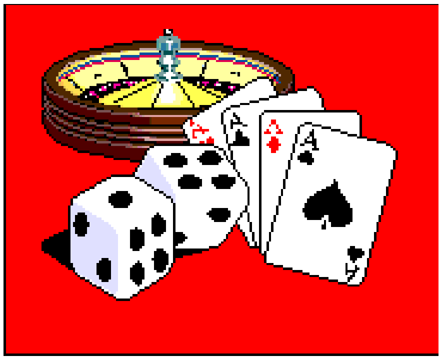
Introduction to Monte Carlo and MCMC Methods

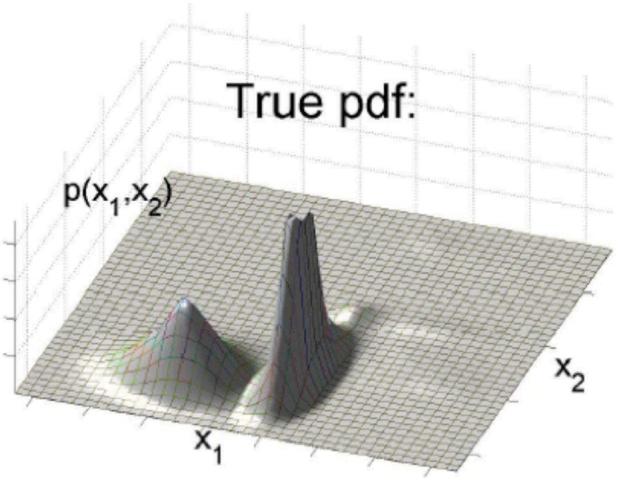


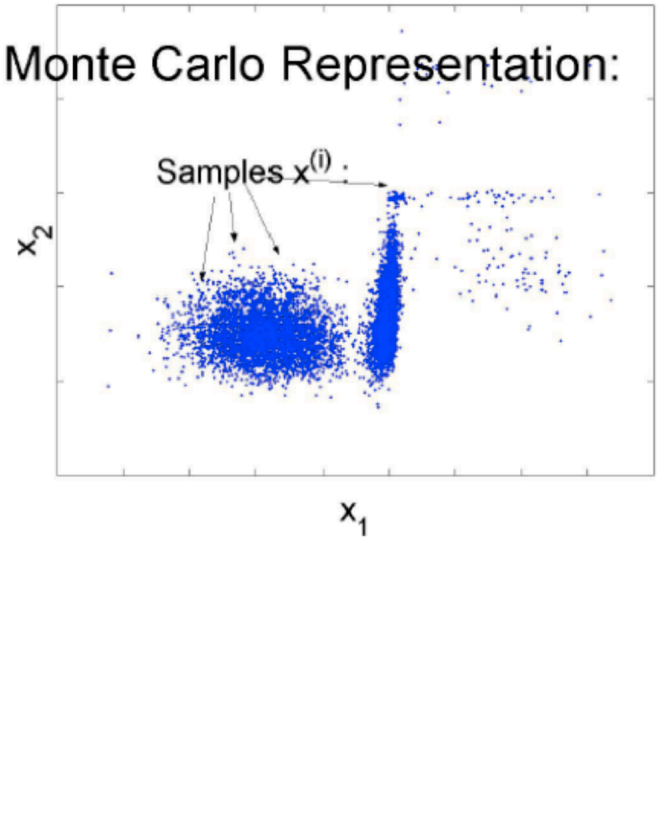
Antonietta Mira

Introduction:

MC MC methods are sophisticated and general algorithms for simulation from complex probability models,
high dimensional,
highly non-Gaussian,
highly non-linear and
possibly multimodal

Given the simulated path of the Markov chain we can compute Monte Carlo expectations for any quantities of interest by averages along the sample path


True pdf:
$p(x_1,x_2)$
x_1
x_2


Monte Carlo Representation:
Samples $x^{(i)}$:
x_2
x_1

INDEX

- 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 and Software

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

HISTORICAL INTRODUCTION

World War II
Los Alamos Scientific Laboratory
J. von Neumann, S. Ulam, E. Fermi
Random neutron diffusion in fissile material

QUESTION: what is the distance covered by a neutron shot trough different materials?

ANSWERS:
Theoretical computation: too complicated
Empirical experiment: too risky
Simulation: approximate but feasible!

They knew, for a single neutron

- average distance with constant velocity
- collision probability with an atomic nucleus
- probability of absorption/repulsion

In physics:

single neutron ⇒ huge number of neutrons
single event ⇒ complicated chain of events

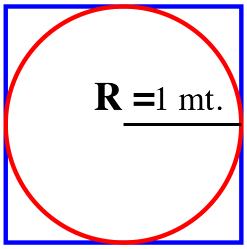
In finance:

single agent ⇒ huge number of actors
single buy/sell action ⇒ many connected events

“The impact of Monte Carlo and Markov Chain Monte Carlo methods on applied statistics has been truly revolutionary” W.S. Kendall

Economics, ecology, climate models, epidemilogy, genetics, ...: similar analogy

Approximation of the value π



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

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

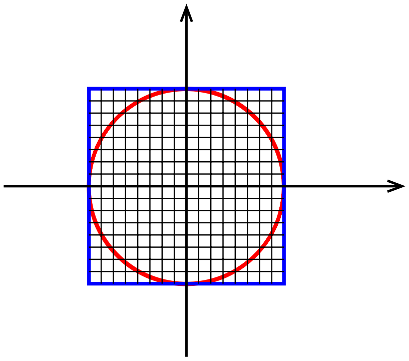
89 successes in 100 tosses $\Rightarrow \widehat{\pi} = 3.57$

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

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, $\cdots$, 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 unit square

The Monte Carlo (MC) estimator of μ is :

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i)$$

where $(X_i, Y_i) \sim \pi, i = 1, \cdots, n$; i.i.d.

Under regularity conditions the LLN + CLT ensure that the MC estimator is asymptotically unbiased and has asymptotic variance

$$\mathsf{V}(\widehat{\mu}; f, \pi) = \frac{1}{n}\sigma^2_{\pi}(f)$$

PROBLEM 1

Increase dimensions

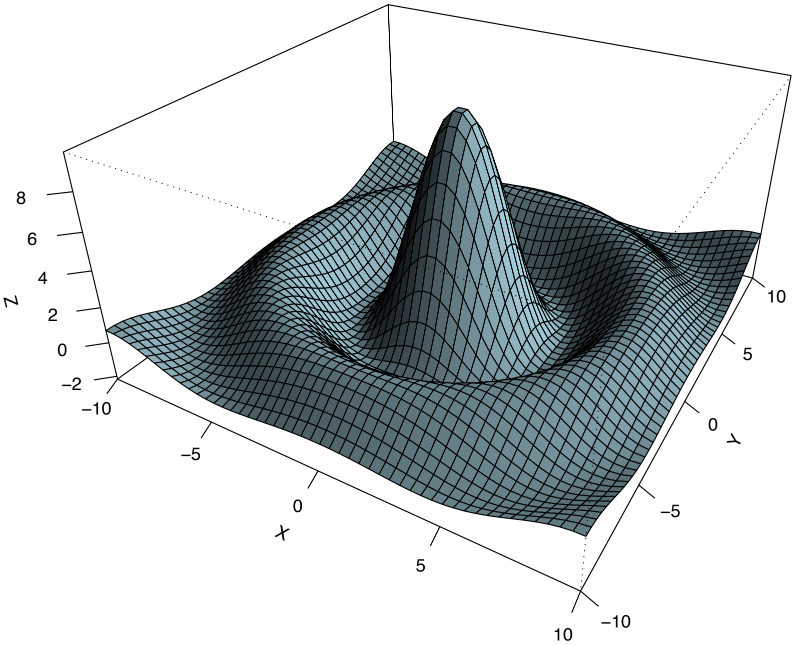
square → cube → $\Re^d$
circle → sphere → $S \subset \Re^d$

as other numerical methods that rely on n-point evaluations we have absolute error of estimators that decreases as $n^{-1/d}$
instead of $n^{-1/2}$

curse of dimensionality

PROBLEM 2

uniform $\pi \rightarrow$ complicated distribution



Often π is so complicated that we are not able to generate i.i.d. samples from it 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
$X_0,\ X_1,\ X_2, \cdots$

And we use the X_i "as if" they were i.i.d. from π

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

REFERENCES

⇒ Ripley
Stochastic simulation, 1987

⇒ Gilks + Richardson + Spiegelhalter
MCMC in Practice, 1998

⇒ Handbook of Markov Chain Monte Carlo
http://www.mcmchandbook.net/

⇒ C. Robert + G. Casella
Monte Carlo Statistical Methods, 2004

⇒ S.P. Meyn + R. Tweedie
Markov Chains and Stochastic Stability, 1993

⇒ Jun S. Liu
M.C. Strategies in Scientific Computing, 2001

- ⇒ David Ardia (2008)
  Financial Risk Management with Bayesian Estimation of GARCH Models

- ⇒Rachev, Hsu, Bagasheva, Fabozzi (2008)
  Bayesian Methods in Finance

- ⇒ Bayesian Analysis (on line journal)

- ⇒ I.S.B.A.
  Internat. Society for Bayesian Analysis
  conferences, events, prices
  J-ISBA

- ⇒ MCMSki
  IMS-ISBA conference

- ⇒ Bayes on the Beach

- ⇒ ISBA World meeting

Examples of applications

- Model estimation and selection:
  GARCH, SV, GLM, Hidden Markov models
- finance: option pricing
- state space models:
  epidemiology and meteorology
- biology - physics - chemistry - genetics
- mixture models for cluster analysis:
  astronomy, population studies
- operational research
  traffic control, quality control,
  production optimization

In general Bayesian models

NOTATION of BAYESIAN INFERENCE

d = data (fixed)
x = parameters (variable)

