Population Monte Carlo Methods

Christian P. Robert
Université Paris Dauphine

1 A Benchmark example

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

Example 1 –Mixture models–

Models with density

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

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.

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.

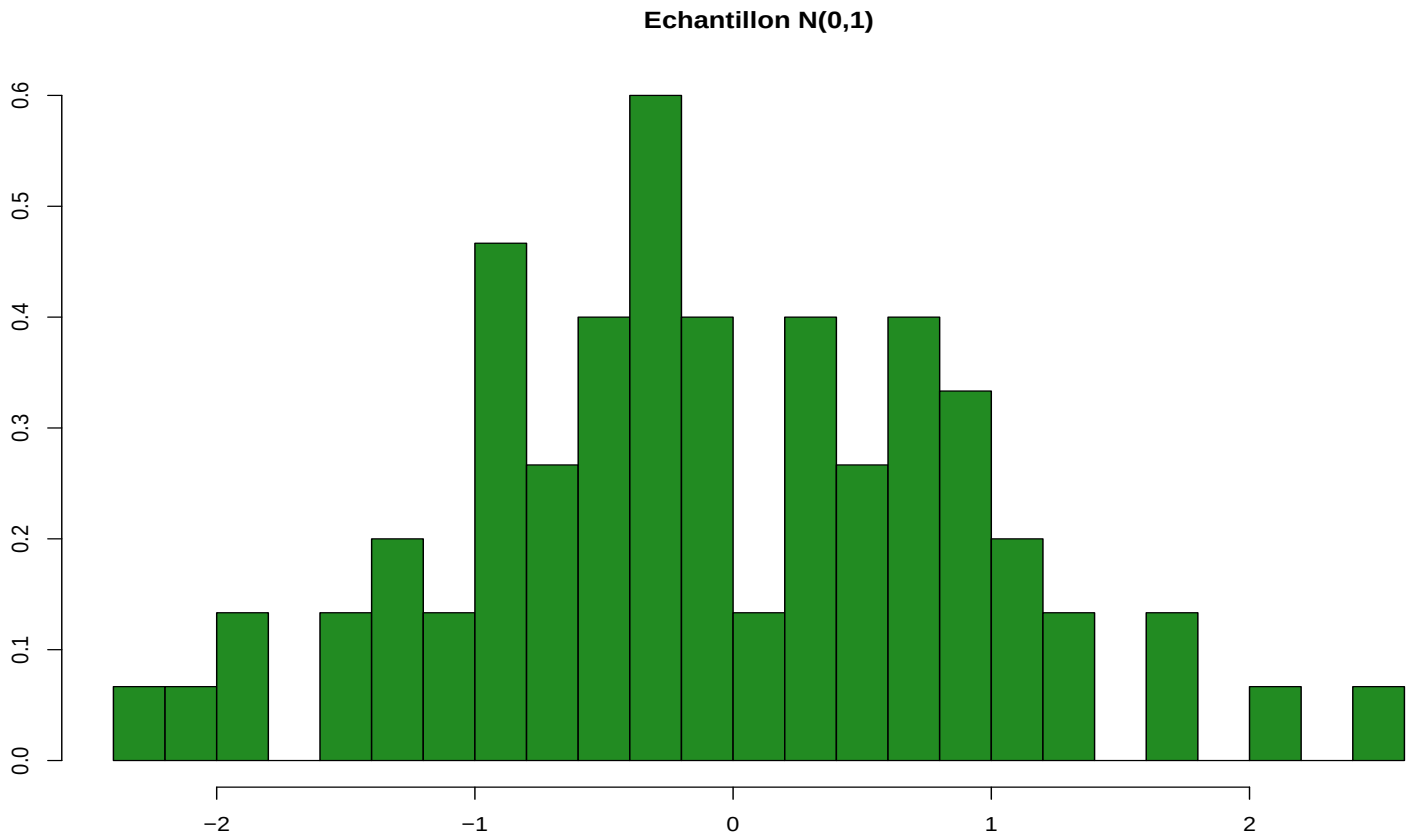
In the special case

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

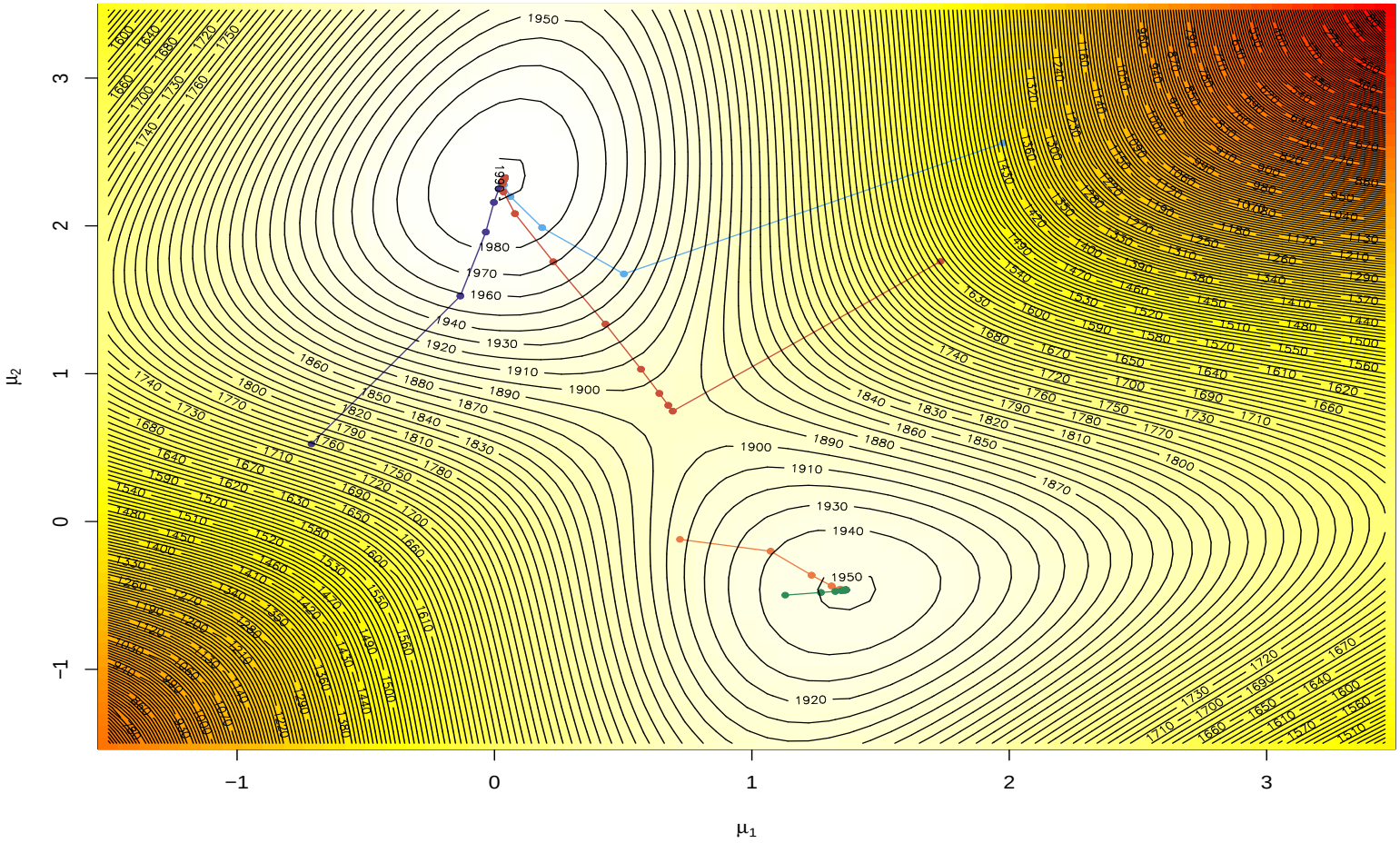
with $\epsilon > 0$ known

Then, whatever n, the likelihood is unbounded:

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



Sample from (1)



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

Similar difficulties with the Bayesian approach:

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

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

where

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

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

2 Monte Carlo Integration

2.1 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 2 –Bayesian decision theory–

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

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

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

2.2 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 3 –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

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

2.3 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.

