sample path of a chain started in 3 with $d = 1$ (proposal), $n = 1000$ steps



Sample path of a chain started in 3 with $d = 4$ (proposal), $n=1000$ steps



Sample path of a chain started in 3 with $d = 8$ (proposal), $n = 1000$ steps



ERROR in MCMC ESTIMATES

Parameter:

$$\mu = \int f(x)\pi(dx) = E_{\pi}f$$

Estimator:

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

Variance of the estimator:

$$\begin{aligned} V(f, P) &= \lim_{n \rightarrow \infty} n \operatorname{Var}_{\pi}[\hat{\mu}_n] \\ &= \sum_{k=-\infty}^{\infty} \operatorname{Cov}_{\pi}[f(X_0), f(X_k)] \\ &= \sum_{k=-\infty}^{\infty} \gamma_k \end{aligned}$$

Estimate the variance of MCMC estimators

- **Empirical covariances:** $\hat{V} = \sum_{-\infty}^{+\infty} \hat{\gamma}_k$

it is well known that this is not a consistent estimator.

Use a truncated version of the above over a proper window: $\hat{V} = \sum_{-M}^{+M} \hat{\gamma}_k$ with M being the smallest integer $\geq 3 \hat{\tau}$ (Sokal, 1989)

- **Blocking:** divide the simulations into b consecutive blocks of length k

$$\bar{g}_{k,i} = \frac{1}{k} \sum_{j=(i-1)k+1}^{ik} f(x_j) = \text{mean of block } i$$

$$\hat{V} = \frac{1}{b(b-1)} \sum_{i=1}^b [\bar{g}_{k,i} - \hat{\mu}]^2$$

- **Use regeneration times**

to improve the blocking estimator of the variance by taking blocks to be **independent tours**

- **Use multiple runs**

compute your statistics on multiple independent runs of your MC

(with different starting points)

and look at the distribution of your statistics (and its variance)

CONVERGENCE RATE

We know the chain converges to π but how fast?

Very rarely we can get numeric bounds for convergence rates

A MC is **GEOMETRICALLY ERGODIC** if there exist $M(x) < \infty$ and $\rho < 1$ s.t.

$$\|P^n(x_n, \cdot) - \pi(\cdot)\| \leq M(x_0)\rho^n$$

A MC is **UNIFORMLY ERGODIC** if, for every x_0

$$\|P^n(x_n, \cdot) - \pi(\cdot)\| \leq M\rho^n$$

The constant ρ is the rate of convergence and coincides with the spectral radius $= \sup_k |\lambda_k|$

Uniform ergodicity \Rightarrow geometric ergodicity

Example of the kind of theorems you can get relative to the convergence rate of MCMC:

The **independence Metropolis-Hastings** algorithm (i.e. when the proposal does not depend on the current position of the MC: $q(x, y) = q(y)$) is uniformly ergodic if and only if

$\frac{\pi(x)}{q(x)}$ is bounded

MORE "HISTORY"

Green (1995)

specifying proposals indirectly
allowing varying dimensions:

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{g(u')}{g(u)} \left| \frac{\partial(y, u')}{\partial(x, u)} \right|$$

Tierney and Mira (1999)

delaying rejection:

$$\alpha(x, y, z) = 1 \wedge \frac{\pi(z)}{\pi(x)} \frac{q_1(z, y)}{q_1(x, y)} \frac{[1 - \alpha(z, y)]q_2(z, y, x)}{[1 - \alpha(x, y)]q_2(x, y, z)}$$

Green and Mira (2001)

delaying rejection specifying proposals
indirectly and allowing varying dimensions:

$$\alpha(x, y, z) = 1 \wedge \left\{ \frac{\pi(z)}{\pi(x)} \frac{g_1(\tilde{u}_1)}{g_1(u_1)} \frac{[1 - \alpha(z, y^*)]}{[1 - \alpha(x, y)]} \right. \\ \left. \frac{g_2(\tilde{u}_2)}{g_2(u_2)} \left| \frac{\partial(z, \tilde{u}_1, \tilde{u}_2)}{\partial(x, u_1, u_2)} \right| \right\}$$

REVERSIBLE JUMP ALGORITHM

(Green, Biometrika '95)

“if the number of things you don't know is one of the things you don't know ...”

EXAMPLES of APPLICATIONS:

- **mixture models** with unknown number of components
- **change points models** with unknown number of changes
- **variable selection** with unknown number of variables
- **graphical models** with unknown number of nodes

Bayesian Model Choice

Typical in model choice settings

- **model construction (nonparametrics)**
- **model checking (goodness of fit)**
- **model improvement (expansion)**
- **model pruning (contraction)**
- **model comparison**
- **hypothesis testing (Science)**
- **prediction (finance)**