$Pr(d|x)$ = likelihood = L

$Pr(x)$ = prior = p

$Pr(x|d)$ = posterior = π

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

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

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

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

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

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):

$$\widehat{\mu}_n \to \mu$$

i.e. the estimator is asymptotically unbiased. The variance of the MC estimator is:

$$\sigma^2(\widehat{\mu}_n) = \frac{1}{n}\sigma^2_\pi(f)$$

and can be estimated by:

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

so that, asymptotically, we have a CLT:

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

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

$\Longrightarrow$ **IMPORTANCE SAMPLING**

$\Longrightarrow$ **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:

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

For any choice of g, as long as
$\text{supp}(\pi) \subset \text{supp}(g)$
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

- the target can be approximated by weighted sum of delta-masses
  the weights compensate for the discrepancy bwn target and importance distribution
  thus select q as close to π as possible

- Sequential importance resampling

CONTRAS

- finite variance only if

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

- 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:

$$T_{(\nu,0,1)} = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_\nu/\nu}}$$

- 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 $T_{(\nu,3,1)}$

MARKOV CHAINS

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

FUTURE indep. PAST $|$ PRESENT

We will assume that the time is discrete and the state space is finite, $S = \{1, 2, \ldots, k\}$
A M.C. is specified by giving

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

- **transition probabilities**: P (a matrix)

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

- 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 $\cdots$

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

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

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

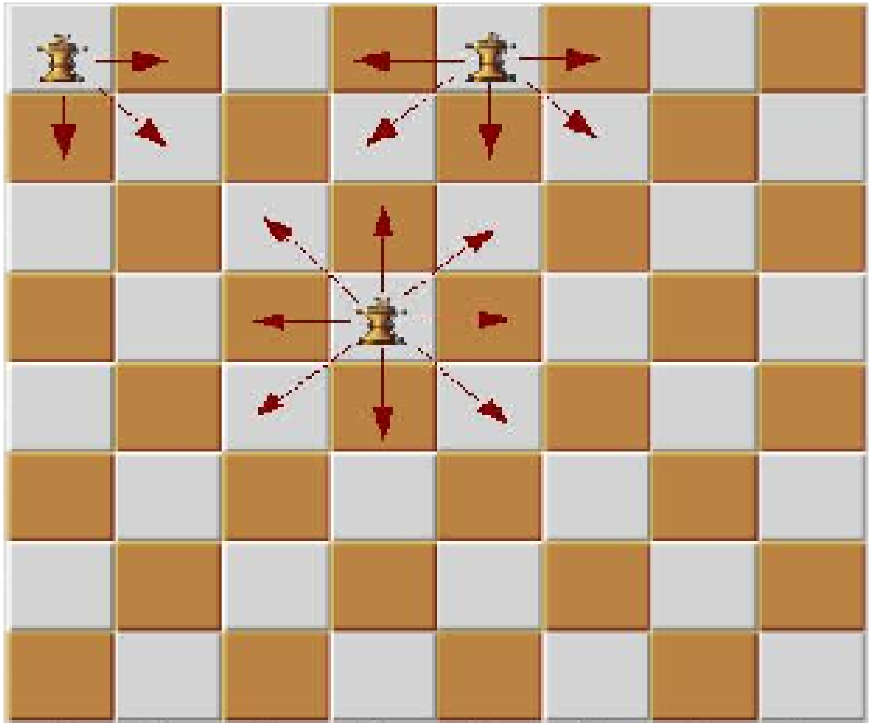
$$\pi P = \pi$$

that is, if

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

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

Imagine moving around the grid as a “**random-king**” on a chess board



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?

Imagine a larger chess board, 2 mt^2

Is the grid coarse enough?

Bigger steps → faster exploration

It depends on how variable/stable are

$$\pi = \text{target}$$

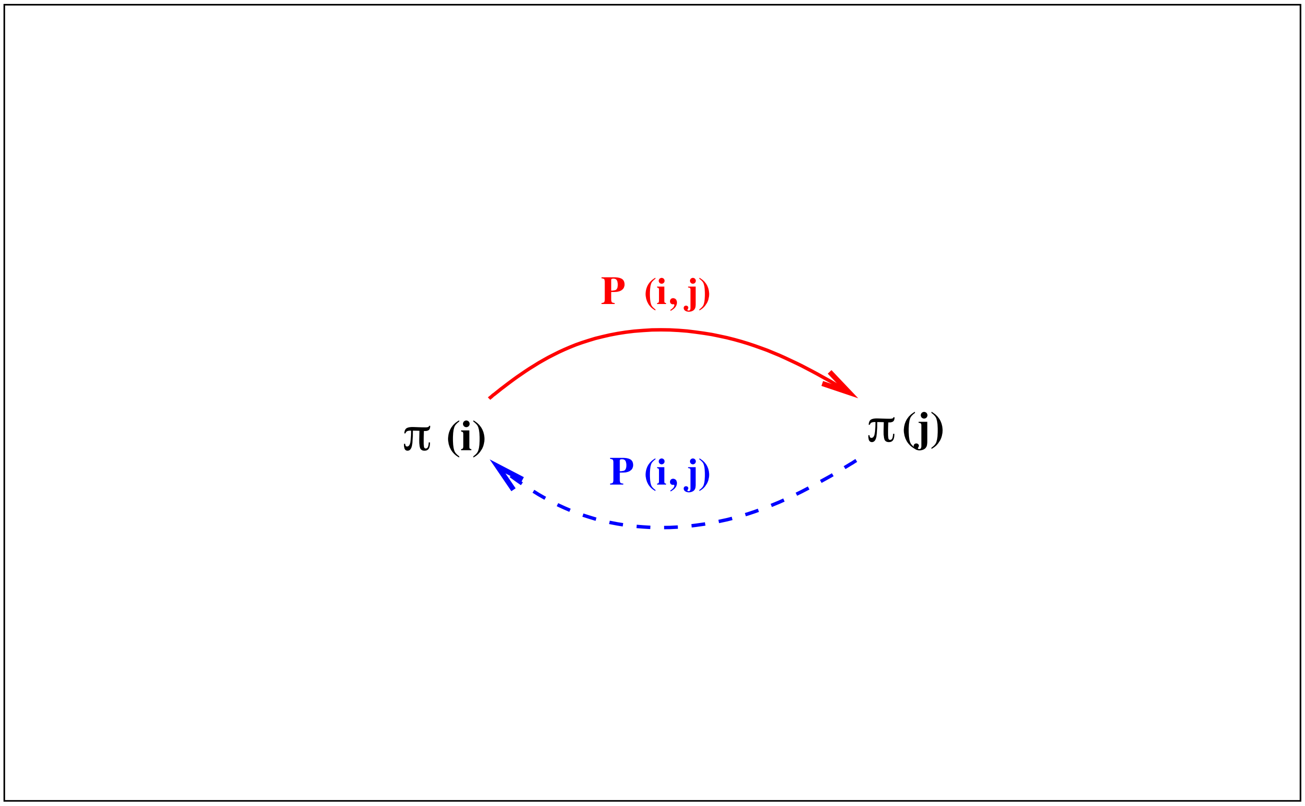
$$f = \text{function}$$

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

P is **reversible** w.r.t. π if

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

(**detailed balance condition**)
i. e. the M. C. looks the same running forward or backward



Prove that: **Reversibility → Stationarity**

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

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

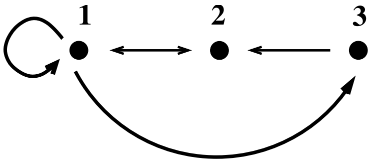
Irreducibility

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

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

$$1 \to 3 \to 2 \to 1$$

this is an **impossible** sequence

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

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

There is a sequence of states for which it is possible to tell in which direction the simulation has occurred and thus 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:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P(X_k = j | X_0 = i) = \pi(j), \qquad \forall i \in S,$$

i.e. regardless of the initial starting value of the chain i, the long run proportion of time the chain spends in state j equals $\pi(j)$, for every possible state j. Then 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

EXAMPLE

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

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

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 visits 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:

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

- **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 $E_\pi f^2 < \infty$ + uniform ergodicity
  or if
  $E_\pi |f|^{2+\delta} < \infty$ + geometric ergodicity
  or if
  $E_\pi f^2 < \infty$ + geometric ergodicity + rev.

EXAMPLES

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

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

Possible transition matrices:

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

periodic - irreducible - non reversible

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

periodic - **reducible** - reversible

EXAMPLES (cont.)

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

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

Possible transition matrices:

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

aperiodic - irreducible - reversible

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

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: $X_0, X_1, X_2 \ldots \sim P$
Estimate μ with

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

How good is the estimate?

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

where

$$\boldsymbol{\sigma}^{\mathbf{2}} = \mathrm{Var}_\pi \, f(X)$$

$$\rho_{\mathbf{k}} = \frac{\mathrm{Cov}_\pi[f(X_0), f(X_k)]}{\sigma^2}$$

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 $X_t = x$

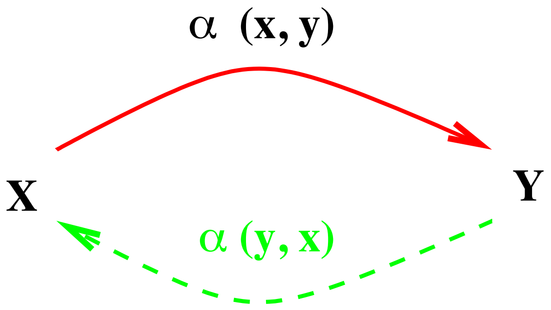
(1) propose a candidate $y \sim q(x, \cdot)$

(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:

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


α (x, y)
X
Y
α (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),

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

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.: $\mathbf{q(x, \cdot)} = N(\mathbf{0}, \sigma^2)$

RANDOM WALK M-H

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

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 $\mathbf{x} = (x_1, \cdots, x_d) \sim \pi$ 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 :

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

Tuning of the proposal is crucial!

