Monte Carlo Methods

Christian P. Robert
Université Paris Dauphine

1 Introduction

Even simple models may lead to computational complications, as in latent variable models:

Example 1 –Mixture models–

Models of *mixtures of distributions*:

$$X \sim f_j \text{ with probability } p_j,$$

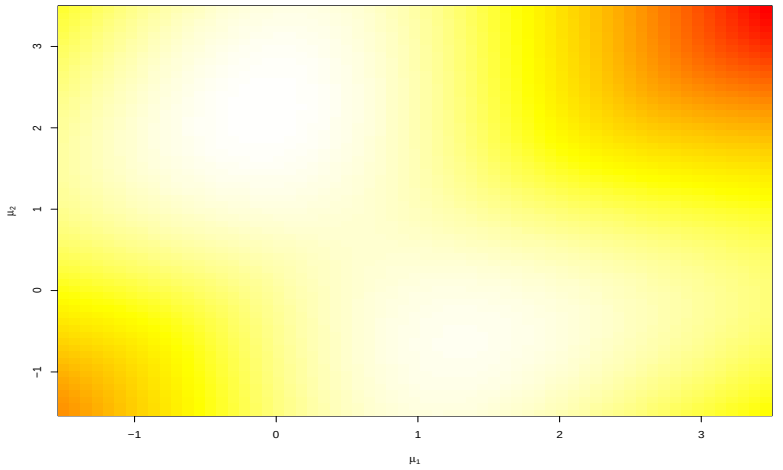
for $j = 1, 2, \ldots, k$, with overall density

$$X \sim p_1 f_1(x) + \cdots + p_k f_k(x) \ .$$

For a sample of independent random variables $(X_1, \cdots, X_n)$, sample density

$$\prod_{i=1}^{n} \{p_1 f_1(x_i) + \cdots + p_k f_k(x_i)\} \ .$$

Expanding this product involves k^n elementary terms: prohibitive to compute in large samples.



1.1 Likelihood Methods

Maximum Likelihood Methods

- For an iid sample $X_1, \ldots, X_n$ from a population with density $f(x|\theta_1, \ldots, \theta_k)$, the *likelihood function* is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) \\ &= \prod_{i=1}^n f(x_i|\theta_1, \ldots, \theta_k). \end{aligned}$$

- Global justifications from asymptotics

Example 2 –Mixtures again–

For a mixture of two normal distributions,

$$p\mathcal{N}(\mu, \tau^2) + (1-p)\mathcal{N}(\theta, \sigma^2)\ ,$$

likelihood proportional to

$$\prod_{i=1}^n \left[p\tau^{-1}\varphi\left(\frac{x_i-\mu}{\tau}\right) + (1-p)\ \sigma^{-1}\ \varphi\left(\frac{x_i-\theta}{\sigma}\right)\right]$$

containing 2^n terms.

Standard maximization techniques often fail to find the global maximum because of multimodality of the likelihood function.

In the special case

$$f(x|\mu, \sigma) = (1-\epsilon)\exp\{(-1/2)x^2\} + \frac{\epsilon}{\sigma}\exp\{(-1/2\sigma^2)(x-\mu)^2\} \quad (1)$$

with $\epsilon > 0$ known

Then, whatever n, the likelihood is unbounded:

$$\lim_{\sigma\to 0} \ell(\mu = x_1, \sigma|x_1, \ldots, x_n) = \infty$$

Missing variable models

$$g(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta)\, dz$$

Completed likelihood

$$L^c(\theta|\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\theta)$$

–EM Algorithm–

Iterate (in m)

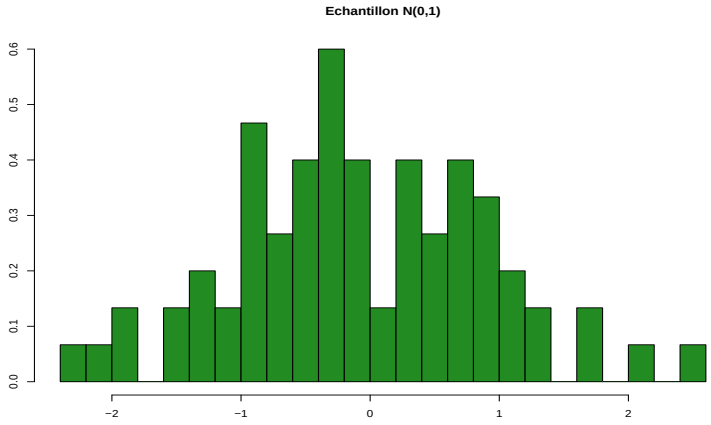
1. (*E step*) Compute

$$Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}) = \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\hat{\theta}_{(m)}, \mathbf{x}]\,,$$

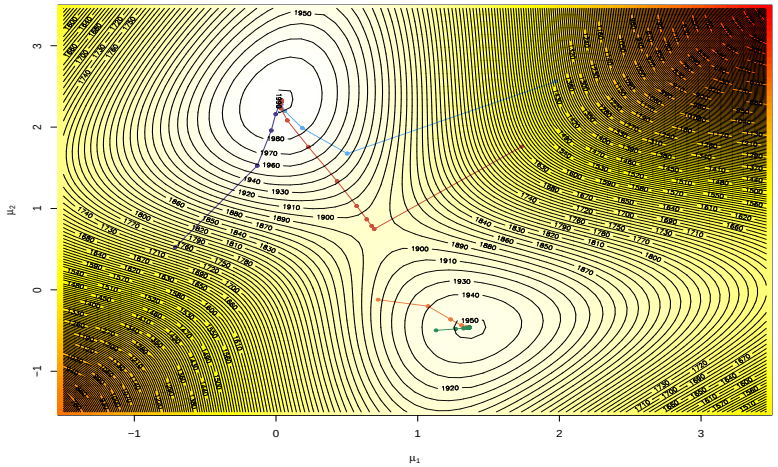
2. (*M step*) Maximise $Q(\theta|\hat{\theta}_{(m)}, \mathbf{x})$ in θ and take

$$\hat{\theta}_{(m+1)} = \arg\max_{\theta}\ Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}).$$

until a fixed point [of Q] is reached



Sample from (1)



Likelihood of $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and EM steps

1.2 Bayesian Methods

In the Bayesian paradigm, information brought by the data x, realization of

$$X \sim f(x|\theta),$$

combined with prior information specified by *prior distribution* with density $\pi(\theta)$

Summary in a probability distribution, $\pi(\theta|x)$, called the **posterior distribution**

Derived from the *joint* distribution $f(x|\theta)\pi(\theta)$, according to

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta},$$

[Bayes Theorem]

where

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta$$

is the *marginal density* of X

Posterior distribution central to Bayesian inference

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information and information brought by x
- Avoids averaging over the unobserved values of x
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected
- Provides a complete inferential scope and an unique motor of inference

Example 3 –Binomial–

For an observation X from the binomial distribution $\mathcal{B}(n,p)$ the (so-called) conjugate prior is the family of beta distributions $\mathcal{B}e(a,b)$

The classical Bayes estimator δ^π is the posterior mean

$$\begin{aligned}\delta^\pi &= \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(n-x+b)} \\ &\quad \times \int_0^1 p\, p^{x+a-1}(1-p)^{n-x+b-1}dp \\ &= \frac{x+a}{a+b+n}.\end{aligned}$$

Example 4 –Normal–

In the normal $\mathcal{N}(\mu,\sigma^2)$ case, when both μ and σ are unknown, there still is a conjugate prior on $\theta = (\mu,\sigma^2)$, of the form

$$(\sigma^2)^{-\lambda_\sigma}\ \exp -\left\{\lambda_\mu(\mu-\xi)^2+\alpha\right\}/\sigma^2$$

since

$$\begin{aligned}\pi((\mu,\sigma^2)|x_1,\ldots,x_n) &\propto (\sigma^2)^{-\lambda_\sigma}\ \exp -\left\{\lambda_\mu(\mu-\xi)^2+\alpha\right\}/\sigma^2 \\ &\quad \times(\sigma^2)^{-n}\ \exp -\left\{n(\mu-\overline{x})^2+s_x^2\right\}/\sigma^2 \\ &\propto (\sigma^2)^{-\lambda_\sigma+n}\ \exp -\Bigg\{(\lambda_\mu+n)(\mu-\xi_x)^2 \\ &\quad +\alpha+s_x^2+\frac{n\lambda_\mu}{n+\lambda_\mu}\Bigg\}/\sigma^2\end{aligned}$$

The curse of conjugate priors

The use of **conjugate priors** for computational reasons

- implies a restriction on the modeling of the available prior information
- may be detrimental to the usefulness of the Bayesian approach
- gives an impression of subjective manipulation of the prior information disconnected from reality.

A typology of Bayes computational problems

(i) use of a complex parameter space, as for instance in constrained parameter sets like those resulting from imposing stationarity constraints in dynamic models;

(ii) use of a complex sampling model with an intractable likelihood, as for instance in missing data and graphical models;

(iii) use of a huge dataset;

(iv) use of a complex prior distribution (which may be the posterior distribution associated with an earlier sample);

(v) use of a complex inferential procedure as for instance, Bayes factors

$$B_{01}^{\pi}(x) = \frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}.$$

Example 5 —Mixture of two normal distributions–

$$x_1, \ldots, x_n \sim f(x|\theta) = p\varphi(x; \mu_1, \sigma_1) + (1-p)\varphi(x; \mu_2, \sigma_2)$$

Prior

$$\mu_i|\sigma_i \sim \mathcal{N}(\xi_i, \sigma_i^2/n_i), \qquad \sigma_i^2 \sim \mathcal{IG}(\nu_i/2, s_i^2/2), \qquad p \sim \mathcal{B}e(\alpha, \beta)$$

Posterior

$$\begin{aligned}\pi(\theta|x_1, \ldots, x_n) &\propto \prod_{j=1}^{n} \{p\varphi(x_j; \mu_1, \sigma_1) + (1-p)\varphi(x_j; \mu_2, \sigma_2)\}\, \pi(\theta) \\ &= \sum_{\ell=0}^{n} \sum_{(k_t)} \omega(k_t)\pi(\theta|(k_t))\end{aligned}$$

[O(2^n)]

For a given permutation (k_t), conditional posterior distribution

$$\begin{aligned}\pi(\theta|(k_t)) = \mathcal{N}&\left(\xi_1(k_t), \frac{\sigma_1^2}{n_1+\ell}\right) \times \mathcal{IG}((\nu_1+\ell)/2, s_1(k_t)/2) \\ \times\mathcal{N}&\left(\xi_2(k_t), \frac{\sigma_2^2}{n_2+n-\ell}\right) \times \mathcal{IG}((\nu_2+n-\ell)/2, s_2(k_t)/2) \\ \times\mathcal{B}&e(\alpha+\ell, \beta+n-\ell)\end{aligned}$$

where

$$\begin{aligned} \bar{x}_1(k_t) &= \tfrac{1}{\ell}\sum_{t=1}^{\ell} x_{k_t}, & \hat{s}_1(k_t) &= \sum_{t=1}^{\ell}(x_{k_t}-\bar{x}_1(k_t))^2, \\ \bar{x}_2(k_t) &= \tfrac{1}{n-\ell}\sum_{t=\ell+1}^{n} x_{k_t}, & \hat{s}_2(k_t) &= \sum_{t=\ell+1}^{n}(x_{k_t}-\bar{x}_2(k_t))^2 \end{aligned}$$

and

$$\begin{aligned} \xi_1(k_t) &= \frac{n_1\xi_1+\ell\bar{x}_1(k_t)}{n_1+\ell}, \qquad \xi_2(k_t) = \frac{n_2\xi_2+(n-\ell)\bar{x}_2(k_t)}{n_2+n-\ell}, \\ s_1(k_t) &= s_1^2+\hat{s}_1^2(k_t)+\frac{n_1\ell}{n_1+\ell}(\xi_1-\bar{x}_1(k_t))^2, \\ s_2(k_t) &= s_2^2+\hat{s}_2^2(k_t)+\frac{n_2(n-\ell)}{n_2+n-\ell}(\xi_2-\bar{x}_2(k_t))^2, \end{aligned}$$

posterior updates of the hyperparameters

Bayes estimator of θ:

$$\delta^\pi(x_1,\ldots,x_n) = \sum_{\ell=0}^{n}\sum_{(k_t)}\omega(k_t)\mathbb{E}^\pi[\theta|\mathbf{x},(k_t)]$$

Too costly: 2^n terms

Example 6 –Poly-t prior–

Normal observation $x\sim\mathcal{N}(\theta,1)$, with conjugate prior

$$\theta\sim\mathcal{N}(\mu,\epsilon)$$

Closed form expression for the posterior mean

$$\int_\Theta \theta\, f(x|\theta)\,\pi(\theta)\,d\theta \;\Big/\; \int_\Theta f(x|\theta)\,\pi(\theta)\,d\theta = \frac{x+\epsilon^{-2}\mu}{1+\epsilon^{-2}}\,.$$

More involved prior distribution:

poly-t distribution

[Bauwens,1985]

$$\pi(\theta) = \prod_{i=1}^{k}\left[\alpha_i+(\theta-\beta_i)^2\right]^{-\nu_i} \qquad \alpha_i,\nu_i>0$$

Computation of $\mathbb{E}[\theta|x]$???

Example 7 **–$AR(p)$ model–**

Auto-regressive representation of a time series,

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + \sigma\varepsilon_t$$

If order p unknown, predictive distribution of x_{t+1} given by

$$\pi(x_{t+1}|x_t,\ldots,x_1) \propto \int f(x_{t+1}|x_t,\ldots,x_{t-p+1})\pi(\theta,p|x_t,\ldots,x_1)dp\,d\theta\,,$$

Integration over the parameters of all models

$$\sum_{p=0}^{\infty} \int f(x_{t+1}|x_t,\ldots,x_{t-p+1})\pi(\theta|p,x_t,\ldots,x_1)\,d\theta\,\pi(p|x_t,\ldots,x_1)\,.$$

Multiple layers of complexity

(i) Complex parameter space within each $AR(p)$ model because of stationarity constraint

(ii) if p unbounded, infinity of models

(iii) θ varies between models $AR(p)$ and $AR(p+1)$, with a different stationarity constraint (except for root reparameterisation).

(iv) if prediction used sequentially, every tick/second/hour/day, posterior distribution $\pi(\theta,p|x_t,\ldots,x_1)$ must be re-evaluated

2 Random Variable Generation

- Rely on the possibility of producing (computer-wise) an endless flow of random variables (usually iid) from well-known distributions
- Given a uniform random number generator, illustration of methods that produce random variables from both standard and nonstandard distributions

2.1 Basic Methods

2.1.1 Introduction

For a function F on $\mathbb{R}$, *generalized inverse* of F, F^-, defined by

$$F^-(u) = \inf \{x;\ F(x) \geq u\} .$$

Probability Integral Transform:

If $U \sim \mathcal{U}_{[0,1]}$, then the random variable $F^-(U)$ has the distribution F.

Consequence:

To generate a random variable $X \sim F$, suffices to generate

$$U \sim \mathcal{U}_{[0,1]}$$

and then make the transform

$$x = F^-(u)$$

2.1.2 Desiderata and Limitations

- Production of a *deterministic* sequence of values in $[0, 1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$.
- Can't use the physical imitation of a "random draw" `[no guarantee of uniformity, no reproducibility]`
- *Random* sequence in the sense: Having generated $(X_1, \cdots, X_n)$, knowledge of X_n [or of $(X_1, \cdots, X_n)$] imparts no discernible knowledge of the value of X_{n+1}.

- Deterministic: Given the initial value X_0, sample $(X_1, \cdots, X_n)$ always the same
- Validity of a random number generator based on a single sample $X_1, \cdots, X_n$ when n tends to $+\infty$, **not** on replications

$$(X_{11}, \cdots, X_{1n}), (X_{21}, \cdots, X_{2n}), \ldots (X_{k1}, \cdots, X_{kn})$$

where n fixed and k tends to infinity.

2.1.3 Uniform pseudo-random number generator

Algorithm starting from an initial value u_0 and a transformation D, which produces a sequence

$$(u_i) = (D^i(u_0))$$

in $[0, 1]$.

For all n,

$$(u_1, \cdots, u_n)$$

reproduces the behavior of an *iid* $\mathcal{U}_{[0,1]}$ sample $(V_1, \cdots, V_n)$ when compared through usual tests

- Validity of the algorithm means that the sequence $U_1, \cdots, U_n$ leads to accept the hypothesis

$$\mathrm{H} : U_1, \cdots, U_n \quad \text{are iid} \quad \mathcal{U}_{[0,1]}.$$

- The set of tests used is generally of some consequence
  - Kolmogorov–Smirnov
  - Time series methods, for correlation between U_i and $(U_{i-1}, \cdots, U_{i-k})$
  - nonparametric tests
  - Marsaglia's battery of tests called *Die Hard* (!)

Usual generators

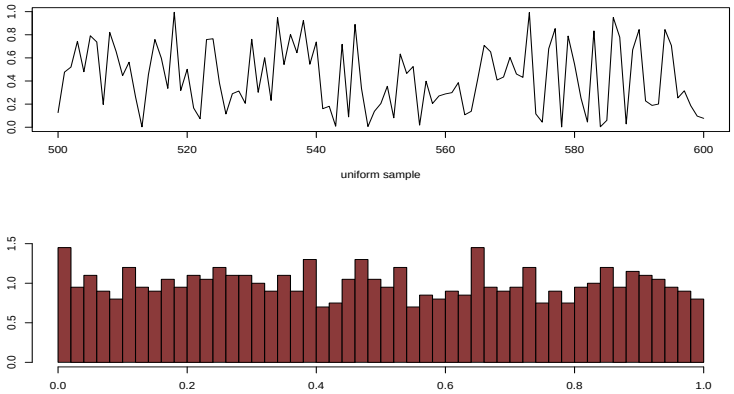
In R and S-plus, procedure `runif()`

```
The Uniform Distribution

Description:
'runif' generates random deviates.

Example:
     u <- runif(20)

'.Random.seed' is an integer vector, containing the
random number generator (RNG) state for random number
generation in R. It can be saved and restored, but
should not be altered by the user.
```



In C, procedure `rand()` or `random()`

```
SYNOPSIS
       #include <stdlib.h>
       long int random(void);
DESCRIPTION
The random() function uses a non-linear additive
feedback random number generator employing a
default table of size 31 long integers to return
successive pseudo-random numbers in the range
from 0 to RAND_MAX. The period of this random
generator is very large, approximately
16*((2**31)-1).
RETURN VALUE
random() returns a value between 0 and RAND_MAX.
```


In Scilab, procedure `rand()`

```
rand() : with no arguments gives a scalar whose
value changes each time it is referenced. By
default, random numbers are uniformly distributed
in the interval (0,1). rand('normal') switches to
a normal distribution with mean 0 and variance 1.

EXAMPLE
  x=rand(10,10,'uniform')
```


2.2 Beyond Uniform Distributions

- Generation of any sequence of random variables can be formally implemented through a uniform generator
  - For distributions with explicit forms of F^- (for instance, exponential, double-exponential or Weibull distributions), the Probability Integral Transform can be implemented.
  - Case specific methods, which rely on properties of the distribution (for instance, normal distribution, Poisson distribution)

- More general (indirect) methods exist, for example the accept-reject and the ratio-of-uniform methods

- Simulation of the standard distributions is accomplished quite efficiently by many statistical programming packages (for instance, `IMSL`, `Gauss`, `Mathematica`, `Matlab/Scilab`, `Splus/R`).

2.2.1 Transformation Methods

Case where a distribution F is linked in a simple way to another distribution easy to simulate.

Example 8 **–Exponential variables–** If $U \sim \mathcal{U}_{[0,1]}$, the random variable

$$X = -\log U/\lambda$$

has distribution

$$\begin{aligned} P(X \leq x) &= P(-\log U \leq \lambda x) \\ &= P(U \geq e^{-\lambda x}) = 1 - e^{-\lambda x}, \end{aligned}$$

the exponential distribution $\mathcal{E}xp(\lambda)$.

Other random variables that can be generated starting from an exponential include

$$Y = -2\sum_{j=1}^{\nu} \log(U_j) \sim \chi^2_{2\nu}$$

$$Y = -\frac{1}{\beta}\sum_{j=1}^{a} \log(U_j) \sim \mathcal{G}a(a, \beta)$$

$$Y = \frac{\sum_{j=1}^{a} \log(U_j)}{\sum_{j=1}^{a+b} \log(U_j)} \sim \mathcal{B}e(a, b)$$

Points to note

- Transformation quite simple to use
- There are more efficient algorithms for gamma and beta random variables
- Cannot generate gamma random variables with a non-integer shape parameter
- For instance, cannot get a χ^2_1 variable, which would get us a $\mathcal{N}(0, 1)$ variable.

Example 9 **–Normal variables–** If r, θ polar coordinates of (X_1, X_2), then,

$$r^2 = X_1^2 + X_2^2 \sim \chi_2^2 = \mathcal{E}xp(1/2)$$

and

$$\theta \sim \text{ uniform distribution on } [0, 2\pi]$$

Consequence: If U_1, U_2 iid $\mathcal{U}_{[0,1]}$,

$$\begin{aligned} X_1 &= \sqrt{-2\log(U_1)}\ \cos(2\pi U_2) \\ X_2 &= \sqrt{-2\log(U_1)}\ \sin(2\pi U_2) \end{aligned}$$

iid $\mathcal{N}(0, 1)$.

–Box-Muller Algorithm–

1 Generate U_1, U_2 iid $\mathcal{U}_{[0,1]}$;

2 Define

$$\begin{cases} x_1 = \sqrt{-2\log(u_1)}\cos(2\pi u_2)\ , \\ x_2 = \sqrt{-2\log(u_1)}\sin(2\pi u_2)\ ; \end{cases}$$

3 Take x_1 and x_2 as two independent draws from $\mathcal{N}(0, 1)$.

- Unlike algorithms based on the CLT, this algorithm is exact
- Get two normals for the price of two uniforms
- Drawback (in speed) in calculating log, cos and sin.

Example 10 **–Poisson generation–**

Poisson–exponential connection:

If $N \sim \mathcal{P}(\lambda)$ and $X_i \sim \mathcal{E}xp(\lambda)$, $i \in \mathbb{N}^*$,

$$P_\lambda(N = k) = P_\lambda(X_1 + \cdots + X_k \leq 1 < X_1 + \cdots + X_{k+1})\ .$$

- A Poisson can be simulated by generating E$xp(1)$ till their sum exceeds 1.
- This method is simple, but is really practical only for smaller values of λ.
- On average, the number of exponential variables required is λ.
- Other approaches are more suitable for large λ's.

–Atkinson's Poisson Simulation–

To generate $N \sim \mathcal{P}(\lambda)$:

0. Define

$$\beta = \pi/\sqrt{3\lambda}, \quad \alpha = \lambda\beta \quad \text{and} \quad k = \log c - \lambda - \log\beta;$$

1. Generate $U_1 \sim \mathcal{U}_{[0,1]}$ and calculate

$$x = \{\alpha - \log\{(1-u_1)/u_1\}\}/\beta$$

until $X > -0.5$;

2. Define $N = \lfloor X + 0.5 \rfloor$ and generate $U_2 \sim \mathcal{U}_{[0,1]}$;

3. Accept N if

$$\alpha - \beta x + \log\left(u_2/\{1+\exp(\alpha-\beta x)\}^2\right) \leq k + N\log\lambda - \log N!\,.$$

- A generator of Poisson random variables can produce negative binomial random variables since,

$$Y \sim \mathcal{G}a(n, (1-p)/p) \quad X|y \sim \mathcal{P}(y)$$

implies

$$X \sim \mathcal{N}eg(n, p)$$

Mixture representation

- The representation of the negative binomial is a particular case of a *mixture distribution*
- The principle of a mixture representation is to represent a density f as the marginal of another distribution, for example

$$f(x) = \sum_{i \in \mathcal{Y}} p_i \; f_i(x) \;,$$

- If the component distributions $f_i(x)$ can be easily generated, X can be obtained by first choosing f_i with probability p_i and then generating an observation from f_i.

Partitioned sampling

Special case of mixture sampling when

$$f_i(x) = f(x)\,\mathbb{I}_{A_i}(x) \Big/ \int_{A_i} f(x)\,dx$$

and

$$p_i = \mathsf{Pr}(X \in A_i)$$

for a partition $(A_i)_i$

2.2.2 Accept-Reject Methods

- Many distributions from which difficult, or even impossible, to **directly** simulate.
- Another class of methods that only require us to know the functional form of the density f of interest **only** up to a multiplicative constant.
- The key to this method is to use a simpler (simulation-wise) density g, the *instrumental density*, from which the simulation from the *target density* f is actually done.

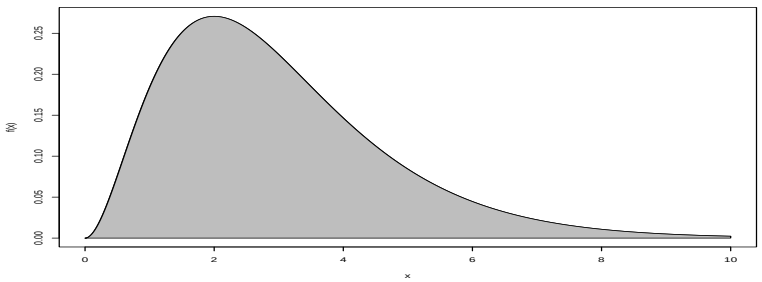
Fundamental theorem of simulation

Simulating

$$X \sim f(x)$$

equivalent to simulating

$$(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f(x)\}$$



Accept-Reject method

Given a density of interest f, find a density g and a constant M such that

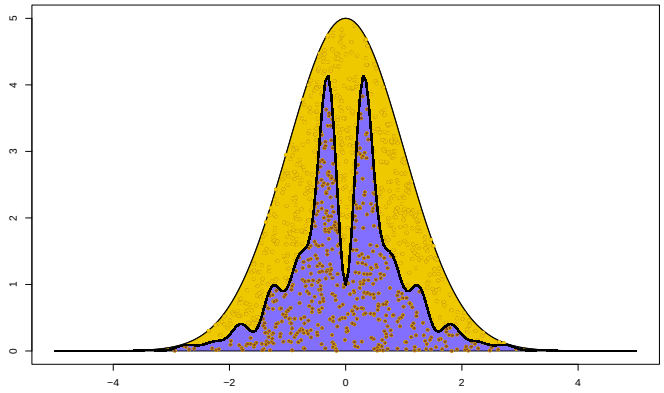
$$f(x) \leq Mg(x)$$

on the support of f.

1. Generate $X \sim g$, $U \sim \mathcal{U}_{[0,1]}$;
2. Accept $Y = X$ if $U \leq f(X)/Mg(X)$;
3. Return to 1. otherwise.

Validation of the Accept-Reject method

This algorithm produces a variable Y distributed according to f



Uniform repartition under the graph of f of accepted points

Two interesting properties:

- First, it provides a generic method to simulate from any density f that is known *up to a multiplicative factor*

  Property particularly important in Bayesian calculations: there, the posterior distribution

  $$\pi(\theta|x) \propto \pi(\theta)\ f(x|\theta)\ .$$

  is specified up to a normalizing constant

- Second, the probability of acceptance in the algorithm is $1/M$, e.g., expected number of trials until a variable is accepted is M

Some intuition

- In cases f and g both probability densities, the constant M is necessarily larger that 1.
- The size of M, and thus the efficiency of the algorithm, functions of how closely g can imitate f, especially in the tails
- For f/g to remain bounded, necessary for g to have tails thicker than those of f.

  It is therefore impossible to use the A-R algorithm to simulate a Cauchy distribution f using a normal distribution g, however the reverse works quite well.

Example 11 –Normal from a Cauchy–

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

and

$$g(x) = \frac{1}{\pi} \frac{1}{1+x^2},$$

densities of the normal and Cauchy distributions.

$$\frac{f(x)}{g(x)} = \sqrt{\frac{\pi}{2}}(1+x^2)\, e^{-x^2/2} \leq \sqrt{\frac{2\pi}{e}} = 1.52$$

attained at $x = \pm 1$.

So probability of acceptance

$$1/1.52 = 0.66,$$

and, on the average, one out of every three simulated Cauchy variables is rejected.

Mean number of trials to success 1.52.

Example 12 –Normal/Double Exponential–

Generate a $\mathcal{N}(0,1)$ by using a double-exponential distribution with density

$$g(x|\alpha) = (\alpha/2)\exp(-\alpha|x|)$$

$$\frac{f(x)}{g(x|\alpha)} \leq \sqrt{\frac{2}{\pi}} \alpha^{-1} e^{-\alpha^2/2}$$

and minimum of this bound (in α) attained for

$$\alpha^\star = 1$$

Probability of acceptance

$$\sqrt{\pi/2e} = .76$$

To produce one normal random variable, this Accept-Reject algorithm requires on the average $1/.76 \approx 1.3$ uniform variables.

To compare with the fixed single uniform required by the Box-Muller algorithm.

Example 13 **–Gamma with non-integer shape parameter–**

Illustrates a real advantage of the Accept-Reject algorithm

The gamma distribution $\mathcal{G}a(\alpha, \beta)$ represented as the sum of α exponential random variables, only if α is an integer

Can use the Accept-Reject algorithm with instrumental distribution

$$\mathcal{G}a(a, b), \text{ with } a = [\alpha], \quad \alpha \geq 0.$$

(Without loss of generality, $\beta = 1$.)

Up to a normalizing constant,

$$f/g_b = b^{-a} x^{\alpha-a} \exp\{-(1-b)x\} \leq b^{-a} \left(\frac{\alpha - a}{(1-b)e}\right)^{\alpha-a}$$

for $b \leq 1$.
The maximum is attained at $b = a/\alpha$.

–Cheng and Feast's Gamma $\mathcal{G}a(\alpha, 1)$, $\alpha > 1$–

Define $c_1 = \alpha - 1$, $c_2 = (\alpha - (1/6\alpha))/c_1$, $c_3 = 2/c_1$, $c_4 = 1 + c_3$, and $c_5 = 1/\sqrt{\alpha}$.

1. Repeat
   generate U_1, U_2
   take $U_1 = U_2 + c_5(1 - 1.86U_1)$ if $\alpha > 2.5$
   until $0 < U_1 < 1$.
2. $W = c_2 U_2/U_1$.
3. If $c_3 U_1 + W + W^{-1} \leq c_4$ or $c_3 \log U_1 - \log W + W \leq 1$,
   take $c_1 W$;
   otherwise, repeat.

Example 14 **–Truncated Normal distributions–**

Truncated Normals appear in many contexts

Constraints $x \geq \underline{\mu}$ produce densities proportional to

$$e^{-(x-\mu)^2/2\sigma^2} \, \mathbb{I}_{x \geq \underline{\mu}}$$

for a bound $\underline{\mu}$ large compared with μ

Alternatives far superior to the naïve method of generating a $\mathcal{N}(\mu, \sigma^2)$ until exceeding $\underline{\mu}$, which requires an average number of $1/\Phi((\mu - \underline{\mu})/\sigma)$ simulations from $\mathcal{N}(\mu, \sigma^2)$ for one acceptance.

Instrumental distribution: translated exponential distribution, $\mathcal{E}xp(\alpha, \underline{\mu})$, with density

$$g_\alpha(z) = \alpha e^{-\alpha(z-\underline{\mu})}\, \mathbb{I}_{z \geq \underline{\mu}}\ .$$

The ratio f/g_α is bounded by

$$f/g_\alpha \leq \begin{cases} 1/\alpha\ \exp(\alpha^2/2 - \alpha\underline{\mu}) & \text{if } \alpha > \underline{\mu}, \\ 1/\alpha\ \exp(-\underline{\mu}^2/2) & \text{otherwise.} \end{cases}$$

2.2.3 Log-concave densities

Densities f whose logarithm is concave, for instance Bayesian posterior distributions such that

$$\log\ \pi(\theta|x) = \log\ \pi(\theta) + \log\ f(x|\theta) + c$$

concave

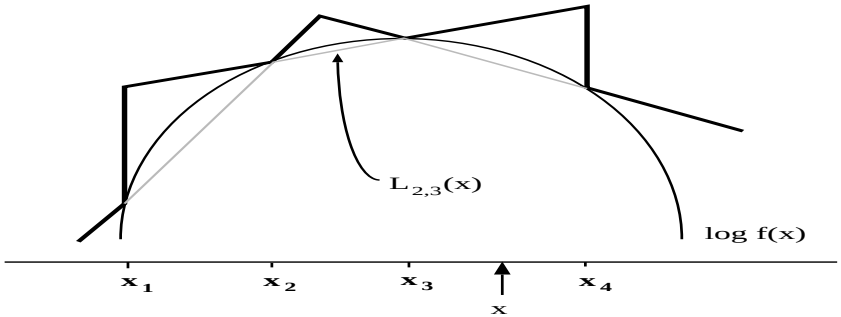
Take

$$\mathcal{S}_n = \{x_i, i = 0, 1, \ldots, n+1\} \subset \text{supp}(f)$$

such that $h(x_i) = \log f(x_i)$ known up to the same constant.

By concavity of h, line $L_{i,i+1}$ through $(x_i, h(x_i))$ and $(x_{i+1}, h(x_{i+1}))$

- below h in $[x_i, x_{i+1}]$ and
- above this graph outside this interval



For $x \in [x_i, x_{i+1}]$, if

$$\overline{h}_n(x) = \min\{L_{i-1,i}(x), L_{i+1,i+2}(x)\} \quad \text{and} \quad \underline{h}_n(x) = L_{i,i+1}(x)\,,$$

the envelopes are

$$\underline{h}_n(x) \leq h(x) \leq \overline{h}_n(x)$$

uniformly on the support of f, with

$$\underline{h}_n(x) = -\infty \quad \text{and} \quad \overline{h}_n(x) = \min(L_{0,1}(x), L_{n,n+1}(x))$$

on $[x_0, x_{n+1}]^c$. Therefore, if

$$\underline{f}_n(x) = \exp \underline{h}_n(x) \text{ and } \overline{f}_n(x) = \exp \overline{h}_n(x)$$

then

$$\underline{f}_n(x) \le f(x) \le \overline{f}_n(x) = \varpi_n \, g_n(x) \, ,$$

where ϖ_n normalizing constant of f_n

Algorithm 15 –ARS Algorithm–

0. Initialize n and $\mathcal{S}_n$.
1. Generate $X \sim g_n(x), U \sim \mathcal{U}_{[0,1]}$.
2. If $U \le \underline{f}_n(X)/\varpi_n \; g_n(X)$, accept X; otherwise, if $U \le f(X)/\varpi_n \; g_n(X)$, accept X and update $\mathcal{S}_n$ to $\mathcal{S}_{n+1} = \mathcal{S}_n \cup \{X\}$.

Example 16 **Northern Pintail ducks**

Ducks captured at time i with both probability p_i and size N of the population unknown.

Dataset

$$(n_1, \ldots, n_{11}) = (32, 20, 8, 5, 1, 2, 0, 2, 1, 1, 0)$$

Number of recoveries over the years 1957–1968 of $N = 1612$ Northern Pintail ducks banded in 1956

Corresponding conditional likelihood

$$L(p_1, \ldots, p_I | N, n_1, \ldots, n_I) = \frac{N!}{(N-r)!} \prod_{i=1}^{I} p_i^{n_i} (1-p_i)^{N-n_i},$$

where I number of captures, n_i number of captured animals during the ith capture, and r is the total number of different captured animals.