REVERSIBLE JUMP ALGORITHM

current state $x \in \mathcal{R}^d$

generate m random variables: $u \sim g(\cdot)$

propose $y = h(x, u) \in \mathcal{R}^{d'}$

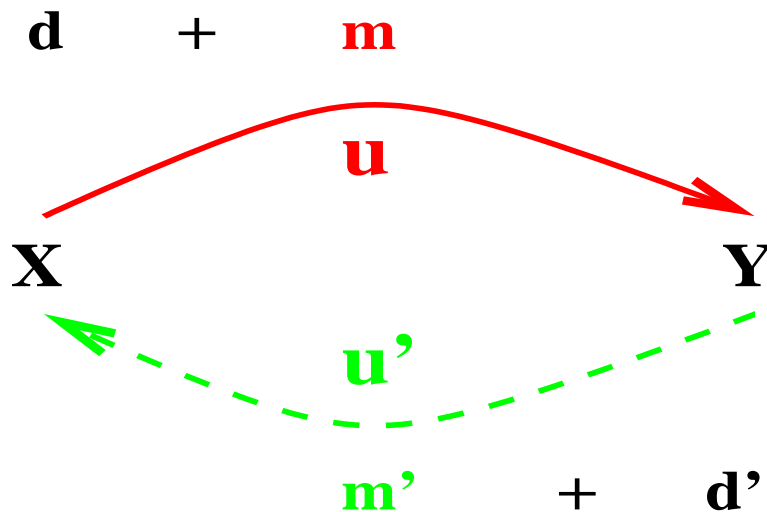
current state $y \in \mathcal{R}^{d'}$

generate m' random variables: $u' \sim g(\cdot)$

propose $x = h'(y, u') \in \mathcal{R}^d$

dimension matching

$$d + m = d' + m'$$



Birth and Death processes

Use of an alternative methodology based on a Birth-&-Death (point) process

Preston, 1976; Ripley, 1977; Geyer & Møller, 1994; Stevens, 1999

Idea: Create a Markov chain in *continuous time*, i.e. a *Markov jump process*, moving between models, by births (to increase the dimension), deaths (to decrease the dimension), and other moves.

QUESTION

If P and Q have stationary distribution π , which one is “better”?

What does “better” mean in this context?

SELECTION CRITERIA

- **speed of convergence** to stationarity
- **asymptotic variance** of MCMC estimates, $V(f, P)$

CONFLICTING BEHAVIORS

- $\{\lambda_{0P} \geq \lambda_{1P} \geq \dots\} =$ ordered eigenvalues
- $\{e_{0P}, e_{1P}, \dots\} =$ corresponding eigenvectors

ASYMPTOTIC VARIANCE:

$$V(f, P) = \sum_j \frac{1 + \lambda_{jP}}{1 - \lambda_{jP}} k_j \sigma_{\pi}^2(f)$$

with $k_j \geq 0$ and $\sum_j k_j = 1$

SPEED of CONVERGENCE:

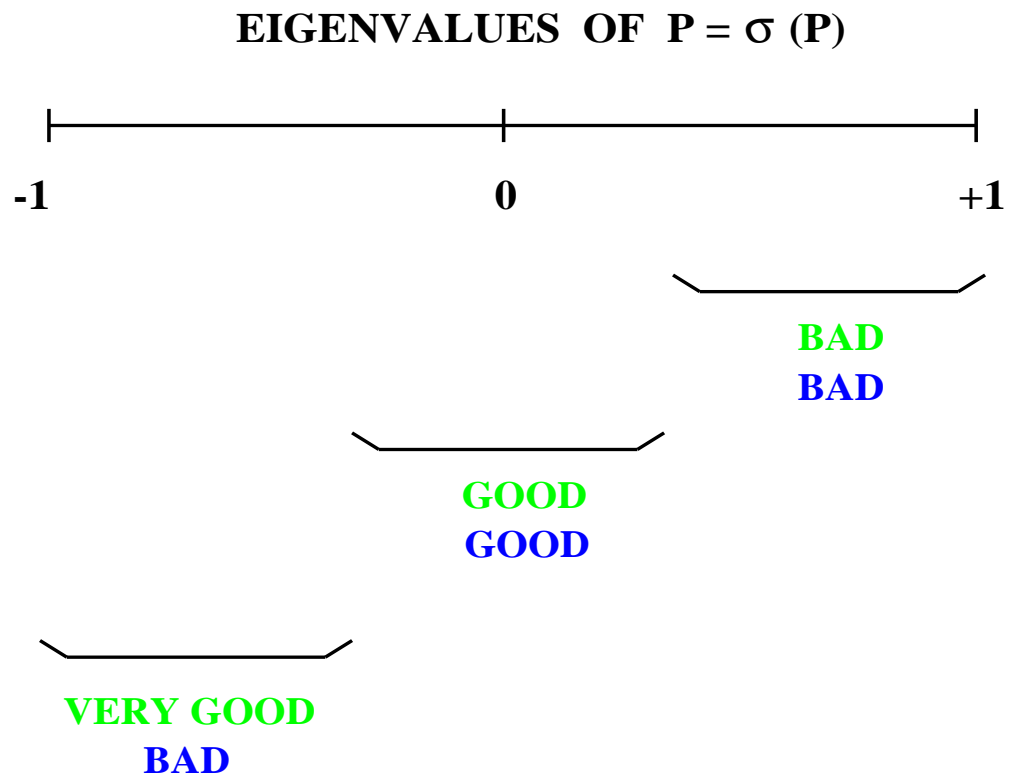
$$P^n(x, y) = \sum_j e_{jP}(x) e_{jP}(y) \lambda_{jP}^n$$

with $e_{0P}(\cdot) = \pi$ and $\lambda_{0P} = 1$

CONFLICTING BEHAVIORS

small variance in CLT \iff small eigenvalues

fast convergence \iff small $|\text{eigenvalues}|$



ORDERING based on EFFICIENCY of ESTIMATES

Given two Markov chains P and Q with the same stationary distribution π

RELATIVE EFFICIENCY

P is more efficient than Q **relative** to f ,

$$\boxed{P \succeq_{E,f} Q}, \text{ if } V(f, P) \leq V(f, Q)$$

ABSOLUTE EFFICIENCY

P is **uniformly** more efficient than Q ,

$$\boxed{P \succeq_f Q}, \text{ if } V(f, P) \leq V(f, Q), \quad \forall f \in L^2(\pi)$$

PESKUN ORDERING

Peskun (1973), Tierney (1995)