– optimal tuning

– adaptive MCMC

(Roberts, Rosenthal and Atchadé)

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)
       accept = exp((x[t-1]^2 - y^2)/2)
       alpha = min(1,accept)
       u = runif(1)
       if (u <= alpha) x[t] = y
       else x[t] = x[t-1]
    }
    plot(1:length(x),x,
          type="l",lty=1,xlab="t",ylab="x")
    x
}
```

Easy example: Gibbs sampler

Consider a single observation $y = (y_1, y_2)$ from a bivariate normal population with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

With a uniform prior on θ, the posterior distribution is

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} |y \sim N(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$$

The full conditional distributions are

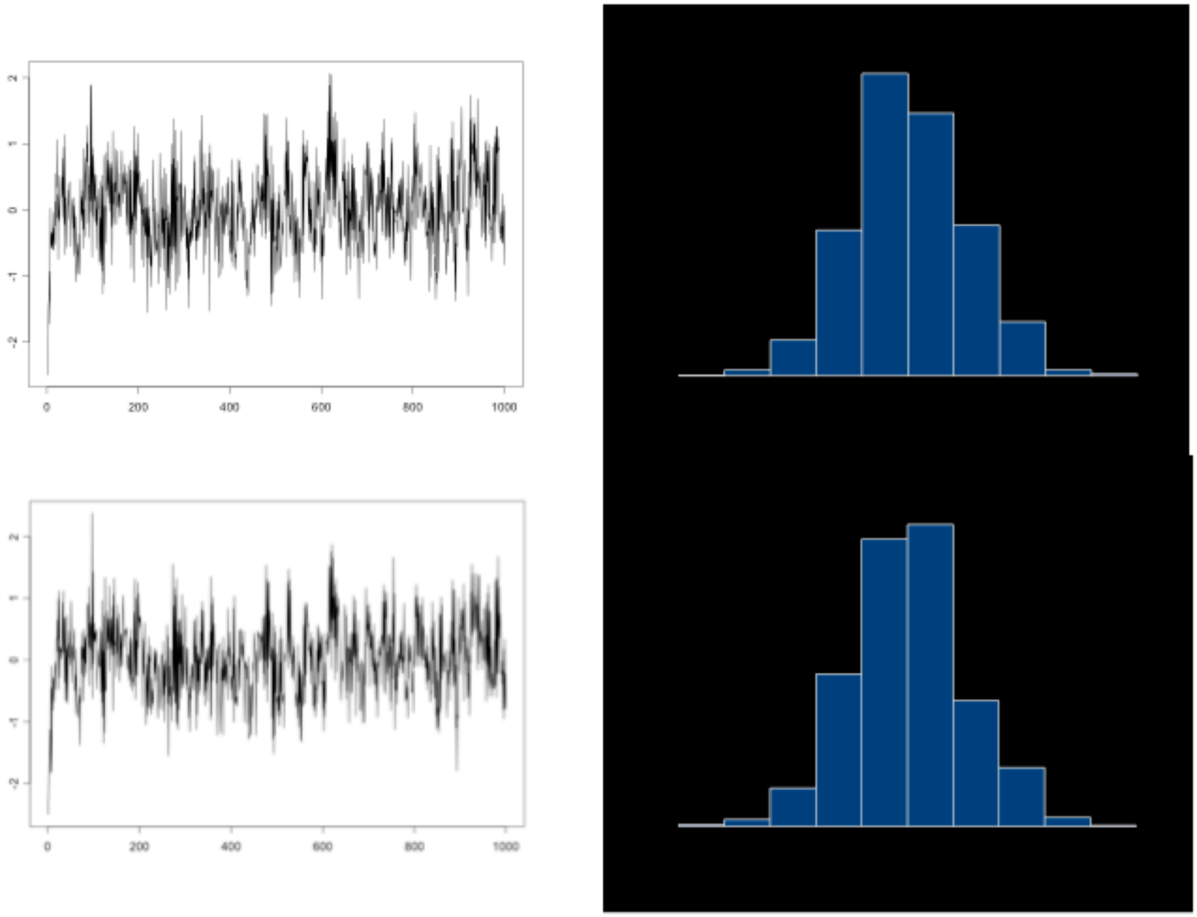
$$\theta_1|\theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2),\, 1 - \rho^2)$$

$$\theta_2|\theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1),\, 1 - \rho^2)$$

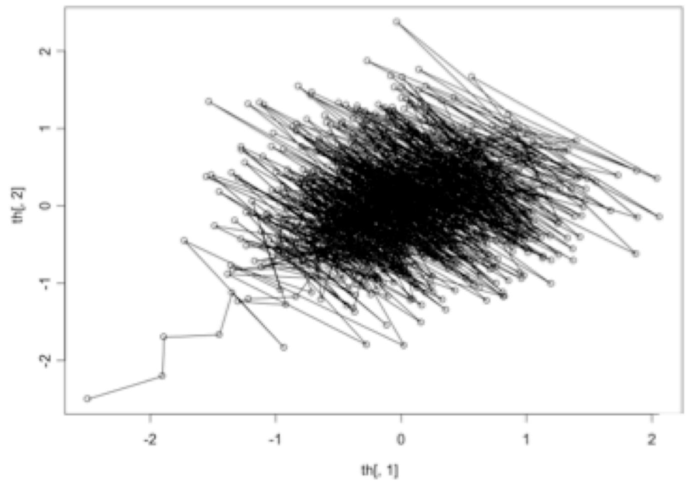
Suppose $(y_1, y_2) = (0, 0)$ and $\rho = 0.8$
Initialize the Markov chain at
$(\theta_1 = -2, \theta_2, = -2)$
Start the simulation:

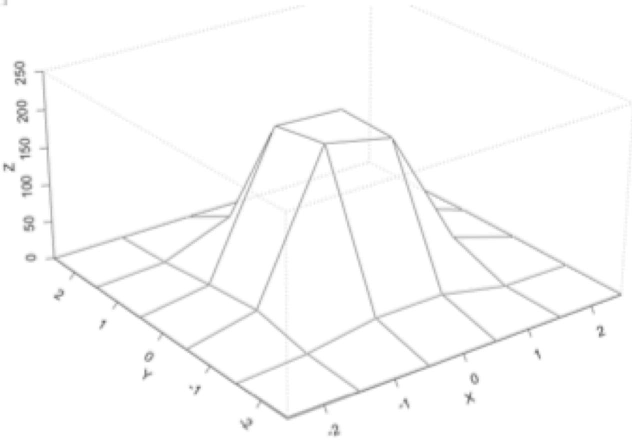
$$\begin{bmatrix} \theta_1 & \theta_2 \\ -2 & -2 \\ -1.902945 & -2.20308 \\ -1.891287 & -1.69893 \\ \ldots\ldots & \ldots\ldots \end{bmatrix}$$

1-dim sample path and histogram of θ_1 and θ_2



2-dim sample path and ”histogram” of (θ_1, θ_2)





More difficult Gibbs sampler

$Y_1, ..., Y_d$ iid $N(\mu, \sigma^2 = \tau^{-1})$

Priors:

$\mu \sim N(\mu_*, \sigma_*^2) \qquad \tau = 1/\sigma^2 \sim G(\alpha_*, \beta_*)$

Posterior:

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

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

Full conditionals:

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

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