Prior selection

If

$$N \sim \mathcal{P}(\lambda)$$

and

$$\alpha_i = \log\left(\frac{p_i}{1-p_i}\right) \sim \mathcal{N}(\mu_i, \sigma^2),$$

[Normal logistic]

Posterior distribution

Then

$$\pi(\alpha, N|, n_1, \ldots, n_I) \propto \frac{N!}{(N-r)!}\frac{\lambda^N}{N!}\prod_{i=1}^{I}(1+e^{\alpha_i})^{-N} \prod_{i=1}^{I}\exp\left\{\alpha_i n_i - \frac{1}{2\sigma^2}(\alpha_i-\mu_i)^2\right\}$$

For the conditional posterior distribution

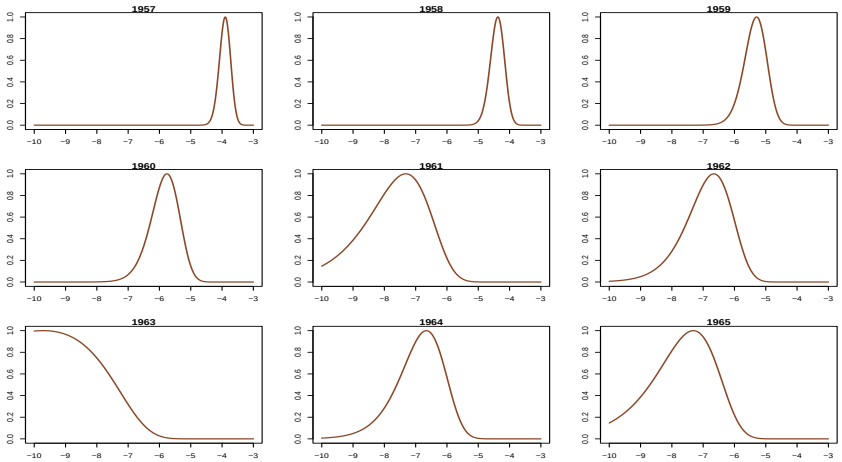
$$\pi(\alpha_i|N, n_1, \ldots, n_I) \propto \exp\left\{\alpha_i n_i - \frac{1}{2\sigma^2}(\alpha_i-\mu_i)^2\right\} \Big/ (1+e^{\alpha_i})^N ,$$

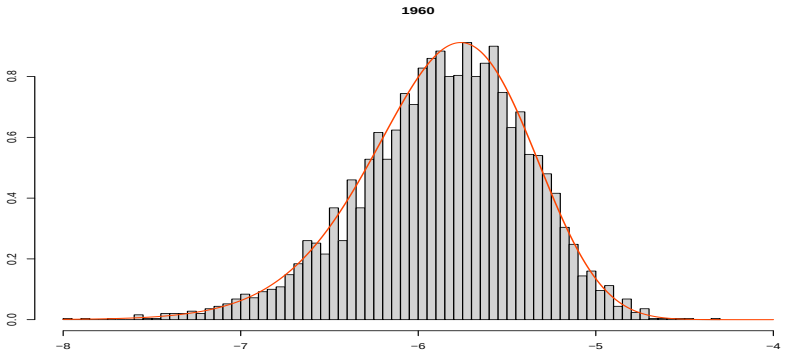
the ARS algorithm can be implemented since

$$\alpha_i n_i - \frac{1}{2\sigma^2}(\alpha_i-\mu_i)^2 - N\ \log(1+e^{\alpha_i})$$

is concave in α_i.

Posterior distributions of capture log-odds ratios for the years 1957–1965.





True distribution versus histogram of simulated sample

3 Monte Carlo Integration

3.1 Introduction

Two major classes of numerical problems that arise in statistical inference

- **optimization** - generally associated with the likelihood approach
- **integration**- generally associated with the Bayesian approach

Example 17 –Bayesian decision theory–

Bayes estimators are not always posterior expectations, but rather solutions of the minimization problem

$$\min_{\delta} \int_{\Theta} \mathrm{L}(\theta, \delta)\ \pi(\theta)\ f(x|\theta)\ d\theta\ .$$

Proper loss:

For $\mathrm{L}(\theta, \delta) = (\theta - \delta)^2$, the Bayes estimator is the **posterior mean**

Absolute error loss:

For $\mathrm{L}(\theta, \delta) = |\theta - \delta|$, the Bayes estimator is the **posterior median**

With no loss function

use the maximum a posteriori (MAP) estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

3.2 Classical Monte Carlo integration

Generic problem of evaluating the integral

$$\mathfrak{I} = \mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x)\, f(x)\, dx$$

where $\mathcal{X}$ is uni- or multidimensional, f is a closed form, partly closed form, or implicit density, and h is a function

First use a sample $(X_1, \ldots, X_m)$ from the density f to approximate the integral $\mathfrak{I}$ by the empirical average

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(x_j)$$

Average

$$\overline{h}_m \longrightarrow \mathbb{E}_f[h(X)]$$

by the **Strong Law of Large Numbers**

Estimate the variance with

$$v_m = \frac{1}{m}\frac{1}{m-1} \sum_{j=1}^{m} [h(x_j) - \overline{h}_m]^2,$$

and for m large,

$$\frac{\overline{h}_m - \mathbb{E}_f[h(X)]}{\sqrt{v_m}} \sim \mathcal{N}(0,1).$$

Note: This can lead to the construction of a convergence test and of confidence bounds on the approximation of $\mathbb{E}_f[h(X)]$.

Example 18 **–Cauchy prior–**

For estimating a normal mean, a *robust* prior is a Cauchy prior

$$X \sim \mathcal{N}(\theta, 1), \quad \theta \sim \mathcal{C}(0, 1).$$

Under squared error loss, posterior mean

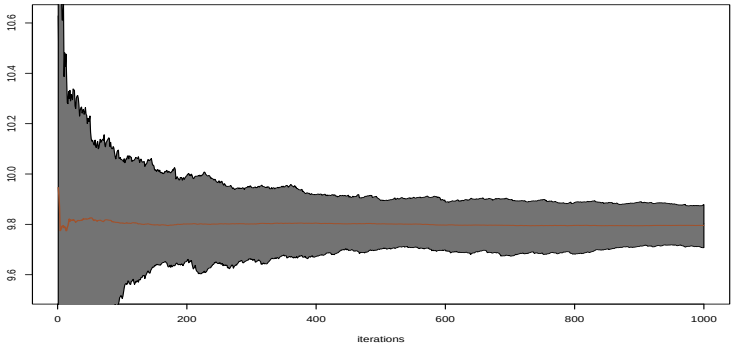
$$\delta^\pi(x) = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{\theta}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}{\displaystyle\int_{-\infty}^{\infty} \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}$$

Form of δ^π suggests simulating iid variables $\theta_1, \cdots, \theta_m \sim \mathcal{N}(x, 1)$ and calculate

$$\hat{\delta}^\pi_m(x) = \frac{\sum_{i=1}^m \dfrac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^m \dfrac{1}{1+\theta_i^2}} \; .$$

The Law of Large Numbers implies

$$\hat{\delta}^\pi_m(x) \longrightarrow \delta^\pi(x) \text{ as } m \longrightarrow \infty.$$



Range of estimators δ^π_m for 100 runs and $x = 10$

3.3 Importance Sampling

Simulation from f (the true density) is not necessarily **optimal**

Alternative to direct sampling from f is **importance sampling**, based on the alternative representation

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} \left[h(x) \, \frac{f(x)}{g(x)} \right] \, g(x) \, dx \; .$$

which allows us to use **other** distributions than f

Evaluation of

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x) \; f(x) \; dx$$

by

1. Generate a sample $X_1, \ldots, X_n$ from a distribution g
2. Use the approximation

$$\frac{1}{m} \sum_{j=1}^{m} \frac{f(X_j)}{g(X_j)} \, h(X_j)$$

Convergence of the estimator

$$\frac{1}{m}\sum_{j=1}^{m}\frac{f(X_j)}{g(X_j)}\,h(X_j)\longrightarrow\int_{\mathcal{X}}h(x)\,f(x)\,dx$$

- Same reason the regular Monte Carlo estimator $\overline{h}_m$ converges
- converges for any choice of the distribution g **[as long as** $supp(g)\supset supp(f)$**]**
- Instrumental distribution g chosen from distributions easy to simulate
- The same sample (generated from g) can be used repeatedly, not only for different functions h, but also for different densities f
- Even dependent proposals can be used, as seen later

Although g can be any density, some choices are better than others:

- Finite variance only when

$$\mathbb{E}_f\left[h^2(X)\frac{f(X)}{g(X)}\right]=\int_{\mathcal{X}}h^2(x)\,\frac{f^2(X)}{g(X)}\,dx<\infty\,.$$

- Instrumental distributions with tails lighter than those of f (that is, with $\sup f/g=\infty$) not appropriate.
- If $\sup f/g=\infty$, the weights $f(x_j)/g(x_j)$ vary widely, giving too much importance to a few values x_j.
- If $\sup f/g=M<\infty$, the accept-reject algorithm can be used as well to simulate f directly.

Example 19 **–Cauchy target–** Case of Cauchy distribution $\mathrm{C}(0,1)$ when importance function is Gaussian $(0,1)$.

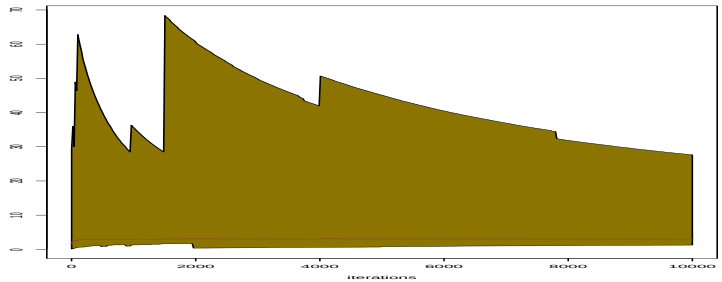
Ratio of the densities

$$\varrho(x)=\frac{p^\star(x)}{p_0(x)}=\sqrt{2\pi}\,\frac{\exp x^2/2}{\pi\,(1+x^2)}$$

very badly behaved: e.g.,

$$\int_{-\infty}^{\infty}\varrho(x)^2p_0(x)dx=\infty\,.$$

Poor performances of the associated importance sampling estimator



Range and average of 500 replications of IS estimate of $\mathbb{E}[\exp -X]$ over $10,000$ iterations.

The choice of g that minimizes the variance of the importance sampling estimator is

$$g^*(x) = \frac{|h(x)|\ f(x)}{\int_{\mathcal{Z}} |h(z)|\ f(z)\ dz}\ .$$

Rather formal optimality result since optimal choice of $g^*(x)$ requires the knowledge of $\mathfrak{I}$, the integral of interest!

Practical alternative

$$\frac{\sum_{j=1}^m h(X_j)\ f(X_j)/g(X_j)}{\sum_{j=1}^m f(X_j)/g(X_j)},$$

where f and g are known up to constants.

- Also converges to $\mathfrak{I}$ by the Strong Law of Large Numbers.
- Biased, but the bias is quite small
- In some settings beats the unbiased estimator in squared error loss.

For ratio estimator

$$\delta_h^n = \sum_{i=1}^n \omega_i\, h(x_i) \Big/ \sum_{i=1}^n \omega_i$$

with $X_i \sim g(y)$ and W_i such that

$$\mathbb{E}[W_i|X_i = x] = \kappa f(x)/g(x)$$

then

$$\mathrm{var}(\delta_h^n) \approx \frac{1}{n^2\kappa^2}\left(\mathrm{var}(S_h^n) - 2\mathbb{E}^\pi[h]\,\mathrm{cov}(S_h^n, S_1^n) + \mathbb{E}^\pi[h]^2\,\mathrm{var}(S_1^n)\right) .$$

for

$$S_h^n = \sum_{i=1}^n W_i h(X_i)\,, \quad S_1^n = \sum_{i=1}^n W_i$$

First approximation

$$\mathrm{var}\delta_h^n \approx \frac{1}{n}\,\mathrm{var}^\pi(h(X))\,\{1 + \mathrm{var}_g(W)\}$$

Example 20 **–Student's t distribution–** $X \sim \mathcal{T}(\nu, \theta, \sigma^2)$, with density

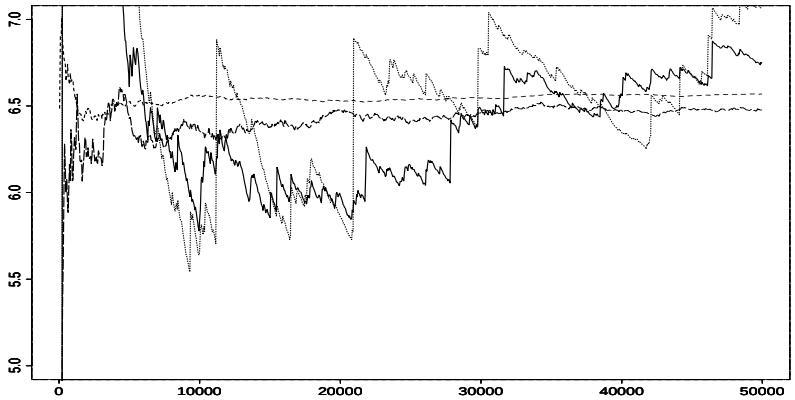
$$f_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1 + \frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} .$$

Without loss of generality, take $\theta = 0$, $\sigma = 1$.

Calculate the integral

$$\int_{2.1}^\infty \left(\frac{\sin(x)}{x}\right)^n f_\nu(x)dx.$$

- Simulation possibilities
  - Directly from f_ν, since $f_\nu = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_\nu}}$
  - Importance sampling using Cauchy $\mathcal{C}(0,1)$
  - Importance sampling using a normal $\mathcal{N}(0,1)$
    (expected to be nonoptimal)
  - Importance sampling using a $\mathcal{U}([0,1/2.1])$
    change of variables



Sampling from f (solid lines), importance sampling with Cauchy instrumental (short dashes), $\mathcal{U}([0, 1/2.1])$ instrumental (long dashes) and normal instrumental (dots).

IS suffers from curse of dimensionality

As dimension increases, discrepancy between importance and target worsens

Explanation:

Take target distribution μ and instrumental distribution ν

Simulation of a sample of iid samples of size n $x_{1:n}$ from $\mu_n = \mu^{\otimes n}$

Importance sampling estimator for $\mu_n(f_n) = \int f_n(x_{1:n})\mu_n(dx_{1:n})$

$$\widehat{\mu_n(f_n)} = \frac{\sum_{i=1}^N f_n(\xi_{1:n}^i) \prod_{j=1}^N W_j^i}{\sum_{j=1}^N \prod_{j=1}^N W_j},$$

where $W_k^i = \frac{d\mu}{d\nu}(\xi_k^i)$, and ξ_j^i are iid with distribution ν.

For $\{V_k\}_{k\geq 0}$, sequence of nonnegative random variables and for $n \geq 1$,

$\mathcal{F}_n = \sigma(V_k; k \leq n)$, set

$$U_n = \prod_{k=1}^n V_k$$

Since $\mathbb{E}[V_{n+1}] = 1$ and V_{n+1} independent from $\mathcal{F}_n$,

$$\mathbb{E}(U_{n+1} \mid \mathcal{F}_n) = U_n \mathbb{E}(V_{n+1} \mid \mathcal{F}_n) = U_n,$$

and thus $\{U_n\}_{n\geq 0}$ **martingale**

Since $x \mapsto \sqrt{x}$ concave, by Jensen's inequality,

$$\mathbb{E}(\sqrt{U_{n+1}} \mid \mathcal{F}_n) \leq \sqrt{\mathbb{E}(U_{n+1} \mid \mathcal{F}_n)} \leq \sqrt{U_n}$$

and thus $\{\sqrt{U_n}\}_{n\geq 0}$ **supermartingale**

Assume $\mathbb{E}(\sqrt{V_{n+1}}) < 1$. Then

$$\mathbb{E}(\sqrt{U_n}) = \prod_{k=1}^n \mathbb{E}(\sqrt{V_k}) \to 0, \quad n \to \infty.$$

But $\{\sqrt{U_n}\}_{n\geq 0}$ is a nonnegative supermartingale and thus $\sqrt{U_n}$ converges a.s. to a random variable $Z \geq 0$. By Fatou's lemma,

$$\mathbb{E}(Z) = \mathbb{E}\left(\lim_{n\to\infty} \sqrt{U_n}\right) \leq \liminf_{n\to\infty} \mathbb{E}(\sqrt{U_n}) = 0.$$

Hence, $Z = 0$ and $U_n \to 0$ a.s., which implies that the martingale $\{U_n\}_{n\geq 0}$ is not regular.

Apply these results to $V_k = \frac{d\mu}{d\nu}(\xi_k^i)$, $i \in \{1, , N\}$:

$$\mathbb{E}\left[\sqrt{\frac{d\mu}{d\nu}(ki)}\right] \leq \mathbb{E}\left[\frac{d\mu}{d\nu}(k)\right] = 1.$$

with equality iff $\frac{d\mu}{d\nu} = 1$, ν-a.e., i.e. $\mu = \nu$.

Thus all importance weights converge to 0

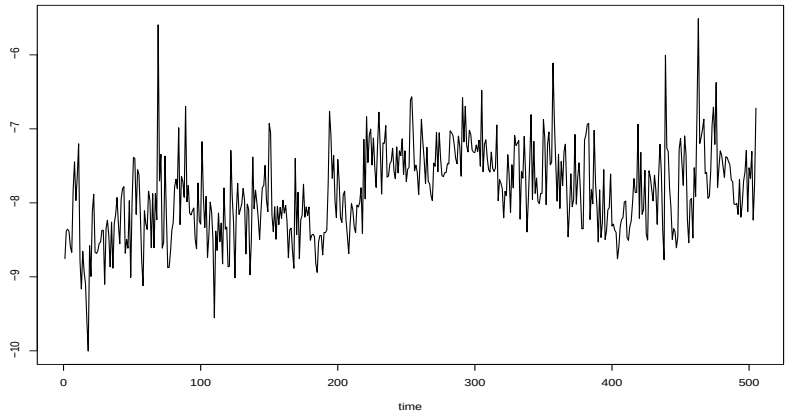
Example 21 **–Stochastic volatility model–**

$$y_t = \beta \exp(x_t/2)\,\epsilon_t\,, \qquad \epsilon_t \sim \mathcal{N}(0,1)$$

with AR(1) log-variance process (or *volatility*)

$$x_{t+1} = \varphi x_t + \sigma u_t\,, \quad u_t \sim \mathcal{N}(0,1)$$

Evolution of IBM stocks (corrected from trend and log-ratio-ed)



Observed likelihood unavailable in closed from.

Joint posterior (or conditional) distribution of the hidden state sequence $\{X_k\}_{1\le k\le n}$ can be evaluated explicitly

$$\prod_{k=2}^{K} \exp - \left\{\sigma^{-2}(x_k - \phi x_{k-1})^2 + \beta^{-2}\exp(-x_k)y_k^2 + x_k\right\}/2\,, \qquad (2)$$

up to a normalizing constant.

Computational problems

Direct simulation from this distribution impossible because of

(a) dependence among the X_k's,

(b) dimension of the sequence $\{X_k\}_{1\le k\le n}$, and

(c) exponential term $\exp(-x_k)y_k^2$ within (2).

Importance sampling

Natural candidate: replace the exponential term with a quadratic approximation to preserve Gaussianity.

E.g., expand $\exp(-x_k)$ around its conditional expectation ϕx_{k-1} as

$$\exp(-x_k) \approx \exp(-\phi x_{k-1}) \left\{ 1 - (x_k - \phi x_{k-1}) + \frac{1}{2}(x_k - \phi x_{k-1})^2 \right\} .$$

Corresponding Gaussian importance distribution with mean

$$\mu_k = \frac{\phi x_{k-1}\{\sigma^{-2} + y_k^2 \exp(-\phi x_{k-1})/2\} - \{1 - y_k^2 \exp(-\phi x_{k-1})\}/2}{\sigma^{-2} + y_k^2 \exp(-\phi x_{k-1})/2},$$

and variance

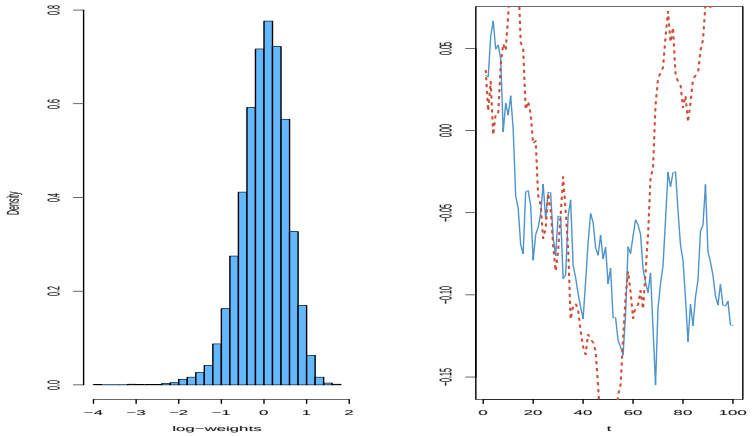
$$\tau_k^2 = (\sigma^{-2} + y_k^2 \exp(-\phi x_{k-1})/2)^{-1} .$$

Prior proposal on X_1, $X_1 \sim \mathcal{N}(0, \sigma^2)$.

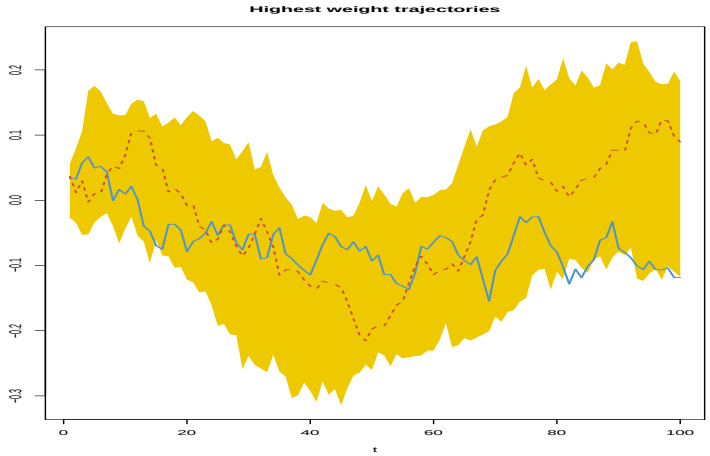
Simulation starts with X_1 and proceeds forward to X_n, each X_k being generated conditional on Y_k and the previously generated X_{k-1}.

Importance weight computed sequentially as the product of

$$\frac{\exp - \left\{\sigma^{-2}(x_k - \phi x_{k-1})^2 + \exp(-x_k) y_k^2 + x_k\right\}/2}{\exp - \left\{\tau_k^{-2}(x_k - \mu_k)^2\right\} \tau_k^{-1}} .$$



Histogram of the logarithms of the importance weights (left) and comparison between the true volatility and the best fit, based on $10,000$ simulated importance samples.



Corresponding range of the simulated $\{X_k\}_{1\leq k\leq 100}$, compared with the true value.

3.4 Acceleration methods

Correlated simulations

Negative correlation reduce variance

Special technique — but efficient when it applies

Two samples $(X_1,\ldots,X_m)$ and $(Y_1,\ldots,Y_m)$ from f to estimate

$$\mathfrak{I} = \int_{\mathbb{R}} h(x)f(x)dx\ .$$

$$\hat{\mathfrak{I}}_1 = \frac{1}{m}\sum_{i=1}^{m} h(X_i) \quad \text{and} \quad \hat{\mathfrak{I}}_2 = \frac{1}{m}\sum_{i=1}^{m} h(Y_i)$$

with mean $\mathfrak{I}$ and variance σ^2

Variance of the average

$$\text{var}\left(\frac{\hat{\mathfrak{I}}_1 + \hat{\mathfrak{I}}_2}{2}\right) = \frac{\sigma^2}{2} + \frac{1}{2}cov(\hat{\mathfrak{I}}_1, \hat{\mathfrak{I}}_2).$$

If the two samples are **negatively correlated**,

$$\text{cov}(\hat{\mathfrak{I}}_1, \hat{\mathfrak{I}}_2) \leq 0\,,$$

they improve on two independent samples of same size

Example: Antithetic variables

- If f symmetric about μ, take $Y_i = 2\mu - X_i$
- If $X_i = F^{-1}(U_i)$, take $Y_i = F^{-1}(1 - U_i)$
- If $(A_i)_i$ partition of $\mathcal{X}$, **partitioned sampling** by sampling X_j's in each A_i (requires to know Pr(A_i))

Control Variates

For

$$\mathfrak{I} = \int h(x)f(x)dx$$

unknown and

$$\mathfrak{I}_0 = \int h_0(x)f(x)dx$$

known,

$\mathfrak{I}_0$ estimated by $\hat{\mathfrak{I}}_0$ (and $\mathfrak{I}$ estimated by $\hat{\mathfrak{I}}$)

Combined estimator

$$\hat{\mathfrak{I}}^* = \hat{\mathfrak{I}} + \beta(\hat{\mathfrak{I}}_0 - I_0)$$

$\hat{\mathfrak{I}}^*$ is unbiased for $\mathfrak{I}$ and

$$\text{var}(\hat{\mathfrak{I}}^*) = \text{var}(\hat{\mathfrak{I}}) + \beta^2\text{var}(\hat{\mathfrak{I}}) + 2\beta cov(\hat{\mathfrak{I}}, \hat{\mathfrak{I}}_0)$$

Optimal choice of β

$$\beta^\star = -\frac{cov(\hat{\mathfrak{I}}, \hat{\mathfrak{I}}_0)}{\text{var}(\hat{\mathfrak{I}}_0)} \ ,$$

with

$$\text{var}(\hat{\mathfrak{I}}^\star) = (1-\rho^2)\ \text{var}(\hat{\mathfrak{I}}) \ ,$$

where ρ correlation between $\hat{\mathfrak{I}}$ and $\hat{\mathfrak{I}}_0$

Usual solution: **regression coefficient of $h(x_i)$ over $h_0(x_i)$**

Example 22 – Quantile Approximation –

Evaluate

$$\varrho = \text{Pr}(X > a) = \int_a^\infty f(x)dx$$

by

$$\hat{\varrho} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i > a),$$

with X_i iid f.

If $\text{Pr}(X > \mu) = \frac{1}{2}$ known

Control variate

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i > a) + \beta\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i > \mu) - \text{Pr}(X > \mu)\right)$$

improves upon $\hat{\varrho}$ if

$$\beta < 0 \quad \text{and} \quad |\beta| < 2\frac{cov(\delta_1, \delta_3)}{\text{var}(\delta_3)} = 2\frac{\text{Pr}(X > a)}{\text{Pr}(X > \mu)}.$$

Integration by conditioning

Use **Rao-Blackwell Theorem**

$$\mathbf{var}(\mathbb{E}[\delta(\mathbf{X})|\mathbf{Y}]) \leq \mathbf{var}(\delta(\mathbf{X}))$$

Consequence : If $\hat{\mathfrak{I}}$ unbiased estimator of $\mathfrak{I} = \mathbb{E}_f[h(X)]$, with X simulated from a joint density $\tilde{f}(x, y)$, where

$$\int \tilde{f}(x, y)dy = f(x),$$

the estimator

$$\hat{\mathfrak{I}}^* = \mathbb{E}_{\tilde{f}}[\hat{\mathfrak{I}}|Y_1, \ldots, Y_n]$$

dominate $\hat{\mathfrak{I}}(X_1, \ldots, X_n)$ variance-wise (and is unbiased)

Example 23 –Student's t expectation–

For

$$\mathbb{E}[h(x)] = \mathbb{E}[\exp(-x^2)] \quad \text{with} \quad X \sim \mathcal{T}(\nu, 0, \sigma^2)$$

a Student's t can be simulated as

$$X|y \sim \mathcal{N}(\mu, \sigma^2 y) \qquad \text{et} \qquad Y^{-1} \sim \chi^2_\nu.$$

Empirical distribution

$$\frac{1}{m}\sum_{j=1}^{m} \exp(-X_j^2)\ ,$$

can be improved from the joint sample

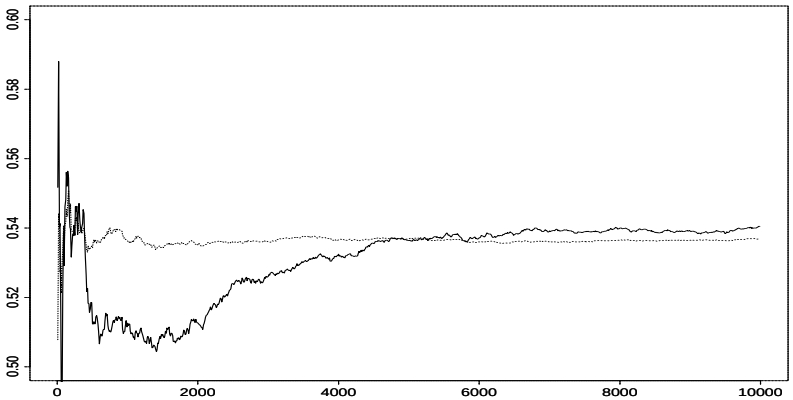
$$((X_1, Y_1), \ldots, (X_m, Y_m))$$

since

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[\exp(-X^2)|Y_j] = \frac{1}{m}\sum_{j=1}^{m} \frac{1}{\sqrt{2\sigma^2 Y_j + 1}}$$

is the conditional expectation.

In this example, precision **ten times** better



Estimators of $\mathbb{E}[\exp(-X^2)]$: empirical average (full) and conditional expectation (dotted) for $(\nu, \mu, \sigma) = (4.6, 0, 1)$.

3.5 Bayesian importance sampling

Recall algorithm:

1. Generate $\theta_1^{(1)}, \cdots, \theta_1^{(T)}$ from $cg(\theta)$ with

$$c^{-1} = \int g(\theta) d\theta$$

2. Take

$$\int f(x|\theta)\pi(\theta)d\theta \approx \frac{1}{T}\sum_{t=1}^{T} f(x|\theta^{(t)})\frac{\pi(\theta^{(t)})}{cg(\theta^{(t)})}$$

$$\approx \frac{\sum_{t=1}^{T} f(x|\theta^{(t)})\frac{\pi(\theta^{(t)})}{g(\theta^{(t)})}}{\sum_{t=1}^{T}\frac{\pi(\theta^{(t)})}{g(\theta^{(t)})}} = m^{IS}(x)$$

[Marginal approximation]

Choice of g

$\boxed{g(\theta) = \pi(\theta)}$

$$m^{IS}(x) = \frac{1}{T}\sum_{t} f(x|\theta^{(t)})$$

- ♢ often inefficient if data informative
- ♢ impossible if π is improper
- ♢ (a) often satisfied
- ♢ (b) always satisfied

$\boxed{g(\theta) = f(x|\theta)\pi(\theta)}$

- ♢ c unknown
- ♢ (a) satisfied
- ♢ $m^{IS}(x) = 1 \Big/ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(x|\theta^{(t)})}$
- ♢ (b) often fails: ex. $\mathcal{N}(0,1)$
- ♢ improper priors allowed

$$\boxed{g(\theta) = \rho\pi(\theta) + (1-\rho)\pi(\theta|x)}$$

[Newton & Raftery, 1994]

♢ defensive mixture

♢ $\rho \ll 1$ Ok

[Hestenberg, 1998]

$$\boxed{g(\theta) = \pi(\theta|x)}$$

♢ $m^h(x) = \dfrac{1}{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \dfrac{h(\theta)}{f(x|\theta)\pi(\theta)}}$

♢ works for any h

♢ finite variance if

$$\int \frac{h^2(\theta)}{f(x|\theta)\pi(\theta)} d\theta < \infty$$

Bridge sampling

[Chen & Shao, 1997]

$$\begin{aligned} \pi_1(\theta_1|x) &= \frac{\pi_1(\theta_1) f_1(x|\theta_1)}{m_1(x)} \\ \pi_2(\theta_2|x) &= \frac{\pi_2(\theta_2) f_2(x|\theta_2)}{m_2(x)} \end{aligned}$$

Bayes factor: $B_{12}(x) = \dfrac{m_1(x)}{m_2(x)}$ ratio of normalising constants

(i)

$$\begin{aligned} \pi_1(\theta_1|x) &\propto \tilde{\pi}_1(\theta_1) \\ \pi_2(\theta_2|x) &\propto \tilde{\pi}_2(\theta_2) \end{aligned}$$

$$B_{12} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{\pi}_1(\theta_i)}{\tilde{\pi}_2(\theta_i)} \qquad \theta_i \sim \pi_2$$

(ii)

$$B_{12} = \frac{\int \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)d\theta}{\int \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)d\theta} \qquad \forall\, \alpha(\cdot)$$

$$\approx \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i})\alpha(\theta_{1i})}{\frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i})\alpha(\theta_{2i})} \qquad \theta_{ji} \sim \pi_j(\theta)$$

Optimal choice

$$\alpha(\theta) = \frac{n_1 + n_2}{n_1\pi_1(\theta) + n_2\pi_2(\theta)} \qquad [?]$$

[Chen, Meng & Wong, 2000]

4 Notions on Markov Chains

4.1 Basics

Markov chain sequence of random variables whose distribution evolves over **time** as a function of past realizations

Chain defined through its **transition kernel**, a function K defined on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ such that

(i) $\forall x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure;

(ii) $\forall A \in \mathcal{B}(\mathcal{X})$, $K(\cdot, A)$ is measurable.