P dominates Q off-diagonally, $\boxed{P \succeq_P Q}$, iff

- finite state spaces

$$P(x, y) \geq Q(x, y) \quad x \neq y$$

- general state spaces

$$P(x, B) \geq Q(x, B) \quad x \notin B$$

Intuition: when $X_{t+1} = X_t$ we fail to explore the state space and increase the covariance along the sample path of the chain

THEOREM

If P and Q are **reversible** w.r.t. π then

$$\boxed{P \succeq_P Q}$$



$$\boxed{P \succeq_E Q}$$

AUXILIARY VARIABLES

Swendsen and Wang (1987) - Edwards and Sokal (1988)
Besag and Green (1993)

- $\pi(x)$, $x \in \mathcal{X}$ → Distribution of interest
- introduce $y \in \mathcal{Y}$ → Auxiliary Variable

so that
$$\pi(x, y) = \underbrace{\pi(x)}_{\text{GIVEN}} \underbrace{\pi(y|x)}_{\text{MADE-UP}}$$

- set up a Markov chain $\{x_t, y_t\}$ with $\pi(x, y)$ as unique stationary dist.
- use the realizations of the marginal X-chain for inference on $\pi(x)$

SIMPLE SLICE SAMPLER

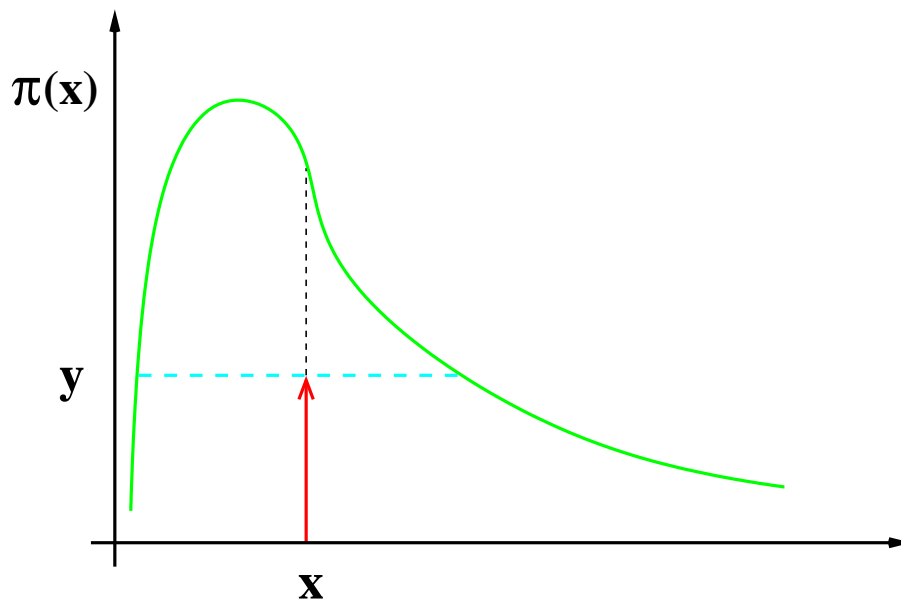
Take $\pi(y|x) \sim \text{Uniform on } [0, \pi(x)]$
so that the joint distribution is

$$\pi(x, y) = \pi(x) \frac{1}{\pi(x)} \mathbf{I}[y < \pi(x)]$$

run a **Gibbs Sampler** with $\pi(x, y)$ as stationary

$$\pi(y|x) \sim \text{Uniform on } [0, \pi(x)]$$

$$\pi(x|y) \sim \text{Uniform on } A_y = \{x : \pi(x) > y\}$$



SLICE SAMPLER

- $\pi(x) \propto \underbrace{q(x)}_{\text{PRIOR SIMPLE}} \underbrace{l(x)}_{\text{LIKELIHOOD COMPLICATED}}$
 $y|x \sim U[0, l(x)]$

- $\pi(x, y) = \pi(x) \frac{1}{l(x)} \mathbf{I}[y < l(x)]$
 $\propto q(x) l(x) \frac{1}{l(x)} \mathbf{I}[y < l(x)]$
 $\propto q(x) \mathbf{I}[y < l(x)]$

- GIBBS SAMPLER

$$\pi(y|x) \sim U[0, l(x)]$$

$$\pi(x|y) \sim q(x) \text{ on } A_y = \{x : l(x) > y\}$$

SLICE SAMPLER IS

- a “positive” Markov chain
- Harris recurrent if $l(x) > 0$
- generalizable with more aux. var.



MULTIPLE SLICE SAMPLER

- stochastically monotone



PERFECT SLICE SAMPLER

(Mira, Møller and Roberts, JRSS B, 2000)

IMPROVING EFFICIENCY of the INDEPENDENCE M-H

Given an independence M-H sampler with $\pi(\cdot)$ = stationary and $q(\cdot)$ = proposal



there exists a corresponding slice sampler that performs better in the Peskun sense



$$\pi(x) \propto q(x) l(x) \quad \text{with} \quad l(x) \propto \frac{\pi(x)}{q(x)}$$

if $l(x)$ is bounded

both samplers are uniformly ergodic

if $l(x)$ is not bounded

I. M. H. is not geometrically ergodic

S. S. can still be geometrically ergodic

use S.S. in mixtures or cycles to make them uniformly ergodic

IMPROVING THE M-H-G ALGORITHM

Peskun says that whenever $X_{t+1} = X_t$
MCMC estimates become less efficient

In the M-H-G algorithm this happens
every time a candidate is rejected

Thus we can beat M-H-G in the Peskun sense
by diminishing the rejection frequency