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

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

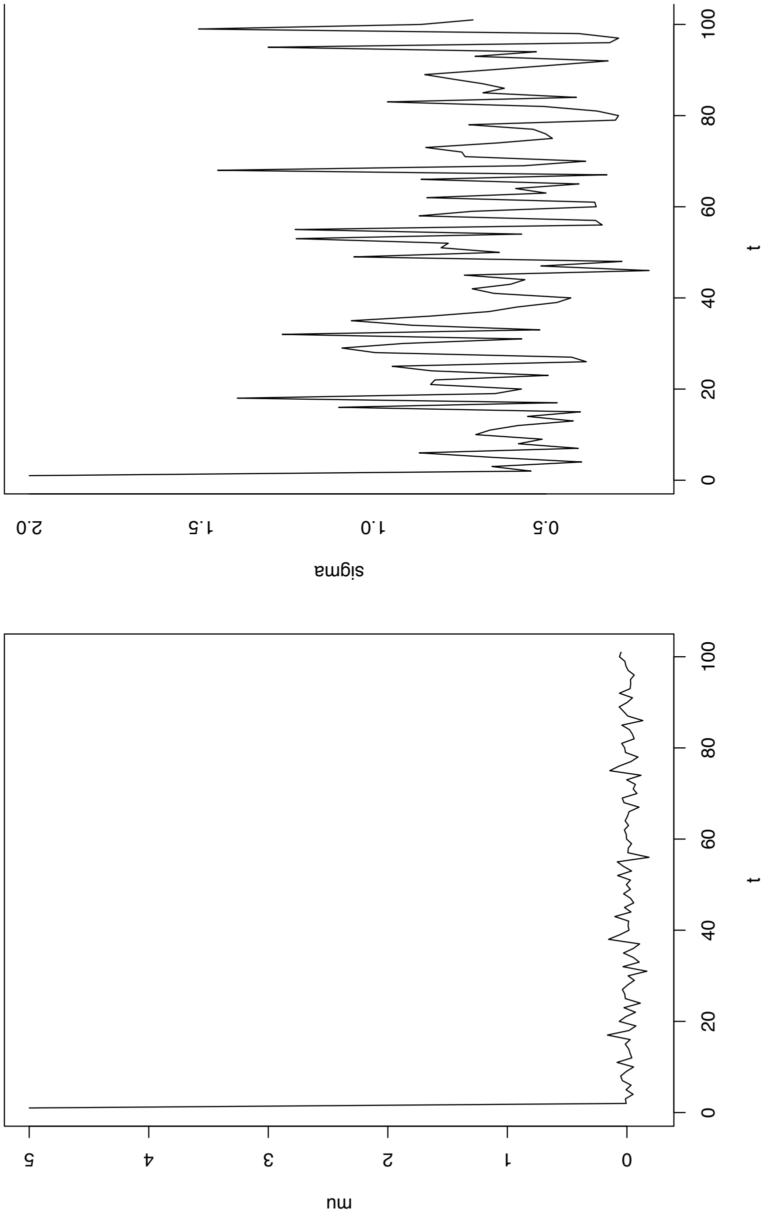
(1) Given $\tau(\mathbf{t})$,
generate $\mu(t+1)$ from $\pi(\mu|\tau(t), y_1, ..., y_d)$

(2) Given $\mu(\mathbf{t+1})$
generate $\tau(t+1)$ from $\pi(\tau|\mu(t{+}1), y_1, ..., y_d)$

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

To avoid working with very small numbers:
use the fact that $X = exp(logX)$ so that:

$$\frac{\pi(y)}{\pi(x)} = exp(log\pi(y) - log\pi(x))$$

Calibration of the proposal
pilot runs - trial and error
aim at acceptance probability $\approx$ 30%
magic number 0.234 (under regularity conditions for the target and in infinite dimensions)
See Roberts Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms*, Statistical Science 16:351-367, 2001.

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. Comp. Graph. 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

$X_0 =$ 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

$P(X_t \in A \; i.o. | X_0 = x) = 1$ 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**
  $\rightsquigarrow$ you get closer to the stationary dist.
  $\rightsquigarrow$ reduce the **variance** of your estimator

- **many short runs**
  $\rightsquigarrow$ you explore the state space better
  $\rightsquigarrow$ reduce the **bias** of your estimator

TRADE OFF

How long is long?

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

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

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

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

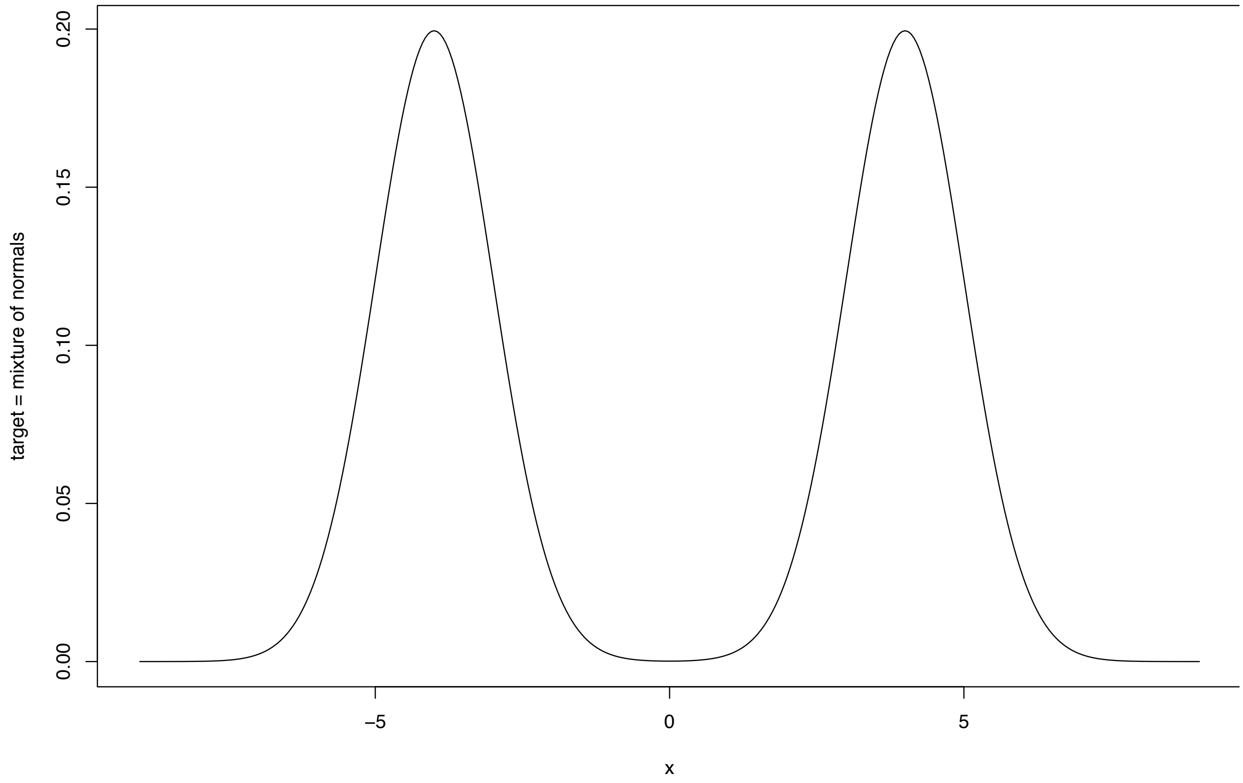
w.l.o.g. take $\sigma^2 = 1$ so that the acceptance probability is

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

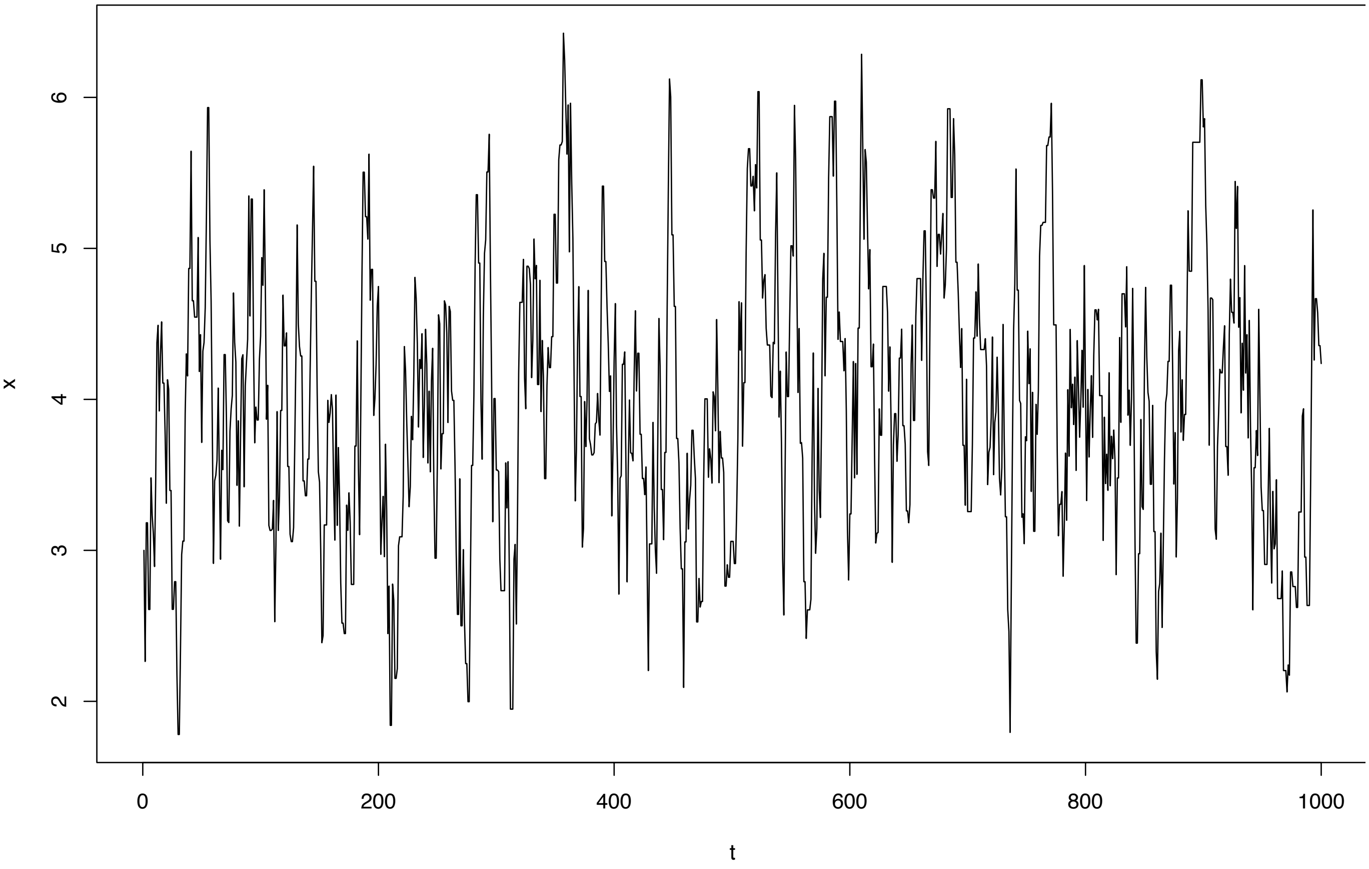
Program in ”R”:

```
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)
     accept = (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))
     alpha = min(1,accept)
     u = runif(1)
     if (u <= alpha) x[t] = y
     else x[t] = x[t-1]
    }
    plot(1:length(x),x,
          type="l",lty=1,xlab="t",ylab="x")
    x
}
```

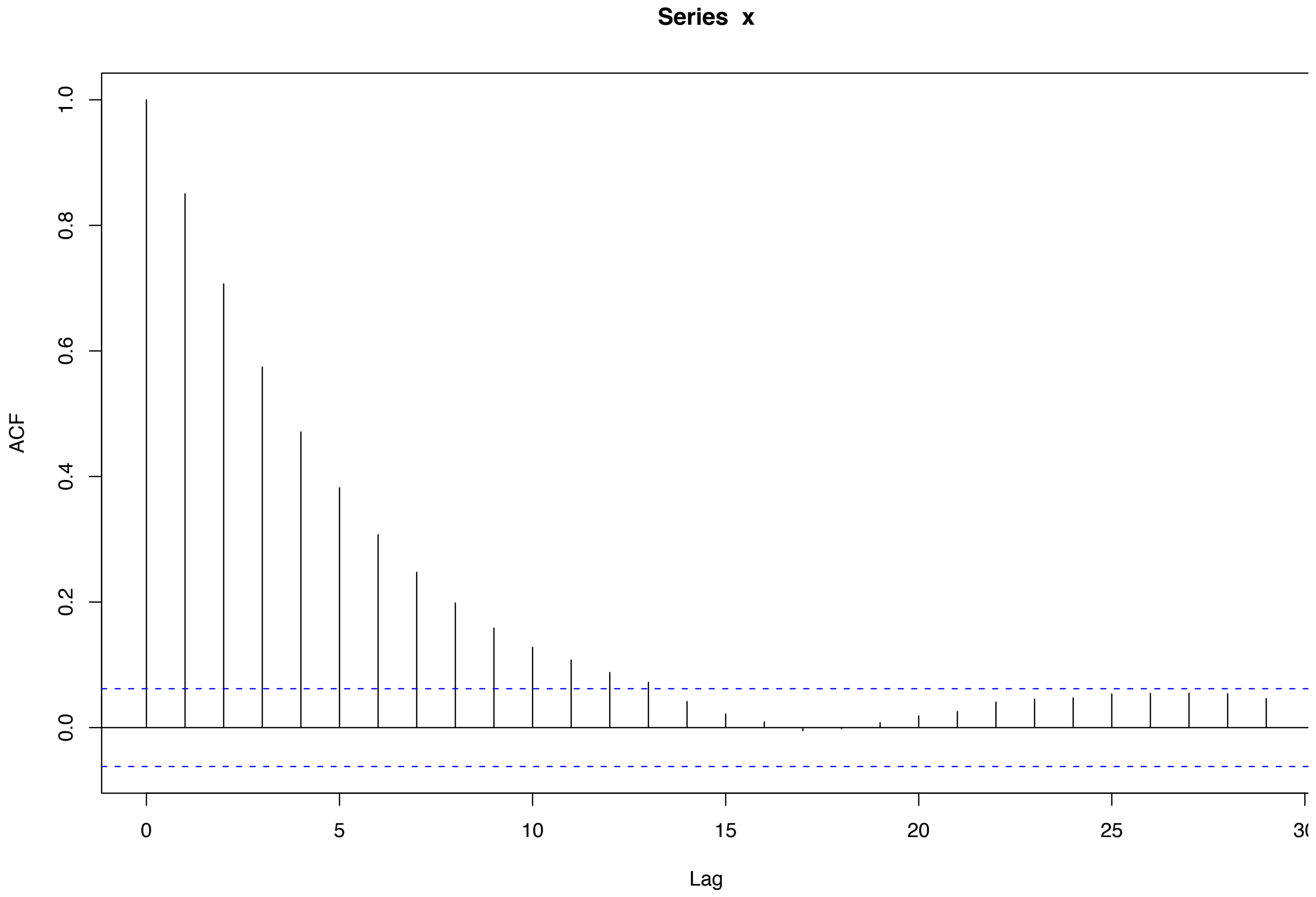
Target distribution with modes in 4 and -4 and standard-deviation =1



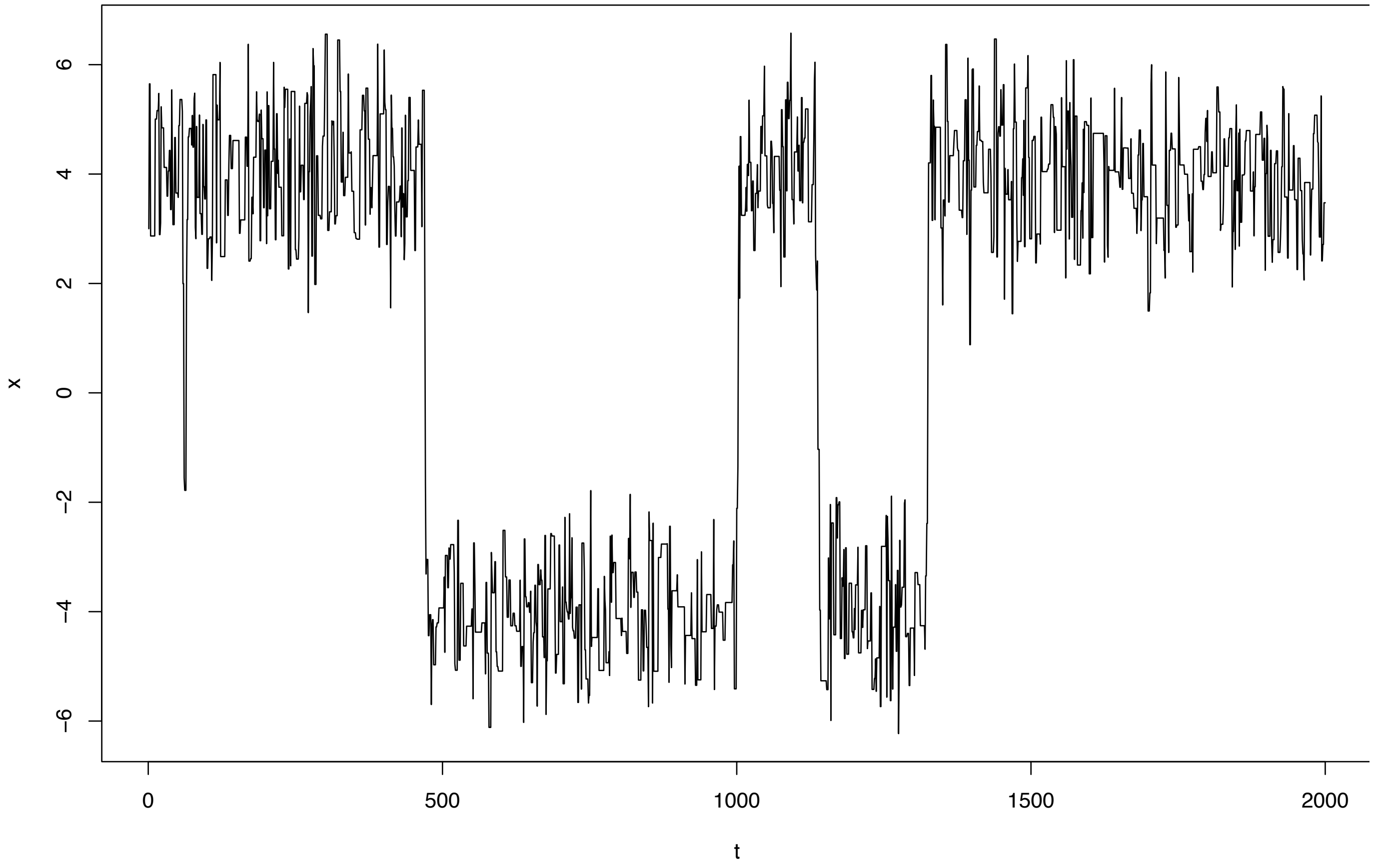
Sample path of a chain started in 3 with
$d = 1$ (proposal), n= 1000 steps:
the chain only visits the first mode
(values taken range from 2 to 6)



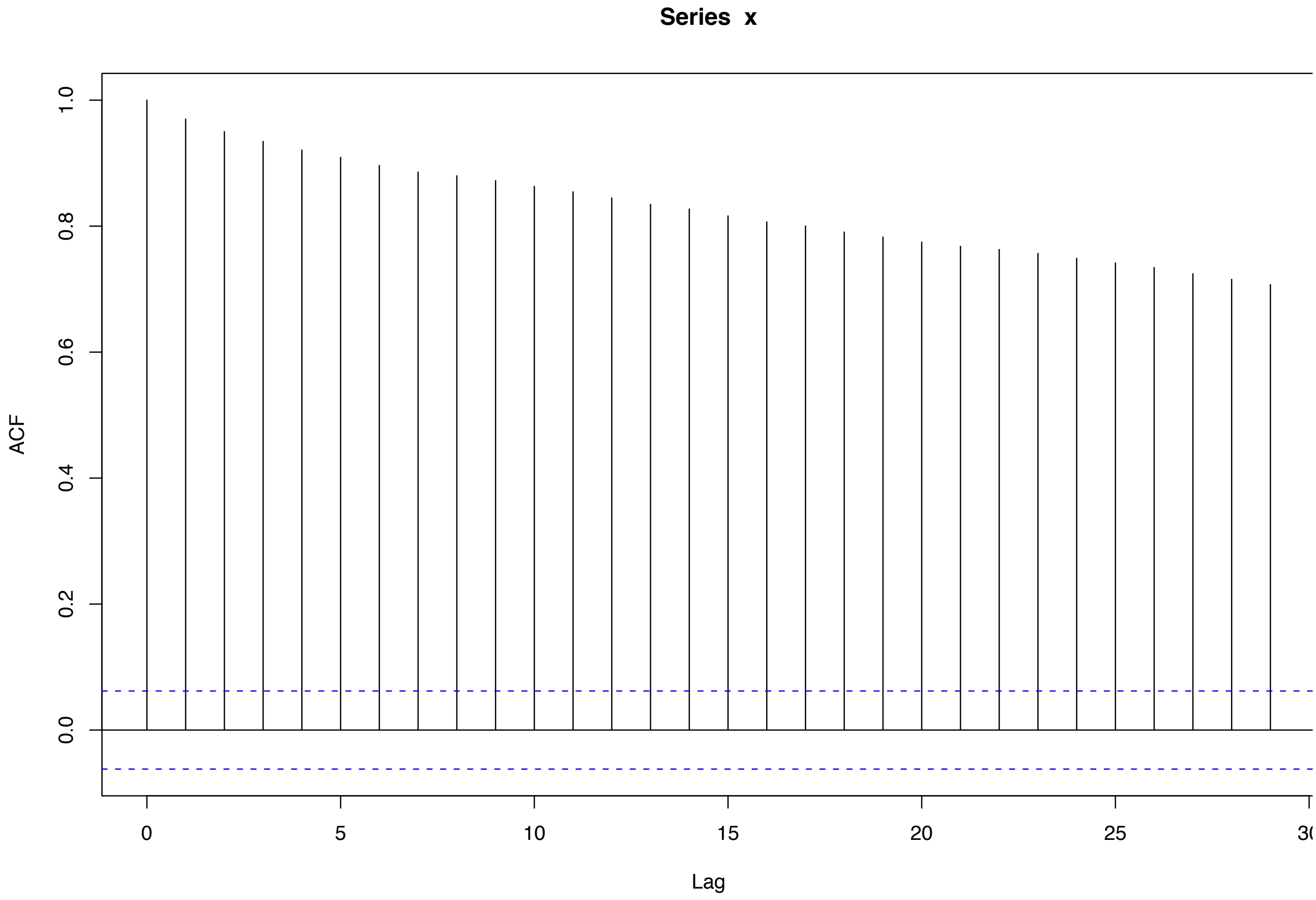
ACF of the Markov chain: the autocorrelation becomes negligible only at lag 15



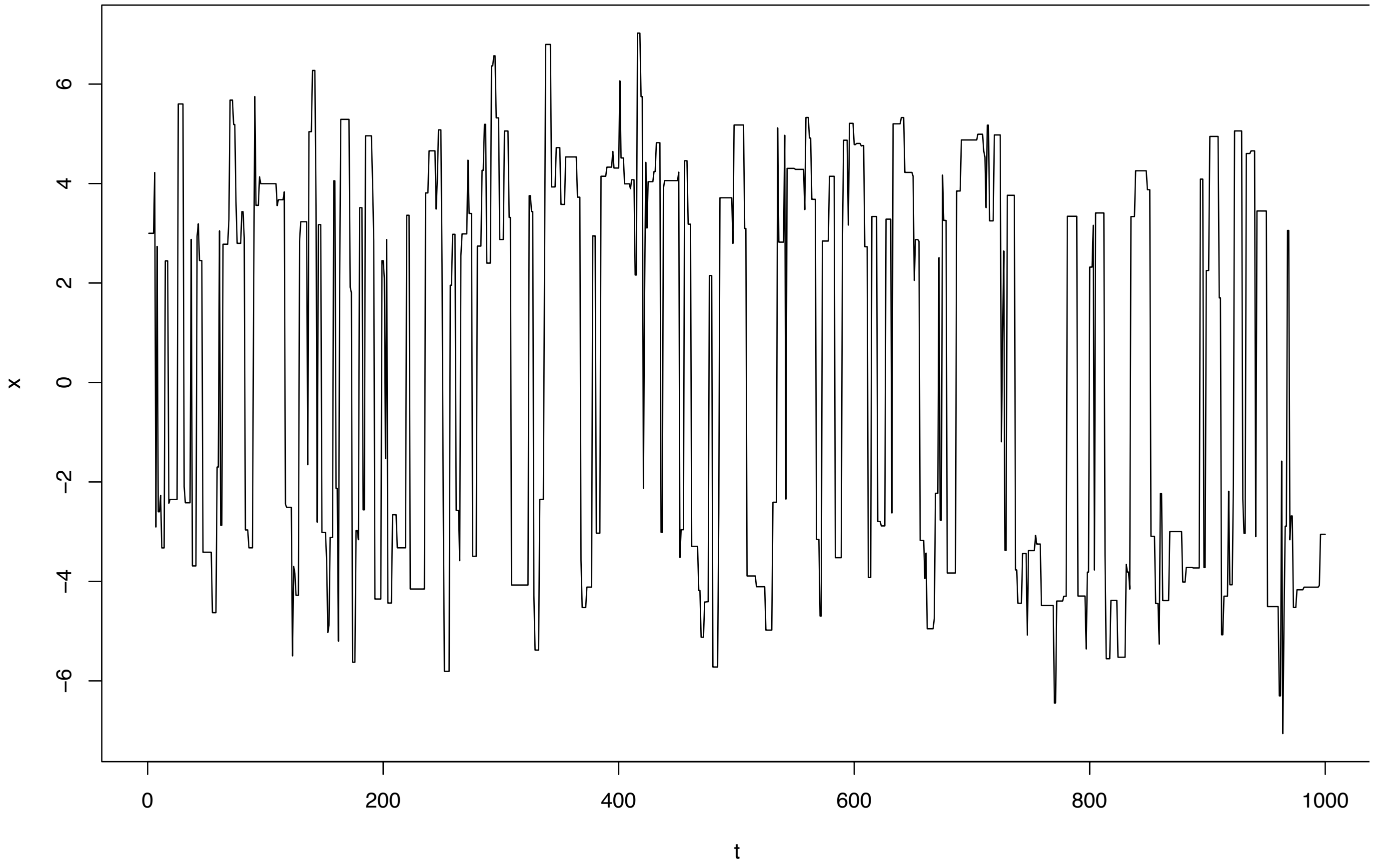
Sample path of a chain started in 3 with $d = 4$ (proposal), n=1000 steps:
the chain visits both modes
(values taken range from -6 to 6)



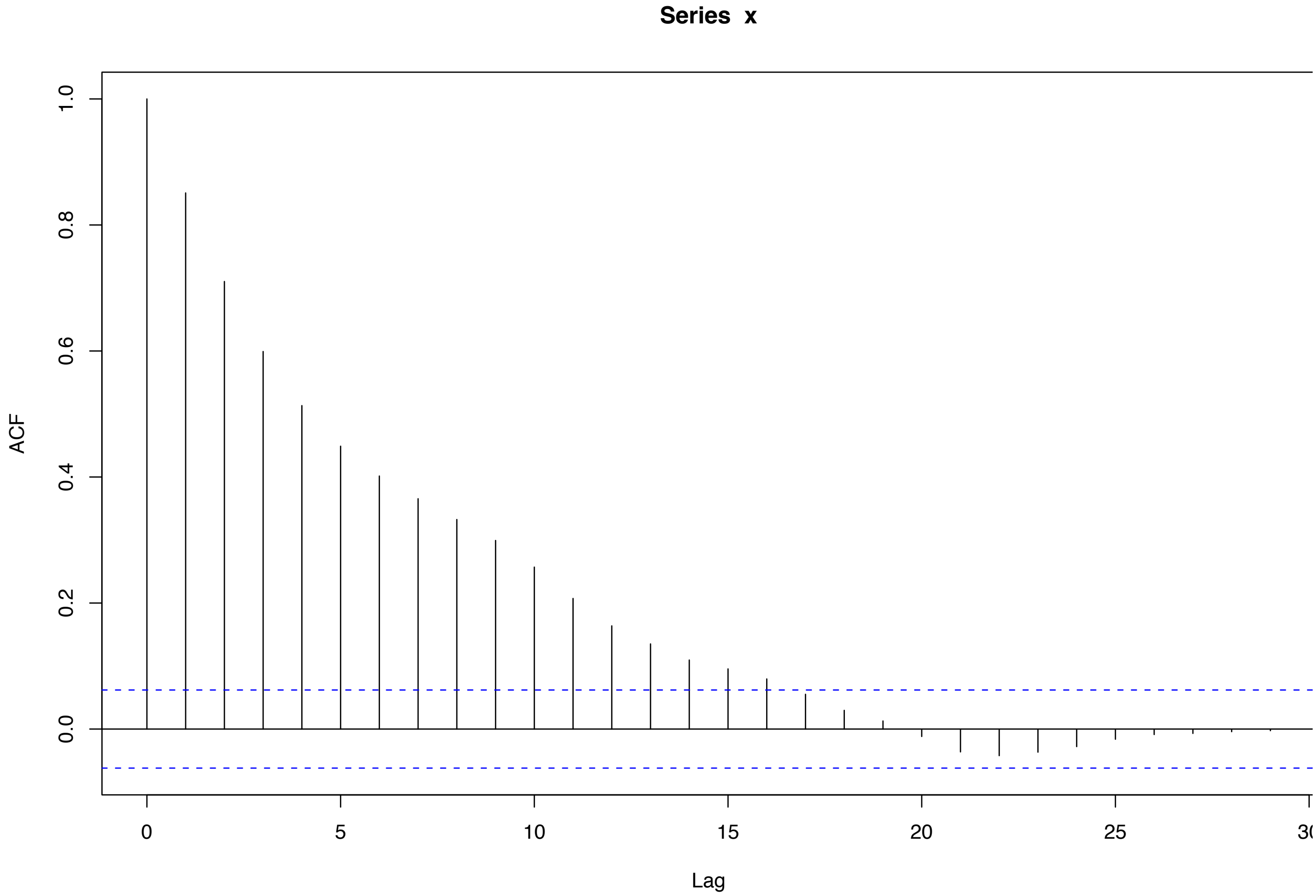
ACF of the Markov chain: the autocorrelation is still very high at lag 30



Sample path of a chain started in 3 with $d = 8$ (proposal), n= 1000 steps:
The Markov chain moves more freely between the two modes



ACF of the Markov chain: the autocorrelation becomes negligible only at lag 20
There is still space for improvement



ERROR in MCMC ESTIMATES

Parameter:

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

Estimator:

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

Variance of the estimator:

$$V(f,P) = \lim_{n\to\infty} n\,\mathrm{Var}_\pi[\widehat{\mu}_n]$$

$$= \sum_{k=-\infty}^{\infty} \mathrm{Cov}_\pi[f(X_0), f(X_k)]$$

$$= \sum_{k=-\infty}^{\infty} \gamma_k$$

Estimate the variance of MCMC estimators

- **Empirical covariances**: $\hat{\mathrm{V}} = \sum_{-\infty}^{+\infty} \widehat{\gamma}_k$
it is well known that this is not a consistent estimator.
Use a truncated version of the above over a proper window: $\hat{\mathrm{V}} = \sum_{-\mathrm{M}}^{+\mathrm{M}} \widehat{\gamma}_k$ with M being the smallest integer $\geq 3\,\widehat{\tau}$ (Sokal, 1989)

- **Blocking**: divide the simulations into $\mathbf{b}$ consecutive blocks of length $\mathbf{k}$

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

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

- **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:

$$\alpha(x,y,z) = 1 \wedge \Big\{ \; \frac{\pi(z)}{\pi(x)} \; \frac{g_1(\widetilde{u_1})}{g_1(u_1)} \; \frac{[1-\alpha(z,y^\star)]}{[1-\alpha(x,y)]} \; \frac{g_2(\widetilde{u_2})}{g_2(u_2)} \; \left|\frac{\partial(z,\widetilde{u_1},\widetilde{u_2})}{\partial(x,u_1,u_2)}\right| \; \Big\}$$

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

Bayesian Model Choice