- When $\mathcal{X}$ is a **discrete** (finite or denumerable) set, the transition kernel simply is a (transition) matrix $\mathbb{K}$ with elements

$$P_{xy} = \Pr(X_n = y | X_{n-1} = x) \ , \qquad x, y \in \mathcal{X}$$

Since, for all $x \in \mathcal{X}$, $K(x, \cdot)$ is a probability, we must have

$$P_{xy} \geq 0 \quad \text{and} \quad K(x, \mathcal{X}) = \sum_{y \in \mathcal{X}} P_{xy} = 1$$

The matrix $\mathbb{K}$ is referred to as a **Markov transition matrix** or a **stochastic matrix**

- In the **continuous** case, the *kernel* also denotes the conditional density $\mathfrak{K}(x, x')$ of the transition $K(x, \cdot)$

$$\Pr(X \in A | x) = \int_A \mathfrak{K}(x, x') dx'.$$

Then, for any bounded ϕ, we may define

$$K\phi(x) = K(x, \phi) = \int_{\mathcal{X}} \mathfrak{K}(x, dy)\phi(y).$$

Note that

$$|K\phi(x)| \leq \int_{\mathcal{X}} \mathfrak{K}(x, dy)|\phi(y)| \leq |\phi|_\infty = \sup_{x \in \mathcal{X}} |\phi(x)|.$$

We may also associate to a probability measure μ the measure μK, defined as

$$\mu K(A) = \int_{\mathcal{X}} \mu(dx) K(x, A).$$

Markov chains

Given a transition kernel K, a sequence $X_0, X_1, \ldots, X_n, \ldots$ of random variables is a **Markov chain** denoted by (X_n), if, for any t, the conditional distribution of X_t given $x_{t-1}, x_{t-2}, \ldots, x_0$ is the same as the distribution of X_t given x_{t-1}. That is,

$$\begin{aligned} \Pr(X_{k+1} \in A | x_0, x_1, x_2, \ldots, x_k) &= \Pr(X_{k+1} \in A | x_k) \\ &= \int_A \mathfrak{K}(x_k, dx) \end{aligned}$$

Note that the entire structure of the chain only depends on

- The transition function K
- The initial state x_0 or initial distribution $X_0 \sim \mu$

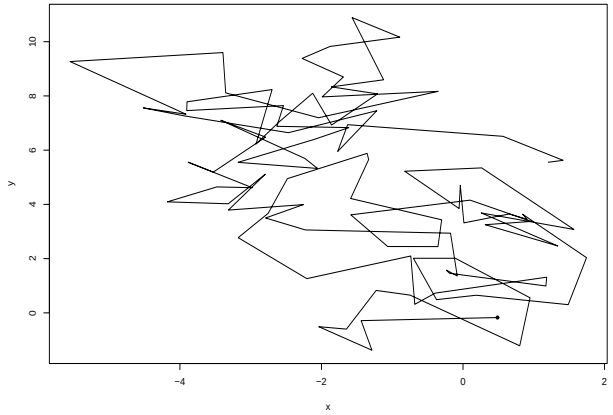
Example 24 **–Random walk–** The normal random walk is the kernel $K(x,\cdot)$ associated with the distribution

$$\mathcal{N}_p(x,\tau^2 I_p)$$

which means

$$X_{t+1} = X_t + \tau\epsilon_t$$

ϵ_t being an iid additional noise



100 consecutive realisations of the random walk in $\mathbb{R}^2$ with $\tau = 1$

On a **discrete state-space** $\mathcal{X} = \{x_0, x_1, \ldots\}$,

- A function ϕ on a discrete state space is uniquely defined by the (column) vector $\phi = (\phi(x_0), \phi(x_1), \ldots,)^{\mathrm{T}}$ and

$$K\phi(x) = \sum_{y\in\mathcal{X}} P_{xy}\phi(y)$$

  can be interpreted as the xth component of the product of the transition matrix $\mathbb{K}$ and of the vector ϕ.

- A probability distribution on $\mathcal{P}(\mathcal{X})$ is defined as a (row) vector $\mu = (\mu(x_0), \mu(x_1), \ldots)$ and the probability distribution μK is defined, for each $y \in \mathcal{X}$ as

$$\mu K(\{y\}) = \sum_{x\in\mathcal{X}} \mu(\{x\})P_{xy}$$

  yth component of the product of the vector μ and of the transition matrix $\mathbb{K}$.

Composition of kernels

Let Q_1 and Q_2 be two probability kernels. Define, for any $x \in \mathcal{X}$ and any $A \in \mathcal{B}(\mathcal{X})$ the **product of kernels** Q_1Q_2 as

$$Q_1Q_2(x,A) = \int_{\mathcal{X}} \mathfrak{Q}_1(x,dy)\mathfrak{Q}_2(y,A)$$

When the state space $\mathcal{X}$ is discrete, the product of Markov kernels coincides with the product of matrices $\mathbb{Q}_1 \times \mathbb{Q}_2$.

4.2 Irreducibility

Irreducibility is one measure of the sensitivity of the Markov chain to initial conditions

It leads to a guarantee of convergence for MCMC algorithms

In the discrete case, the chain is *irreducible* if all states communicate, namely if

$$P_x(\tau_y < \infty) > 0 \ , \qquad \forall x, y \in \mathcal{X} \ ,$$

τ_y being the first (positive) time y is visited

In the continuous case, the chain is φ-*irreducible* for some measure φ if for some n,

$$K^n(x, A) > 0$$

- for all $x \in \mathcal{X}$
- for every $A \in \mathcal{B}(\mathcal{X})$ with $\varphi(A) > 0$

Minoration condition

Assume there exist a probability measure ν and $\epsilon > 0$ such that, for all $x \in \mathcal{X}$ and all $A \in \mathcal{B}(\mathcal{X})$,

$$K(x, A) \geq \epsilon\nu(A)$$

This is called a **minoration condition**.

When K is a Markov chain on a discrete state space, this is equivalent to saying that $P_{xy} > 0$ for all $x, y \in \mathcal{X}$.

Small sets

If there exist $C \in \mathcal{B}(\mathcal{X})$, $\varphi(C) > 0$, a probability measure ν and $\epsilon > 0$ such that, for all $x \in C$ and all $A \in \mathcal{B}(\mathcal{X})$,

$$K(x, A) \geq \epsilon\nu(A)$$

C is called a **small set**

For discrete state space, **atoms** are small sets.

4.3 Transience and Recurrence

- Irreducibility ensures that every set A will be visited by the Markov chain (X_n)
- This property is too weak to ensure that the trajectory of (X_n) will enter A often enough.
- A Markov chain must enjoy good *stability* properties to guarantee an acceptable approximation of the simulated model.
  - Formalizing this stability leads to different notions of *recurrence*
  - For discrete chains, the *recurrence of a state* equivalent to probability one of sure return.
  - Always satisfied for irreducible chains on finite spaces

In a finite state space $\mathcal{X}$, denote the average number of visits to a state ω by

$$\eta_\omega = \sum_{i=1}^{\infty} \mathbb{I}_\omega(X_i)$$

If $\mathbb{E}_\omega[\eta_\omega] = \infty$, the state is *recurrent*

If $\mathbb{E}_\omega[\eta_\omega] < \infty$, the state is *transient*

For irreducible chains, recurrence/transience **property of the chain**, not of a particular state

Similar definitions for the continuous case.

Stronger form of recurrence: **Harris recurrence**

A set A is *Harris recurrent* if

$$P_x(\eta_A = \infty) = 1 \text{ for all } x \in A.$$

The chain (X_n) is Ψ–*Harris recurrent* if it is

- ψ–irreducible
- for every set A with $\psi(A) > 0$, A is Harris recurrent.

Note that

$$P_x(\eta_A = \infty) = 1 \text{ implies } \mathbb{E}_x[\eta_A] = \infty$$

4.4 Invariant Measures

Stability increases for the chain (X_n) if marginal distribution of X_n independent of n

Requires the existence of a probability distribution π such that

$$X_{n+1} \sim \pi \quad \text{if} \quad X_n \sim \pi$$

A measure π is **invariant** for the transition kernel $K(\cdot, \cdot)$ if

$$\pi(B) = \int_{\mathcal{X}} K(x, B)\, \pi(dx)\,, \qquad \forall B \in \mathcal{B}(\mathcal{X})\,.$$

- The chain is **positive recurrent** if π is a probability measure.
- Otherwise it is **null recurrent** or **transient**

- If π probability measure, π also called *stationary distribution* since

$$X_0 \sim \pi \text{ implies that } X_n \sim \pi \text{ for every } n$$

- The stationary distribution is unique

Insights

Invariant probability measures are important not merely because they define stationary processes, but also because they turn out to be the measures which define the long-term or ergodic behavior of the chain.

To understand why this is so, consider $P_\mu(X_n \in \cdot)$ for a starting distribution μ. If a limiting measure γ_μ exists such as

$$P_\mu(X_n \in A) \to \gamma_\mu(A)$$

for all $A \in \mathcal{B}(\mathcal{X})$, then

$$\begin{aligned}
\gamma_\mu(A) &= \lim_{n\to\infty} \int \mu(dx) P^n(x, A) \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} \int P^{n-1}(x, dw) K(w, A) \\
&= \int_{\mathcal{X}} \gamma_\mu(dw) K(w, A)
\end{aligned}$$

since setwise convergence of $\int \mu P^n(x, \cdot)$ implies convergence of integrals of bounded measurable functions. Hence, if a limiting distribution exists, it is an invariant probability measure; and obviously, if there is a unique invariant probability measure, the limit γ_μ will be independent of μ whenever it exists.

4.5 Ergodicity and convergence

We finally consider: *to what is the chain converging?*

The invariant distribution π natural candidate for the *limiting distribution*

A fundamental property is **ergodicity**, or independence of initial conditions.

In the discrete case, a state ω is *ergodic* if

$$\lim_{n\to\infty} |K^n(\omega, \omega) - \pi(\omega)| = 0 \ .$$

In general , we establish convergence using the *total variation norm*

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \sup_A \ |\mu_1(A) - \mu_2(A)|$$

and we want

$$\left\| \int K^n(x, \cdot)\mu(dx) - \pi \right\|_{\mathrm{TV}}$$

$$= \sup_A \left| \int K^n(x, A)\mu(dx) - \pi(A) \right|$$

to be small.

Harris recurrence and ergodicity

If (X_n) Harris positive recurrent and aperiodic, then

$$\lim_{n\to\infty} \left\| \int K^n(x, \cdot)\mu(dx) - \pi \right\|_{TV} = 0$$

for every initial distribution μ.

We thus take "*Harris positive recurrent and aperiodic*" as equivalent to "*ergodic*"

[Meyn & Tweedie, 1993]

Convergence in total variation implies

$$\lim_{n\to\infty} \ |\mathbb{E}_\mu[h(X_n)] - \mathbb{E}^\pi[h(X)]| = 0$$

for every *bounded function* h.

There are difference speeds of convergence

- ergodic (fast enough)
- *geometrically* ergodic (faster)
- *uniformly* ergodic (fastest)

Geometric ergodicity

A ϕ-irreducible aperiodic Markov kernel P with invariant distribution π is **geometrically ergodic** if there exist $V \geq 1$, and constants $\rho < 1$, $R < \infty$ such that $(n \geq 1)$

$$\|P^n(x, .) - \pi(.)\|_V \leq RV(x)\rho^n \,,$$

on $\{V < \infty\}$ which is full and absorbing.

Geometric ergodicity implies a lot of important results

- CLT for additive functionals $n^{-1/2}\sum g(X_k)$ and functions $|g| < V$
- Rosenthal's type inequalities

$$\mathbb{E}_x\left|\sum_{k=1}^{n} g(X_k)\right|^p \leq C(p)n^{p/2}, \qquad |g|^p \leq 2$$

- exponential inequalities (for bounded functions and α small enough)

$$\mathbb{E}_x\left\{\exp\left(\alpha\sum_{k=1}^{n} g(X_k)\right)\right\} < \infty$$

Minoration condition and uniform ergodicity

Under the minoration condition, the kernel K is thus contractant and standard results in functional analysis shows the existence and the unicity of a fixed point π. The previous relation implies that, for all $x \in \mathcal{X}$.

$$\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq (1-\epsilon)^n$$

Such Markov chains are called **uniformly ergodic**.

The following conditions are equivalent:

- $(X_n)_n$ **is uniformly ergodic,**
- **there exist** $\rho < 1$ **and** $R < \infty$ **such that, for all** $x \in \mathcal{X}$**,**

$$\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq R\rho^n.$$

- **for some** $n > 0$**,**

$$\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} < 1.$$

[Meyn and Tweedie, 1993]

4.6 Limit theorems

Ergodicity determines the probabilistic properties of **average** behavior of the chain.

But also need of *statistical inference*, made by induction from the observed sample.

If $\|P_x^n - \pi\|$ close to 0, no direct information about

$$X_n \sim P_x^n$$

We need LLN's and CLT's!!!

Classical LLN's and CLT's not directly applicable due to:

- Markovian dependence structure between the observations X_i
- Non-stationarity of the sequence

Ergodic Theorem

If the Markov chain (X_n) is Harris recurrent, then for any function h with $\mathrm{E}|h| < \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_i h(X_i) = \int h(x) d\pi(x),$$

Central Limit Theorem

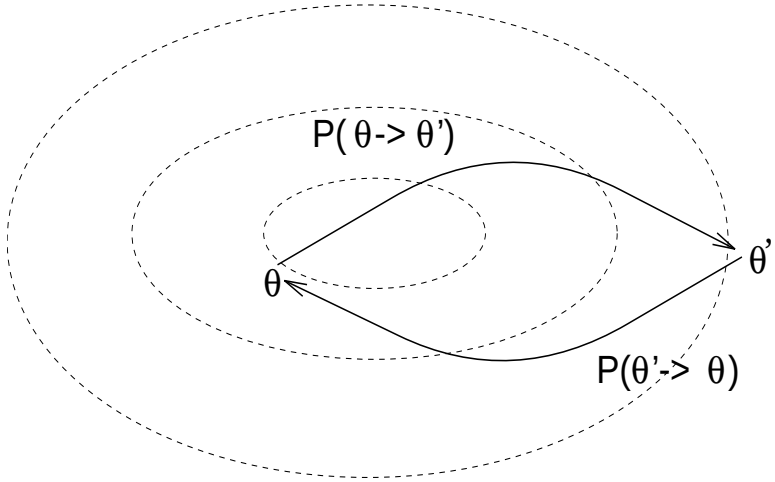
To get a CLT, we need more assumptions.

For MCMC, the easiest is **reversibility**:

A Markov chain (X_n) is *reversible* if for all n

$$X_{n+1}|X_{n+2} = x \sim X_{n+1}|X_n = x$$

The direction of time does not matter



[Green, 1995]

If the Markov chain (X_n) is Harris recurrent and reversible,

$$\frac{1}{\sqrt{N}}\left(\sum_{n=1}^{N}\left(h(X_n)-\mathbb{E}^{\pi}[h]\right)\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\gamma_h^2)\ .$$

where

$$\begin{aligned}0<\gamma_h^2 &= \mathbb{E}_\pi[\overline{h}^2(X_0)]\\ &\quad +2\sum_{k=1}^{\infty}\mathbb{E}_\pi[\overline{h}(X_0)\overline{h}(X_k)]<+\infty.\end{aligned}$$

[Kipnis & Varadhan, 1986]

4.7 Quantitative convergence rates

Let P a Markov transition kernel on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$, with P positive recurrent and π its stationary distribution

Convergence rate Determine, from the kernel, a sequence $B(\nu,n)$, such that

$$\|\nu P^n-\pi\|_V\le B(\nu,n)$$

where $V:\mathcal{X}\to[1,\infty)$ and for any signed measure μ,

$$\|\mu\|_V=\sup_{|\phi|\le V}|\mu(\phi)|$$

In the 90's, a wealth of contributions on quantitative bounds triggered by MCMC algorithms to answer questions like: what is the appropriate *burn in*? or how long should the sampling continue after burn in?

[Douc, Moulines and Rosenthal, 2001]

[Jones and Hobert, 2001]

For MCMC algorithms, kernels are "explicitly" known.

Type of quantities (more or less directly) available:

- Minoration constants

$$K^s(x,A)\ge\epsilon\nu(A),\quad\text{for all}\quad x\in C,$$

- Foster-Lyapunov Drift conditions,

$$KV\le\lambda V+b\mathbb{I}_C$$

and goal is to obtain a bound depending explicitly upon ϵ, λ, b, &c...

Coupling

If $X \sim \mu$ and $X' \sim \mu'$ and $\mu \wedge \mu' \geq \epsilon\nu$, one can construct two random variables $\tilde{X}$ and $\tilde{X}'$ such that

$$\tilde{X} \sim \mu, \tilde{X}' \sim \mu' \quad \text{and} \quad \tilde{X} = \tilde{X}' \quad \text{with probability } \epsilon$$

The **basic coupling construction**

- with probability ϵ, draw Z according to ν and set $\tilde{X} = \tilde{X}' = Z$.
- with probability $1 - \epsilon$, draw $\tilde{X}$ and $\tilde{X}'$ under distributions

$$(\mu - \epsilon\nu)/(1 - \epsilon) \quad \text{and} \quad (\mu' - \epsilon\nu)/(1 - \epsilon),$$

respectively.

[Thorisson, 2000]

X, X' r.v.'s with probability distribution $K(x, .)$ and $K(x', .)$, respectively, can be coupled with probability ϵ if:

$$K(x, \cdot) \wedge K(x', \cdot) \geq \epsilon\nu_{x,x'}(.)$$

where $\nu_{x,x'}$ is a probability measure, or, equivalently,

$$\|K(x, \cdot) - K(x', \cdot)\|_{\text{TV}} \leq (1 - \epsilon)$$

Define an **ϵ-coupling set** as a set $\bar{C} \subset \mathcal{X} \times \mathcal{X}$ satisfying :

$$\forall (x, x') \in \bar{C},\ \forall A \in \mathcal{B}(\mathcal{X}), \quad K(x, A) \wedge K(x', A) \geq \epsilon\nu_{x,x'}(A)$$

Small set and coupling sets

$C \subseteq \mathcal{X}$ **small set** if there exist $\epsilon > 0$ and a probability measure ν such that, for all $A \in \mathcal{B}(\mathcal{X})$

$$K(x, A) \geq \epsilon\nu(A), \qquad \forall x \in C. \tag{3}$$

Small sets always exist when the MC is φ-irreducible

[Jain and Jamieson, 1967]

For MCMC kernels, small sets in general easy to find.

If C is a small set, then $\bar{C} = C \times C$ is a coupling set:

$$\forall (x, x') \in \bar{C}, \forall A \in \mathcal{B}(\mathcal{X}), \quad K(x, A) \wedge K(x', A) \geq \epsilon\nu(A).$$

Coupling for Markov chains

$\bar{P}$ Markov transition kernel on $\mathcal{X} \times \mathcal{X}$ such that, for all $(x, x') \notin \bar{C}$ (where $\bar{C}$ is an ϵ-coupling set) and all $A \in \mathcal{B}(\mathcal{X})$:

$$\bar{P}(x, x'; A \times \mathcal{X}) = K(x, A) \quad \text{and} \quad \bar{P}(x, x'; \mathcal{X} \times A) = K(x', A)$$

For example,

- for $(x, x') \notin \bar{C}$, $\bar{P}(x, x'; A \times A') = K(x, A)K(x', A')$.
- For all $(x, x') \in \bar{C}$ and all $A, A' \in \mathcal{B}(\mathcal{X})$, define the **residual kernel**

$$\bar{R}(x, x'; A \times \mathcal{X}) = (1 - \epsilon)^{-1}(K(x, A) - \epsilon\nu_{x,x'}(A))$$
$$\bar{R}(x, x'; \mathcal{X} \times A') = (1 - \epsilon)^{-1}(K(x', A) - \epsilon\nu_{x,x'}(A')).$$

Coupling algorithm

- Initialisation Let $X_0 \sim \xi$ and $X'_0 \sim \xi'$ and set $d_0 = 0$.
- After coupling If $d_n = 1$, then draw $X_{n+1} \sim K(X_n, \cdot)$, and set $X'_{n+1} = X_{n+1}$.
- Before coupling If $d_n = 0$ and $(X_n, X'_n) \in \bar{C}$,
  - with probability ϵ, draw $X_{n+1} = X'_{n+1} \sim \nu_{X_n, X'_n}$ and set $d_{n+1} = 1$.
  - with probability $1 - \epsilon$, draw $(X_{n+1}, X'_{n+1}) \sim \bar{R}(X_n, X'_n; \cdot)$ and set $d_{n+1} = 0$.
  - If $d_n = 0$ and $(X_n, X'_n) \notin \bar{C}$, then draw $(X_{n+1}, X'_{n+1}) \sim \bar{P}(X_n, X'_n; \cdot)$.

(X_n, X'_n, d_n) [where d_n is the **bell variable** which indicates whether the chains have coupled or not] **is a Markov chain on** $(\mathcal{X} \times \mathcal{X} \times \{0, 1\})$**.**

Coupling inequality

Define the coupling time T as

$$T = \inf\{k \geq 1, d_k = 1\}$$

Coupling inequality

$$\sup_A |\xi P^k(A) - \xi' P^k(A)| \leq P_{\xi,\xi',0}[T > k]$$

[Pitman, 1976; Lindvall, 1992]

Drift conditions

To exploit the coupling construction, we need to control the hitting time

Moments of the return time to a set C are most often controlled using **Foster-Lyapunov drift condition**:

$$PV \leq \lambda V + b\mathbb{I}_C, \quad V \geq 1$$

$M_k = \lambda^{-k} V(X_k)\mathbb{I}(\tau_C \geq k), k \geq 1$ is a supermartingale and thus

$$\mathbb{E}_x[\lambda^{-\tau_C}] \leq V(x) + b\lambda^{-1}\mathbb{I}_C(x).$$

Conversely, if there exists a set C such that $\mathbb{E}_x[\lambda^{-\tau_C}] < \infty$ for all x (in a full and absorbing set), then there exists a drift function verifying the Foster-Lyapunov conditions.

[Meyn and Tweedie, 1993]

If the drift condition is imposed directly on the joint transition kernel $\bar{P}$, there exist $V \geq 1, 0 < \lambda < 1$ and a set $\bar{C}$ such that :

$$\bar{P}V(x, x') \leq \lambda V(x, x') \quad \forall (x, x') \notin \bar{C}$$

When $\bar{P}(x, x'; A \times A') = K(x, A)K(x', A')$, one may consider

$$\bar{V}(x, x') = (1/2)\,(V(x) + V(x'))$$

where V drift function for P (but not necessarily the best choice)

DMR'01 result

For any distributions ξ and ξ', and any $j \leq k$, then:

$$\|\xi P^k(\cdot) - \xi' P^k(\cdot)\|_{TV} \;\leq\; (1-\epsilon)^j + \lambda^k B^{j-1}\, \mathbb{E}_{\xi,\xi',0}[V(X_0, X_0')]$$

where

$$B = 1 \vee \lambda^{-1}(1-\epsilon) \sup_{\bar{C}} \overline{R}V.$$

4.8 Renewal and CLT

Given a Markov chain $(X_n)_n$, how good an approximation of

$$\mathfrak{I} = \int g(x)\pi(x)dx$$

is

$$\overline{g}_n := \frac{1}{n}\sum_{i=0}^{n-1} g(X_i) \;?$$

Standard MC if CLT

$$\sqrt{n}\,(\overline{g}_n - \mathbb{E}_\pi[g(X)]) \xrightarrow{d} \mathcal{N}(0, \gamma_g^2)$$

and there exists an easy-to-compute, consistent estimate of γ_g^2...

Minoration

Assume that the kernel density $\mathfrak{K}$ satisfies, for some density $\mathfrak{q}(\cdot)$, $\varepsilon \in (0,1)$ and a small set $C \subseteq \mathcal{X}$,

$$\mathfrak{K}(y|x) \geq \varepsilon\, \mathfrak{q}(y) \quad \text{for all} \;\; y \in \mathcal{X} \;\; \text{and} \;\; x \in C$$

Then split $\mathfrak{K}$ into a **mixture**

$$\mathfrak{K}(y|x) = \varepsilon\, \mathfrak{q}(y) + (1-\varepsilon)\, \mathfrak{R}(y|x)$$

where $\mathfrak{R}$ is **residual kernel**

Split chain

Let $\delta_0, \delta_1, \delta_2, \ldots$ be iid Ber(ε). Then the *split chain*

$$\{(X_0, \delta_0), (X_1, \delta_1), (X_2, \delta_2), \ldots\}$$

is such that, when $X_i \in C$, δ_i determines X_{i+1}:

$$X_{i+1} \sim \begin{cases} \mathfrak{q}(x) & \text{if } \delta_i = 1, \\ \mathfrak{R}(x|X_i) & \text{otherwise} \end{cases}$$

[Regeneration] **When** $(X_i, \delta_i) \in C \times \{1\}$, $X_{i+1} \sim \mathfrak{q}$

Renewals

For $X_0 \sim q$ and R successive renewals, define by $\tau_1 < \ldots < \tau_R$ the renewal times.

Then

$$\sqrt{R}\left(\overline{g}_{\tau_R} - \mathbb{E}_\pi[g(X)]\right) = \frac{\sqrt{R}}{\overline{N}}\left[\frac{1}{R}\sum_{t=1}^{R}(S_t - N_t\,\mathbb{E}_\pi[g(X)])\right]$$

where N_t length of the t th tour, and S_t sum of the $g(X_j)$'s over the t th tour.

Since (N_t, S_t) are iid and $\mathbb{E}_{\mathfrak{q}}[S_t - N_t\,\mathbb{E}_\pi[g(X)]] = 0$, if N_t and S_t have finite 2nd moments,

- $\sqrt{R}\left(\overline{g}_{\tau_R} - \mathbb{E}_\pi g\right) \xrightarrow{d} \mathcal{N}(0, \gamma_g^2)$
- there is a simple, consistent estimator of γ_g^2

[Mykland & al., 1995; Robert, 1995]

Moment conditions

We need to show that, for the minoration condition, $\mathbb{E}_{\mathfrak{q}}[N_1^2]$ and $\mathbb{E}_{\mathfrak{q}}[S_1^2]$ are finite.

If

1. **the chain is geometrically ergodic, and**
2. $\mathbb{E}_\pi[|g|^{2+\alpha}] < \infty$ **for some** $\alpha > 0$**,**

then $\mathbb{E}_{\mathfrak{q}}[N_1^2] < \infty$ **and** $\mathbb{E}_{\mathfrak{q}}[S_1^2] < \infty$**.**

[Hobert & al., 2002]

Note that drift + minoration ensures geometric ergodicity

[Rosenthal, 1995; Roberts & Tweedie, 1999]

5 The Metropolis-Hastings Algorithm

5.1 Monte Carlo Methods based on Markov Chains

Unnecessary to use a sample from the distribution f to approximate the integral

$$\int h(x)f(x)dx\ ,$$

Now we obtain $X_1, \ldots, X_n \sim f$ **(approx)** without directly simulating from f, **using an ergodic Markov chain with stationary distribution** f

Idea For an arbitrary starting value $x^{(0)}$, an ergodic chain $(X^{(t)})$ is generated using a transition kernel with stationary distribution f

- Insures the convergence in distribution of $(X^{(t)})$ to a random variable from f.
- For a "large enough" T_0, $X^{(T_0)}$ can be considered as distributed from f
- Produce a *dependent* sample $X^{(T_0)}, X^{(T_0+1)}, \ldots$, which is generated from f, sufficient for most approximation purposes.

5.2 The Metropolis–Hastings algorithm

5.2.1 Basics

The algorithm starts with the **objective (target) density**

$$f$$

A conditional density

$$q(y|x)$$

called the **instrumental (or proposal) distribution**, is then chosen.

Algorithm 25 –Metropolis–Hastings–

Given $x^{(t)}$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \text{with prob. } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

where

$$\rho(x, y) = \min \left\{ \frac{f(y)}{f(x)} \, \frac{q(x|y)}{q(y|x)} \, , 1 \right\} .$$

Features

- Always accept upwards moves
- Independent of normalizing constants for both f and $q(\cdot|x)$ (constants independent of x)
- Never move to values with $f(y) = 0$
- The chain $(x^{(t)})_t$ may take the same value several times in a row, even though f is a density wrt Lebesgue measure
- The sequence $(y_t)_t$ is usually **not** a Markov chain

5.2.2 Convergence properties

1 The M-H Markov chain is **reversible**, with invariant/stationary density f since it satisfies the **detailed balance condition**

$$\boxed{f(y)\,K(y,x) = f(x)\,K(x,y)}$$

2 As f is a probability measure, the chain is **positive recurrent**

3 If

$$\Pr\left[\frac{f(Y_t)\,q(X^{(t)}|Y_t)}{f(X^{(t)})\,q(Y_t|X^{(t)})} \geq 1\right] < 1. \qquad (1)$$

that is, the event $\{X^{(t+1)} = X^{(t)}\}$ is possible, then the chain is **aperiodic**

4 If

$$q(y|x) > 0 \text{ for every } (x,y), \qquad (2)$$

the chain is **irreducible**

5 For M-H, f-irreducibility implies **Harris recurrence**

6 Thus, for M-H satisfying (1) and (2)

(a) For h, with $\mathbb{E}_f|h(X)| < \infty$,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) = \int h(x)df(x) \qquad \text{a.e. } f.$$

(b) and

$$\lim_{n\to\infty} \left\| \int K^n(x,\cdot)\mu(dx) - f \right\|_{TV} = 0$$

for every initial distribution μ, where $K^n(x,\cdot)$ denotes the kernel for n transitions.

5.3 A Collection of Metropolis-Hastings Algorithms

5.3.1 The Independent Case

The instrumental distribution q is independent of $X^{(t)}$, and is denoted g by analogy with Accept-Reject.

Algorithm 26 –Independent Metropolis-Hastings–

Given $x^{(t)}$,

1 Generate $Y_t \sim g(y)$

2 Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \min\left\{\dfrac{f(Y_t)\,g(x^{(t)})}{f(x^{(t)})\,g(Y_t)}, 1\right\}, \\ x^{(t)} & \text{otherwise.} \end{cases}$$

The resulting sample is **not** iid

There can be strong convergence properties:

The algorithm produces a uniformly ergodic chain if there exists a constant M such that

$$f(x) \leq Mg(x)\,, \quad x \in \text{supp}\, f.$$

In this case,

$$\|K^n(x,\cdot) - f\|_{TV} \leq \left(1 - \frac{1}{M}\right)^n .$$

and the expected acceptance probability is at least $\frac{1}{M}$.

[Mengersen & Tweedie, 1996]

Example 27 **Noisy AR(1)**

Hidden Markov chain from a regular AR(1) model,

$$x_{t+1} = \varphi x_t + \epsilon_{t+1} \qquad \epsilon_t \sim \mathcal{N}(0, \tau^2)$$

and observables

$$y_t | x_t \sim \mathcal{N}(x_t^2, \sigma^2)$$

The distribution of x_t given x_{t-1}, x_{t+1} and y_t is

$$\exp \frac{-1}{2\tau^2} \left\{ (x_t - \varphi x_{t-1})^2 + (x_{t+1} - \varphi x_t)^2 + \frac{\tau^2}{\sigma^2} (y_t - x_t^2)^2 \right\} .$$

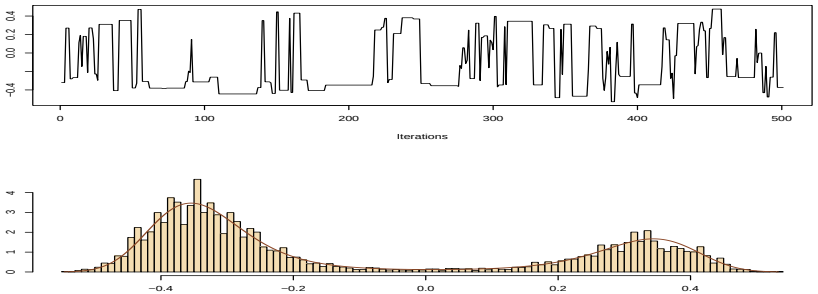
Use for proposal the $\mathcal{N}(\mu_t, \omega_t^2)$ distribution, with

$$\mu_t = \varphi \frac{x_{t-1} + x_{t+1}}{1 + \varphi^2} \quad \text{and} \quad \omega_t^2 = \frac{\tau^2}{1 + \varphi^2} .$$

Ratio

$$\pi(x)/q_{\text{ind}}(x) = \exp -(y_t - x_t^2)^2/2\sigma^2$$

is bounded



***(top)* Last 500 realisations of the chain $\{X_k\}_k$ out of $10,000$ iterations;**
***(bottom)* histogram of the chain, compared with the target distribution.**

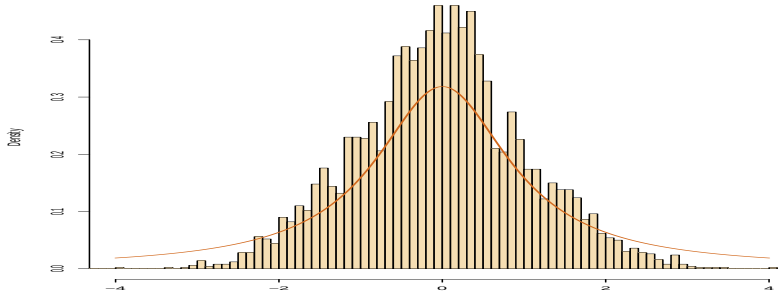
Example 28 –Cauchy by normal–

Given a Cauchy $\mathscr{C}(0,1)$ distribution, consider a normal $\mathscr{N}(0,1)$ proposal u.

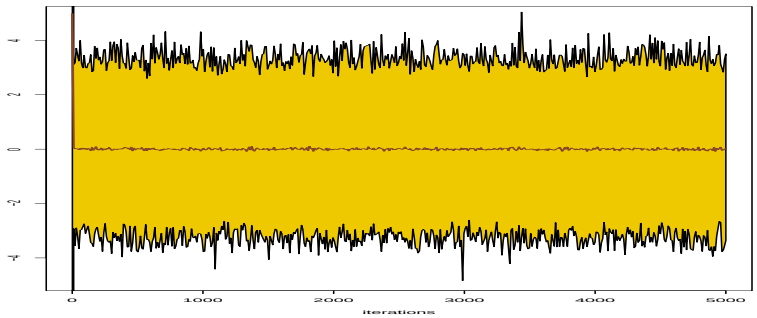
The Metropolis–Hastings acceptance ratio is

Poor perfomances: the proposal distribution has lighter tails than the target Cauchy and convergence to the stationary distribution is not even geometric!

[Mengersen & Tweedie, 1996]



Histogram of Markov chain $(\xi_t)_{1\leq t\leq 5000}$ against target $\mathscr{C}(0,1)$ distribution.



Range and average of 1000 parallel runs when initialized with a normal $(0,100^2)$ distribution.

5.3.2 Random walk Metropolis–Hastings

Use the proposal

$$Y_t = X^{(t)} + \varepsilon_t,$$

where $\varepsilon_t \sim g$, independent of $X^{(t)}$.

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

Algorithm 29 –Random walk Metropolis–

Given $x^{(t)}$

1 Generate $Y_t \sim g(y - x^{(t)})$

2 Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \min\left\{1, \dfrac{f(Y_t)}{f(x^{(t)})}\right\}, \\ x^{(t)} & \text{otherwise.} \end{cases}$$

Example 30 –Random walk normal–

Generate $\mathcal{N}(0,1)$ based on the uniform proposal $[-\delta, \delta]$

[Hastings (1970)]

The probability of acceptance is then

$$\rho(x^{(t)}, y_t) = \exp\{({x^{(t)}}^2 - y_t^2)/2\} \wedge 1.$$

Sample statistics

δ	0.1	0.5	1.0
mean	0.399	-0.111	0.10
variance	0.698	1.11	1.06

As $\delta \uparrow$, we get better histograms and a faster exploration of the support of f.

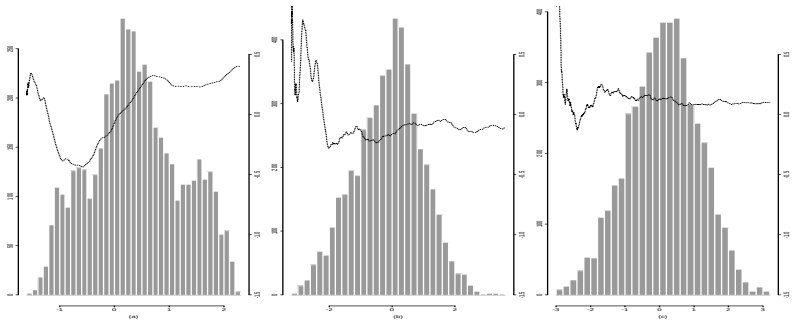


Figure 1: Three samples based on $\mathcal{U}[-\delta, \delta]$ with (a) $\delta = 0.1$, (b) $\delta = 0.5$ and (c) $\delta = 1.0$, superimposed with the convergence of the means ($15,000$ simulations).