Delaying Rejection $\succeq_{\mathbf{E}}$ Metropolis-Hastings
in Metropolis-Hastings Algorithm

Delaying Rejection $\succeq_{\mathbf{E}}$ Reversible Jump
in Reversible Jump Algorithm

Whether delaying rejection is useful in practice
depends on whether the **reduction in variance**
compensates the additional **computational cost**

DELAYING REJECTION IN METROPOLIS-HASTINGS ALGORITHMS

Current position $X_t = x$

(1) propose a candidate move $y \sim q_1(x, \cdot)$

(2) with probability $\alpha(x, y)$ let $X_{t+1} = y$

(3) if y is rejected propose a new candidate
move $z \sim q_2(x, y, \cdot)$

(4) with probability $\alpha(x, y, z)$ let $X_{t+1} = z$

(5) keep proposing candidates until acceptance

(5') interrupt the delaying process and set $X_{t+1} = x$

The acceptance probabilities are computed
so that **reversibility** w.r.t. π is preserved
separately at each stage

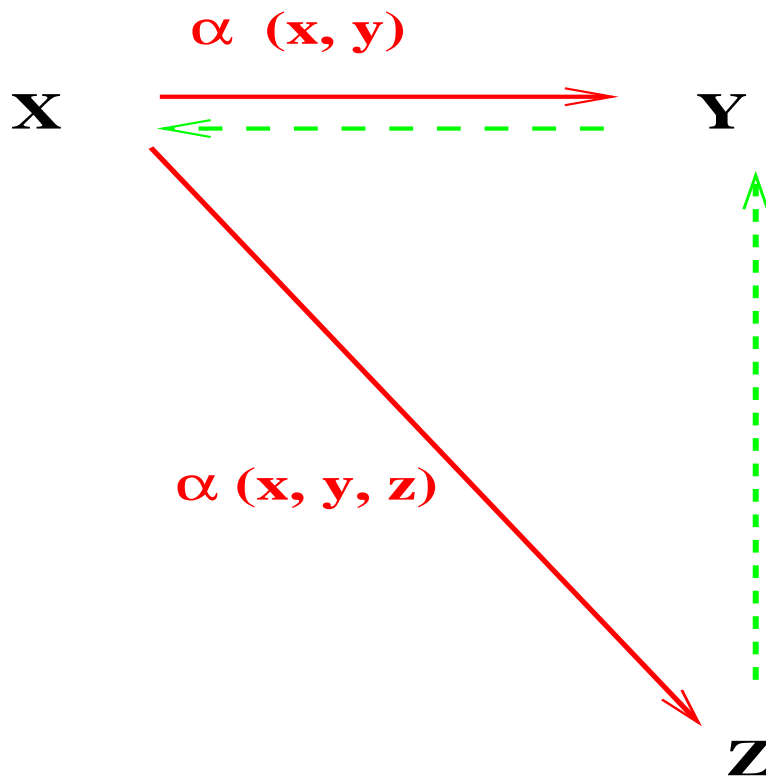
- First stage acceptance probability:

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{q_1(y, x)}{q_1(x, y)}$$

same as in std Metropolis-Hastings

- Second stage acceptance probability:

$$\alpha(x, y, z) = 1 \wedge \frac{\pi(z)}{\pi(x)} \frac{q_1(z, y)}{q_1(x, y)} \frac{[1 - \alpha(z, y)]}{[1 - \alpha(x, y)]} \frac{q_2(z, y, x)}{q_2(x, y, z)}$$



ADJUSTING THE PROPOSAL DIST.

One possible reason for rejection in M-H-G algorithms is that the proposal is **locally badly calibrated** to the target

With the delaying strategy you have **freedom to use intuition** in designing the way proposals at later stages “learn” from previous mistakes

Validity is ensured by

- using the correct acceptance probability
- matching dimensions
(in a Rev. Jump setting)

ADJUSTING THE PROPOSAL DIST.

- **independence + rnd walk** proposals: the rnd walk gives protection against the potentially poor behavior of an independence chain with bad proposal distribution
- **trust region** based proposals: start with a local quadratic approximation of $\log(\pi)$ and gradually reduce the region supporting the proposal
- **griddy proposals**: select a point from the previously rejected ones with probability $\propto \pi(y_j)$ and add to the point a random increment

A PHARMACOKINETIC EXAMPLE

Plasma concentration of the drug Caldralazine in cardiac failure patients is modeled with a single compartment pharmacokinetic model:

$$\eta(t) = 30\alpha^{-1} \exp(-\beta t/\alpha)$$

$\eta(t)$ = plasma concentration t hours after adm.

α = volume of distribution

β = clearance for the patient

Measurements y_i are modeled as:

$$\log y_i = \log \eta(t_i) + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

PRIORS:

$\alpha \sim$ Uniform $[a, b]$

$\beta \sim$ Uniform $[c, d]$

$\sigma^2 \sim$ Improper Prior



integrated out analytically

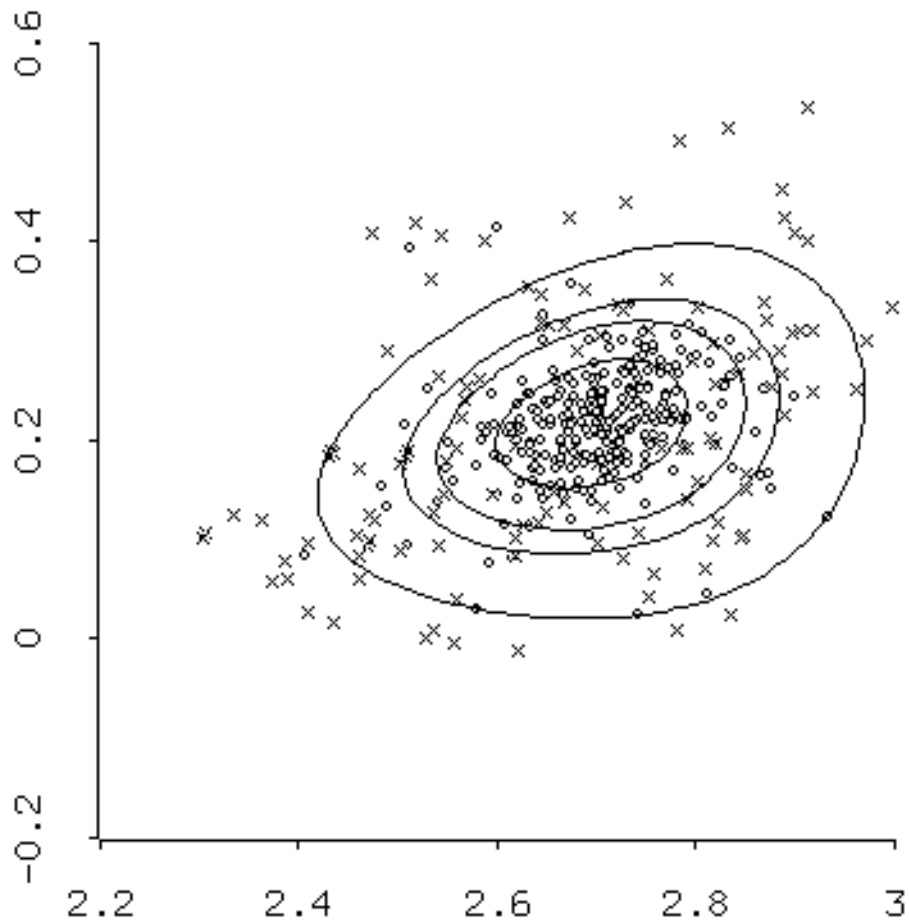
POSTERIOR:

two-dimensional distribution for α and β

SAMPLER 1

First stage: independence proposal

Second stage: random walk proposal



Circles = points accepted at the **first** stage (70 %)

Crosses = points accepted at **second** stage (80 %)

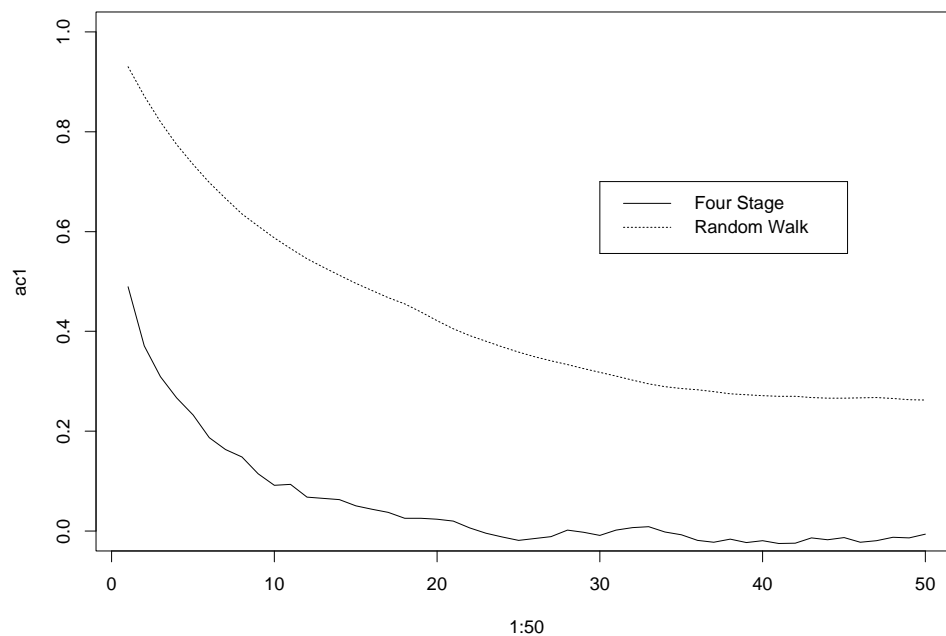
SAMPLER 2

Four stages model-trust region sampler

Acceptance proportions:

Stage 1	Stage 2	Stage 3	Stage 4	Rejected
0.78	0.09	0.03	0.01	0.09

Autocorrelations



Sample autocorrelation functions

BAYESIAN CREDIT SCORING

estimate the default probability of companies that apply to banks for loan

DIFFICULTIES

- default events are **rare events**
- analysts may have **strong prior** opinions
- observations are **exchangeable** within sectors
- different sectors might present **similar behaviors** relative to risk

THE DATA

7520 companies

1.6 % of which defaulted

7 macro-sectors (identified by experts)

4 performance indicators (derived by experts from balance sheet)

	Dimension	% Default
Sector 1	63	0%
Sector 2	638	1.41%
Sector 3	1343	1.49%
Sector 4	1164	1.63%
Sector 5	1526	1.51%
Sector 6	315	9.52%
Sector 7	2471	0.93%

THE MODEL

Bayesian hierarchical logistic regression model

Notation:

- n_j : number of companies belonging to sector j , $j = 1, \dots, 7$
- $y(i_j)$: binary response of company i in sector j . $y = 1 \Leftrightarrow$ default
- $\underline{x}(i_j)$: 4×1 vector of covariates (performance indicators) for company i in sector j
- $\underline{\alpha}$: 7×1 vector of intercepts one for each sector
- $\underline{\beta}$: 4×1 vector of slopes one for each performance indicator