Typical in model choice settings

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

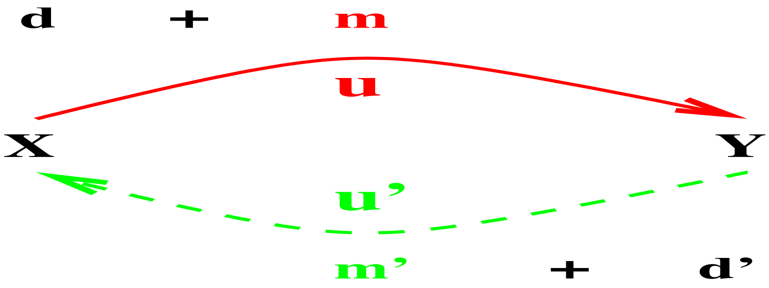
REVERSIBLE JUMP ALGORITHM

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

$d + m = d' + m'$



Example of RJ:
Han and Carlin, JASA 2001, ex. 3.1
Non nested linear regression

Data = 42 specimens of radiata pine

Y = maximum compressive strength parallel to the grain

X = the density

Z = the resin-adjusted density

Model 1: $y_1 = \alpha + \beta(x_i - \overline{x}) + \epsilon_i$
$\epsilon_i \sim N(0, \sigma^2), i = 1, \cdots n$
$\theta_1 = (\alpha, \beta, \sigma^2)$
Priors:
$(\alpha, \beta) \sim N((3000, 185), Diag(10^6, 10^4))$
$\sigma^2 \sim IG(3, (2 * 300^2)^{-1})$

Model 2: $y_1 = \gamma + \delta(z_i - \overline{z}) + \eta_i$
$\eta_i \sim N(0, \tau^2), i = 1, \cdots n$
$\theta_2 = (\gamma, \delta, \tau^2)$
Priors:
$(\gamma, \delta) \sim N((3000, 185), Diag(10^6, 10^4))$
$\tau^2 \sim IG(3, (2 * 300^2)^{-1})$

i.e. both have prior mean and standard deviation equal to 300^2
These priors are roughly centered on the corresponding least squares solutions, but they are rather vague

We assume **prior independence** among all the parameters given the corresponding model indicators

The **full conditional distributions** of the model specific parameters are also bivariate normal and inverse gamma

Prior model probabilities:
$p_1 = 0.9995$ and $p_2 = 0.0005$

use log-transformation for variances:
$\lambda = log\, \sigma^2$
$\omega = log\, \tau^2$

Model-switching probabilities = 0.5

When a **move bwn models** is proposed set $(\alpha, \beta, \lambda) = (\gamma, \delta, \omega)$
(and viceversa)

the **dimension-matching** requirement is automatically satisfied without generating an additional random vector and the Jacobian = 1

$$\alpha_{12} = 1 \wedge \frac{L(y|\gamma,\delta,\omega,M=2)p_2}{L(y|\alpha,\beta,\lambda,M=1)p_1}$$

$$\alpha_{21} = 1 \wedge \frac{L(y|\alpha,\beta,\lambda,M=1)p_1}{L(y|\gamma,\delta,\omega,M=2)p_2}$$

when a **move within model** is proposed use a MH with

$(\alpha, \beta, \lambda) \sim N(\text{current values}, Diag(500, 250, 1))$
accept w.p.

$$\alpha_{11} = 1 \wedge \frac{\text{prior Lhd(proposed)}}{\text{prior Lhd (current)}}$$

similarly for $\alpha 22$
note: symmetric proposal cancels

Alternative: use gibbs steps
(no need to log-transform)

Exercice: Try example 4.1 of Han and Carlin (JASA)
hierarchical longitudinal model

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

- $\{\lambda_{0P} \geq \lambda_{1P} \geq \ldots\}$= ordered eigenvalues

- $\{e_{0P}, e_{1P}, \ldots\}$ = corresponding eigenvectors

ASYMPTOTIC VARIANCE:

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