Example 31 —Mixture models—

$$\pi(\theta|x) \propto \prod_{j=1}^{n} \left(\sum_{\ell=1}^{k} p_\ell f(x_j|\mu_\ell, \sigma_\ell) \right) \pi(\theta)$$

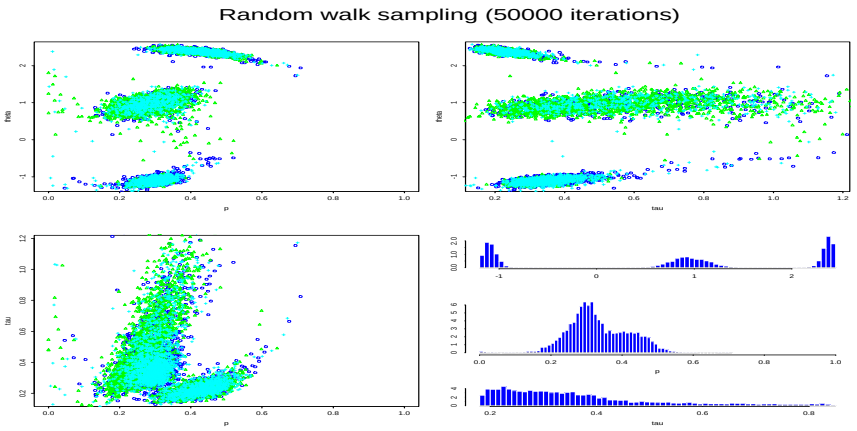
Metropolis-Hastings proposal:

$$\theta^{(t+1)} = \begin{cases} \theta^{(t)} + \omega\varepsilon^{(t)} & \text{if } u^{(t)} < \rho^{(t)} \\ \theta^{(t)} & \text{otherwise} \end{cases}$$

where

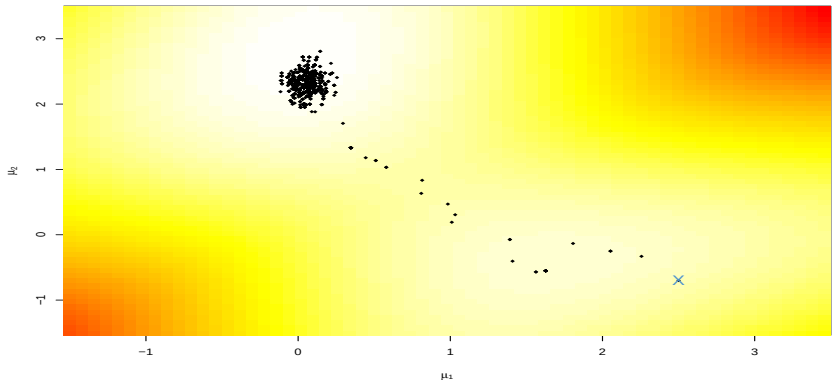
$$\rho^{(t)} = \frac{\pi(\theta^{(t)} + \omega\varepsilon^{(t)}|x)}{\pi(\theta^{(t)}|x)} \wedge 1$$

and ω scaled for good acceptance rate



[Celeux & al., 2000]

Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$



Example 32 Likelihood of the **probit model**

$$\prod_{i=1}^{n} \Phi(y_i^{\mathsf{T}}\beta)^{x_i}\, \Phi(-y_i^{\mathsf{T}}\beta)^{1-x_i}$$

[Observed likelihood]

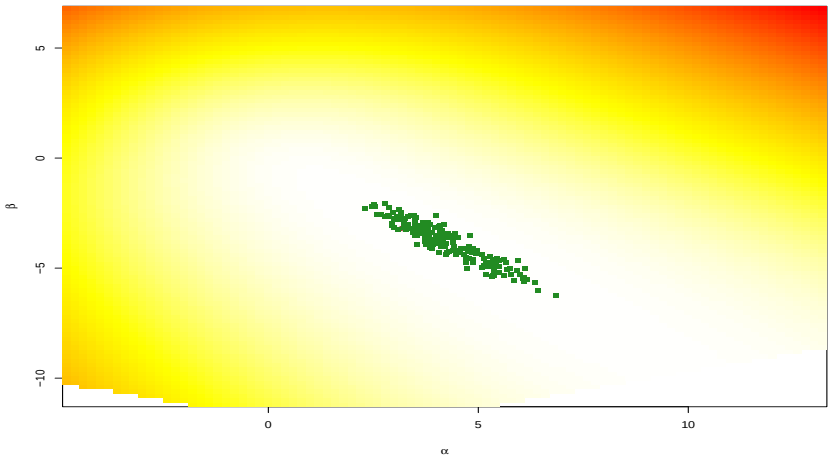
Random walk proposal

$$\beta^{(t+1)} = \beta^{(t)} + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}_p(0, \Sigma)$$

where, for instance,

$$\Sigma = \alpha(YY^{\mathsf{T}})^{-1}$$

Likelihood surface and random walk Metropolis-Hastings steps



Convergence properties

Uniform ergodicity prohibited by random walk structure

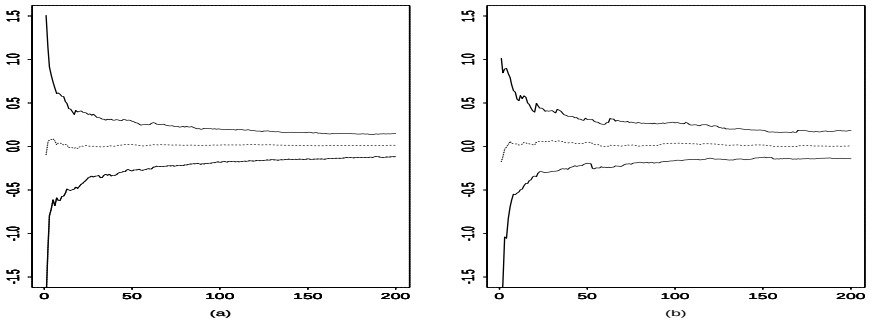
At best, geometric ergodicity:

For a symmetric density f, log-concave in the tails, and a positive and symmetric density g, the chain $(X^{(t)})$ is geometrically ergodic.

[Mengersen & Tweedie, 1996]

Example 33 Comparison of tail effects

Random-walk Metropolis–Hastings algorithms based on a $\mathcal{N}(0,1)$ instrumental for the generation of (a) a $\mathcal{N}(0,1)$ distribution and (b) a distribution with density $\psi(x) \propto (1+|x|)^{-3}$



90% confidence envelopes of the means, derived from 500 parallel independent chains

Example 34 –Cauchy by normal–

Again, Cauchy $\mathscr{C}(0,1)$ target and Gaussian random walk proposal, $\xi' \sim \mathscr{N}(\xi, \sigma^2)$, with acceptance probability

$$\frac{1+\xi^2}{1+(\xi')^2} \wedge 1\,,$$

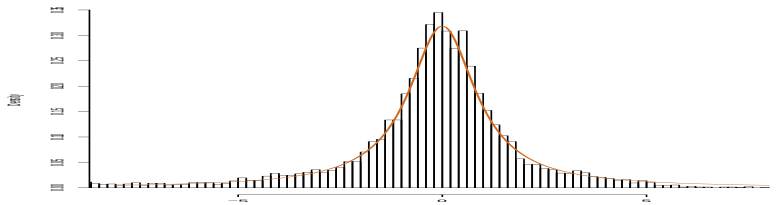
Overall fit of the Cauchy density by the histogram satisfactory, but poor exploration of the tails: 99% quantile of $\mathscr{C}(0,1)$ equal to 3, but no simulation exceeds 14 out of $10,000$!

[Roberts & Tweedie, 2004]

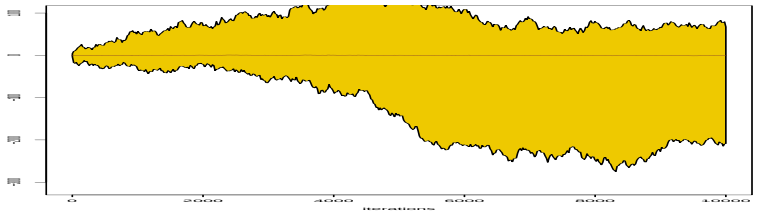
Again, lack of geometric ergodicity!

[Mengersen & Tweedie, 1996]

Slow convergence shown by the non-stable range after $10,000$ iterations.



Histogram of the $10,000$ first steps of a random walk Metropolis–Hastings algorithm using a $\mathcal{N}(\xi, 1)$ proposal



Range of 500 parallel runs for the same setup

Further convergence properties

Under assumptions

- **(A1)** f is super-exponential, *i.e.* it is positive with positive continuous first derivative such that $\lim_{|x|\to\infty} n(x)'\nabla \log f(x) = -\infty$ where $n(x) := x/|x|$.

  In words : exponential decay of f in every direction with rate tending to ∞

- **(A2)** $\limsup_{|x|\to\infty} n(x)'m(x) < 0$, where $m(x) = \nabla f(x)/|\nabla f(x)|$.

  In words: non degeneracy of the countour manifold $\mathcal{C}_{f(y)} = \{y : f(y) = f(x)\}$

Q is geometrically ergodic, and
$V(x) \propto f(x)^{-1/2}$ verifies the drift condition

[Jarner & Hansen, 2000]

Further [further] convergence properties

If P ψ-irreducible and aperiodic, for $r = (r(n))_{n\in\mathbb{N}}$ real-valued non decreasing sequence, such that, for all $n, m \in \mathbb{N}$,

$$r(n+m) \leq r(n)r(m),$$

and $r(0) = 1$, for C a small set, $\tau_C = \inf\{n \geq 1, X_n \in C\}$, and $h \geq 1$, assume

$$\sup_{x\in C} \mathbb{E}_x \left[\sum_{k=0}^{\tau_C - 1} r(k)h(X_k) \right] < \infty,$$

then,

$$S(f, C, r) := \left\{ x \in X, \mathbb{E}_x \left\{ \sum_{k=0}^{\tau_C - 1} r(k)h(X_k) \right\} < \infty \right\}$$

is full and absorbing and for $x \in S(f, C, r)$,

$$\lim_{n\to\infty} r(n) \|P^n(x, .) - f\|_h = 0.$$

[Tuominen & Tweedie, 1994]

Comments

[CLT, Rosenthal's inequality...] h-ergodicity implies CLT for additive (possibly unbounded functionals) of the chain (under additional conditions, guaranteeing the integrability of the limit), Rosenthal's inequality (also for functions whose growth at infinity is controlled properly) and so on...

[Control of the moments of the return-time] The condition implies (because $h \geq 1$) that

$$\sup_{x\in C} \mathbb{E}_x[r_0(\tau_C)] \leq \sup_{x\in C} \mathbb{E}_x \left\{ \sum_{k=0}^{\tau_C - 1} r(k)h(X_k) \right\} < \infty, \text{ where } r_0(n) = \sum_{l=0}^{n} r(l)$$

Can be used to derive bounds for the coupling time, an essential step to determine computable bounds, using coupling inequalities

[Roberts & Tweedie, 1998; Fort & Moulines, 2000]

Alternative conditions

The condition is not really easy to work with...

[Possible alternative conditions]

(a) **[Tuominen, Tweedie, 1994]** There exists a sequence $(V_n)_{n\in\mathbb{N}}$, $V_n \geq r(n)h$, such that

(i) $\sup_C V_0 < \infty$,

(ii) $\{V_0 = \infty\} \subset \{V_1 = \infty\}$ and

(iii) $PV_{n+1} \leq V_n - r(n)h + br(n)\mathbb{I}_C$.

(b) **[Fort 2000]** $\exists V \geq f \geq 1$ and $b < \infty$, such that $\sup_C V < \infty$ and

$$PV(x) + \mathbb{E}_x \left\{ \sum_{k=0}^{\sigma_C} \Delta r(k) f(X_k) \right\} \leq V(x) + b\mathbb{I}_C(x)$$

where σ_C is the hitting time on C and

$$\Delta r(k) = r(k) - r(k-1), k \geq 1 \text{ and } \Delta r(0) = r(0).$$

Result (a) $\Leftrightarrow$ (b) $\Leftrightarrow$ $\sup_{x \in C} \mathbb{E}_x \left\{ \sum_{k=0}^{\tau_C - 1} r(k) f(X_k) \right\} < \infty.$

5.4 Extensions

There are many other algorithms

- *Adaptive Rejection Metropolis Sampling*
- *Reversible Jump (later!)*
- *Langevin algorithms*

to name a few...

5.4.1 Langevin Algorithms

Proposal based on the *Langevin diffusion* L_t is defined by the stochastic differential equation

$$dL_t = dB_t + \frac{1}{2} \nabla \log f(L_t) dt,$$

where B_t is the standard *Brownian motion*

The Langevin diffusion is the only non-explosive diffusion which is reversible with respect to f.

Discretization:

$$x^{(t+1)} = x^{(t)} + \frac{\sigma^2}{2} \nabla \log f(x^{(t)}) + \sigma \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}_p(0, I_p)$$

where σ^2 corresponds to the discretization

Unfortunately, the discretized chain may be be transient, for instance when

$$\lim_{x \to \pm\infty} \left| \sigma^2 \nabla \log f(x) |x|^{-1} \right| > 1$$

MH correction

Accept the new value Y_t with probability

$$\frac{f(Y_t)}{f(x^{(t)})} \cdot \frac{\exp\left\{-\left\|Y_t - x^{(t)} - \frac{\sigma^2}{2}\nabla \log f(x^{(t)})\right\|^2 \Big/ 2\sigma^2\right\}}{\exp\left\{-\left\|x^{(t)} - Y_t - \frac{\sigma^2}{2}\nabla \log f(Y_t)\right\|^2 \Big/ 2\sigma^2\right\}} \wedge 1 \,.$$

Choice of the scaling factor σ

Should lead to an acceptance rate of 0.574 to achieve optimal convergence rates (when the components of x are uncorrelated)

[Roberts & Rosenthal, 1998]

5.4.2 Optimizing the Acceptance Rate

Problem of choice of the transition kernel from a practical point of view

Most common alternatives:

(a) a fully automated algorithm like ARMS;

(b) an instrumental density g which approximates f, such that f/g is bounded for uniform ergodicity to apply;

(c) a random walk

In both cases (b) and (c), the choice of g is critical,

Case of the independent Metropolis–Hastings algorithm

Choice of g that maximizes the average acceptance rate

$$\begin{aligned}\rho &= \mathbb{E}\left[\min\left\{\frac{f(Y)\,g(X)}{f(X)\,g(Y)}, 1\right\}\right] \\ &= 2P\left(\frac{f(Y)}{g(Y)} \geq \frac{f(X)}{g(X)}\right), \qquad X \sim f,\ Y \sim g,\end{aligned}$$

Related to the speed of convergence of

$$\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$$

to $\mathbb{E}_f[h(X)]$ and to the ability of the algorithm to explore any complexity of f

Practical implementation

Choose a parameterized instrumental distribution $g(\cdot|\theta)$ and adjusting the corresponding parameters θ based on the evaluated acceptance rate

$$\hat{\rho}(\theta) = \frac{2}{m}\sum_{i=1}^{m} \mathbb{I}_{\{f(y_i)g(x_i) > f(x_i)g(y_i)\}} \,,$$

where $x_1, \ldots, x_m$ sample from f and $y_1, \ldots, y_m$ iid sample from g.

Example 35 **Inverse Gaussian distribution.**

Simulation from

$$f(z|\theta_1,\theta_2) \propto z^{-3/2} \exp\left\{-\theta_1 z - \frac{\theta_2}{z} + 2\sqrt{\theta_1\theta_2} + \log \sqrt{2\theta_2}\right\} \mathbb{I}_{\mathbb{R}_+}(z)$$

based on the Gamma distribution $\mathcal{G}a(\alpha,\beta)$ with $\alpha = \beta\sqrt{\theta_2/\theta_1}$

Since

$$\frac{f(x)}{g(x)} \propto x^{-\alpha-1/2} \exp\left\{(\beta-\theta_1)x - \frac{\theta_2}{x}\right\} ,$$

the maximum is attained at

$$x^*_\beta = \frac{(\alpha+1/2) - \sqrt{(\alpha+1/2)^2 + 4\theta_2(\theta_1-\beta)}}{2(\beta-\theta_1)} .$$

The analytical optimization (in β) of

$$M(\beta) = (x^*_\beta)^{-\alpha-1/2} \exp\left\{(\beta-\theta_1)x^*_\beta - \frac{\theta_2}{x^*_\beta}\right\}$$

is impossible

β	0.2	0.5	0.8	0.9	1	1.1	1.2	1.5
$\hat{\rho}(\beta)$	0.22	0.41	0.54	0.56	0.60	0.63	0.64	0.71
$\mathbb{E}[Z]$	1.137	1.158	1.164	1.154	1.133	1.148	1.181	1.148
$\mathbb{E}[1/Z]$	1.116	1.108	1.116	1.115	1.120	1.126	1.095	1.115

($\theta_1 = 1.5, \theta_2 = 2$, and $m = 5000$).

Case of the random walk

Different approach to acceptance rates

A high acceptance rate does not indicate that the algorithm is moving correctly since it indicates that the random walk is moving too slowly on the surface of f.

If $x^{(t)}$ and y_t are close, i.e. $f(x^{(t)}) \simeq f(y_t)$ y is accepted with probability

$$\min\left(\frac{f(y_t)}{f(x^{(t)})}, 1\right) \simeq 1 .$$

For multimodal densities with well separated modes, the negative effect of limited moves on the surface of f clearly shows.

If the average acceptance rate is low, the successive values of $f(y_t)$ tend to be small compared with $f(x^{(t)})$, which means that the random walk moves quickly on the surface of f since it often reaches the "borders" of the support of f

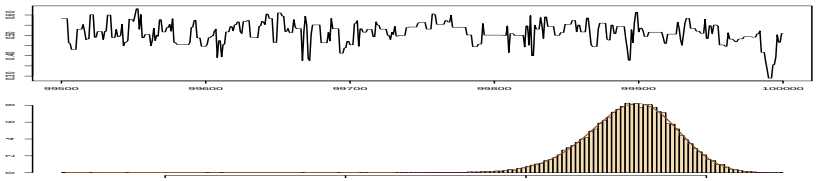
Rule of thumb

In small dimensions, aim at an average acceptance rate of 50%. In large dimensions, at an average acceptance rate of 25%.

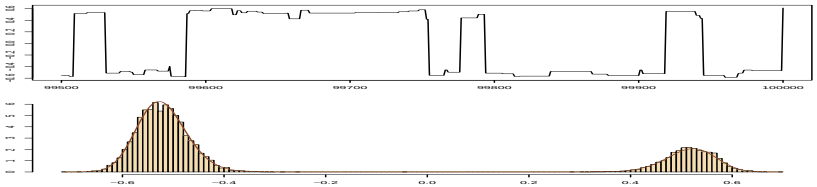
[Gelman,Gilks and Roberts, 1995]

Example 36 **Noisy AR(1) cont'd**

For a Gaussian random walk with scale ω small enough, the random walk never jumps to the other mode. But if the scale ω is sufficiently large, the Markov chain explores both modes and give a satisfactory approximation of the target distribution.



Markov chain based on a random walk with scale $\omega = .1$.



Markov chain based on a random walk with scale $\omega = .5$.

6 The Gibbs Sampler

6.1 General Principles

A very **specific** simulation algorithm based on the target distribution f:

1 Uses the conditional densities $f_1, \ldots, f_p$ from f

2 Start with the random variable $\mathbf{X} = (X_1, \ldots, X_p)$

3 Simulate from the conditional densities,

$$X_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \sim f_i(x_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

for $i = 1, 2, \ldots, p$.

Algorithm 37 –The Gibbs sampler–

Given $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, generate

1. $X_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \ldots, x_p^{(t)})$;
2. $X_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$,

   $\ldots$

p. $X_p^{(t+1)} \sim f_p(x_p | x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$

Then $\mathbf{X}^{(t+1)} \to \mathbf{X} \sim f$

Properties

The full conditionals densities $f_1, \ldots, f_p$ are the only densities used for simulation. Thus, even in a high dimensional problem, all of the simulations may be univariate

The Gibbs sampler **is not reversible** with respect to f. However, each of its p components is. Besides, it can be turned into a reversible sampler, either using the *Random Scan Gibbs sampler* (see below) or running instead the (double) sequence

$$f_1 \cdots f_{p-1} f_p f_{p-1} \cdots f_1$$

Example 38 –Bivariate Gibbs sampler–

$$(X, Y) \sim f(x, y)$$

Generate a sequence of observations by

Set $X_0 = x_0$

For $t = 1, 2, \ldots,$ generate

$$\begin{aligned} Y_t &\sim f_{Y|X}(\cdot | x_{t-1}) \\ X_t &\sim f_{X|Y}(\cdot | y_t) \end{aligned}$$

where $f_{Y|X}$ and $f_{X|Y}$ are the conditional distributions

- $(X_t, Y_t)_t$, is a Markov chain
- $(X_t)_t$ and $(Y_t)_t$ individually **are Markov chains**
- For example, the chain $(X_t)_t$ has transition density

$$K(x, x^*) = \int f_{Y|X}(y|x) f_{X|Y}(x^*|y) dy,$$

with invariant density $f_X(\cdot)$

For the special case

$$(X, Y) \sim \mathcal{N}_2 \left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) ,$$

the Gibbs sampler is

Given y_t, generate

$$\begin{aligned} X_{t+1} \mid y_t &\sim \mathcal{N}(\rho y_t,\, 1 - \rho^2) \,, \\ Y_{t+1} \mid x_{t+1} &\sim \mathcal{N}(\rho x_{t+1},\, 1 - \rho^2). \end{aligned}$$

Properties of the Gibbs sampler

Formally, a special case of a sequence of 1-D M-H kernels, all with acceptance rate uniformly equal to 1.

The Gibbs sampler

1. limits the choice of instrumental distributions
2. requires some knowledge of f
3. is, by construction, multidimensional
4. does not apply to problems where the number of parameters varies as the resulting chain is not irreducible.

6.1.1 Completion

The Gibbs sampler can be generalized in much wider generality

A density g is a completion of f if

$$\int_{\mathcal{Z}} g(x, z)\, dz = f(x)$$

Purpose g should have full conditionals that are easy to simulate for a Gibbs sampler to be implemented with g rather than f

For $p > 1$, write $y = (x, z)$ and denote the conditional densities of $g(y) = g(y_1, \ldots, y_p)$ by

$$\begin{aligned} Y_1|y_2, \ldots, y_p &\sim g_1(y_1|y_2, \ldots, y_p), \\ Y_2|y_1, y_3, \ldots, y_p &\sim g_2(y_2|y_1, y_3, \ldots, y_p), \\ &\ldots, \\ Y_p|y_1, \ldots, y_{p-1} &\sim g_p(y_p|y_1, \ldots, y_{p-1}). \end{aligned}$$

The move from $Y^{(t)}$ to $Y^{(t+1)}$ is defined as follows:

Algorithm 39 –Completion Gibbs sampler–

Given $(y_1^{(t)}, \ldots, y_p^{(t)})$, simulate

1 $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t)}, \ldots, y_p^{(t)})$,

2 $Y_2^{(t+1)} \sim g_2(y_2|y_1^{(t+1)}, y_3^{(t)}, \ldots, y_p^{(t)})$,

...

p. $Y_p^{(t+1)} \sim g_p(y_p|y_1^{(t+1)}, \ldots, y_{p-1}^{(t+1)})$.

Example 40 —Mixtures all over again—

Hierarchical missing data structure

If

$$X_1, \ldots, X_n \sim \sum_{i=1}^{k} p_i f(x|\theta_i),$$

then

$$X|Z \sim f(x|\theta_Z), \quad Z \sim p_1 \mathbb{I}(z = 1) + \ldots + p_k \mathbb{I}(z = k),$$

and Z is the component indicator associated with observation x

Conditionally on $(Z_1, \ldots, Z_n) = (z_1, \ldots, z_n)$:

$$\begin{aligned} &\pi(p_1, \ldots, p_k, \theta_1, \ldots, \theta_k|x_1, \ldots, x_n, z_1, \ldots, z_n) \\ &\propto p_1^{\alpha_1+n_1-1} \ldots p_k^{\alpha_k+n_k-1} \\ &\quad \times \pi(\theta_1|y_1 + n_1\bar{x}_1, \lambda_1 + n_1) \ldots \pi(\theta_k|y_k + n_k\bar{x}_k, \lambda_k + n_k), \end{aligned}$$

with

$$n_i = \sum_j \mathbb{I}(z_j = i) \quad \text{et} \quad \bar{x}_i = \sum_{j;\, z_j=i} x_j/n_i.$$

Corresponding Gibbs sampler

1. Simulate

$$\theta_i \sim \pi(\theta_i|y_i + n_i\bar{x}_i, \lambda_i + n_i) \quad (i = 1, \ldots, k)$$
$$(p_1, \ldots, p_k) \sim D(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$$

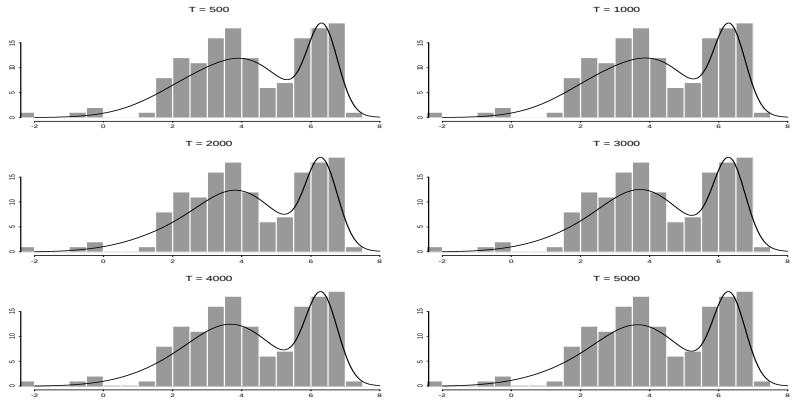
2. Simulate $(j = 1, \ldots, n)$

$$Z_j|x_j, p_1, \ldots, p_k, \theta_1, \ldots, \theta_k \sim \sum_{i=1}^{k} p_{ij}\mathbb{I}(z_j = i)$$

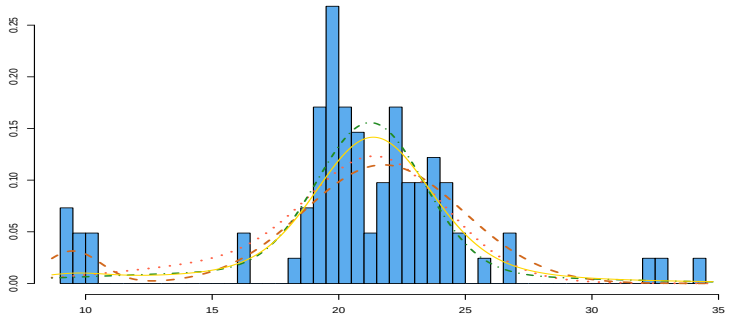
with $(i = 1, \ldots, k)$

$$p_{ij} \propto p_i f(x_j|\theta_i)$$

and update n_i and $\bar{x}_i$ $(i = 1, \ldots, k)$.



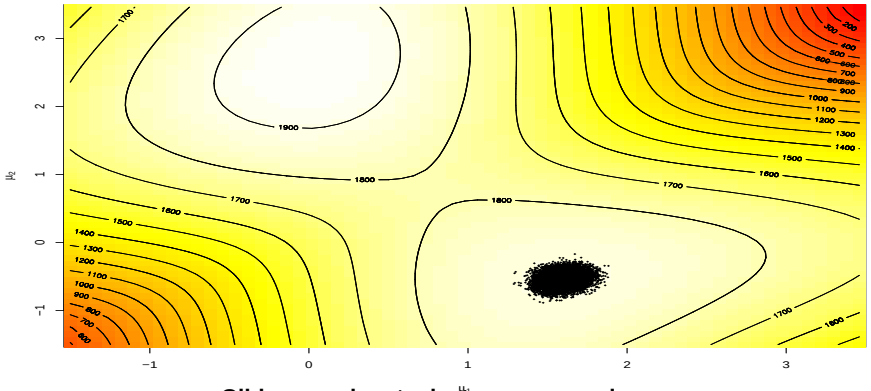
Estimation of the pluggin density for 3 components and T iterations for 149 observations of acidity levels in lakes in the American North-East



Galaxy dataset (82 observations) with $k = 2$ components

average density (yellow), and pluggins:

average (tomato), marginal MAP (green), MAP (marroon)



Gibbs sampler stuck at a wrong mode

6.1.2 Random Scan Gibbs sampler

Modification of the above Gibbs sampler where, with probability $1/p$, the i-th component is drawn from $f_i(x_i|X_{-i})$

The Random Scan Gibbs sampler is reversible.

6.1.3 Slice sampler

If $f(\theta)$ can be written as a product

$$\prod_{i=1}^{k} f_i(\theta),$$

it can be completed

$$\prod_{i=1}^{k} \mathbb{I}_{0\le\omega_i\le f_i(\theta)},$$

leading to the following Gibbs algorithm:

Algorithm 41 –Slice sampler–

Simulate

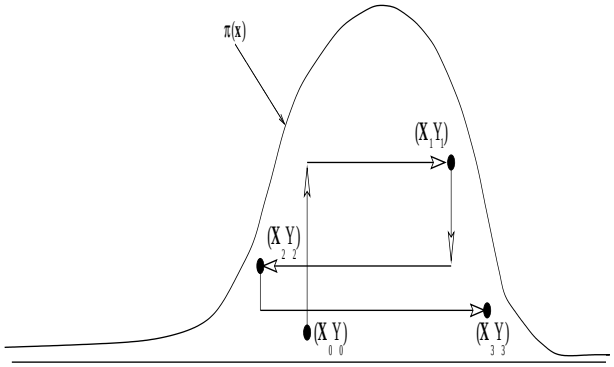
1. $\omega_1^{(t+1)} \sim \mathcal{U}_{[0,f_1(\theta^{(t)})]}$;

   ...

k. $\omega_k^{(t+1)} \sim \mathcal{U}_{[0,f_k(\theta^{(t)})]}$;

k+1. $\theta^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$, with

$$A^{(t+1)} = \{y;\ f_i(y) \ge \omega_i^{(t+1)},\ i = 1,\dots,k\}.$$



Representation of a few steps of the slice sampler

[Roberts & Rosenthal, 1998]

The slice sampler usually enjoys good theoretical properties (like geometric ergodicity).

As k increases, the determination of the set $A^{(t+1)}$ may get increasingly complex.

Example 42 **Stochastic volatility core distribution**

Difficult part of the stochastic volatility model

$$\pi(x) \propto \exp - \left\{\sigma^2(x-\mu)^2 + \beta^2 \exp(-x) y^2 + x\right\}/2\,,$$

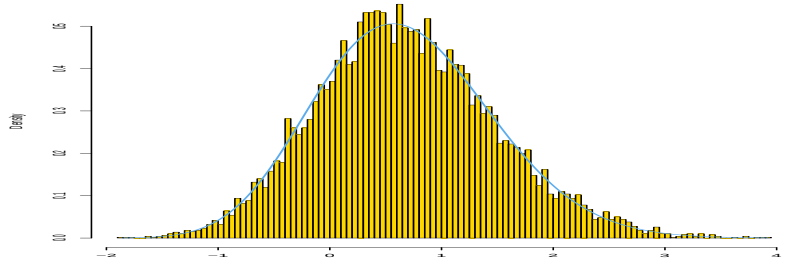
simplified in $\exp - \left\{x^2 + \alpha \exp(-x)\right\}$

Slice sampling means simulation from a uniform distribution on

$$\mathfrak{A} = \left\{x; \exp - \left\{x^2 + \alpha \exp(-x)\right\}/2 \geq u\right\} = \left\{x; x^2 + \alpha \exp(-x) \leq \omega\right\}$$

if we set $\omega = -2 \log u$.

Inversion of $x^2 + \alpha \exp(-x) = \omega$ needs to be done by trial-and-error.



Histogram of a Markov chain produced by a slice sampler and target distribution in overlay.

6.1.4 Properties of the Gibbs sampler

$$(Y_1, Y_2, \cdots, Y_p) \sim g(y_1, \ldots, y_p)$$

If either

(i) $g^{(i)}(y_i) > 0$ for every $i = 1, \cdots, p$, implies that $g(y_1, \ldots, y_p) > 0$, where $g^{(i)}$ denotes the marginal distribution of Y_i, or

[Positivity condition]

(ii) the transition kernel is absolutely continuous with respect to g,

then the chain is *irreducible* and *positive Harris recurrent*.

(i) If $\int h(y)g(y)dy < \infty$, then

$$\lim_{nT\to\infty} \frac{1}{T}\sum_{t=1}^{T} h_1(Y^{(t)}) = \int h(y)g(y)dy \text{ a.e. } g.$$

(ii) If, in addition, $(Y^{(t)})$ is aperiodic, then

$$\lim_{n\to\infty} \left\| \int K^n(y,\cdot)\mu(dx) - f \right\|_{TV} = 0$$

for every initial distribution μ.

Slice sampler

Properties of X_t and of $f(X_t)$ identical

If f is bounded and $\text{supp} f$ is bounded, the simple slice sampler is uniformly ergodic.