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.

Example 4 –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

3 The Metropolis-Hastings Algorithm

3.1 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 $\left(X^{(t)}\right)$ is generated using a transition kernel with stationary distribution f

- Insures the convergence in distribution of $\left(X^{(t)}\right)$ 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.

3.2 The Metropolis–Hastings algorithm

3.2.1 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 5 –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

3.2.2 Convergence properties

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

$$f(y)\,K(y,x) = f(x)\,K(x,y)$$

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

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

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.

3.3 A Collection of Metropolis-Hastings Algorithms

3.3.1 The Independent Case

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

Algorithm 6 –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 7 –Generating gamma variables–

Generate the $\mathcal{G}a(\alpha, \beta)$ distribution using a gamma $\mathcal{G}a([\alpha], b = [\alpha]/\alpha)$ candidate

Algorithm 8 –Gamma accept-reject–

---

1. Generate $Y \sim \mathcal{G}a([\alpha], [\alpha]/\alpha)$
2. Accept $X = Y$ with prob.

$$\left(\frac{e\ y\ \exp(-y/\alpha)}{\alpha} \right)^{\alpha - [\alpha]} .$$

---

and

Algorithm 9 –Gamma Metropolis-Hastings–

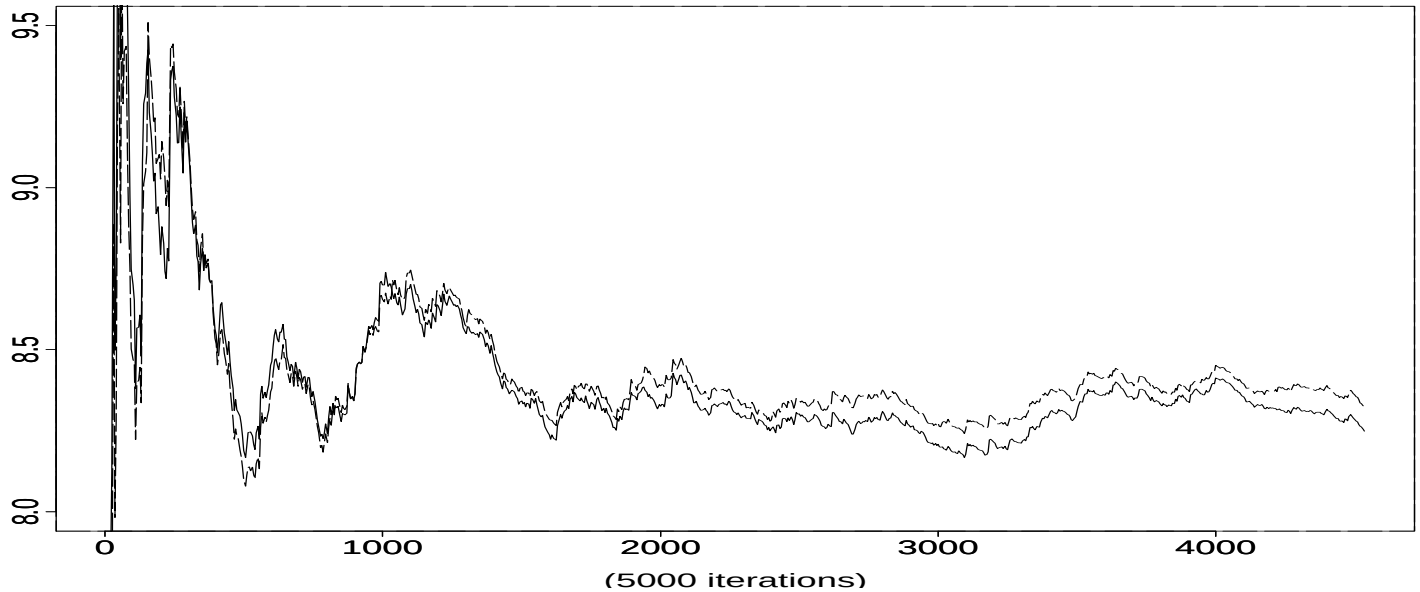
---

1. Generate $Y_t \sim \mathcal{G}a([\alpha], [\alpha]/\alpha)$
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \left(\dfrac{Y_t}{x^{(t)}} \exp\left\{\dfrac{x^{(t)} - Y_t}{\alpha}\right\}\right)^{\alpha - [\alpha]} \\ x^{(t)} & \text{otherwise.} \end{cases},$$

Comparison

Close agreement in M-H and A-R, with a slight edge to M-H.