PARAMETERS of INTEREST: $\underline{\alpha}$ and $\underline{\beta}$

PRIORS:

$$\alpha_j | \mu_\alpha, \sigma_\alpha \sim N_1(\mu_\alpha, \sigma_\alpha^2) \quad \forall j$$

$$\mu_\alpha \sim N_1(0, 64)$$

$$\sigma_\alpha^2 \sim IG(25/9, 5/9)$$

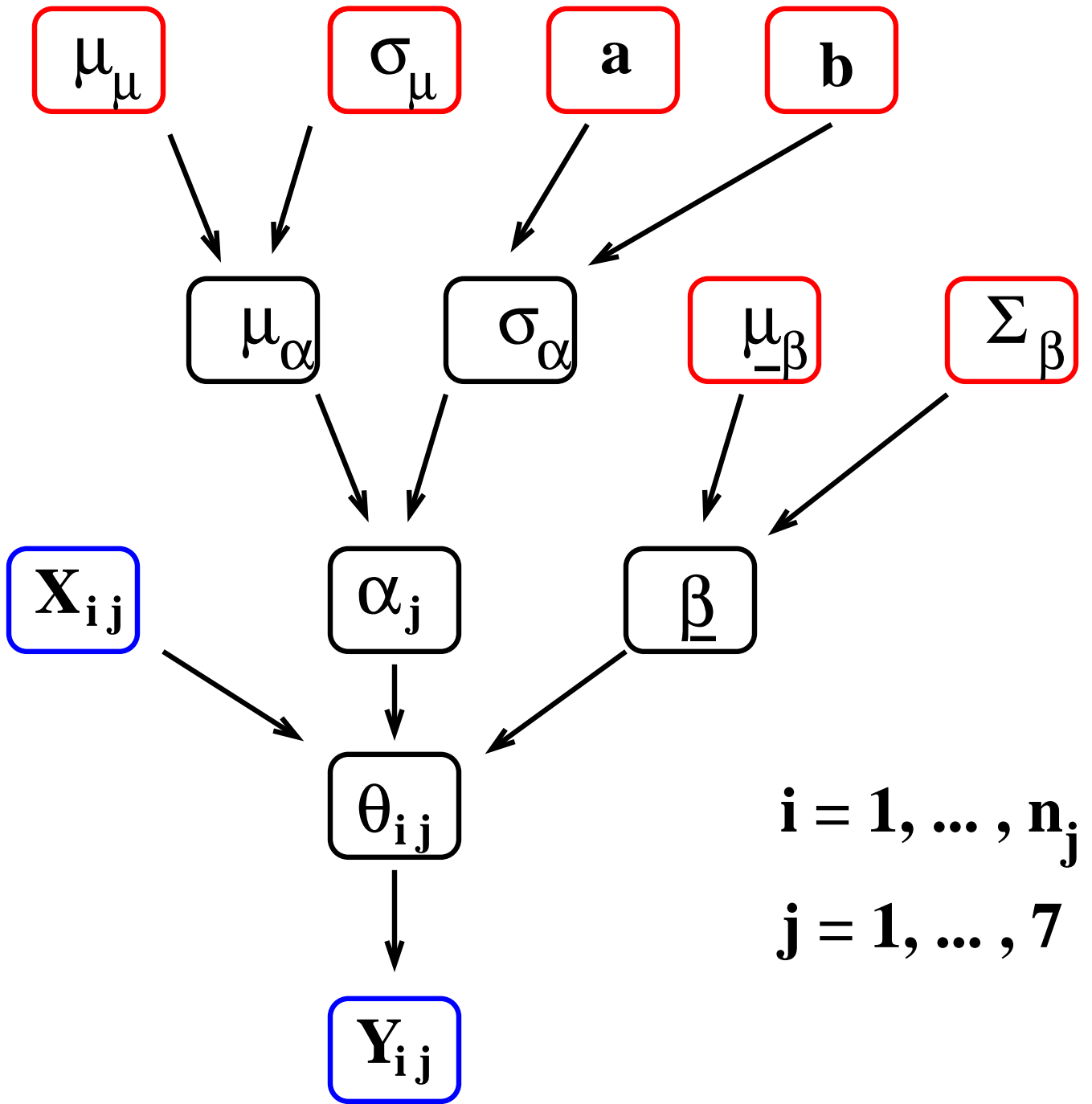
$$\underline{\beta} \sim N_4(\underline{0}, 64 \times I_4)$$

POSTERIOR:

$$\pi(\underline{\alpha}, \underline{\beta}, \mu_\alpha, \sigma_\alpha | y, x) \propto \prod_j \prod_i \theta_{ij}^{y(i_j)} (1 - \theta_{ij})^{1-y(i_j)} \prod_j p(\alpha_j | \mu_\alpha, \sigma_\alpha) p(\mu_\alpha) p(\sigma_\alpha) p(\underline{\beta})$$

where

$$\theta_{ij} = \frac{\exp[\alpha_j + \underline{x}'(i_j)\underline{\beta}]}{1 + \exp[\alpha_j + \underline{x}'(i_j)\underline{\beta}]}$$



TUNING of the PROPOSALS

Joint updates of all, or groups, of variables result in very low acceptance probabilities and thus slowly mixing sampler

thus we update each one of 13 parameters of interest separately in a **fixed scan**

D.R.: σ_1 as in table below, $\sigma_2 = \frac{\sigma_1}{2}$

M.H.: $\sigma = \frac{\sigma_1 + \sigma_2}{2}$

α_1	1.2
$\alpha_2, \dots, \alpha_7, \mu_\alpha, b_2$	0.4
σ_α	3
b_1, b_4	0.15
b_3	0.3