[Mira & Tierney, 1997]

For $\epsilon^\star > \epsilon_\star$,

$$C = \{x \in \mathcal{X};\ \epsilon_\star < f(x) < \epsilon^\star\}$$

is a **small set**:

$$\Pr(x,\cdot) \geq \frac{\epsilon_\star}{\epsilon^\star}\,\mu(\cdot)$$

where

$$\mu(A) = \frac{1}{\epsilon_\star}\int_0^{\epsilon_\star} \frac{\lambda(A \cap L(\epsilon))}{\lambda(L(\epsilon))} d\epsilon$$

if $L(\epsilon) = \{x \in \mathcal{X}; f(x) > \epsilon\}$‘

[Roberts & Rosenthal, 1998]

Slice sampler: drift

Under some differentiability and monotonicity conditions, the slice sampler also verifies a drift condition with $V(x) = f(x)^{-\beta}$, is geometrically ergodic, and there exist explicit bounds on the total variation distance

[Roberts & Rosenthal, 1998]

Example 43 —Exponential $\mathcal{E}xp(1)$—

For $n > 23$,

$$||K^n(x,\cdot) - f(\cdot)||_{TV} \leq .054865\,(0.985015)^n\,(n - 15.7043)$$

For *any density* such that

$$\epsilon \frac{\partial}{\partial \epsilon} \lambda\left(\{x \in \mathcal{X};\, f(x) > \epsilon\}\right) \quad \text{is non-increasing}$$

then

$$||K^{523}(x,\cdot) - f(\cdot)||_{TV} \leq .0095$$

[Roberts & Rosenthal, 1998]

Example 44 —A poor slice sampler—

Consider

$$f(x) = \exp\{-||x||\} \qquad x \in \mathbb{R}^d$$

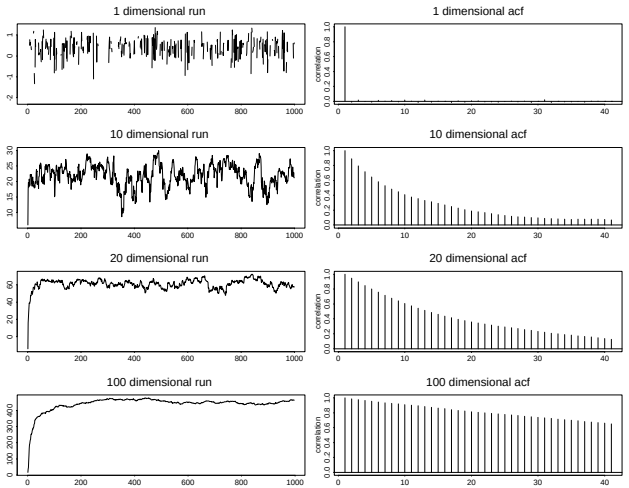
Slice sampler equivalent to one-dimensional slice sampler on

$$\pi(z) = z^{d-1}\, e^{-z} \qquad z > 0$$

or on

$$\pi(u) = e^{-u^{1/d}} \qquad u > 0$$

Poor performances when d large (heavy tails)



Sample runs of $\log(u)$ and ACFs for $\log(u)$ (Roberts & Rosenthal, 1999)

6.1.5 Hammersley-Clifford Theorem

An illustration that conditionals determine the joint distribution

If the joint density $g(y_1, y_2)$ have conditional distributions $g_1(y_1|y_2)$ and $g_2(y_2|y_1)$, then

$$g(y_1, y_2) = \frac{g_2(y_2|y_1)}{\int g_2(v|y_1)/g_1(y_1|v)\, dv}.$$

General case

Under the positivity condition, the joint distribution g satisfies

$$g(y_1, \ldots, y_p) \propto \prod_{j=1}^{p} \frac{g_{\ell_j}(y_{\ell_j}|y_{\ell_1}, \ldots, y_{\ell_{j-1}}, y'_{\ell_{j+1}}, \ldots, y'_{\ell_p})}{g_{\ell_j}(y'_{\ell_j}|y_{\ell_1}, \ldots, y_{\ell_{j-1}}, y'_{\ell_{j+1}}, \ldots, y'_{\ell_p})}$$

for every permutation ℓ on $\{1, 2, \ldots, p\}$ and every $y' \in \mathcal{Y}$.

6.1.6 Hierarchical models

The Gibbs sampler is particularly well suited to *hierarchical models*

Example 45 **–Hierarchical models in animal epidemiology–**

Counts of the number of cases of clinical mastitis in 127 dairy cattle herds over a one year period.

Number of cases in herd i

$$X_i \sim \mathcal{P}(\lambda_i) \qquad i = 1, \cdots, m$$

where λ_i is the underlying rate of infection in herd i

Lack of independence might manifest itself as overdispersion.

Modified model

$$\begin{aligned} X_i &\sim \mathcal{P}(\lambda_i) \\ \lambda_i &\sim \mathcal{G}a(\alpha, \beta_i) \\ \beta_i &\sim \mathcal{IG}(a, b), \end{aligned}$$

The Gibbs sampler corresponds to conditionals

$$\begin{aligned} \lambda_i &\sim \pi(\lambda_i|\mathbf{x}, \alpha, \beta_i) = \mathcal{G}a(x_i + \alpha, [1 + 1/\beta_i]^{-1}) \\ \beta_i &\sim \pi(\beta_i|\mathbf{x}, \alpha, a, b, \lambda_i) = \mathcal{IG}(\alpha + a, [\lambda_i + 1/b]^{-1}) \end{aligned}$$

Example 46 Experiment where rats are intoxicated by a substance, then treated by either a placebo or a drug:

$$\begin{aligned} x_{ij} &\sim \mathcal{N}(\theta_i, \sigma_c^2), & 1 \le j \le J_i^c\,, & \quad \text{control} \\ y_{ij} &\sim \mathcal{N}(\theta_i + \delta_i, \sigma_a^2), & 1 \le j \le J_i^a\,, & \quad \text{intoxication} \\ z_{ij} &\sim \mathcal{N}(\theta_i + \delta_i + \xi_i, \sigma_t^2), & 1 \le j \le J_i^t\,, & \quad \text{treatment} \end{aligned}$$

Additional variable w_i, equal to 1 if the rat is treated with the drug, and 0 otherwise.

Prior distributions $(1 \le i \le I)$,

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \qquad \delta_i \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2),$$

and

$$\xi_i \sim \mathcal{N}(\mu_P, \sigma_P^2) \quad \text{or} \quad \xi_i \sim \mathcal{N}(\mu_D, \sigma_D^2),$$

depending on whether the ith rat is treated with a placebo or a drug.

Hyperparameters of the model,

$$\mu_\theta, \mu_\delta, \mu_P, \mu_D, \sigma_c, \sigma_a, \sigma_t, \sigma_\theta, \sigma_\delta, \sigma_P, \sigma_D\,,$$

associated with Jeffreys' noninformative priors.

Alternative prior

$$\delta_i \sim p\mathcal{N}(\mu_{\delta 1}, \sigma_{\delta 1}^2) + (1-p)\mathcal{N}(\mu_{\delta 2}, \sigma_{\delta 2}^2),$$

allows for two possible levels of intoxication.

Conditional decompositions

Easy decomposition of the posterior distribution

For instance, if

$$\theta|\theta_1 \sim \pi_1(\theta|\theta_1), \qquad \theta_1 \sim \pi_2(\theta_1),$$

then

$$\pi(\theta|x) = \int_{\Theta_1} \pi(\theta|\theta_1, x)\pi(\theta_1|x)\, d\theta_1,$$

where

$$\begin{aligned} \pi(\theta|\theta_1, x) &= \frac{f(x|\theta)\pi_1(\theta|\theta_1)}{m_1(x|\theta_1)}, \\ m_1(x|\theta_1) &= \int_\Theta f(x|\theta)\pi_1(\theta|\theta_1)\, d\theta, \\ \pi(\theta_1|x) &= \frac{m_1(x|\theta_1)\pi_2(\theta_1)}{m(x)}, \\ m(x) &= \int_{\Theta_1} m_1(x|\theta_1)\pi_2(\theta_1)\, d\theta_1. \end{aligned}$$

Moreover, this decomposition works for the posterior moments, that is, for every function h,

$$\mathbb{E}^{\pi}[h(\theta)|x] = \mathbb{E}^{\pi(\theta_1|x)}\left[\mathbb{E}^{\pi_1}\left[h(\theta)|\theta_1, x\right]\right],$$

where

$$\mathbb{E}^{\pi_1}[h(\theta)|\theta_1, x] = \int_{\Theta} h(\theta)\pi(\theta|\theta_1, x)\, d\theta.$$

Example 47 **(Example 46 continued)** The posterior distribution of the complete parameter vector is given by

$$\begin{aligned}
&\pi((\theta_i, \delta_i, \xi_i)_i, \mu_\theta, \ldots, \sigma_c, \ldots|\mathcal{D}) \propto \\
&\prod_{i=1}^{I}\left\{\exp-\{(\theta_i-\mu_\theta)^2/2\sigma_\theta^2 + (\delta_i-\mu_\delta)^2/2\sigma_\delta^2\}\right. \\
&\prod_{j=1}^{J_i^c}\exp-\{(x_{ij}-\theta_i)^2/2\sigma_c^2\}\prod_{j=1}^{J_i^a}\exp-\{(y_{ij}-\theta_i-\delta_i)^2/2\sigma_a^2\} \\
&\left.\prod_{j=1}^{J_i^t}\exp-\{(z_{ij}-\theta_i-\delta_i-\xi_i)^2/2\sigma_t^2\}\right\} \\
&\prod_{\ell_i=0}\exp-\{(\xi_i-\mu_P)^2/2\sigma_P^2\}\prod_{\ell_i=1}\exp-\{(\xi_i-\mu_D)^2/2\sigma_D^2\} \\
&\sigma_c^{-\sum_i J_i^c-1}\sigma_a^{-\sum_i J_i^a-1}\sigma_t^{-\sum_i J_i^t-1}(\sigma_\theta\sigma_\delta)^{-I-1}\sigma_D^{-I_D-1}\sigma_P^{-I_P-1},
\end{aligned}$$

Local conditioning property

For the hierarchical model

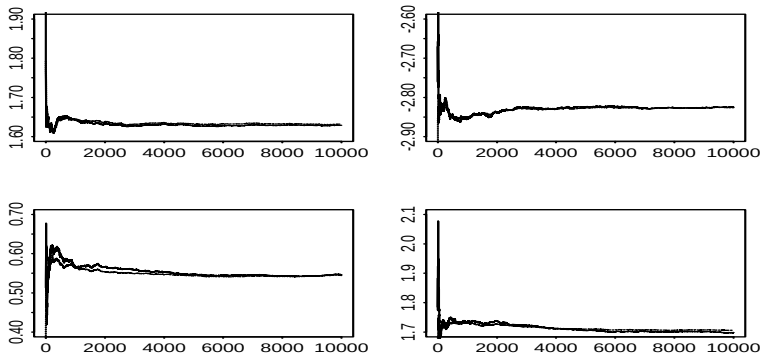
$$\pi(\theta) = \int_{\Theta_1\times\ldots\times\Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\cdots\pi_{n+1}(\theta_n)\, d\theta_1\cdots d\theta_{n+1}.$$

we have

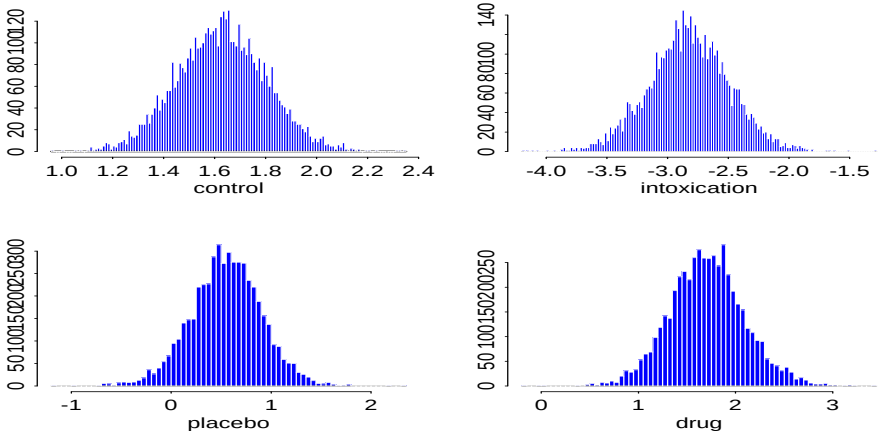
$$\pi(\theta_i|x, \theta, \theta_1, \ldots, \theta_n) = \pi(\theta_i|\theta_{i-1}, \theta_{i+1})$$

with the convention $\theta_0 = \theta$ and $\theta_{n+1} = 0$.

Example 48 **(Example 46 continued)** The full conditional distributions correspond to standard distributions and Gibbs sampling applies.



Convergence of the posterior means



Posteriors of the effects

	μ_δ	μ_D	μ_P	$\mu_D - \mu_P$
Probability	1.00	0.9998	0.94	0.985
Confidence	[-3.48,-2.17]	[0.94,2.50]	[-0.17,1.24]	[0.14,2.20]

Posterior probabilities of significant effects

6.2 Data Augmentation

The Gibbs sampler with only two steps is particularly useful

Algorithm 49 –Data Augmentation–

Given $y^{(t)}$,

1. Simulate $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t)})$;
2. Simulate $Y_2^{(t+1)} \sim g_2(y_2|y_1^{(t+1)})$.

Convergence is ensured

$$\begin{aligned}(Y_1, Y_2)^{(t)} &\rightarrow (Y_1, Y_2) \sim g \\ Y_1^{(t)} &\rightarrow Y_1 \sim g_1 \\ Y_2^{(t)} &\rightarrow Y_2 \sim g_2\end{aligned}$$

Example 50 –Grouped counting data–

360 consecutive records of the number of passages per unit time.

Number of passages	0	1	2	3	4 or more
Number of observations	139	128	55	25	13

Feature Observations with 4 passages and more are grouped

If observations are Poisson $\mathcal{P}(\lambda)$, the likelihood is

$$\begin{aligned}&\ell(\lambda|x_1, \ldots, x_5) \\ &\qquad \propto e^{-347\lambda} \lambda^{128+55\times 2+25\times 3} \left(1 - e^{-\lambda} \sum_{i=0}^{3} \frac{\lambda^i}{i!}\right)^{13},\end{aligned}$$

which can be difficult to work with.

Idea With a prior $\pi(\lambda) = 1/\lambda$, complete the vector $(y_1, \ldots, y_{13})$ of the 13 units larger than 4

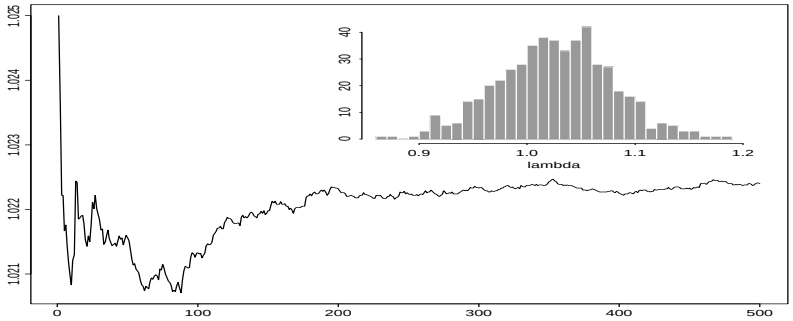
Algorithm 51 –Poisson-Gamma Gibbs–

1. Simulate $Y_i^{(t)} \sim \mathcal{P}(\lambda^{(t-1)})\, \mathbb{I}_{y \geq 4} \quad i = 1, \ldots, 13$
2. Simulate
$$\lambda^{(t)} \sim \mathcal{G}a\left(313 + \sum_{i=1}^{13} y_i^{(t)},\ 360\right).$$

The Bayes estimator

$$\delta^\pi = \frac{1}{360T} \sum_{t=1}^{T} \left(313 + \sum_{i=1}^{13} y_i^{(t)}\right)$$

converges quite rapidly



6.2.1 Rao-Blackwellization

If $(y_1, y_2, \ldots, y_p)^{(t)}, t = 1, 2, \ldots T$ is the output from a Gibbs sampler

$$\delta_0 = \frac{1}{T} \sum_{t=1}^{T} h\left(y_1^{(t)}\right) \rightarrow \int h(y_1) g(y_1) dy_1$$

and is unbiased. The Rao-Blackwellization replaces δ_0 with its conditional expectation

$$\delta_{rb} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[h(Y_1)|y_2^{(t)}, \ldots, y_p^{(t)}\right].$$

Then

- Both estimators converge to $\mathbb{E}[h(Y_1)]$
- Both are unbiased,
- and

$$\text{var}\left(\mathbb{E}\left[h(Y_1)|Y_2^{(t)}, \ldots, Y_p^{(t)}\right]\right) \leq \text{var}(h(Y_1)),$$

so δ_{rb} is uniformly better (for Data Augmentation)

Some examples of Rao-Blackwellization

- For the bivariate normal

$$(X, Y)' \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

the Gibbs sampler is based upon

$$\begin{aligned} X \mid y &\sim \mathcal{N}(\rho y,\ 1 - \rho^2) \\ Y \mid x &\sim \mathcal{N}(\rho x,\ 1 - \rho^2). \end{aligned}$$

To estimate $\mu = \mathbb{E}(X)$ we could use

$$\delta_0 = \frac{1}{T}\sum_{i=1}^{T} X^{(i)}$$

or its Rao-Blackwellized version

$$\delta_1 = \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}[X^{(i)}|Y^{(i)}] = \frac{1}{T}\sum_{i=1}^{T} \varrho Y^{(i)},$$

which satisfies $\sigma^2_{\delta_0}/\sigma^2_{\delta_1} = \frac{1}{\rho^2} > 1$.

- For the Poisson-Gamma Gibbs sampler, we could estimate λ with

$$\delta_0 = \frac{1}{T}\sum_{t=1}^{T} \lambda^{(t)},$$

but we instead used the Rao-Blackwellized version

$$\begin{aligned}\delta^\pi &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\lambda^{(t)}|x_1, x_2, \ldots, x_5, y_1^{(i)}, y_2^{(i)}, \ldots, y_{13}^{(i)}] \\ &= \frac{1}{360T}\sum_{t=1}^{T}\left(313 + \sum_{i=1}^{13} y_i^{(t)}\right),\end{aligned}$$

Another substantial benefit of Rao-Blackwellization is in the approximation of densities of different components of y without nonparametric density estimation methods.

The estimator

$$\frac{1}{T}\sum_{t=1}^{T} g_i(y_i|y_j^{(t)}, j \neq i) \to g_i(y_i),$$

and is unbiased.

6.2.2 The Duality Principle

Ties together the properties of the two Markov chains in Data Augmentation

Consider a Markov chain $(X^{(t)})$ and a sequence $(Y^{(t)})$ of random variables generated from the conditional distributions

$$\begin{aligned}X^{(t)}|y^{(t)} &\sim \pi(x|y^{(t)}) \\ Y^{(t+1)}|x^{(t)}, y^{(t)} &\sim f(y|x^{(t)}, y^{(t)})\ .\end{aligned}$$

Properties

- If the chain $(Y^{(t)})$ is ergodic then so is $(X^{(t)})$
- The conclusion holds for geometric or uniform ergodicity.
- The chain $(Y^{(t)})$ can be discrete, and the chain $(X^{(t)})$ can be continuous.

6.3 Improper Priors

Unsuspected danger resulting from careless use of MCMC algorithms: It can happen that

- all conditional distributions are well defined,
- all conditional distributions may be simulated from, but...
- the system of conditional distributions may not correspond to any joint distribution

Warning The problem is due to careless use of the Gibbs sampler in a situation for which the underlying assumptions are violated

Example 52 **–Conditional exponential distributions–**

For the model

$$X_1|x_2 \sim \mathcal{E}xp(x_2)\ , \quad X_2|x_1 \sim \mathcal{E}xp(x_1)$$

the only candidate $f(x_1, x_2)$ for the joint density is

$$f(x_1, x_2) \propto \exp(-x_1 x_2),$$

but

$$\int f(x_1, x_2)dx_1 dx_2 = \infty$$

(C) These conditionals do not correspond to a joint probability distribution

Example 53 **–Improper random effects–**

For a random effect model,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

where

$$\alpha_i \sim \mathcal{N}(0, \sigma^2) \text{ and } \varepsilon_{ij} \sim \mathcal{N}(0, \tau^2),$$

the Jeffreys (improper) prior for the parameters μ, σ and τ is

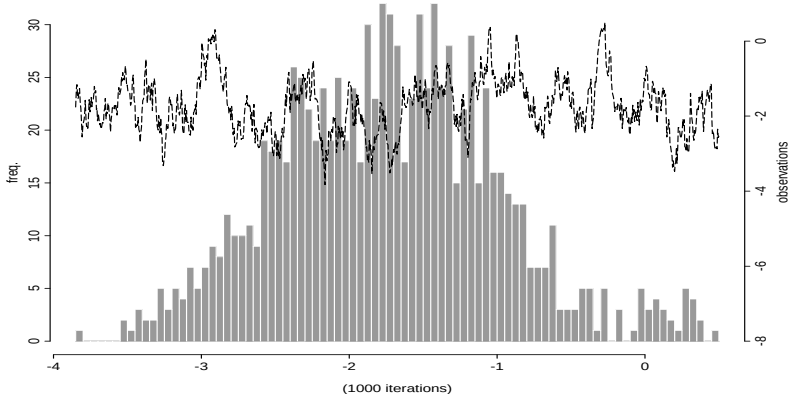
$$\pi(\mu, \sigma^2, \tau^2) = \frac{1}{\sigma^2\tau^2}\ .$$

The conditional distributions

$$\begin{aligned}
\alpha_i|y, \mu, \sigma^2, \tau^2 &\sim \mathcal{N}\left(\frac{J(\bar{y}_i - \mu)}{J + \tau^2\sigma^{-2}}, (J\tau^{-2} + \sigma^{-2})^{-1}\right)\ , \\
\mu|\alpha, y, \sigma^2, \tau^2 &\sim \mathcal{N}(\bar{y} - \bar{\alpha}, \tau^2/JI)\ , \\
\sigma^2|\alpha, \mu, y, \tau^2 &\sim \mathcal{IG}\left(I/2, (1/2)\sum_i \alpha_i^2\right)\ , \\
\tau^2|\alpha, \mu, y, \sigma^2 &\sim \mathcal{IG}\left(IJ/2, (1/2)\sum_{i,j} (y_{ij} - \alpha_i - \mu)^2\right)\ ,
\end{aligned}$$

are well-defined and a Gibbs sampling can be easily implemented in this setting.

Evolution of $(\mu^{(t)})$ and corresponding histogram



The figure shows the sequence of the $\mu^{(t)}$ and the corresponding histogram for 1000 iterations. The trend of the sequence and the histogram **do not** indicate that the corresponding "joint distribution" **does not exist**

Final notes on impropriety

The improper posterior Markov chain cannot be positive recurrent

The major task in such settings is to find indicators that flag that something is wrong. However, the output of an "improper" Gibbs sampler may not differ from a positive recurrent Markov chain.

Example The random effects model was initially treated in Gelfand *et al.* (1990) as a legitimate model

7 MCMC tools for variable dimension problems

7.1 Introduction

There exist setups where

> **One of the things we do not know is the number of things we do not know**
>
> [Peter Green]

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)**

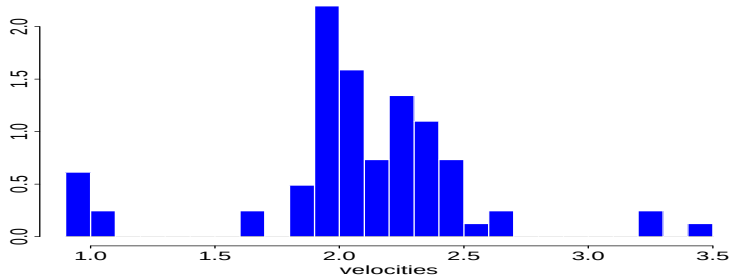
Many areas of application

- *variable* **selection**
- **change point(s) determination**
- **image analysis**
- **graphical models and expert systems**
- *variable* **dimension models**
- **causal inference**

Example 54 **—Mixture modelling–**

Benchmark dataset: Speed of galaxies

[Roeder, 1990; Richardson & Green, 1997]



Modelling by a mixture model

$$\mathfrak{M}_i : x_j \sim \sum_{\ell=1}^{i} p_{\ell i} \mathcal{N}(\mu_{\ell i}, \sigma_{\ell i}^2) \qquad (j = 1, \ldots, 82)$$

i?

Bayesian variable dimension model

A variable dimension model is defined as a collection of models $(k = 1. \ldots, K)$,

$$\mathfrak{M}_k = \{ f(\cdot|\theta_k);\ \theta_k \in \Theta_k \}\ ,$$

associated with a collection of priors on the parameters of these models,

$$\pi_k(\theta_k)\,,$$

and a prior distribution on the indices of these models,

$$\{ \varrho(k)\,, k = 1, \ldots, K \}\ .$$

Alternative notation:

$$\pi(\mathfrak{M}_k, \theta_k) = \varrho(k)\, \pi_k(\theta_k)$$

Formally over:

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \displaystyle\int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\displaystyle\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}$$

2. Take largest $p(\mathfrak{M}_i|x)$ to determine model, or use

$$\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j$$

as predictive

[Different decision theoretic perspectives]

Difficulties

Not at

- (formal) inference level [see above]
- parameter space representation

$$\Theta = \bigoplus_k \Theta_k\,,$$

[even if there are parameters common to several models]

Rather at

- (practical) inference level:
  model separation, interpretation, overfitting, prior modelling, prior coherence
- computational level:
  infinity of models, moves between models, predictive computation

7.2 Green's method

Setting up a proper measure–theoretic framework for designing moves *between* models $\mathfrak{M}_k$

[Green, 1995]

Create a **reversible kernel** $\mathfrak{K}$ on $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ such that

$$\int_A \int_B \mathfrak{K}(x, dy)\pi(x)dx = \int_B \int_A \mathfrak{K}(y, dx)\pi(y)dy$$

for the invariant density π [x is of the form $(k, \theta^{(k)})$]

Write $\mathfrak{K}$ as

$$\mathfrak{K}(x, B) = \sum_{m=1}^{\infty} \int \rho_m(x, y)\mathfrak{q}_m(x, dy) + \omega(x)\mathbb{I}_B(x)$$

where $\mathfrak{q}_m(x, dy)$ is a transition measure to model $\mathfrak{M}_m$ and $\rho_m(x, y)$ the corresponding acceptance probability.

Introduce a **symmetric** measure $\xi_m(dx, dy)$ on $\mathfrak{H}^2$ and impose on $\pi(dx)\mathfrak{q}_m(x, dy)$ to be absolutely continuous wrt ξ_m,

$$\frac{\pi(dx)\mathfrak{q}_m(x, dy)}{\xi_m(dx, dy)} = g_m(x, y)$$

Then

$$\rho_m(x, y) = \min\left\{1, \frac{g_m(y, x)}{g_m(x, y)}\right\}$$

ensures reversibility

Special case

When contemplating a move between two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the Markov chain being in state $\theta_1 \in \mathfrak{M}_1$, denote by $\mathfrak{K}_{1\to2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

$$\pi(d\theta_1)\, \mathfrak{K}_{1\to2}(\theta_1, d\theta) = \pi(d\theta_2)\, \mathfrak{K}_{2\to1}(\theta_2, d\theta)\,,$$

and take, wlog, $\dim(\mathfrak{M}_2) > \dim(\mathfrak{M}_1)$.

Proposal expressed as

$$\theta_2 = \Psi_{1\to2}(\theta_1, v_{1\to2})$$

where $v_{1\to2}$ is a random variable of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to2} \sim \varphi_{1\to2}(v_{1\to2})\,.$$

In this case, $\mathfrak{q}_{1\to2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to2}(v_{1\to2}) \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|^{-1},$$

by the Jacobian rule.

If probability $\varpi_{1\to2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})} \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|.$$

Interpretation (1)

The representation puts us back in a fixed dimension setting:

- $\mathfrak{M}_1 \times \mathfrak{V}_{1\to 2}$ and $\mathfrak{M}_2$ in one-to-one relation.
- *regular* Metropolis–Hastings move from the couple $(\theta_1, v_{1\to 2})$ to θ_2 when stationary distributions are $\pi(\mathfrak{M}_1, \theta_1) \times \varphi_{1\to 2}(v_{1\to 2})$ and $\pi(\mathfrak{M}_2, \theta_2)$, and when proposal distribution is *deterministic* (??)

Consider, instead, that the proposals

$$\theta_2 \sim \mathcal{N}(\Psi_{1\to 2}(\theta_1, v_{1\to 2}), \varepsilon) \quad \text{and} \quad \Psi_{1\to 2}(\theta_1, v_{1\to 2}) \sim \mathcal{N}(\theta_2, \varepsilon)$$

Reciprocal proposal has density

$$\frac{\exp\left\{-(\theta_2 - \Psi_{1\to 2}(\theta_1, v_{1\to 2}))^2/2\varepsilon\right\}}{\sqrt{2\pi\varepsilon}} \times \left|\frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})}\right|$$

by the Jacobian rule.

Thus Metropolis–Hastings acceptance probability is

$$1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)}{\pi(\mathfrak{M}_1, \theta_1)\, \varphi_{1\to 2}(v_{1\to 2})} \left|\frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})}\right|$$

Does not depend on ε: **Let ε go to 0**

Interpretation (2): saturation

[Brooks, Giudici, Roberts, 2003]

Consider series of models $\mathfrak{M}_i$ $(i = 1, \ldots, k)$ such that

$$\max_i \dim(\mathfrak{M}_i) = n_{\max} < \infty$$

Parameter of model $\mathfrak{M}_i$ then completed with an auxiliary variable U_i such that

$$\dim(\theta_i, u_i) = n_{\max} \quad \text{and} \quad U_i \sim q_i(u_i)$$

Posit the following joint distribution for [augmented] model $\mathfrak{M}_i$

$$\pi(\mathfrak{M}_i, \theta_i)\, q_i(u_i)$$

Saturation: no varying dimension anymore since (θ_i, u_i) of fixed dimension.

Three stage MCMC update:

1. Update the current value of the parameter, θ_i;
2. Update u_i conditional on θ_i;
3. Update the current model from $\mathfrak{M}_i$ to $\mathfrak{M}_j$ using the bijection

$$(\theta_j, u_j) = \Psi_{i\to j}(\theta_i, u_i)$$

Example 55 —Mixture of normal distributions—

$$\mathfrak{M}_k : \sum_{j=1}^{k} p_{jk}\mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$$

[Richardson & Green, 1997]

Moves:

(i) Split

$$\begin{cases} p_{jk} &= p_{j(k+1)} + p_{(j+1)(k+1)} \\ p_{jk}\mu_{jk} &= p_{j(k+1)}\mu_{j(k+1)} + p_{(j+1)(k+1)}\mu_{(j+1)(k+1)} \\ p_{jk}\sigma_{jk}^2 &= p_{j(k+1)}\sigma_{j(k+1)}^2 + p_{(j+1)(k+1)}\sigma_{(j+1)(k+1)}^2 \end{cases}$$

(ii) Merge (*reverse*)

Additional **Birth and Death** moves for empty components (created from the prior distribution)

Equivalent

(i) Split

$$(T) \begin{cases} u_1, u_2, u_3 &\sim \mathcal{U}(0,1) \\ p_{j(k+1)} &= u_1 p_{jk} \\ \mu_{j(k+1)} &= u_2 \mu_{jk} \\ \sigma_{j(k+1)}^2 &= u_3 \sigma_{jk}^2 \end{cases}$$

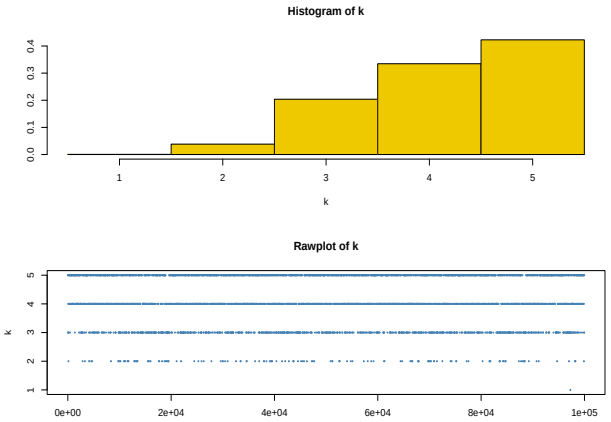
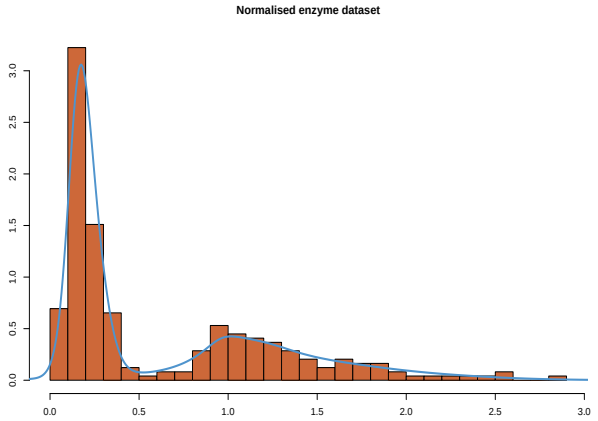


Figure 2: Histogram and rawplot of $100,000$ k's produced by RJMCMC under the imposed constraint $k \leq 5$.



Example 56 **—Hidden Markov model—**

$$\begin{aligned} P(X_t + 1 = j | X_t = i) &= w_{ij}, \\ w_{ij} &= \omega_{ij} / \sum_{\ell} \omega_{i\ell}, \\ Y_t | X_t = i &\sim \mathcal{N}(\mu_i, \sigma_i^2). \end{aligned}$$

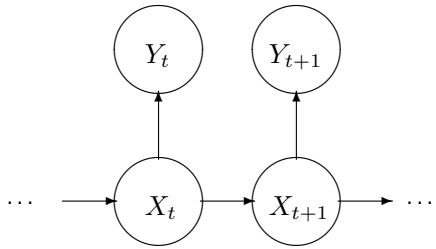


Figure 3: DAG representation of a simple hidden Markov model

Move to split component $j_\star$ into j_1 and j_2:

$$\omega_{ij_1} = \omega_{ij_\star}\varepsilon_i, \quad \omega_{ij_2} = \omega_{ij_\star}(1 - \varepsilon_i), \quad \varepsilon_i \sim \mathcal{U}(0, 1);$$

$$\omega_{j_1 j} = \omega_{j_\star j}\xi_j, \quad \omega_{j_2 j} = \omega_{j_\star j}/\xi_j, \quad \xi_j \sim \log \mathcal{N}(0, 1);$$

similar ideas give $\omega_{j_1 j_2}$ etc.;

$$\mu_{j_1} = \mu_{j_\star} - 3\sigma_{j_\star}\varepsilon_\mu, \quad \mu_{j_2} = \mu_{j_\star} + 3\sigma_{j_\star}\varepsilon_\mu, \quad \varepsilon_\mu \sim \mathcal{N}(0, 1);$$

$$\sigma_{j_1}^2 = \sigma_{j_*}^2\xi_\sigma, \quad \sigma_{j_2}^2 = \sigma_{j_*}^2/\xi_\sigma, \quad \xi_\sigma \sim \log \mathcal{N}(0, 1).$$

[Robert & al., 2000]

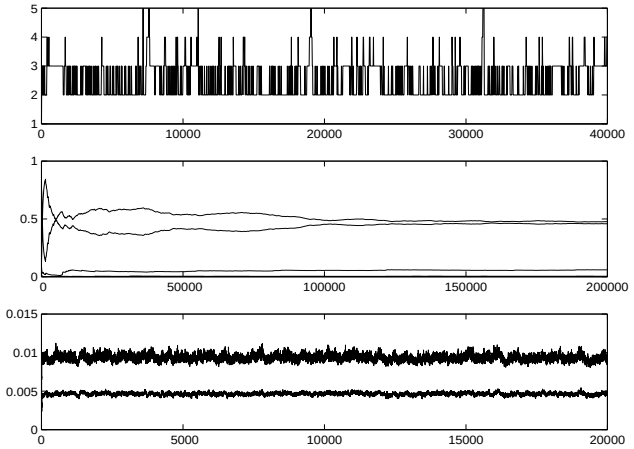


Figure 4: Upper panel: First 40,000 values of k for S&P 500 data, plotted every 20th sweep. Middle panel: estimated posterior distribution of k for S&P 500 data as a function of number of sweeps. Lower panel: σ_1 and σ_2 in first 20,000 sweeps with $k = 2$ for S&P 500 data.

Example 57 **—Autoregressive model—**

Typical setting for model choice: determine order p of $AR(p)$ model

Consider the (less standard) representation

$$\prod_{i=1}^{p}(1-\lambda_i B)\ X_t = \epsilon_t\,, \quad \epsilon_t \sim \mathcal{N}(0,\sigma^2)$$

where the λ_i's are within the unit circle if complex and within $[-1,1]$ if real.