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

SPEED of CONVERGENCE:

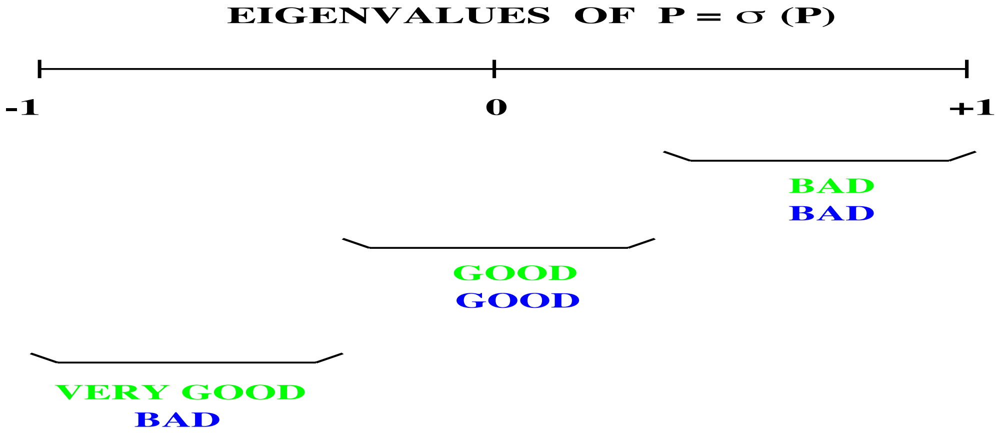
$$P^n(x,y) = \sum_j e_{jP}(x) e_{jP}(y) \, \lambda^n_{jP}$$

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

CONFLICTING BEHAVIORS

small variance in CLT $\Longleftrightarrow$ small eigenvalues

fast convergence $\Longleftrightarrow$ small |eigenvalues|



EIGENVALUES OF $P = \sigma(P)$

-1

0

+1

BAD
BAD

GOOD
GOOD

VERY GOOD
BAD

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,

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

ABSOLUTE EFFICIENCY

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

$\boxed{P \succeq_{E} Q}$, if $V(f,P) \leq V(f,Q), \qquad \forall f \in L^2(\pi)$

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

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

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

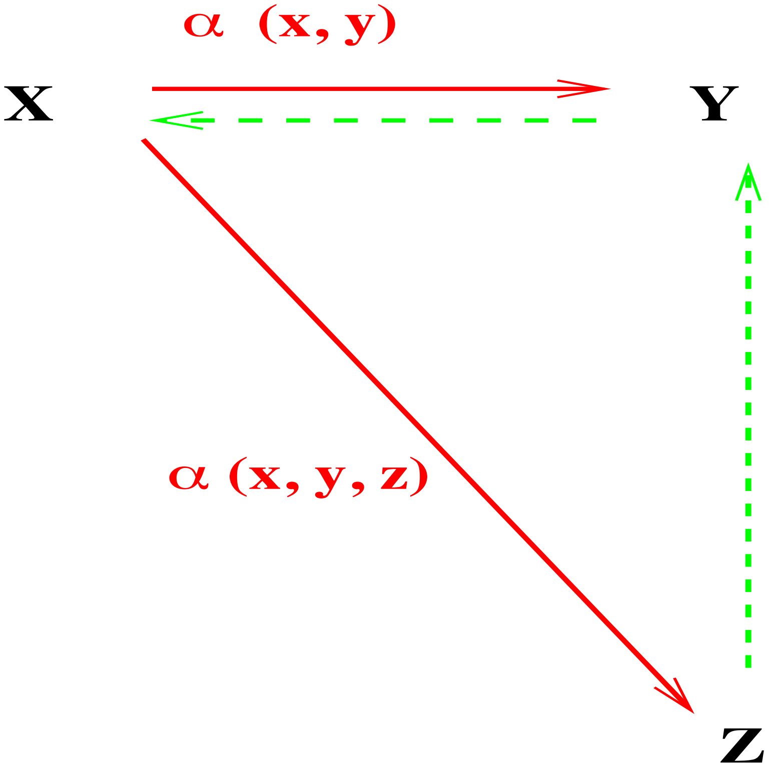
- First stage acceptance probability:

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

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

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 $\boxed{j, \;\; j = 1, \cdots, 7}$

- $y(i_j)$: binary response of company i $\boxed{i = 1, \cdots, n_j}$ in sector j. $\boxed{y = 1 \Leftrightarrow \text{default}}$

- $\underline{x}(i_j)$: 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:

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

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

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

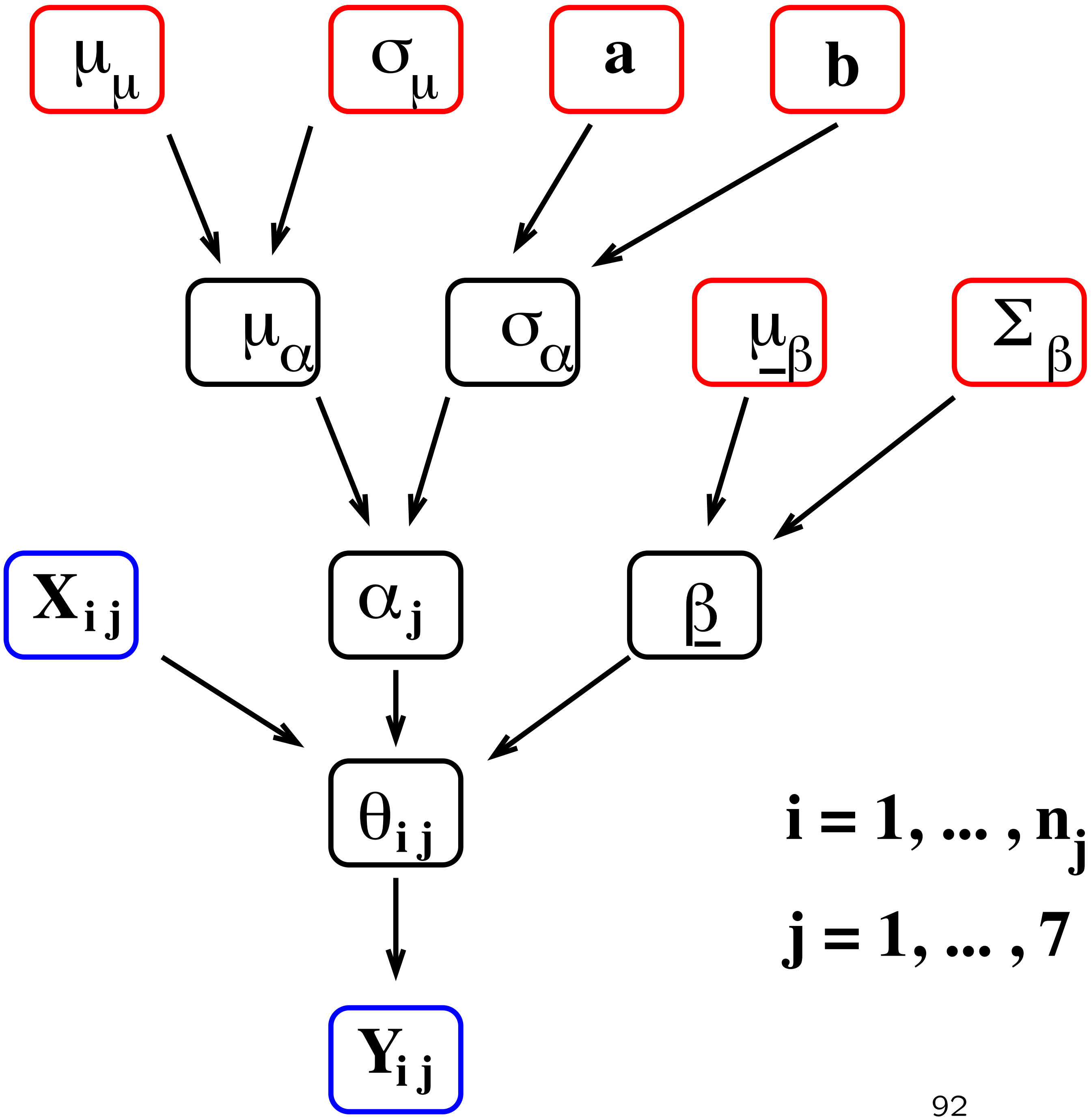
$$\underline{\beta} \sim N_4(\underline{0}, 64 \times I_4)$$

POSTERIOR:

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

where

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



$\mathbf{i = 1, \ldots, n_j}$

$\mathbf{j = 1, \ldots, 7}$

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
$\alpha_2, \cdots, \alpha_7, \mu_\alpha, b_2$	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:

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

Asymptotic variance:

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

$$\Downarrow$$

$\tau =$ integrated autocor. time

$$\Downarrow$$

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

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

$\widehat{\tau}$	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

$\widehat{\tau}$	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

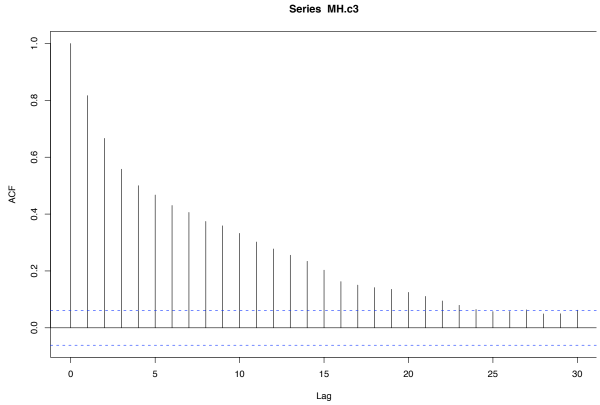
$\widehat{\tau}$	β_1	β_2	β_3	β_4	μ_α	σ_α
MH	10.0	64.5	23.4	5.6	15.9	20.2
DR	7.2	38.1	20.9	4.2	14.6	15.6

$\widehat{\tau}$	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