**PARAMETER ESTIMATES:
a comparison**

par.	MH Estimates	MH Cred. Int.
α_1	-5.87 (0.11)	-7.63; -4.36
α_2	-5.15 (0.05)	-5.99; -4.60
α_3	-4.94 (0.03)	-5.66; -4.50
α_4	-4.72 (0.05)	-5.45; -4.32
α_5	-5.03 (0.04)	-5.70; -4.64
α_6	-3.75 (0.04)	-4.45; -3.33
α_7	-6.08 (0.07)	-6.93; -5.64
β_1	-0.09 (0.01)	-0.19; 0.031
β_2	-1.16 (0.04)	-1.83; -0.74
β_3	-1.30 (0.04)	-1.70; -1.02
β_4	0.06 (0.001)	-0.05; 0.15
μ_c	-5.06 (0.07)	-6.04; -4.35
σ_c^2	-5.06 (0.07)	-6.04; -4.35

**PARAMETER ESTIMATES:
a comparison**

par.	DR Estimates	DR Cred. Int.
α_1	-5.95 (0.05)	-7.67; -4.48
α_2	-5.2 (0.02)	-5.97; -4.67
α_3	-4.98 (0.02)	-5.55; -4.60
α_4	-4.77 (0.01)	-5.39; -4.36
α_5	-5.08 (0.02)	-5.66; -4.68
α_6	-3.79 (0.02)	-4.41; -3.36
α_7	-6.14 (0.03)	-6.79; -5.75
β_1	-0.1 (0.002)	-0.19; 0.008
β_2	-1.19 (0.02)	-1.76; -0.74
β_3	-1.32 (0.02)	-1.67; -1.09
β_4	0.07 (0.002)	-0.023; 0.14
μ_c	-5.13 (0.02)	-6.07; -4.40
σ_c^2	-5.13 (0.02)	-6.07; -4.40

PERFORMANCE COMPARISON: MH VS DR

based on the asymptotic variance of the resulting MCMC estimates obtained by averaging along the chain sample path

MCMC estimate:

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

Asymptotic variance:

$$V(f, P) = \sigma^2 \sum_{k=-\infty}^{\infty} \rho_k = \lim_{n \rightarrow \infty} n \operatorname{Var}_P[\hat{\mu}_n]$$

↓

τ = integrated autocor. time

↓

$\hat{\tau}$ = Sokal's adaptive truncated correlogram estimate

With $\sigma_\alpha^2 = 1$ and empirical Bayes approach

$\hat{\tau}$	β_1	β_2	β_3	β_4	μ_α
MH	6.5	22.9	21.1	5.7	17.2
DR	4.1	18.9	12.7	3.9	11.6

$\hat{\tau}$	α_1	α_2	α_3	α_4	α_5	α_6	α_7
MH	9.8	15.3	20.0	18.7	20.7	23.7	25.6
DR	6.7	10.8	9.4	14.9	12.5	17.1	14.7

With Gamma prior on σ_α^2 and diffuse priors centered at zero

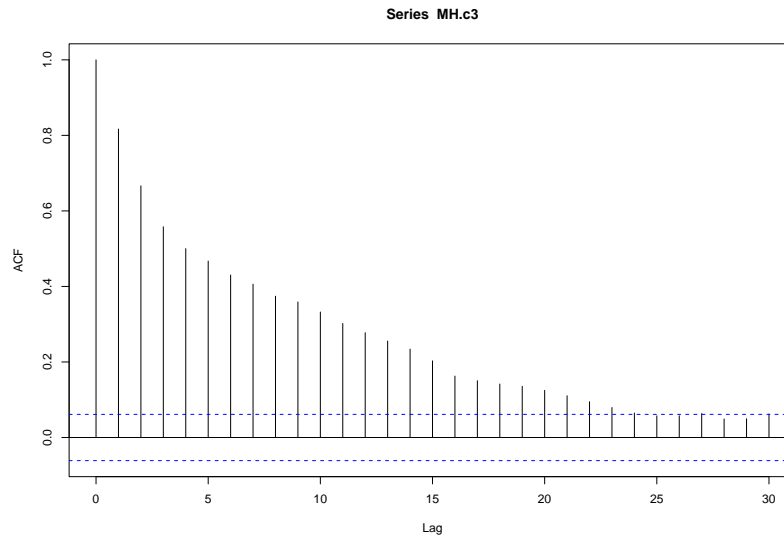
$\hat{\tau}$	β_1	β_2	β_3	β_4	μ_α	σ_α
MH	10.0	64.5	23.4	5.6	15.9	20.2
DR	7.2	38.1	20.9	4.2	14.6	15.6

$\hat{\tau}$	α_1	α_2	α_3	α_4	α_5	α_6	α_7
MH	26.9	50.1	43.2	50.3	54.6	60.6	60.2
DR	17.0	18.4	28.1	28.4	30.1	32.3	35.1