[Huerta and West, 1998]

Roots [may] change drastically from one p to the other.

$AR(p)$ reversible jump algorithm

Uniform priors for the real and complex roots λ_j,

$$\frac{1}{\lfloor k/2\rfloor+1}\prod_{\lambda_i\in\mathbb{R}}\frac{1}{2}\mathbb{I}_{|\lambda_i|<1}\prod_{\lambda_i\notin\mathbb{R}}\frac{1}{\pi}\mathbb{I}_{|\lambda_i|<1}$$

and (purely birth-and-death) proposals based on these priors

- k $\longrightarrow$ k+1 [Creation of real root]
- k $\longrightarrow$ k+2 [Creation of complex root]
- k $\longrightarrow$ k-1 [Deletion of real root]
- k $\longrightarrow$ k-2 [Deletion of complex root]

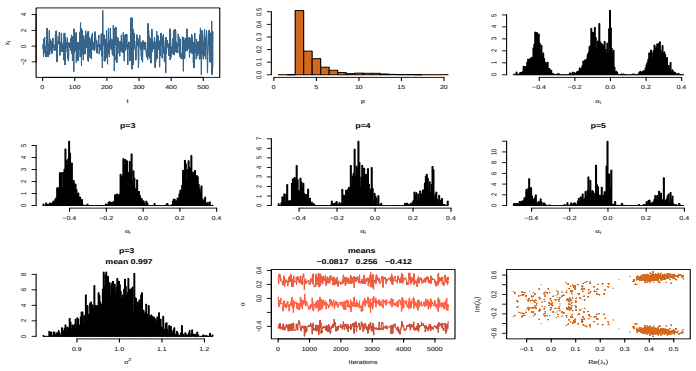


Figure 5: Reversible jump algorithm based on an $AR(3)$ simulated dataset of 530 points *(upper left)* with true parameters α_i $(-0.1, 0.3, -0.4)$ and $\sigma = 1$. First histogram associated with p, the following histograms with the α_i's, for different values of p, and of σ^2. Final graph: scatterplot of the complex roots. One before last: evolution of $\alpha_1, \alpha_2, \alpha_3$.

7.3 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 $\mathfrak{M}_k$, by births (to increase the dimension), deaths (to decrease the dimension), and other moves.

Time till next modification **(jump)** is exponentially distributed with rate depending on current state

Remember: if $\xi_1, \ldots, \xi_v$ are exponentially distributed, $\xi_i \sim \mathcal{E}(\lambda_i)$,

$$\min \xi_i \sim \mathcal{E}\left(\sum_i \lambda_i\right)$$

Difference with MH-MCMC: Whenever a jump occurs, the corresponding move *is always accepted*. Acceptance probabilities replaced with holding times.

Implausible configurations

$$L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ll 1$$

die quickly.

Balance condition

Sufficient to have **detailed balance**

$$L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\theta}') = L(\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}', \boldsymbol{\theta}) \quad \textbf{for all } \boldsymbol{\theta}, \boldsymbol{\theta}'$$

for $\tilde{\pi}(\boldsymbol{\theta}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ to be stationary.

Here $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ rate of moving from state $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$.

Possibility to add split/merge and fixed-k processes if balance condition satisfied.

Example 58 **—Mixture modelling (cont'd)—**

Stephen's original modelling:

- Representation as a (marked) point process

$$\Phi = \left\{\{p_j, (\mu_j, \sigma_j)\}\right\}_j$$

- Birth rate λ_0 (constant)
- Birth proposal from the prior
- Death rate $\delta_j(\Phi)$ for removal of point j
- Death proposal removes component and modifies weights
- Overall death rate

$$\sum_{j=1}^{k} \delta_j(\Phi) = \delta(\Phi)$$

- Balance condition

$$(k+1)\, d(\Phi \cup \{p, (\mu, \sigma)\})\, L(\Phi \cup \{p, (\mu, \sigma)\}) = \lambda_0 L(\Phi) \frac{\pi(k)}{\pi(k+1)}$$

with

$$d(\Phi \setminus \{p_j, (\mu_j, \sigma_j)\}) = \delta_j(\Phi)$$

- Case of Poisson prior $k \sim \mathcal{P}oi(\lambda_1)$

$$\delta_j(\Phi) = \frac{\lambda_0}{\lambda_1} \frac{L(\Phi \setminus \{p_j, (\mu_j, \sigma_j)\})}{L(\Phi)}$$

Stephen's original algorithm:

For $v = 0, 1, \cdots, V$

$t \leftarrow v$

Run till $t > v + 1$

1. Compute $\delta_j(\Phi) = \dfrac{L(\Phi|\Phi_j)}{L(\Phi)} \dfrac{\lambda_0}{\lambda_1}$
2. $\delta(\Phi) \leftarrow \displaystyle\sum_{j=1}^{k} \delta_j(\Phi_j), \xi \leftarrow \lambda_0 + \delta(\Phi), u \sim \mathcal{U}([0,1])$
3. $t \leftarrow t - u\log(u)$

4. With probability $\delta(\Phi)/\xi$

   Remove component j with probability $\delta_j(\Phi)/\delta(\Phi)$

   $k \leftarrow k - 1$

   $p_\ell \leftarrow p_\ell/(1 - p_j) \ \ (\ell \neq j)$

   Otherwise,

   Add component j from the prior $\pi(\mu_j, \sigma_j)$

   $p_j \sim \mathcal{B}e(\gamma, k\gamma)$

   $p_\ell \leftarrow p_\ell(1 - p_j) \ \ (\ell \neq j)$

   $k \leftarrow k + 1$

5. Run I MCMC(k, β, p)

Rescaling time

In discrete-time RJMCMC, let the time unit be $1/N$, put

$$\beta_k = \lambda_k/N \quad \text{and} \quad \delta_k = 1 - \lambda_k/N$$

As $N \to \infty$, each birth proposal will be accepted, and having k components births occur according to a Poisson process with rate λ_k while component (w, ϕ) dies with rate

$$\begin{aligned}
&\lim_{N\to\infty} N\delta_{k+1} \times \frac{1}{k+1} \times \min(A^{-1}, 1) \\
&\quad = \lim_{N\to\infty} N\frac{1}{k+1} \times \text{likelihood ratio}^{-1} \times \frac{\beta_k}{\delta_{k+1}} \times \frac{b(w,\phi)}{(1-w)^{k-1}} \\
&\quad = \text{likelihood ratio}^{-1} \times \frac{\lambda_k}{k+1} \times \frac{b(w,\phi)}{(1-w)^{k-1}}.
\end{aligned}$$

Hence **"RJMCMC→BDMCMC"**. This holds more generally.

Example 59 **—HMM models (cont'd)—**

Implementation of the split-and-combine rule of Richardson and Green (1997) in continuous time

Move to split component j_* into j_1 and j_2:

$$\omega_{ij_1} = \omega_{ij_*}\epsilon_i, \quad \omega_{ij_2} = \omega_{ij_*}(1 - \epsilon_i), \quad \epsilon_i \sim \mathcal{U}(0,1);$$

$$\omega_{j_1 j} = \omega_{j_* j}\xi_j, \quad \omega_{j_2 j} = \omega_{j_* j}/\xi_j, \quad \xi_j \sim \log\mathcal{N}(0,1);$$

similar ideas give $\omega_{j_1 j_2}$ etc.;

$$\mu_{j_1} = \mu_{j_*} - 3\sigma_{j_*}\epsilon_\mu, \quad \mu_{j_2} = \mu_{j_*} + 3\sigma_{j_*}\epsilon_\mu, \quad \epsilon_\mu \sim \mathcal{N}(0,1);$$

$$\sigma^2_{j_1} = \sigma^2_{j_*}\xi_\sigma, \quad \sigma^2_{j_2} = \sigma^2_{j_*}/\xi_\sigma, \quad \xi_\sigma \sim \log\mathcal{N}(0,1).$$

[Cappé & al, 2001]

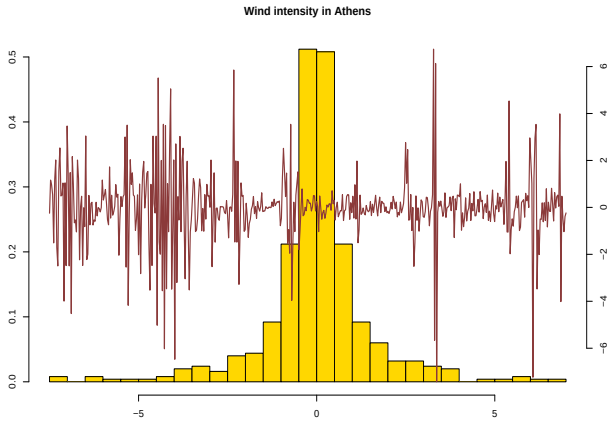


Figure 6: Histogram and rawplot of 500 wind intensities in Athens

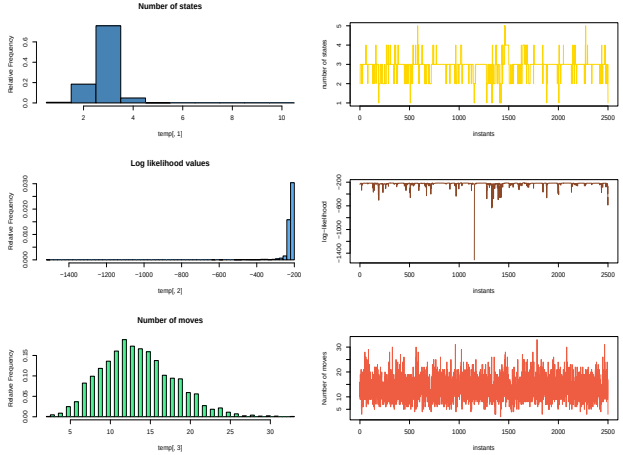


Figure 7: MCMC output on k (histogram and rawplot), corresponding loglikelihood values (histogram and rawplot), and number of moves (histogram and rawplot)

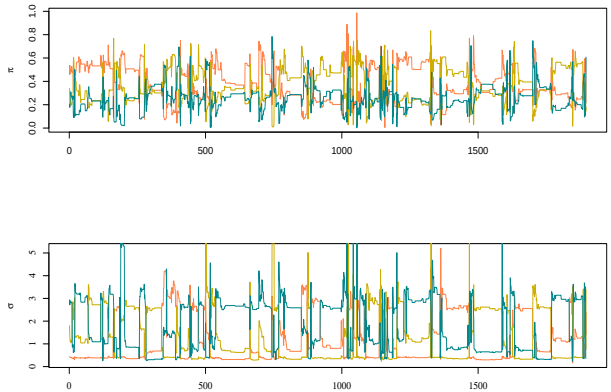


Figure 8: MCMC sequence of the probabilities π_j of the stationary distribution (top) and the parameters σ (bottom) of the three components when conditioning on $k = 3$

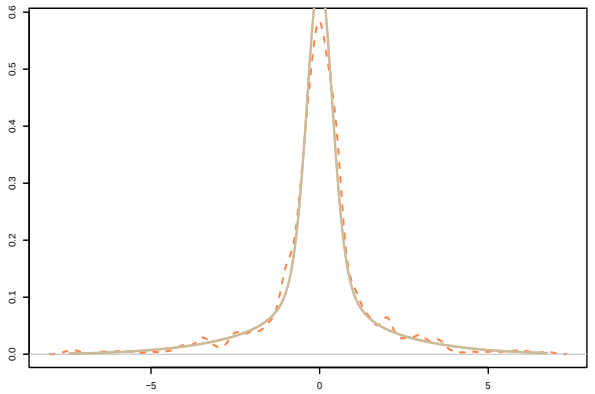


Figure 9: MCMC evaluation of the marginal density of the dataset (dashes), compared with R nonparametric density estimate (solid lines).

Even closer to RJMCM

Exponential (random) sampling is not necessary, nor is continuous time!

Estimator of

$$\mathfrak{I} = \int g(\theta)\pi(\theta)d\theta$$

by

$$\hat{\mathfrak{I}} = \frac{1}{N}\sum_{1}^{N} g(\theta(\tau_i))$$

where $\{\theta(t)\}$ continuous time MCMC process and $\tau_1, \ldots, \tau_N$ sampling instants.

New notations:

1. T_n time of the n-th jump of $\{\theta(t)\}$ with $T_0 = 0$
2. $\{\widetilde{\theta}_n\}$ jump chain of states visited by $\{\theta(t)\}$
3. $\lambda(\theta)$ total rate of $\{\theta(t)\}$ leaving state θ

Then holding time $T_n - T_{n-1}$ of $\{\theta(t)\}$ in its n-th state $\widetilde{\theta}_n$ exponential rv with rate $\lambda(\widetilde{\theta}_n)$

Rao–Blackwellisation

If sampling interval goes to 0, limiting case

$$\hat{\mathfrak{I}}_\infty = \frac{1}{T_N}\sum_{n=1}^{N} g(\widetilde{\theta}_{n-1})(T_n - T_{n-1})$$

Rao–Blackwellisation argument: replace $\hat{\mathfrak{I}}_\infty$ with

$$\tilde{\mathfrak{I}} = \frac{1}{T_N}\sum_{n=1}^{N} \frac{g(\widetilde{\theta}_{n-1})}{\lambda(\widetilde{\theta}_{n-1})} = \frac{1}{T_N}\sum_{n=1}^{N} E[T_n - T_{n-1} \mid \widetilde{\theta}_{n-1}]\, g(\widetilde{\theta}_{n-1})\,.$$

Conclusion: Only simulate jumps and store average holding times!

Example 60 —Mixture modelling (cont'd)—

Comparison of RJMCMC and CTMCMC in the Galaxy dataset

[Cappé & al., 2001]

Experiment:

- Same proposals (same C code)
- Moves proposed in equal proportions by both samplers (setting the probability P^F of proposing a fixed k move in RJMCMC equal to the rate η^F at which fixed k moves are proposed in CTMCMC, and likewise $P^B = \eta^B$ for the birth moves)
- Rao–Blackwellisation
- `Number of jumps (number of visited configurations) in CTMCMC == number of iterations of RJMCMC`

Results:

- **If one algorithm performs poorly, so does the other.** (For RJMCMC manifested as small A's—birth proposals are rarely accepted—while for BDMCMC manifested as large δ's—new components are indeed born but die again quickly.)
- No significant difference between samplers for birth and death only
- CTMCMC slightly better than RJMCMC with split-and-combine moves
- Marginal advantage in accuracy for split-and-combine addition
- For split-and-combine moves, computation time associated with one step of continuous time simulation is about 5 times longer than for reversible jump simulation.

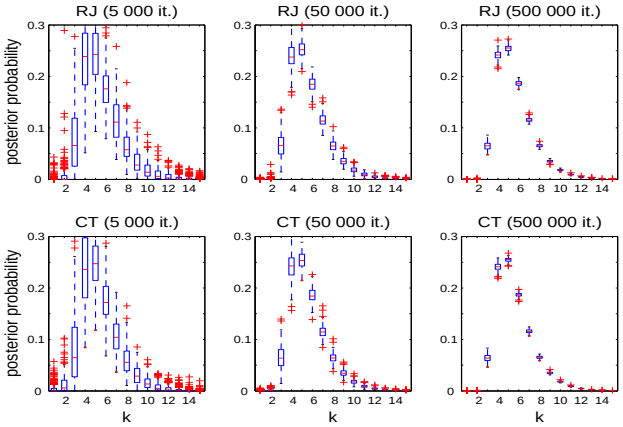


Figure 10: Galaxy dataset, box plot for the estimated posterior on k obtained from 200 independent runs: RJMCMC (top) and BDMCMC (bottom). The number of iterations varies from 5 000 (left), to 50 000 (middle) and 500 000 (right).

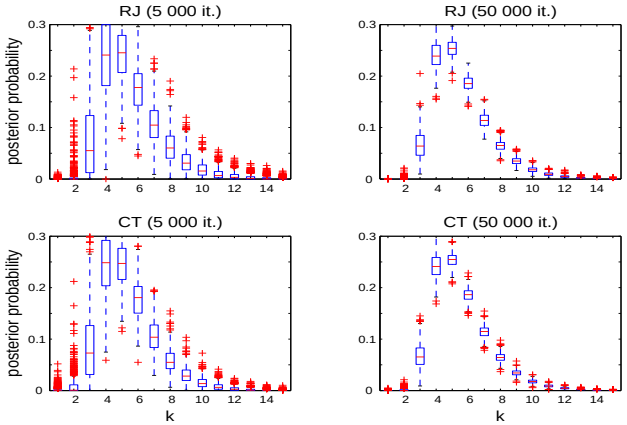


Figure 11: Galaxy dataset, box plot for the estimated posterior on k obtained from 500 independent runs: Top RJMCMC and bottom, CTMCMC. The number of iterations varies from 5 000 (left plots) to 50 000 (right plots).

8 Population Monte Carlo

8.1 Adaptive MCMC

Algorithms trained on-line usually invalid:

using the whole past of the “chain” implies that this is not a Markov chain any longer!

Example 61 **–Poly t distribution–**

t-distribution $\mathcal{T}(3,\theta,1)$ sample $(x_1,\ldots,x_n)$ and flat prior $\pi(\theta)=1$

Try fit a normal proposal from empirical mean and variance of the chain so far,

$$\mu_t=\frac{1}{t}\sum_{i=1}^{t}\theta^{(i)} \quad \text{and} \quad \sigma_t^2=\frac{1}{t}\sum_{i=1}^{t}(\theta^{(i)}-\mu_t)^2\,.$$

Metropolis–Hastings algorithm with acceptance probability

$$\prod_{j=2}^{n}\left[\frac{\nu+(x_j-\theta^{(t)})^2}{\nu+(x_j-\xi)^2}\right]^{-(\nu+1)/2}\frac{\exp-(\mu_t-\theta^{(t)})^2/2\sigma_t^2}{\exp-(\mu_t-\xi)^2/2\sigma_t^2}\,,$$

where $\xi\sim\mathcal{N}(\mu_t,\sigma_t^2)$.

Invalid scheme:

- when range of initial values too small, the $\theta^{(i)}$'s cannot converge to the target distribution and concentrates on too small a support.
- long-range dependence on past values modifies the distribution of the sequence.
- using past simulations to create a non-parametric approximation to the target distribution does not work either

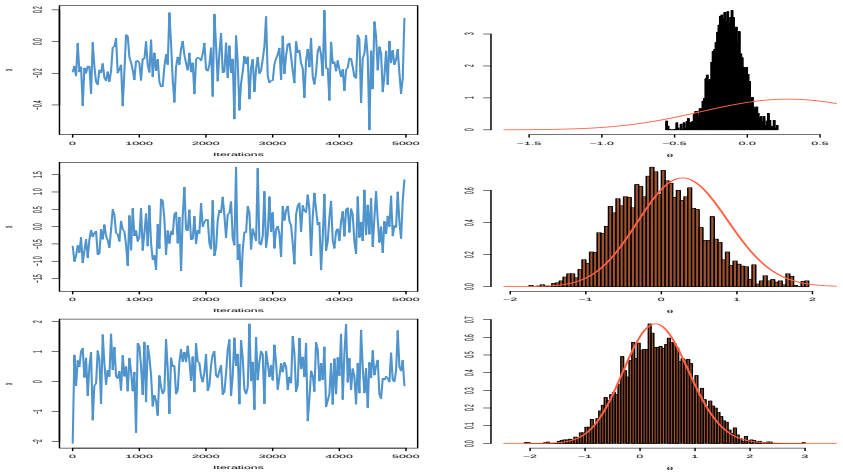


Figure 12: Adaptive scheme for a sample of 10 $x_j\sim\mathcal{T}_3$ and initial variances of *(top)* 0.1, *(middle)* 0.5, and *(bottom)* 2.5.

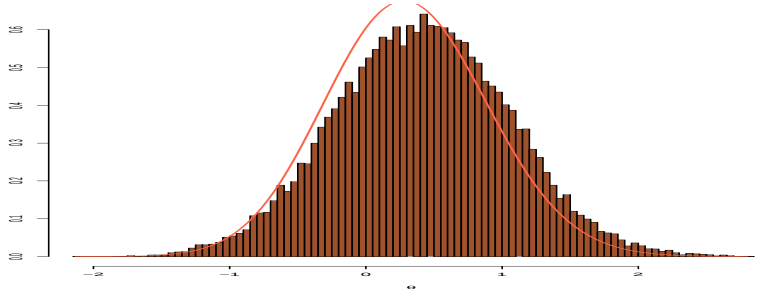


Figure 13: Comparison of the distribution of an adaptive scheme sample of $25,000$ points with initial variance of 2.5 and of the target distribution.

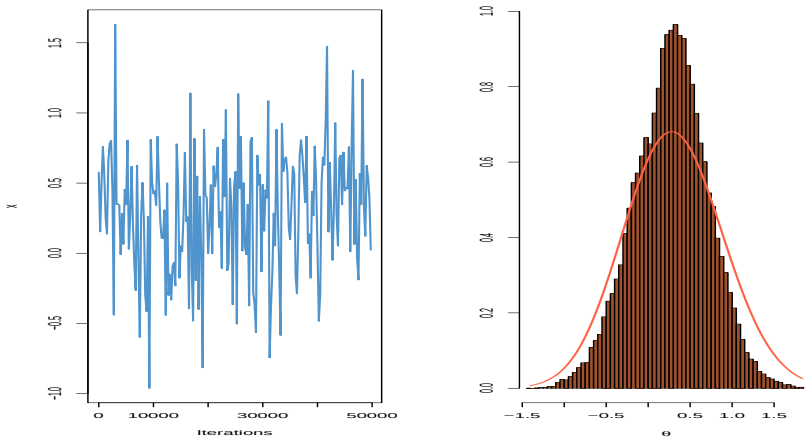


Figure 14: Sample produced by $50,000$ iterations of a nonparametric adaptive MCMC scheme and comparison of its distribution with the target distribution.

Warning

One should not *constantly* adapt the proposal on past performances of the chain.

Either adaptation ceases after a period of *burnin*
or the adaptive scheme must be theoretically assess on its own right.

8.2 Importance sampling revisited

Approximation of integrals

$$\mathfrak{I} = \int h(x)\pi(x)dx$$

by *unbiased estimators*

$$\hat{\mathfrak{I}} = \frac{1}{n}\sum_{i=1}^{n} \varrho_i h(x_i)$$

when

$$x_1, \ldots, x_n \overset{iid}{\sim} q(x) \qquad \text{and} \qquad \varrho_i \overset{\text{def}}{=} \frac{\pi(x_i)}{q(x_i)}$$

Dependent extension

For densities f and g, and importance weight

$$\omega(x) = f(x)/g(x)\,,$$

for any kernel $K(x, x')$ with stationary distribution f,

$$\int \omega(x)\, K(x, x')\, g(x) dx = f(x')\,.$$

[McEachern, Clyde, and Liu, 1999]

Consequence: An importance sample transformed by MCMC transitions keeps its weights

Unbiasedness preservation:

$$\begin{aligned}\mathbb{E}\left[\omega(X)h(X')\right] &= \int \omega(x)\, h(x')\, K(x, x')\, g(x)\, dx\, dx' \\ &= \mathbb{E}_f\left[h(X)\right]\end{aligned}$$

Drawback The weights do not change!

If x has small weight

$$\omega(x) = f(x)/g(x)\,,$$

then

$$x' \sim K(x, x')$$

keeps this small weight.

Dynamic extension

As in Markov Chain Monte Carlo (MCMC) algorithms, introduction of a *temporal dimension* :

$$x_i^{(t)} \sim q_t(x|x_i^{(t-1)}) \qquad i = 1, \ldots, n, \quad t = 1, \ldots$$

and

$$\hat{\mathfrak{I}}_t = \frac{1}{n} \sum_{i=1}^{n} \varrho_i^{(t)} h(x_i^{(t)})$$

is still unbiased for

$$\varrho_i^{(t)} = \frac{\pi(x_i^{(t)})}{q_t(x_i^{(t)}|x_i^{(t-1)})}\,, \qquad i = 1, \ldots, n$$

Reason why:

$$\begin{aligned}\mathbb{E}\left[h(X^{(t)})\ \frac{\pi(X^{(t)})}{q_t(X^{(t)}|X^{(t-1)})}\right] &= \int h(x)\,\frac{\pi(x)}{q_t(x|y)}\,q_t(x|y)\,g(y)\,dx\,dy \\ &= \int h(x)\,\pi(x)\,dx\end{aligned}$$

for **any distribution** g on $X^{(t-1)}$

Variance decomposition

Furthermore,

$$\text{var}\left(\hat{\mathfrak{I}}_t\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}\left(\varrho_i^{(t)}h(x_i^{(t)})\right)\ ,$$

if $\text{var}\left(\varrho_i^{(t)}\right)$ exists, because the $x_i^{(t)}$'s are conditionally uncorrelated

Note: Decomposition still valid for correlated $x_i^{(t)}$'s when incorporating weights $\varrho_i^{(t)}$

8.3 Dynamic sampling

More global dynamic schemes with random weights

(a) Self-regenerative chains

[Sahu & Zhigljavsky, 1998; Gasemyr, 2002]

Proposal

$$Y \sim p(y) \propto \tilde{p}(y)$$

and target distribution $\pi(x) \propto \tilde{\pi}(y)$

Ratios

$$\omega(x) = \pi(x)/p(x) \quad \text{and} \quad \tilde{\omega}(x) = \tilde{\pi}(x)/\tilde{p}(x)$$

Unknown **Known**

Acceptance function

$$\alpha(x) = \frac{1}{1+\kappa\tilde{\omega}(x)} \qquad \kappa > 0$$

Geometric jumps

If

$$Y \sim p(y)$$

and

$$W|Y = y \sim \mathscr{G}(\alpha(y))\,,$$

then

$$X_t = \cdots = X_{t+W-1} = Y \neq X_{t+W}$$

defines a Markov chain with stationary distribution π

- Valid for any choice of κ [κ small = large variance and κ large = slow convergence]
- Only depends on current value [Difference with Metropolis]
- Random integer weight W [Similarity with Metropolis]
- Saves on the rejections: always accept [Difference with Metropolis]
- Introduces geometric noise compared with importance sampling

$$\sigma^2_{SZ} = 2\,\sigma^2_{IS} + (1/\kappa)\sigma^2_\pi$$

- Can be used with a sequence of proposals p_k and constants κ_k [Adaptativity]

Generalisation

[Gasemyr, 2002]

Proposal density $p(y)$ and probability $q(y)$ of accepting a jump.

Sequence of random weights W_n such that

$$\begin{aligned} Y_n &\sim p(y) \\ V_n &\sim \mathcal{B}(q(y_n)) \\ S_n &\sim \mathcal{G}eo(\alpha(y_n)) \\ W_n &= V_n S_n \end{aligned}$$

Validation: If

$$\phi(y) = \frac{p(y)q(y)}{\int p(y)q(y)dy}\,,$$

chain (X_t) associated with the sequence (Y_n, W_n) by

$$Y_1 = X_1 = \cdots = X_{1+W_1-1}, Y_2 = X_{1+W_1} = \cdots$$

is a Markov chain with transition

$$K(x, y) = \alpha(x)\phi(y)$$

and point mass at $y = x$ with weight $1 - \alpha(x)$.

Necessary and sufficient condition

π is stationary for (X_t) if

$$\alpha(y) = q(y)/(\kappa\pi(y)/p(y)) = q(y)/(\kappa w(y)) \tag{4}$$

for some constant κ.

Implies that

$$\mathbb{E}[W^n|Y^n = y] = \kappa w(y)\,.$$

[Importance sampling]

Special case: $\alpha(y) = 1/(1 + \kappa w(y))$ of Sahu and Zhigljavski (2001)

Constraint on κ: for $\alpha(y) \leq 1$, κ must be such that

$$\frac{p(y)q(y)}{\pi(y)} \leq \kappa$$

Reverse of accept-reject conditions

Variance of

$$\sum_n W_n h(Y_n)/\sum_n W_n \tag{5}$$

is

$$2\int \frac{(h(y)-\mu)^2}{q(y)} w(y)\pi(y)dy - (1/\kappa)\sigma_\pi^2\,,$$

[Cramer-Wold/Slutsky]

Still worse than importance sampling.

(b) Dynamic weighting

[Wong & Liang, 1997; Liu, Liang & Wong, 2001; Liang, 2002]

Simultaneous generation of points and weights, (θ_t, ω_t), under the constraint

$$\mathbb{E}[\omega_t|\theta_t] \propto \pi(\theta_t) \tag{6}$$

Same use as importance sampling weights

Liang's Dynamic Importance Sampling:

1 Generate $y \sim K(x, y)$ and compute

$$\varrho = \omega \, \frac{\pi(y)K(y,x)}{\pi(x)K(x,y)}$$

2 Generate $u \sim \mathcal{U}(0,1)$ and take

$$(x', \omega') = \begin{cases} (y, (1+\delta)\varrho/a) & \text{if } u < a, \\ (x, (1+\delta)\omega/(1-a)) & \text{otherwise} \end{cases}$$

where $a = \varrho/(\varrho+\theta)$, $\theta = \theta(x,\omega)$, and $\delta > 0$ constant or independent rv

Preserves the equilibrium equation (6):

If g_- and g_+ denote the distributions of the augmented variable (X, W) before the step and after the step, respectively, then

$$\int_0^\infty \omega' \, g_+(x', \omega') \, d\omega' =$$
$$\int (1+\delta) \, [\varrho(\omega, x, x') + \theta] \; g_-(x, \omega) \, K(x, x') \frac{\varrho(\omega, x, x')}{\varrho(\omega, x, x') + \theta} \, dx \, d\omega$$
$$+ \int (1+\delta) \frac{\omega(\varrho(\omega, x', z) + \theta)}{\theta} \, g_-(x', \omega) \, K(x, z) \frac{\theta}{\varrho(\omega, x', z) + \theta} \, dz \, d\omega$$
$$= (1+\delta) \left\{ \int \omega \, g_-(x, \omega) \, \frac{\pi(x')K(x', x)}{\pi(x)} \, dx \, d\omega \right.$$
$$\left. + \int \omega \, g_-(x', \omega) \, K(x', z) \, dz \, d\omega \right\}$$

$$= (1+\delta) \left\{ \pi(x') \int c_0 \, K(x', x) \, dx + c_0 \pi(x') \right\}$$
$$= 2(1+\delta) c_0 \pi(x') \,,$$

where c_0 proportionality constant

Expansion phenomenon

$$\mathbb{E}[\omega_{t+1}] = 2(1+\delta)\mathbb{E}[\omega_t]$$

R-move

[Liang, 2002]

$\delta = 0$ and $\theta \equiv 1$, and thus

$$(x', \omega') = \begin{cases} (y, \varrho + 1) & \text{if } u < \varrho/(\varrho+1), \\ (x, \omega(\varrho+1)) & \text{otherwise,} \end{cases}$$

[Importance sampling]

W-move

$\theta \equiv 0$, thus $a = 1$ and

$$(x', \omega') = (y, \varrho)\,.$$

Q-move

[Liu & al, 2001]

$$(x', \omega') = \begin{cases} (y, \theta \vee \varrho) & \text{if } u < 1 \wedge \varrho/\theta, \\ (x, a\omega) & \text{otherwise,} \end{cases}$$

with $a \geq 1$ either a constant or an independent random variable.

Notes:

- Updating step in Q and R schemes written as

$$(x_{t+1}, \omega_{t+1}) = \{x_t, \omega_t / \mathsf{Pr}(R_t = 0)\}$$

with probability $\mathsf{Pr}(R_t = 0)$ and

$$(x_{t+1}, \omega_{t+1}) = \{y_{t+1}, \omega_t r(x_t, y_{t+1}) / \mathsf{Pr}(R_t = 1)\}$$

with probability $\mathsf{Pr}(R_t = 1)$, where R_t is the move indicator and

$$y_{t+1} \sim K(x_t, y)$$

- Geometric structure of the weights

$$\mathsf{Pr}(R_t = 0) = \frac{\omega_t}{\omega_{t+1}}\,.$$

and

$$\mathsf{Pr}(R_t = 0) = \frac{\omega_t\, r(x_t, y_t)}{\omega_t\, r(x_t, y_t) + \theta}, \quad \theta > 0\,,$$

for the R scheme

- Number of steps T before an acceptance (a jump) such that

$$\begin{aligned} \mathsf{Pr}\,(T \geq t) &= P(R_1 = 0, \ldots, R_{t-1} = 0) \\ &= \mathbb{E}\left[\prod_{j=0}^{t-1} \frac{\omega_j}{\omega_{j+1}}\right] \propto \mathbb{E}[1/\omega_t]\,. \end{aligned}$$

Alternative scheme

Preservation of weight expectation:

$$(x_{t+1}, \omega_{t+1}) = (x_t, \alpha_t \omega_t / \mathsf{Pr}(R_t = 0))$$

with probability $\mathsf{Pr}(R_t = 0)$ and

$$(x_{t+1}, \omega_{t+1}) = (y_{t+1}, (1 - \alpha_t)\omega_t r(x_t, y_{t+1}) / \mathsf{Pr}(R_t = 1))$$

with probability $\mathsf{Pr}(R_t = 1)$.

Then

$$\begin{aligned}\Pr(T=t) &= P(R_1=0,\ldots,R_{t-1}=0,R_t=1)\\ &= \mathbb{E}\left[\prod_{j=0}^{t-1}\alpha_j\frac{\omega_j}{\omega_{j+1}}(1-\alpha_t)\frac{\omega_{t-1}r(x_0,Y_t)}{\omega_t}\right]\end{aligned}$$

which is equal to

$$\alpha^{t-1}(1-\alpha)\mathbb{E}[\omega_o\, r(x,Y_t)/\omega_t]$$

when α_j constant and deterministic.

Example

Choose a function $0<\beta(\cdot,\cdot)<1$ and to take, while in (x_0,ω_0),

$$(x_1,\omega_1)=\left(y_1,\frac{\omega_0 r(x_0,y_1)}{\alpha(x_0,y_1)}(1-\beta(x_0,y_1))\right)$$

with probability

$$\min(1,\omega_0 r(x_0,y_1)) \stackrel{\Delta}{=} \alpha(x_0,y_1)$$

and

$$(x_1,\omega_1)=\left(x_0,\frac{\omega_0}{1-\alpha(x_0,y_1)}\times\beta(x_0,y_1)\right)$$

with probability $1-\alpha(x_0,y_1)$.