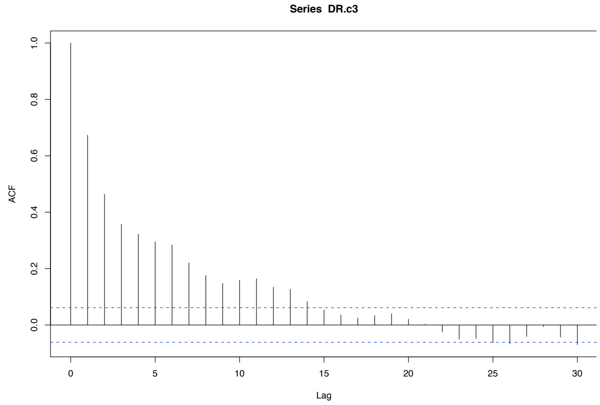
Values obtained averaging over 5 simulations of length 1024 after a burn-in of 150 steps

AUTOCORRELATION FCT for α_3

Metropolis-Hastings sampler



Delaying rejection sampler



RESULTS

Estimates of DP:

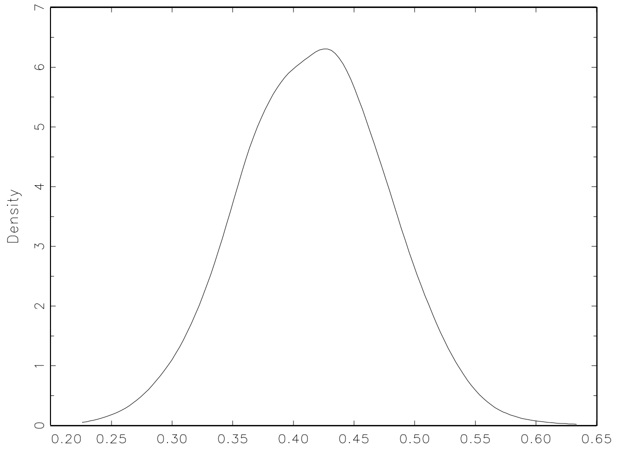
- company 30 in sector 6
- company 20 in sector 2

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

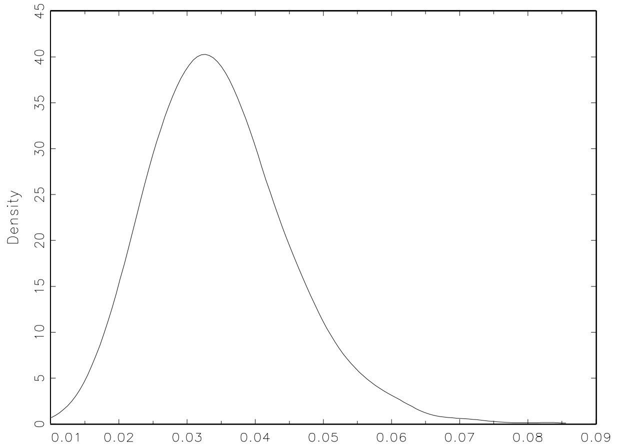
	plug in $\widehat{\alpha_j}$ and $\widehat{\underline{\beta}}$	$\frac{1}{N}\sum_{n=150}^{1174}\theta_{i,j}^{n}$	MLE
$\widehat{\theta}_{30,6}$	0.431	0.434	0.372
$\widehat{\theta}_{20,2}$	0.032	0.034	0.026

Posterior kernel density estimate of DP

- company 30 in sector 6



- company 20 in sector 2



Comparison of Bayesian vs MLE in terms of prediction

Cross Validation Analysis:
70 % of the obs used to estimate the model
30 % of the obs used to validate the model
(test and training samples are "balanced": same proportion of default in different sectors)

Root mean squared error of classification:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\theta}_i)^2}$$

where $y_i = 0, 1$ and $\widehat{\theta}_i$ = estimated def. prob.

	MLE	Bayesian
all	0.1282	0.1003
not defaulted	0.0280	0.0137
defaulted	0.8646	0.6531

CONCLUSIONS

- 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

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

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

RESEARCH CONTRIBUTIONS
(Theory)

Ordering MCMC:

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

Ways of improving MCMC algorithms:

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

Financial/Economical applications

- Estimate of default probabilities
  Mira and Tenconi, Stoch. Analysis,
  Random Fields and Appl. IV, 2004

- Latent class models for credit-scoring
  Scaccia, Mira, Bartolucci, ISI Proc., 2003

- Detection of structural change points
  Mira and Green, Biometrika, 2001

- Stability of factor models of interest rates
  Audrino, Barone-Adesi, Mira, J.Fin.Ec.2005

Available software for MCMC:

- R routines to run MCMC and to detect convergence
  - BOA
  - CODA
  - MCMCpack
  - mcmc
  - MCMCglmm
  - mcclust
  - AMCMC
- WinBugs

Conclusions:

MCMC methods are general methods for simulation from a complex probability distribution, may be high dimensional, highly non-Gaussian and multimodal

Given a set of samples from the target we can compute Monte Carlo expectations for any quantities of interest by ergodic averages

Rates of convergence to stationarity are hard to compute - lots of theory, but not typically applicable in practice

However, many models can be proven to have geometric convergence rates

antonietta.mira@usi.ch