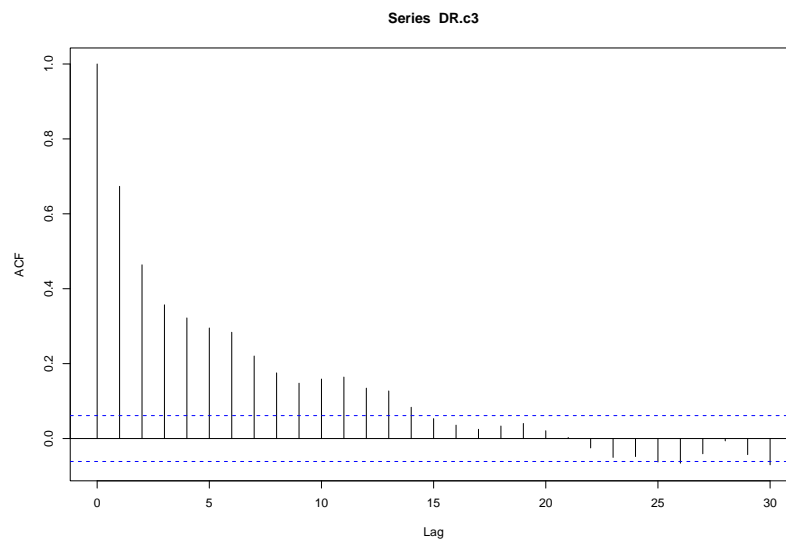
Values obtained averaging over 5 simulations of length 1024 after a burn-in of 150 steps

AUTOCORRELATION FCT for α_3

Metropolis-Hastings sampler



Delaying rejection sampler



RESULTS

Estimates of DP:

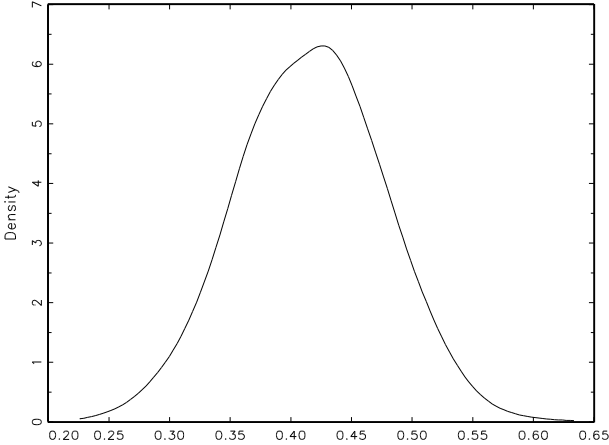
- company 30 in sector 6
- company 20 in sector 2

$$DP = \theta_{i,j} = \frac{\exp(\alpha_j + \underline{x}'(i_j)\underline{\beta})}{1 + \exp(\alpha_j + \underline{x}'(i_j)\underline{\beta})}$$

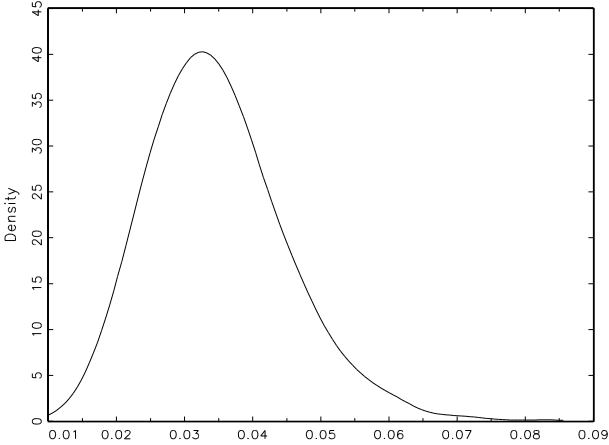
	plug in $\hat{\alpha}_j$ and $\hat{\underline{\beta}}$	$\frac{1}{N} \sum_{n=150}^{1174} \theta_{i,j}^n$	MLE
$\hat{\theta}_{30,6}$	0.431	0.434	0.372
$\hat{\theta}_{20,2}$	0.032	0.034	0.026

Posterior kernel density estimate of DP

- company 30 in sector 6



- company 20 in sector 2



CONCLUSIONS

- **no bias** induced by rare events
- can incorporate **experts prior** opinions
- sectors with low or no default events
borrow information from other sectors
- having the joint posterior of all DP
can compute **risk of a portfolio** of loans

EXTENSIONS

- allow **different slopes** for different sectors
- select the sectors based on default probabilities via **partition models**
- include **economic cycle indicators** among the covariates
- include **time** in the analysis:
dynamic MCMC, particle filters

CONCLUSIONS on Delaying Rejection

the **Delaying Rejection** strategy improves
Metropolis-Hastings-Green algorithms

modulo extra computational and
programming effort

RECENT DEVELOPMENTS

Langevin diffusions:

if you have information on the gradient of the log target use it to construct better proposals:

$$q(\theta, \theta') = \frac{1}{(2\pi\sigma)^{d/2}} \exp \left\{ \frac{-\|\theta' - \theta - \sigma^2 \nabla \log \pi(x)/2\|^2}{2\sigma^2} \right\}$$

Adaptive MCMC:

- use sampled path to calibrate the proposal
- loose the Markovian property
- need to prove ergodicity from first principles

EXAMPLE: DR+AM

AM uses a Gaussian proposal with covariance matrix calibrated via sample path of the MC

$$\text{Cov}(X_0, \dots, X_k) = \frac{1}{k} \left(\sum_{i=0}^k X_i X_i^T - (k+1) \bar{X}_k \bar{X}_k^T \right)$$

AM = global adaptive strategy

DR = local adaptive strategy

AM → protects from under calibration of q

DR → protects from over calibration of q

Particle filters:

for target distributions that evolve over time as in target tracking, patients monitoring or financial applications

General advices:

- when using improper priors check that your posterior is integrable otherwise your MC becomes transient eventually but you might not realize this if you do not run your MC long enough
- when possible try to integrate out what you can, do not run your MC blindly

- to compute acceptance probabilities work on the log scale
- debugging: take special cases where you know the answer and you only need to change the code a little to get to that special case