8.4 Population Monte Carlo

Pros and cons of Imp'Samp. vs. MCMC

- Production of a sample (IS) vs. of a Markov chain (MCMC)
- Dependence on importance function (IS) vs. on previous value (MCMC)
- Unbiasedness (IS) vs. convergence to the true distribution (MCMC)
- Variance control (IS) vs. learning costs (MCMC)
- Recycling of past simulations (IS) vs. progressive adaptability (MCMC)
- Processing of moving targets (IS) vs. handling large dimensional problems (MCMC)
- **Non-asymptotic validity (IS) vs. difficult asymptotia for adaptive algorithms (MCMC)**

Population Monte Carlo

Idea Simulate from the product distribution

$$\pi^{\otimes n}(x_1,\ldots,x_n)=\prod_{i=1}^{n}\pi(x_i)$$

and apply dynamic importance sampling to the sample

$$\mathbf{x}^{(t)}=(x_1^{(t)},\ldots,x_n^{(t)})$$

The importance distribution of the sample $\mathbf{x}^{(t)}$

$$q_t(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$$

can depend on the previous sample $\mathbf{x}^{(t-1)}$ in any possible way as long as marginal distributions

$$q_{it}(x) = \int q_t(\mathbf{x}^{(t)})\, d\mathbf{x}_{-i}^{(t)}$$

can be expressed to build importance weights

$$\varrho_{it} = \frac{\pi(x_i^{(t)})}{q_{it}(x_i^{(t)})}$$

Note: Using the *marginal* distributions creates correlation terms in the variance of $\hat{\mathfrak{I}}_t$ but reduces the overall variance $\mathrm{var}\hat{\mathfrak{I}}_t$ by a Rao–Blackwellisation argument

Special case

$$q_t(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \prod_{i=1}^{n} q_{it}(x_i^{(t)}|\mathbf{x}^{(t-1)})$$

[Independent proposals]

In that case,

$$\mathrm{var}\left(\hat{\mathfrak{I}}_t\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\left(\varrho_i^{(t)} h(x_i^{(t)})\right)\,,$$

because

$$\begin{aligned}
&\mathbb{E}\left[\varrho_i^{(t)} h(X_i^{(t)})\ \varrho_j^{(t)} h(X_j^{(t)})\right] \\
&\quad = \int h(x_i)\frac{\pi(x_i)}{q_{it}(x_i|\mathbf{x}^{(t-1)})}\ \frac{\pi(x_j)}{q_{jt}(x_j|\mathbf{x}^{(t-1)})}\, h(x_j) \\
&\qquad q_{it}(x_i|\mathbf{x}^{(t-1)})\, q_{jt}(x_j|\mathbf{x}^{(t-1)})\, dx_i\, dx_j\, g(\mathbf{x}^{(t-1)}) d\mathbf{x}^{(t-1)} \\
&\quad = \mathbb{E}_\pi\left[h(X)\right]^2
\end{aligned}$$

whatever the distribution g on $\mathbf{x}^{(t-1)}$

Normalising constants

In general, π is unscaled and

$$\varrho_i^{(t)} \propto \frac{\pi(x_i^{(t)})}{q_{it}(x_i^{(t)})}\,, \qquad i = 1, \ldots, n\,,$$

scaled so that

$$\sum_i \varrho_i^{(t)} = 1$$

- Loss of the unbiasedness property and the variance decomposition
- Normalising constant can be estimated by

$$\varpi_t = \frac{1}{tn} \sum_{\tau=1}^{t} \sum_{i=1}^{n} \frac{\pi(x_i^{(\tau)})}{q_{i\tau}(x_i^{(\tau)})}$$

- Variance decomposition (approximately) recovered if ϖ_{t-1} used instead

Resampling

Over iterations (in t), weights are multiplied, resulting in *degeneracy* of the sample

$\varrho_1 \equiv 1$, while $\varrho_2, \ldots$ negligible

Use instead Rubin's (1987) *systematic resampling:* at each iteration resample the $x_i^{(t)}$'s according to their weight $\varrho_i^{(t)}$ and reset the weights to 1

PMCA: Population Monte Carlo Algorithm

For $t = 1, \ldots, T$

For $i = 1, \ldots, n$,

(a) Select the generating distribution $q_{it}(\cdot)$

(b) Generate $x_i^{(t)} \sim q_{it}(x)$

(c) Compute $\varrho_i^{(t)} = \pi(x_i^{(t)})/q_{it}(x_i^{(t)})$

Normalise the $\varrho_i^{(t)}$'s to sum up to 1

Resample n values from the $x_i^{(t)}$'s with replacement, using the weights $\varrho_i^{(t)}$, to create the sample $(x_1^{(t)}, \ldots, x_n^{(t)})$

Links with particle filters

- Usually setting where $\pi = \pi_t$ changes with t: Population Monte Carlo also adapts to this case
- Gilks and Berzuini (2001) produce iterated samples with (SIR) resampling steps, and add an MCMC step: this step must use a π_t invariant kernel
- Chopin (2001) uses iterated importance sampling to handle large datasets: this is a special case of PMC where the q_{it}'s are the posterior distributions associated with a portion k_t of the observed dataset
- Stavropoulos and Titterington's (1999) *smooth bootstrap*, and Warnes' (2001) *kernel coupler* use nonparametric kernels on the previous importance sample to build an improved proposal: this is a special case of PMC

8.5 Mixtures of distributions

Observation of an iid sample $\mathbf{x} = (x_1, \ldots, x_n)$ from

$$p\mathcal{N}(\mu_1, \sigma^2) + (1-p)\mathcal{N}(\mu_2, \sigma^2),$$

with $p \neq 1/2$ and $\sigma > 0$ known.

Usual $\mathcal{N}(\theta, \sigma^2/\lambda)$ prior on μ_1 and μ_2:

$$\pi(\mu_1, \mu_2|\mathbf{x}) \propto f(\mathbf{x}|\mu_1, \mu_2)\, \pi(\mu_1, \mu_2)$$

Population Monte Carlo Algorithm

Step 0: Initialisation

For $j = 1, \ldots, n = pm$, choose $(\mu_1)_j^{(0)}, (\mu_2)_j^{(0)}$

For $k = 1, \ldots, p$, set $r_k = m$

Step i: Update $(i = 1, \ldots, I)$

For $k = 1, \ldots, p$,

1 generate a sample of size r_k as

$$(\mu_1)_j^{(i)} \sim \mathcal{N}\left((\mu_1)_j^{(i-1)}, v_k\right) \quad \text{and} \quad (\mu_2)_j^{(i)} \sim \mathcal{N}\left((\mu_2)_j^{(i-1)}, v_k\right)$$

2 compute the weights

$$\varrho_j \propto \frac{f\left(\mathbf{x}\left|(\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right.\right) \pi\left((\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right)}{\varphi\left((\mu_1)_j^{(i)}\left|(\mu_1)_j^{(i-1)}, v_k\right.\right) \varphi\left((\mu_2)_j^{(i)}\left|(\mu_2)_j^{(i-1)}, v_k\right.\right)}$$

Resample the $\left((\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right)_j$ using the weights ϱ_j,

For $k = 1, \ldots, p$,

update r_k as the number of elements generated with variance v_k which have been resampled.

Details

After an arbitrary initialisation, use of the previous (importance) sample (after resampling) to build random walk proposals,

$$\mathcal{N}((\mu)_j^{(i-1)}, v_j)$$

with a multiscale variance v_j within a predetermined set of p scales ranging from 10^3 down to 10^{-3}, whose importance is proportional to its survival rate in the resampling step.

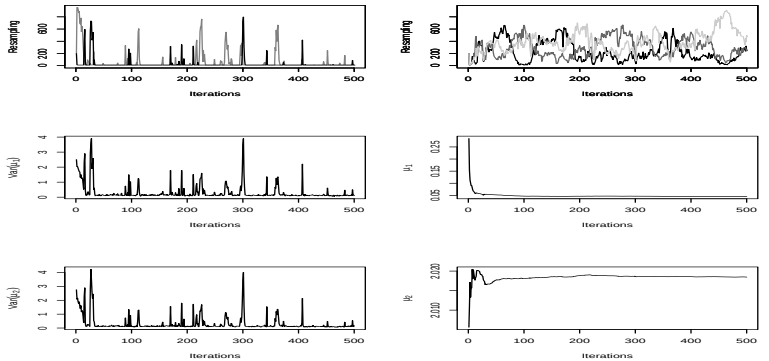


Figure 15: *(u.left)* Number of resampled points for $v_1 = 5$ (darker) and $v_2 = 2$; *(u.right)* Number of resampled points for the other variances; *(m.left)* Variance of the μ_1's along iterations; *(m.right)* Average of the μ_1's over iterations; *(l.left)* Variance of the μ_2's along iterations; *(l.right)* Average of the simulated μ_2's over iterations.

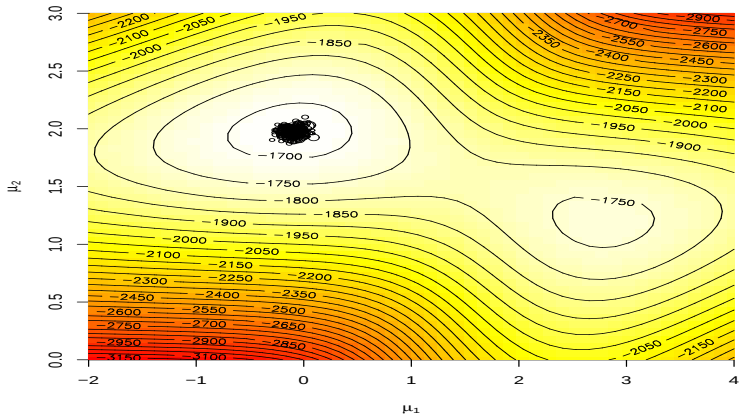


Figure 16: Log-posterior distribution and sample of means

8.6 Ion Channel Modelling

Formalised representation of ion exchanges between neurons as neurotransmission regulators.

Ion channel can be in one of several states, each state corresponding to a given electric intensity.

Indirect observation of these intensities: *patch clamp recordings*, ie intensity variations.

A hidden semi-Markov model

Observables $\mathbf{y} = (y_t)_{1 \le t \le T}$ directed by a hidden Gamma (indicator) process $\mathbf{x} = (x_t)_{1 \le t \le T}$

$$y_t | x_t \sim \mathcal{N}(\mu_{x_t}, \sigma^2)\,,$$

Hidden process such that

$$d_{j+1} = t_{j+1} - t_j \sim \mathcal{G}a(s_i, \lambda_i)$$

if $x_t = i$ for $t_j \le t < t_{j+1}$

[Ball et al., Carpenter et al., Hodgson, 1999]

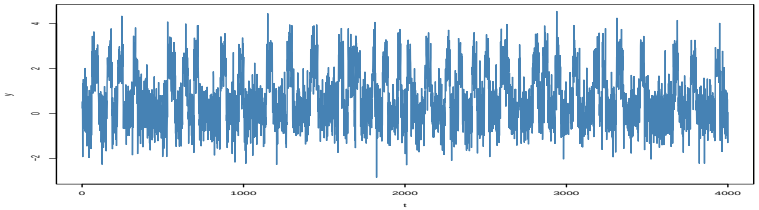


Figure 17: Simulated sample of size 4000

Our assumptions

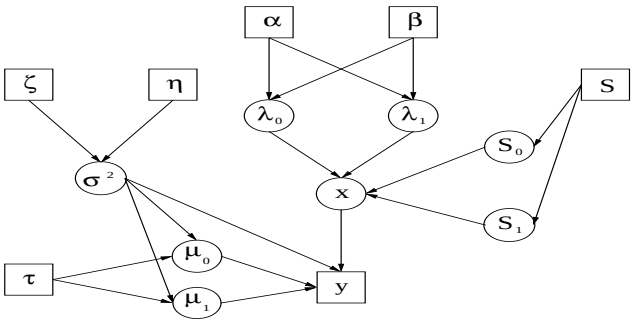
- The durations d_j are **integers**
  - generalisation of HMM: geometric *vs.* negative binomial
  - identifiability issue
- s_0 and s_1 are **integers** and uniform on $\{1, \ldots, S\}$
  - generalisation of HMM: exponential *vs.* sum of exponentials
  - alternative to duplicate states

    [Carpenter et al., Hodgson and Green, 1999]

  - alternative to variable dimension modelling
- Observables **independent**, given the x_t's

Prior modelling

$$\begin{aligned} \mu_0, \mu_1 &\sim \mathcal{N}(\theta_0, \tau\sigma^2) \\ \sigma^{-2} &\sim \mathcal{G}(\zeta, \eta) \\ \lambda_0, \lambda_1 &\sim \mathcal{G}(\alpha, \beta) \end{aligned}$$



Instrumental distribution

$\pi(\omega^{(j)}|\mathbf{y}, \mathbf{x}_-^{(j)})$ Gibbs like:

$$\sigma^{-2}\,|y, x \sim \mathcal{G}\left(\frac{T}{2}+\eta, \left(\frac{1}{2}\right)\left[2\nu + 1_T' y^2 - \frac{\left((1_T - x)' y\right)^2}{v_0} - \frac{(x' y)^2}{v_1} + \frac{\tau v_0 \left(\theta_0 - \frac{(1_T - x)' y}{v_0}\right)^2}{v_0+\tau} + \frac{\tau v_1 \left(\theta_0 - \frac{x' y}{v_1}\right)^2}{v_1+\tau}\right]\right)$$

$$\mu_0\,|y, x, \sigma^2 \sim \mathcal{N}\left(\frac{(1_T - x)' y + \theta_0 \tau}{v_0+\tau}, \frac{\sigma^2}{v_0+\tau}\right)$$

$$\mu_1\,|y, x, \sigma^2 \sim \mathcal{N}\left(\frac{x' y + \theta_0 \tau}{v_1+\tau}, \frac{\sigma^2}{v_1+\tau}\right)$$

if $v_0 = (1_T - x)'\, 1_T$ and $v_1 = x' 1_T$ and

π_H conditional distribution of a (pseudo) hidden Markov chain $\mathbf{x}$ given the observables $\mathbf{y}$ and constructed via the forward–backward formula for the pseudo transition

$$\mathbb{P} = \begin{pmatrix} 1 - \frac{\lambda_0}{s_0} & \frac{\lambda_0}{s_0} \\ \frac{\lambda_1}{s_1} & 1 - \frac{\lambda_1}{s_1} \end{pmatrix}$$

Motivations

1. Importance sampling bypasses exact simulation of the hidden process $(x_t)_{1\le t\le T}$, using a pseudo-HMM, and avoids recourse to variable dimension models
2. Provides unrestricted moves between configurations of the process $(x_t)_{1\le t\le T}$.
3. Iterated importance sampling provides progressive selection of the most relevant particles [Berzuini et al., 1997]
4. Metropolis–Hastings scheme based on the same proposal does not work so well
5. Produces a sample in the parameter space close to an iid sample from the true posterior distribution
6. Can be tuned on-the-run while remaining valid.

Population Monte Carlo Algorithm

Step 0. Generate $(j = 1, \ldots, J)$

1 $\omega^{(j)} \sim \pi(\omega)$

2 $\mathbf{x}_-^{(j)} = (x_t^{(j)})_{1 \le t \le T} \sim \pi_H(\mathbf{x}|\mathbf{y}, \omega^{(j)})$

compute the weights $(j = 1, \ldots, J)$

$$\varrho_j \propto \frac{\pi(\omega^{(j)}, \mathbf{x}_-^{(j)}|\mathbf{y})}{\pi(\omega^{(j)})\pi_H(\mathbf{x}_-^{(j)}|\mathbf{y}, \omega^{(j)})}$$

resample the $(\omega^{(j)}, \mathbf{x}_-^{(j)})_j$ using the weights ϱ_j

Step i. $(i = 1, \ldots)$ Generate $(j = 1, \ldots, J)$

1 $\omega^{(j)} \sim \pi(\omega|\mathbf{y}, \mathbf{x}_-^{(j)})$

2 $\mathbf{x}_+^{(j)} = (x_t^{(j)})_{1 \le t \le T} \sim \pi_H(\mathbf{x}|\mathbf{y}, \omega^{(j)})$

compute the weights $(j = 1, \ldots, J)$

$$\varrho_j \propto \frac{\pi(\omega^{(j)}, \mathbf{x}_+^{(j)}|\mathbf{y})}{\pi(\omega^{(j)}|\mathbf{y}, \mathbf{x}_-^{(j)})\pi_H(\mathbf{x}_+^{(j)}|\mathbf{y}, \omega^{(j)})}$$

resample the $(\omega^{(j)}, \mathbf{x}_+^{(j)})_j$ using the weights ϱ_j, and take $\mathbf{x}_-^{(j)} = \mathbf{x}_+^{(j)}$ $(j = 1, \ldots, J)$.

Results

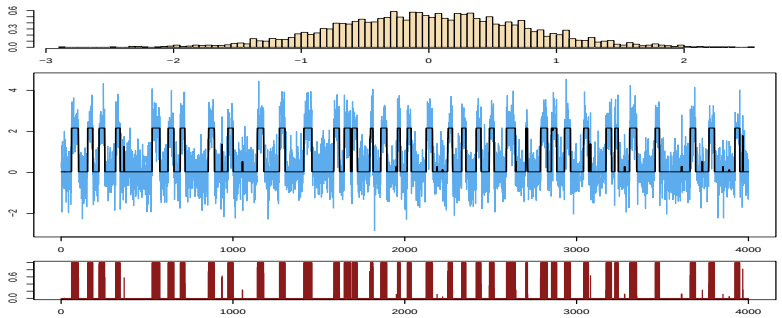


Figure 18: (top) Histograms of residuals after fit by averaged μ_{x_t}; (middle) Simulated sample of size 4000 against fitted averaged μ_{x_t}; (bottom) Probability of allocation to first state for each observation

Population Monte Carlo

Adaptive algorithm: self-improvement of the importance sampler

Long-term behaviour of the algorithm?

stopping rule?

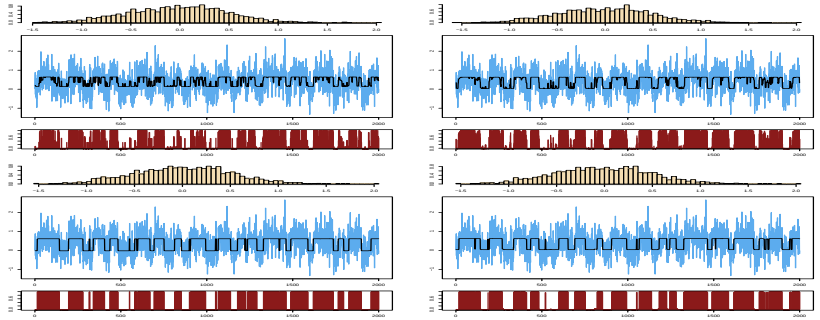


Figure 19: Successive fits for 2000 observations 2000 particles and 1, 2, 5 and 10 iterations.

Degeneracy

Percentage of relevant particles less than 10% on average

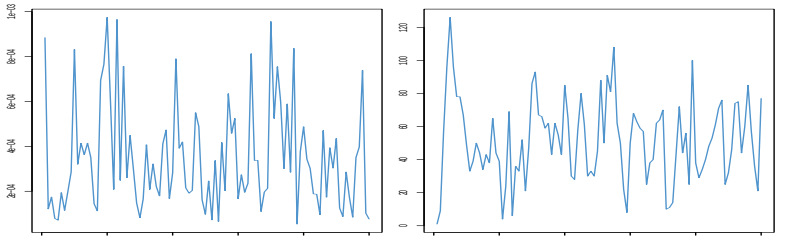


Figure 20: (left) Variance of the weights ϱ_j and (right) Number of particles with descendants along 100 iterations, for 4000 observations and 1000 particles.

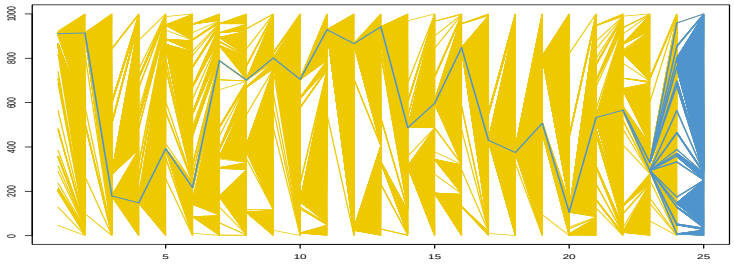


Figure 21: Representation of the sequence of descendents (yellow) and ancestors (blue) for 4000 observations and 1000 particles.

Comparison with Hastings–Metropolis

Uses **exactly** the same proposal in an HM framework

MCMC Algorithm

Step i ($i = 1, \ldots, J$)

- Generate $\omega^{(i)} \sim \pi(\omega|\mathbf{y}, \mathbf{x}^{(i-1)})$
- Generate $\mathbf{x}^\star \sim \pi_H(\mathbf{x}|\mathbf{y}, \omega^{(i)})$, $u \sim \mathcal{U}([0,1])$ and take

$$\mathbf{x}^{(i)} = \begin{cases} \mathbf{x}^\star & \text{if } u \leq \dfrac{\pi(\mathbf{x}^\star|\omega^{(i)}\mathbf{y})}{\pi_H(\mathbf{x}^\star|\mathbf{y},\omega^{(i)})} \Big/ \dfrac{\pi(\mathbf{x}^{(i-1)}|\omega^{(i)}\mathbf{y})}{\pi_H(\mathbf{x}^{(i-1)}|\mathbf{y},\omega^{(i)})}, \\ \mathbf{x}^{(i-1)} & \text{otherwise} \end{cases}$$

Performances

- Poor overall performances/mixing abilities
- Degenerates (to single state) if started at random
- Requires a sequential burnin $(n = 100, 200, \ldots)$ *and even...*
- No visible improvement over population Monte Carlo

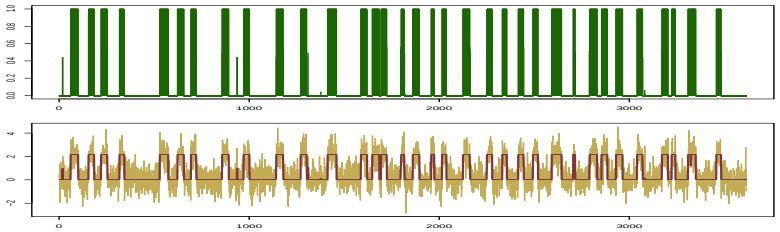


Figure 22: 5000 MCMC iterations

8.7 Stochastic volatility

Simplest model

$$y_t = \beta \exp(z_t/2)\, \epsilon_t\,, \qquad \epsilon_t \sim \mathcal{N}(0,1)$$

with AR(1) log-variance process

$$z_{t+1} = \varphi z_t + \sigma u_t\,, \quad u_t \sim \mathcal{N}(0,1)$$

Observed likelihood unavailable in closed from.

Necessary MCMC completion by the missing data z (of the same dimension as the data)

[Jacquier, Polson & Rossi, 1994; Kim, Chib and Shephard, 1998]

Lack of robustness:

the MCMC algorithm may fail to converge for long series or extreme values of the parameters β and φ.

Very sensitive to the generation of the missing data

May well fail to converge even when initialized at the true parameter values.

Posterior distributions for β^2 and σ^2 conditional on the x_t's and z_t's both inverse Gamma distributions with $(n-1)/2$ shape parameters and

$$\sum_{t=1}^{n} y_t^2 \exp(-z_t)/2 \quad \text{and} \quad \sum_{t=2}^{n} (z_t - \varphi z_{t-1})^2 /2$$

scales

Conditional distribution of φ proportional to

$$\sqrt{1-\varphi^2} \exp - \left(\varphi^2 \sum_{t=2}^{n-1} z_t^2 - 2\varphi \sum_{t=2}^{n} z_t z_{t-1} \right) /2\sigma^2 \, \mathbb{I}_{]-1,1[}(\varphi) \,,$$

and standard proposal truncated normal on $]-1,1[$ with mean and variance

$$\sum_{t=2}^{n} z_t z_{t-1} \Big/ \sum_{t=2}^{n-1} z_t^2 \quad \text{and} \quad \sigma^2 \Big/ \sum_{t=2}^{n-1} z_t^2$$

Simulation of $z_t|y, \varphi, \sigma^2$ $(2 \leq t \leq n-1)$, proportional to

$$\exp\left\{ -0.5 \left(1+\varphi^2\right) (z_t - \mu_t)^2 /\sigma^2 - 0.5 \exp(-z_t) \, y_t^2/\beta^2 - 0.5 z_t \right\} \,,$$

where $\mu_t = \varphi \left(z_{t-1} + z_{t+1}\right)/(1+\varphi^2)$.

Solution:

Expand $\exp(-z_t)$ by a Taylor expansion around μ_t.

Normal proposal with mean

$$\frac{\left(1+\varphi^2\right) \mu_t/\sigma^2 + 0.5 \exp(-\mu_t) \, y_t^2 \, (1+\mu_t) /\beta^2 - 0.5}{\left(1+\varphi^2\right)/\sigma^2 + 0.5 \exp(-\mu_t) \, y_t^2/\beta^2}$$

and variance

$$1 \Big/ \left\{ \left(1+\varphi^2\right)/\sigma^2 + 0.5 \exp(-\mu_t) \, y_t^2/\beta^2 \right\} .$$

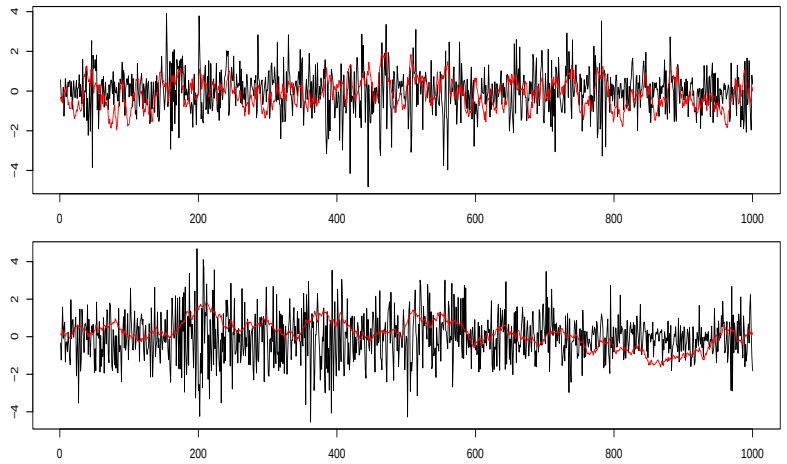


Figure 23: Weekly *(upper)* and daily *(lower)* simulated datasets with $n = 1000$ observations y_t *(black)* and volatilities z_t *(red)*.

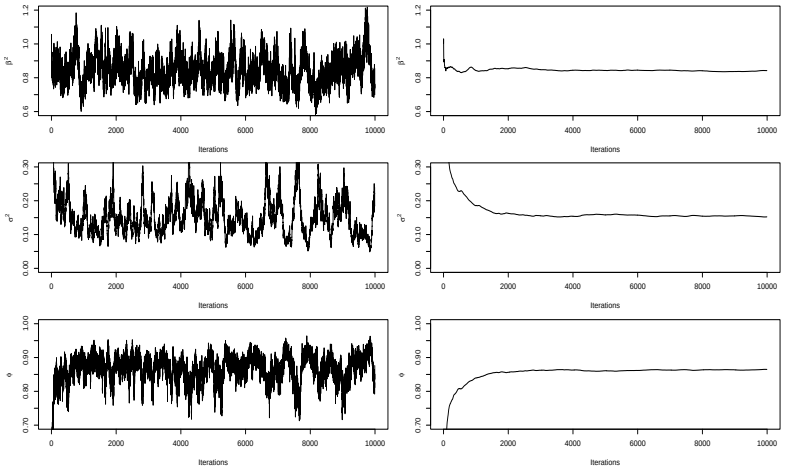


Figure 24: Weekly dataset: evolution of the MCMC samples for the three parameters *(left)* and convergence of the MCMC estimators *(right)*.

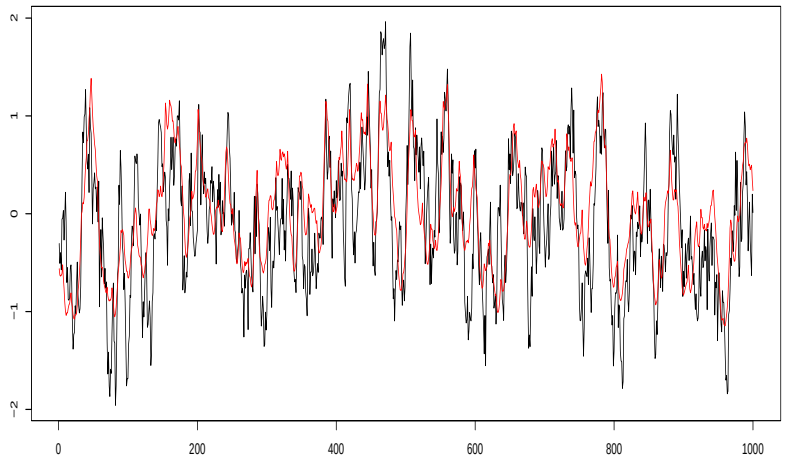


Figure 25: Weekly dataset: estimation of the stochastic volatility (in black the true volatility and in red the MCMC estimation based on the last 5000 iterations).

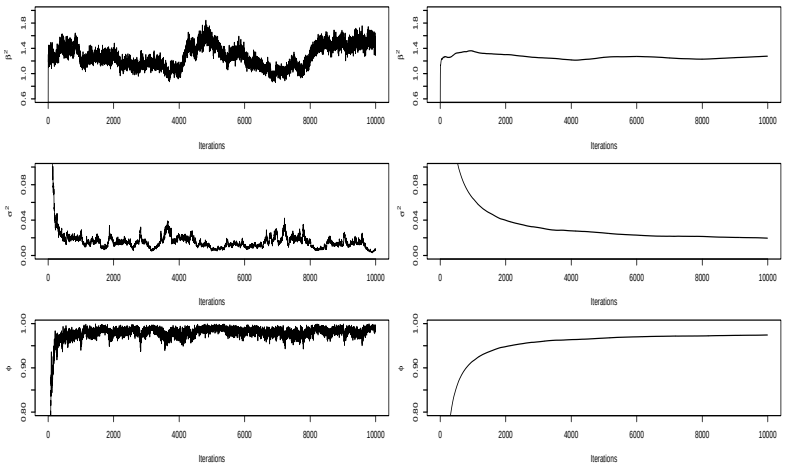


Figure 26: Daily dataset: same legend as Figure 24.

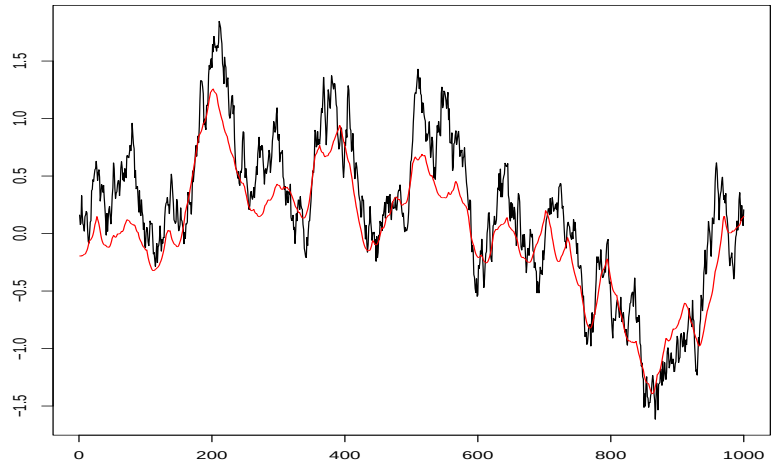


Figure 27: Daily dataset: same legend as Figure 25.

Run a PMC sampler with exactly the same proposals as completion MCMC

Results:

Better parameter estimates *and* volatility reconstruction for PMC than for MCMC

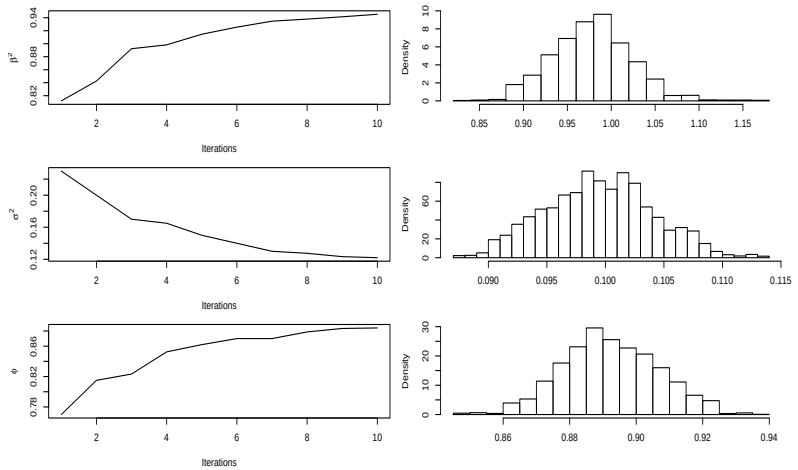


Figure 28: Weekly dataset: evolution over iterations of the Rao–Blackwellised PMC approximation *(left)* and 10th iteration weighted PMC sample *(right)*.

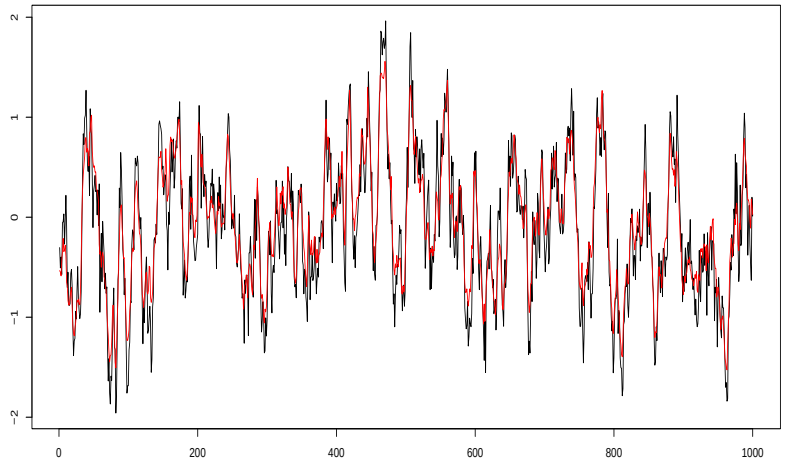


Figure 29: Weekly dataset: estimation of the stochastic volatility (in black the true volatility and in red the PMC estimation based on the 10th iteration weighted PMC sample).

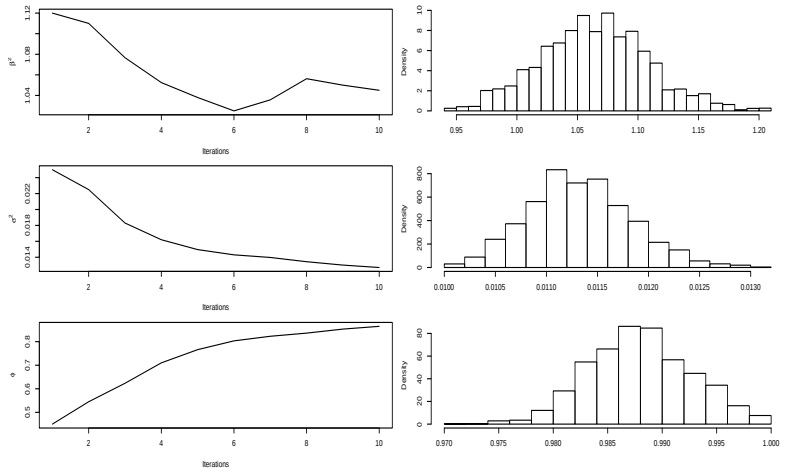


Figure 30: Daily dataset: same legend as Figure 28..

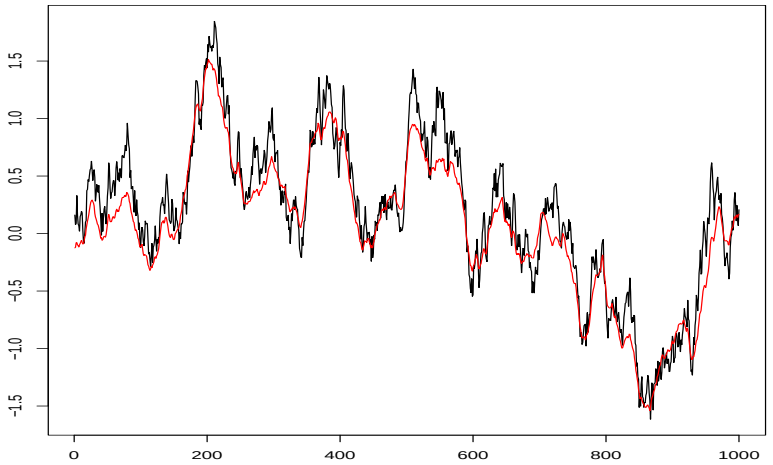


Figure 31: Daily dataset: same legend as Figure 29.

9 Perfect simulation

9.1 Propp and Wilson's

Difficulty devising MCMC stopping rules:
when should one **stop** an MCMC algorithm?!

[Robert, 1995, 1998]

Coupling from the past (CFTP): rather than start at $t = 0$ and wait till $t = +\infty$, start at $t = -\infty$ and wait till $t = 0$

[Propp & Wilson, 1996]

CFTP Algorithm

1 Start from the m possible values at time $-t$

2 Run the m chains till time 0 *(coupling allowed)*

3 Check if the chains are equal at time 0

4 If not, start further back: $t \leftarrow 2 * t$, using the *same* random numbers at time already simulated

Random mappings

Equivalent formulation

For $t = -1, -2, \ldots,$

1 Simulate a random mapping ψ_t from each state to its successor

2 Compose with the more recent random mappings, $\psi_{t'}$ $t' > t$

$$\Psi_t = \Psi_{t+1} \circ \psi_t$$

3 Check if Ψ_t is constant

Example 62 —Beta-Binomial—

$$\theta \sim \text{Beta}(\alpha, \beta) \quad \text{and} \quad X|\theta \sim \text{Bin}(n, \theta),$$

with joint density

$$\pi(x, \theta) \propto \binom{n}{x} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

and posterior density

$$\theta|x \sim \text{Beta}(\alpha + x, \beta + n - x)$$

Gibbs sampler

1 $\theta_{t+1} \sim \text{Beta}(\alpha + x_t, \beta + n - x_t)$

2 $X_{t+1} \sim \text{Bin}(n, \theta_{t+1})$.

Transition kernel

$$f((x_{t+1}, \theta_{t+1})|(x_t, \theta_t)) \propto \binom{n}{x_{t+1}} \theta^{x_{t+1}+\alpha+x_t-1} (1-\theta)^{\beta+2n-x_t-x_{t+1}-1}.$$

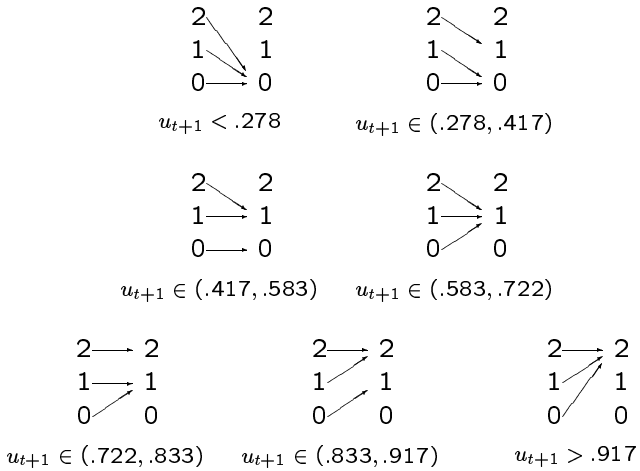
$n = 2$, $\alpha = 2$ **and** $\beta = 4$.

State space

$$\mathcal{X} = \{0, 1, 2\}\,.$$

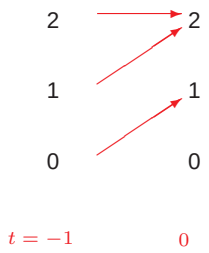
Transition probabilities

$$\begin{aligned}
\Pr(0 \mapsto 0) &= .583, \quad \Pr(0 \mapsto 1) = .333, \quad \Pr(0 \mapsto 2) = .083, \\
\Pr(1 \mapsto 0) &= .417, \quad \Pr(1 \mapsto 1) = .417, \quad \Pr(1 \mapsto 2) = .167, \\
\Pr(2 \mapsto 0) &= .278, \quad \Pr(2 \mapsto 1) = .444, \quad \Pr(2 \mapsto 2) = .278
\end{aligned}$$

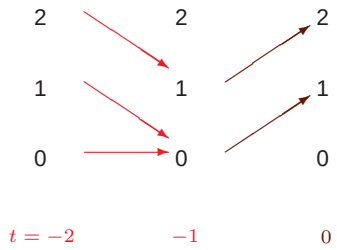


All possible transitions for the Beta-Binomial(2,2,4) example

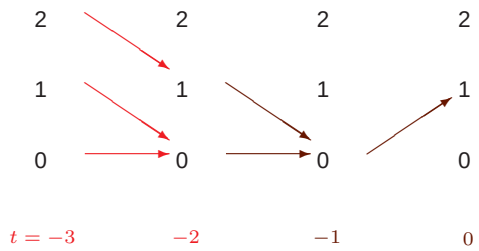
Begin at time $t = -1$ and draw U_0. Suppose $U_0 \in (.833, .917)$.



The chains have not coalesced, so go to time $t = -2$ and draw U_{-1}. Suppose $U_{-1} \in (.278, 417)$.



The chains have still not coalesced so go to time $t = -3$. Suppose $U_{-2} \in (.278, .417)$.



All chains have coalesced into $X_0 = 1$. We accept X_0 as a draw from π . Note that even though the chains have coalesced at $t = -1$, we do not accept $X_{-1} = 0$ as a draw from π.

Extension to continuous chains

[Murdoch & Green, 1998]

- **Multigamma coupling**
- Find a discretization of the continuum of states (renewal, small set, accept-reject, &tc...)
- Run CFTP for a finite number of chains

Example 63 **—Mixture models—**

Simplest possible mixture structure

$$pf_0(x) + (1 - p)f_1(x),$$

with uniform (or Beta) prior on p.

Data Augmentation Gibbs sampler:

At iteration t:

1 Generate n iid $\mathcal{U}(0, 1)$ rv's $u_1^{(t)}, \ldots, u_n^{(t)}$.

2 Derive the indicator variables $z_i^{(t)}$ as $z_i^{(t)} = 0$ iff

$$u_i^{(t)} \leq \frac{p^{(t-1)} f_0(x_i)}{p^{(t-1)} f_0(x_i) + (1 - p^{(t-1)}) f_1(x_i)}$$

and compute

$$m^{(t)} = \sum_{i=1}^{n} z_i^{(t)} .$$

3 Simulate $p^{(t)} \sim \mathcal{B}e(n + 1 - m^{(t)}, 1 + m^{(t)})$.

Corresponding CFTP :

At iteration $-t$:

1 Generate n iid uniform rv's $u_1^{(-t)}, \ldots, u_n^{(-t)}$.

2 Partition $[0,1)$ into intervals $[q_{[j]}, q_{[j+1]})$.

3 For each $[q_{[j]}^{(-t)}, q_{[j+1]}^{(-t)})$, generate

$$p_j^{(-t)} \sim \mathcal{B}e(n-j+1, j+1).$$

4 For each $j = 0, 1, \ldots, n$, $r_j^{(-t)} \leftarrow p_j^{(-t)}$

5 For $(\ell = 1,\ \ell < T,\ \ell + +)$ $r_j^{(-t+\ell)} \leftarrow p_k^{(-t+\ell)}$ with k such that

$$r_j^{(-t+\ell-1)} \in [q_{[k]}^{(-t+\ell)}, q_{[k+1]}^{(-t+\ell)}]$$

6 Stop if the $r_j^{(0)}$'s $(0 \le j \le n)$ are all equal. Otherwise, $t \leftarrow 2 * t$.

Duality Principle and marginalisation

Finite number of starting chains more obvious in the finite state space!

Equivalent version based on the simulations of the $(n+1)$ chains $m^{(t)}$ started from all possible values $m = 0, \ldots, n$

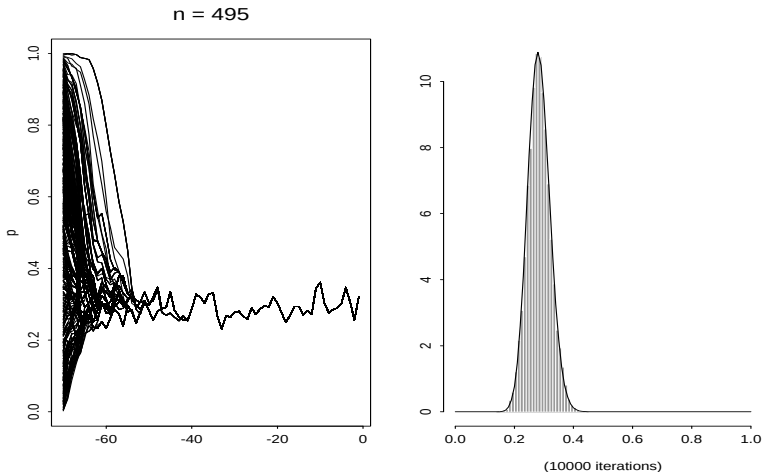


Figure 32: Simulation of $n = 495$ iid rv's from $.33\mathcal{N}(3.2, 3.2) + .67\mathcal{N}(1.4, 1.4)$ and coalescence at $t = -73$.

Coupling between chains

Follows from the $\mathcal{B}e(m+1, n-m+1)$ representation:

1 Generate $n+2$ iid exponential $\mathcal{E}xp(1)$ rv's $\omega_1, \ldots, \omega_{n+2}$.

2 Take

$$p = \frac{\sum_{i=1}^{m+1} \omega_i}{\sum_{i=1}^{n+2} \omega_i}$$

Explanation: Pool of exponentials ω_i common to all chains

Monotonicity & CFTP

Assumption of a partial or total **ordering** on the states

- Quest: maximal/majorizing and minimal/minorizing elements, $\tilde{0}$ and $\tilde{1}$
- Request: Monotone transitions *(Stochastic versus effective)*
- Conquest: Run only the chains that start from $\tilde{0}$ and $\tilde{1}$

Reduces the number of chains to examine to 2 (or more) Often delicate to implement in continuous settings

[Kendall & Møller, 1999a,b,...]

Works in the 2 component mixture case *(thanks to Beta representation trick!)*

Case $k = 3$

Gibbs sampler:

1 Generate $u_1, \ldots, u_n \sim \mathcal{U}(0,1)$.

2 Take

$$\begin{aligned} n_1 &= \sum_{i=1}^{n} \mathbb{I}\left(u_i \leq \frac{p_1 f_1(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i) + p_3 f_3(x_i)}\right), \\ n_2 &= \sum_{i=1}^{n} \left\{ \mathbb{I}\left(u_i > \frac{p_1 f_1(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i) + p_3 f_3(x_i)}\right) \right. \\ &\qquad \left. \times \mathbb{I}\left(u_i \leq \frac{p_1 f_1(x_i) + p_2 f_2(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i) + p_3 f_3(x_i)}\right) \right\}, \end{aligned}$$

and $n_3 = n - n_1 - n_2$.

3 Generate $(p_1, p_2, p_3) \sim \mathcal{D}(n_1 + 1, n_2 + 1, n_3 + 1)$.

CFTP can be implemented as for $k = 2$

But $(n+2)(n+1)/2$ different values of (n_1, n_2, n_3) to consider

No obvious monotone structure

Towards coupling

Representation of the Dirichlet $\mathcal{D}(n_1 + 1, n_2 + 1, n_3 + 1)$ distribution : if

$$\omega_{11}, \ldots, \omega_{1(n+1)}, \omega_{21}, \ldots, \omega_{3(n+1)} \sim \mathcal{E}xp(1)\,,$$

then

$$\left(\frac{\sum_{i=1}^{n_1+1} \omega_{1i}}{\sum_{j=1}^{3} \sum_{i=1}^{n_j+1} \omega_{ji}}, \frac{\sum_{i=1}^{n_2+1} \omega_{2i}}{\sum_{j=1}^{3} \sum_{i=1}^{n_j+1} \omega_{ji}}, \frac{\sum_{i=1}^{n_3+1} \omega_{3i}}{\sum_{j=1}^{3} \sum_{i=1}^{n_j+1} \omega_{ji}} \right)$$

is a $\mathcal{D}(n_1 + 1, n_2 + 1, n_3 + 1)$ rv..

Common pool of $3(n+1)$ exponential rv's.

Lozenge monotonicity

The image of the triangle

$$\mathcal{T} = \{(n_1, n_2); n_1 + n_2 \leq n\}$$

by Gibbs is contained in the lozenge

$$\mathcal{L} = \{(n_1, n_2); \underline{n}_1 \leq n_1 \leq \overline{n}_1,\ n_2 \geq 0,\ \underline{n}_3 \leq n - n_1 - n_2 \leq \overline{n}_3\}\,,$$

where

- $\underline{n}_1$ is $\min n_1$ over the images of the left border of $\mathcal{T}$
- $\overline{n}_3$ is the n_3 coordinate of the image of $(0, 0)$,
- $\overline{n}_1$ is the n_1 coordinate of the image of $(n, 0)$,
- $\underline{n}_3$ is $\min n_3$ over the images of the diagonal of $\mathcal{T}$.

[Hobert & al., 1999]

Lozenge monotonicity (explained)

For a fixed n_2,

$$\frac{p_2}{p_1} = \sum_{i=1}^{n_2+1} w_{2i} \Big/ \sum_{i=1}^{n_1+1} w_{1i} \quad \text{and} \quad \frac{p_3}{p_1} = \sum_{i=1}^{n-n_1-n_2+1} w_{3i} \Big/ \sum_{i=1}^{n_1+1} w_{1i}$$

are both decreasing in n_1.

So is

$$m_1 = \sum_{i=1}^{n} \mathbb{I}\left(u_i \leq \left[1 + \frac{p_2 f_2(x_i) + p_3 f_3(x_i)}{p_1 f_1(x_i)}\right]^{-1}\right)\,.$$

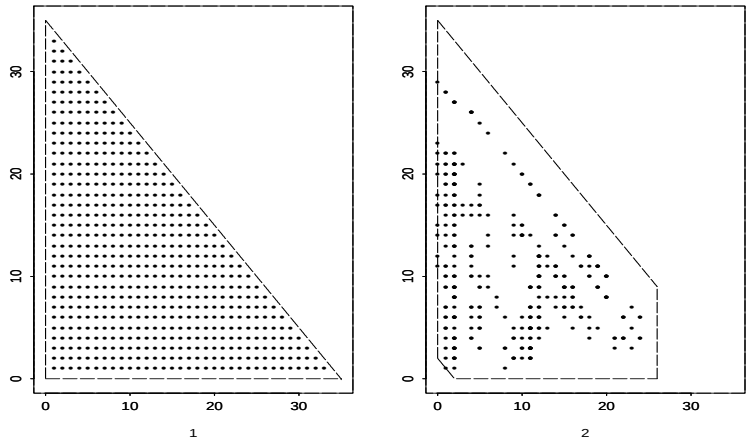


Figure 33: Sample of $n = 35$ observations from $.23\mathcal{N}(2.2, 1.44) + .62\mathcal{N}(1.4, 0.49) + .15\mathcal{N}(0.6, 0.64)$

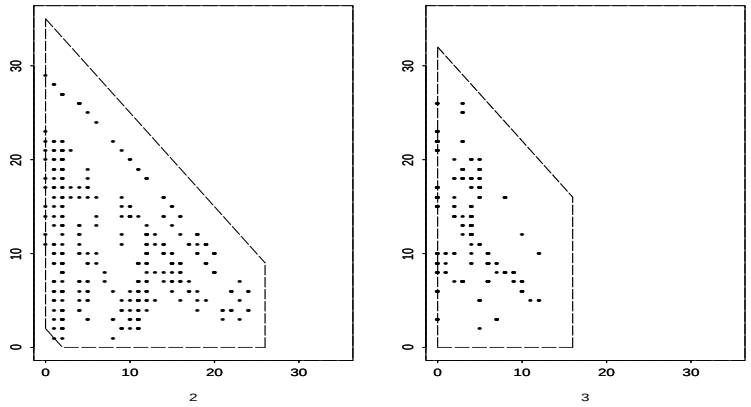
Lozenge monotonicity (preserved)

The image of $\mathcal{L}$ is contained in

$$\mathcal{L}' = \{(m_1, m_2); \underline{m}_1 \leq m_1 \leq \overline{m}_1\,, m_2 \geq 0,\ \underline{m}_3 \leq m_3 \leq \overline{m}_3\},$$

where

- $\underline{m}_1$ is $\min n_1$ over the images of the left border $\{n_1 = \underline{n}_1\}$
- $\overline{m}_1$ is $\max n_1$ over the images of the right border $\{n_1 = \overline{n}_1\}$
- $\underline{m}_3$ is $\min n_3$ over the images of the upper border $\{n_3 = \underline{n}_3\}$
- $\overline{m}_3$ is $\max n_3$ of the images of the lower border $\{n_3 = \overline{n}_3\}$



Lozenge monotonicity (completed)

- Envelope result: generation of the images of all points on the borders of $\mathcal{L}$

[Kendall, 1998]

- $O(n)$ complexity versus $O(n^2)$ for brute force CFTP
- Checking for coalescence of the borders only : *almost perfect* !
- Extension to $k = 4$ underway

[Machida, 1999]

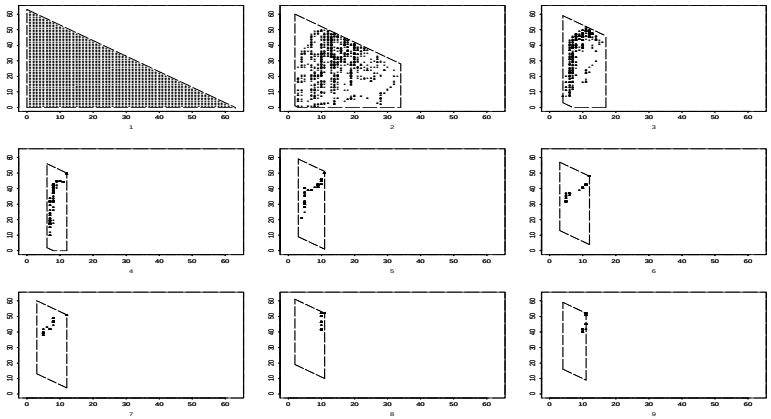


Figure 34: $n = 63$ observations from $.12\mathcal{N}(1.1, 0.49) + .76\mathcal{N}(3.2, 0.25) + .12\mathcal{N}(2.5, 0.09)$

Interruptable version

For impatient users: if we just stop runs that take "too long", *this gives biased results*

Fill's algorithm:

1. Choose arbitrary time T and set $x_T = z$
2. Generate $X_{T-1}|x_T$, $X_{T-2}|x_{T-1}, \ldots, X_0|x_1$ from the reversed chain
3. Generate $[U_1|x_0, x_1], \ldots, [U_T|x_{T-1}, x_T]$
4. Begin chains in all states at $T = 0$ and use common $U_1, \ldots, U_T$ to update all chains
5. If the chains have coalesced in z by T, accept x_0 as a draw from π
6. Otherwise begin again, possibly with new T and z.

[Fill, 1996]

Proof

Need to prove $\Pr[X_0 = x | C_T(z)] = \pi(x)$

$$\Pr[X_0 = x | C_T(z)] = \frac{\Pr[z \to x] \Pr[C_T(z) | x \to z]}{\sum_{x'} \Pr[z \to x'] \Pr[C_T(z) | x' \to z]}.$$

Now for every x'

$$\Pr[C_T(z) | x' \to z] = \frac{\Pr[C_T(z) \text{ and } x' \to z]}{\Pr[x' \to z]} = \frac{\Pr[C_T(z)]}{\Pr[x' \to z]},$$

and, since $\Pr[x' \to z] = K^T(x', z)$,

$$\Pr[X_0 = x | C_T(z)] = \frac{K^T(z, x) \Pr[C_T(z)] / K^T(x, z)}{\sum_{x'} K^T(z, x') \Pr[C_T(z)] / K^T(x', z)}$$

$$= \frac{K^T(z, x) / K^T(x, z)}{\sum_{x'} K^T(z, x') / K^T(x', z)},$$

Using detailed balance,

$$K^T(z, x) / K^T(x, z) = \pi(x) / \pi(z)\,,$$

and thus,

$$\Pr[X_0 = x | C_T(z)] = \frac{\pi(x)/\pi(z)}{\sum_{x'} \pi(x')/\pi(z)} = \pi(x).$$

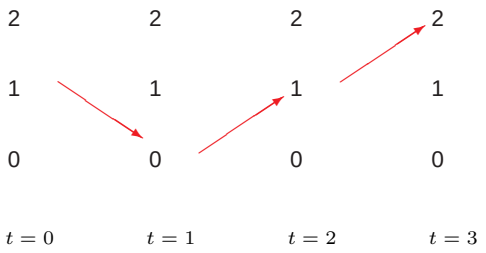
Example 64 **—Beta-Binomial—**

Choose $T = 3$ and $X_T = 2$.

Reversible chain, so

$$\begin{aligned} X_2 | X_3 = 2 &\sim \text{BetaBin}(2, 4, 4) \\ X_1 | X_2 = 1 &\sim \text{BetaBin}(2, 3, 5) \\ X_0 | X_1 = 2 &\sim \text{BetaBin}(2, 4, 4) \end{aligned}$$

Suppose



$$X_0 = 1, \quad X_1 = 0, \quad X_2 = 1 \quad \text{and} \quad X_3 = 2$$

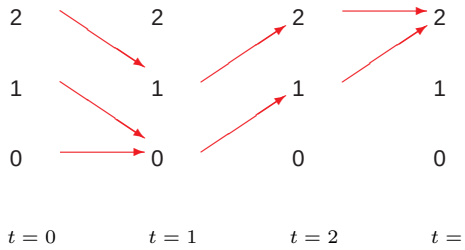
imply

$$U_1 \sim \mathsf{U}(0, .417), \quad U_2 \sim \mathsf{U}(.583, .917), \quad U_3 \sim \mathsf{U}(.833, 1)$$

Suppose

$$U_1 \in (.278, .417) \quad U_2 \in (.833, .917) \quad U_3 > .917$$

Begin chains in states 0, 1 and 2.



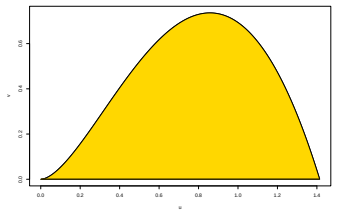
The chains coalesce in $X_3 = 2$; so we accept $X_0 = 1$ as a draw from π.

9.2 Slice sampling

Remember that slice sampling associated with π amounts to simulation from

$$\mathcal{U}\left(\{\omega\,;\, \pi(\omega) \geq u\pi(\omega_0)\}\right)$$

and $u \sim \mathcal{U}([0,1])$



Properties

Slice samplers do not require normalising constants

Slice samplers induce a natural order

If $\pi(\omega_1) \leq \pi(\omega_2)$

$$\mathcal{A}_2 = \{\omega\,;\, \pi(\omega) \geq u\pi(\omega_2)\} \subset \mathcal{A}_1 = \{\omega\,;\, \pi(\omega) \geq u\pi(\omega_1)\}$$

Slice samplers induce a natural discretization of continuous state space

[Mira, Møller & Roberts, 2001]

Slice samplers preserve monotonicity

1 Start from $\tilde{0} = \arg\min \pi(\omega)$ and $\tilde{1} = \arg\max \pi(\omega)$

2 Generate $u_{-t}, \ldots, u_0$

3 Get the successive images of $\tilde{0}$ for $t = -T, \ldots, 0$

4 Check if those are acceptable as successive images of $\tilde{1}$
If not, generate the corresponding images

But slice samplers are real hard to implement: for instance,

$$\mathcal{U}\left(\left\{\theta\,;\,\prod_{i=1}^{n}\sum_{j=1}^{k} p_j f(x_i \mid \theta_j) \geq \epsilon\right\}\right)$$

is impossible to simulate

Duality principle

Dual marginalization: integrate out the parameters (θ, p) in

$$\mathbf{z}, \theta \mid \mathbf{x} \sim \pi(\theta, p) \prod_{i=1}^{n} p_{z_i} f(x_i \mid \theta_{z_i})$$

Easily done in conjugate (exponential) settings.

Use the slice sampler on the marginal posterior of $\mathbf{z}$

- Finite state space
- Link with Rao–Blackwellisation
- Perfect sampling on $\mathbf{z}$ equivalent to perfect sampling on θ

Example 65 —Exponential example ($k = 2, p$ known)

Joint distribution

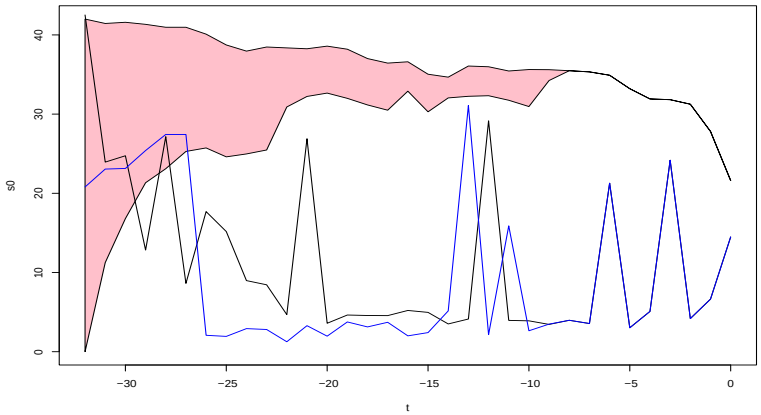
$$\prod_{i=1}^{n} p^{(1-z_i)}(1-p_i)^{z_i}\lambda_{z_i}\exp(-\lambda_{z_i}x_i)\prod_{j=1}^{k}\lambda_j^{\alpha_j-1}\exp(-\lambda_j\beta_j)$$

leads to

$$\mathbf{z}\mid\mathbf{x}\sim p^{n_0}(1-p)^{n_1}\frac{\Gamma(\alpha_0+n_0-1)\Gamma(\alpha_1+n_1-1)}{(\beta_0+s_0)^{\alpha_0+n_0}(\beta_1+s_1)^{\alpha_1+n_1}}\,.$$

- Closed form computable expression (up to constant)
- Factorises through (n_0, s_0), sufficient statistic
- Maximum $\tilde{1}$ and minimum $\tilde{0}$ can be derived

But... slice sampler still difficult to implement
because of number of values of s_0 : $\binom{n}{n_0}$
Still, feasible for small values of n $(n \leq 40)$



Fixed n_0, 40 observations

Perfect sampling is possible!

Idea: Use Breyer and Roberts' (1999) automatic coupling:

If

$$x_1^{(t+1)} = \begin{cases} y_t \sim q(y|x_1^{(t)}) & \text{if } u_t \leq \frac{\pi(y_t)\,q(x_1^{(t)}|y_t)}{\pi(x_1^{(t)})\,q(y_t|x_1^{(t)})}, \\ x_1^{(t)} & \text{otherwise.} \end{cases}$$

generate

$$x_2^{(t+1)} = \begin{cases} y_t & \text{if } u_t \leq \frac{\pi(y_t)\,q(x_2^{(t)}|x_1^{(t)})}{\pi(x_2^{(t)})\,q(y_t|x_1^{(t)})}, \\ x_2^{(t)} & \text{otherwise.} \end{cases} \tag{7}$$

Theorem In the special case

$$q(y|x) = h(y),$$

if $(x_1^{(t)})$ starts from

$$\tilde{0} = \arg\min \pi/h,$$

if $(x_2^{(t)})$ starts from

$$\tilde{1} = \arg\max \pi/h,$$

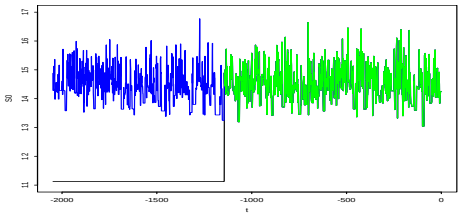
the coupling (7) preserves the ordering.

[Now, this is a result from Corcoran and Tweedie!!!]

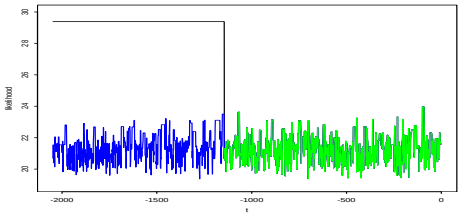
Example When state space $\mathcal{X}$ compact,
use for h the uniform distribution on $\mathcal{X}$.

Extremal elements $\tilde{0}$ and $\tilde{1}$ then induced by π only.

Implementation: start from arbitrary value for $x_1^{(0)}$ and keep proposing for $x_2^{(0)} = \tilde{1}$



Coupling history



Corresponding likelihoods

Back to Basics!

When $\mathcal{X}$ compact, and $\pi(x) \leq \pi(\tilde{1})$, independent Metropolis–Hasting coupling **is accept–reject**, based on uniform proposals

Reason:

When coupling occurs, $x_2^{(t)} = y_t$,

$$u_t \leq \frac{\pi(y_t)}{\pi(\tilde{1})} = \frac{\pi(y_t)}{\max \pi}$$

and therefore the chain is in stationnary regime at coupling time.

This extends to the general case, with accept–reject based on proposal h.

In this case, the accept–reject algorithm could have been conceived independently from perfect sampling (?)

while Fill's (1998) algorithm is an accept–reject algorithm in disguise, but it could not have been conceived independently from perfect sampling

9.3 Kacs' formula

Consider two Markov kernels K_1 and K_2

What of the mixture

$$K_3 = pK_1 + (1-p)K_2\,?$$

Stability (1)

If K_1 and K_2 are recurrent kernels, the mixture kernel K_3 is recurrent.

Stability (2)

If K_1 and K_2 define positive recurrent chains with the same potential function V, that is, there exist a small set C, $\lambda < 1$, $V \geq 1$ and V bounded on C such that

$$\mathbb{E}_{K_i}[V(x)|y] = \lambda V(y) + b\mathbb{I}_C(y)$$

then the mixture kernel K_3 is also positive recurrent.

Stationary measure

If $\pi_1 = \pi_2$ and K_3 is positive recurrent, π_1 is its stationary distribution.

Otherwise...

Special case: K_1 is an iid kernel π_1. Then

$$K_3 = p\pi_1 + (1-p)K_2$$

No assumption on K_2 (it can even be transient!) but, still,

Theorem 3 **K_3 is positive recurrent with stationary distribution**

$$\pi_3 = \sum_{i=0}^{+\infty} (1-p)^i p\, P_2^i \pi_1 \ ,$$

when $P_2^i\pi_1$ is the transform of π_1 under i transitions using K_2.

Special special case: K_3 is uniformly ergodic:

$$K_3(x,y) \geq \varepsilon\nu(y), \qquad \forall\, x \in \mathcal{X} \ ,$$

Mixture decomposition:

$$\begin{aligned} K_3(x,y) &= \varepsilon\nu(y) + (1-\varepsilon)\frac{K_3(x,y) - \varepsilon\nu(y)}{1-\varepsilon} \\ &= \varepsilon\nu(y) + (1-\varepsilon)K_2(x,y) \end{aligned}$$

Representation of the stationary distribution:

$$\sum_{i=0}^{+\infty} \varepsilon(1-\varepsilon)^i P_2^i \nu \ ,$$

where P_2 is associated with K_2

1 Simulate $x_0 \sim \nu, \omega \sim \mathcal{G}eo(\varepsilon)$.

2 Run the transition $x_{t+1} \sim K_2(x_t, y)$ $t = 0, \cdots, \omega - 1$, and take x_ω.

[Murdoch and Green, 1998]

General case

Minorizing condition

$$K_3(x, y) \geq \varepsilon\nu(y)\mathbb{I}_C(x) \qquad [MNRZ]$$

Splitting decomposition

$$\begin{aligned} K_3(x,y) &= \left\{\varepsilon\nu(y) + (1-\varepsilon)\frac{K_3(x,y) - \varepsilon\nu(y)}{1-\epsilon}\right\}\mathbb{I}_C(y) + K_3(x,y)\mathbb{I}_{C^c}(y) \\ &= \{\varepsilon\nu(y) + (1-\varepsilon)K_2(x,y)\}\mathbb{I}_C(y) + K_3(x,y)\mathbb{I}_{C^c}(y) \end{aligned}$$

[Nummelin, 1984]

K_2 is the *depleted measure* of K_3

Introduction of the *split chain* $\Phi^\star = \{(X_n, \delta_n)\}_n$, on $\mathcal{X} \times \{0, 1\}$, with transition kernel

$$P'\left[(x,0), A \times \delta\right] = \begin{cases} \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] K_3(x, A) & x \notin C \\ \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] K_2(x, A) & x \in C \end{cases}$$

and

$$P'\left[(x,1), A \times \delta\right] = \begin{cases} \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] K_3(x, A) & x \notin C \\ \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] \nu(A) & x \in C \end{cases}$$

where $\delta \in \{0, 1\}$ *(renewal indicator)*

[Athreya and Ney, 1984]

Then $\alpha := C \times \{1\}$ is an *accessible atom*

1 Simulate $X_n \sim K_3(x_{n-1}, \cdot)$

2 Simulate δ_{n-1} conditional on (x_{n-1}, x_n)

$$\Pr(\delta_{n-1} = 1 | x_{n-1}, x_n) = \frac{\varepsilon\nu(x_n)}{K_3(x_{n-1}, x_n)}$$

[Mykland, Tierney and Yu, 1995]

General Mixture Representation

Let τ_α be the first return time to α

$$\tau_\alpha = \min\{n \geq 1 : (X_n, \delta_n) \in \alpha\} \ .$$

and

$$\Pr_\alpha(\cdot) \quad \text{and} \quad \mathbb{E}_\alpha(\cdot)\,,$$

probability and expectation conditional on $(X_0, \delta_0) \in \alpha$

Tail renewal time T^*

$$\Pr(T^* = t) = \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\mathbb{E}_\alpha(\tau_\alpha)}$$

If the chain is recurrent, $\mathbb{E}_\alpha(\tau_\alpha) < \infty$

Theorem 4 **If $(X_n)_n$ is μ-irreducible, aperiodic, and Harris recurrent with invariant probability distribution π, with a minorization condition [MNRZ], then**

$$\pi(A) = \sum_{t=1}^{\infty} \Pr(N_t \in A) \Pr(T^* = t)$$

where N_t is equal in distribution to X_t given $X_1 \sim \nu(\cdot)$ and given *no regenerations before time t***.**

Follows from Kac's theorem

$$\pi(A) = \frac{1}{\mathbb{E}_\alpha(\tau_\alpha)} \sum_{t=1}^{\infty} \Pr_\alpha(X_t \in A, \tau_\alpha \geq t)$$

Can be extended to stationary measures