Accept-reject (solid line) vs. Metropolis–Hastings (dotted line) estimators of $\mathbb{E}_f[X^2] = 8.33$**, for** $\alpha = 2.43$ **based on** $\mathcal{G}a(2, 2/2.43)$

3.3.2 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 10 –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 11 –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.

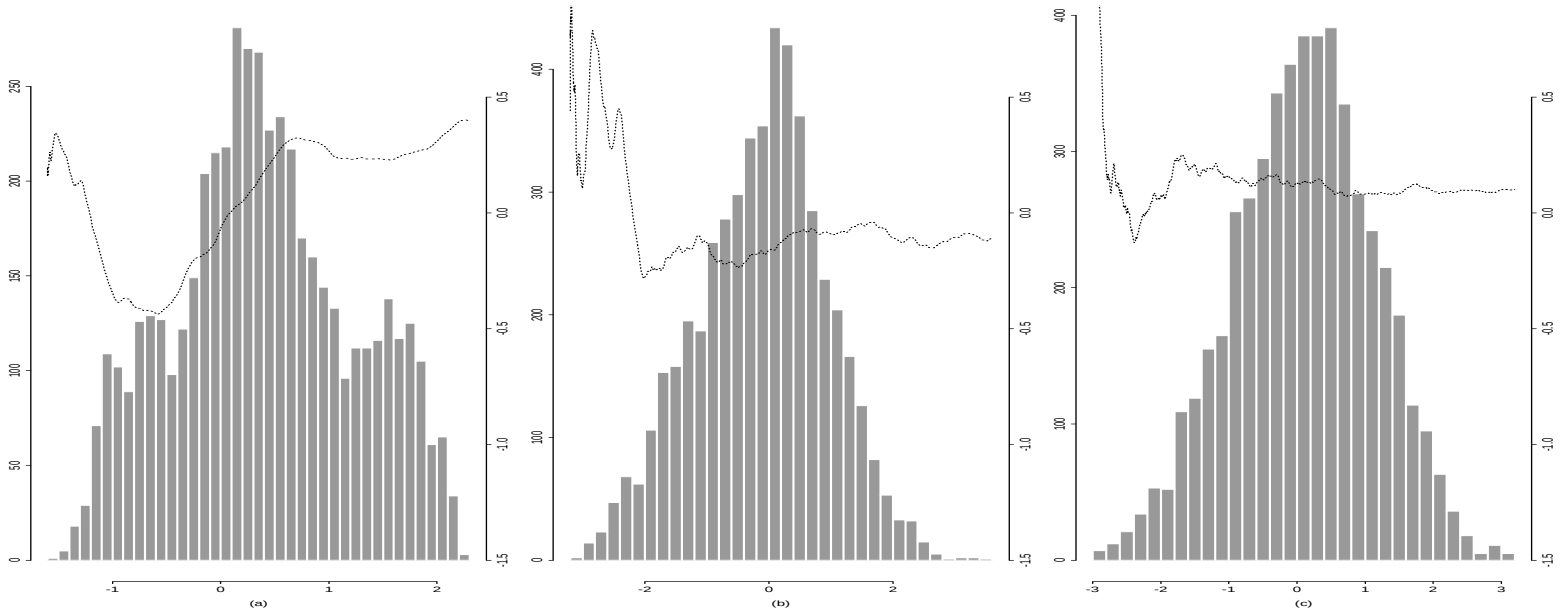


Figure 1: Three samples based on $\mathcal{U}[-\delta, \delta]$ with (a) $\delta = 0.1$, (b) $\delta = 0.5$ and (c) $\delta = 1.0$, superimposed with the convergence of the means ($15,000$ simulations).

Example 12 —Mixture models—

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

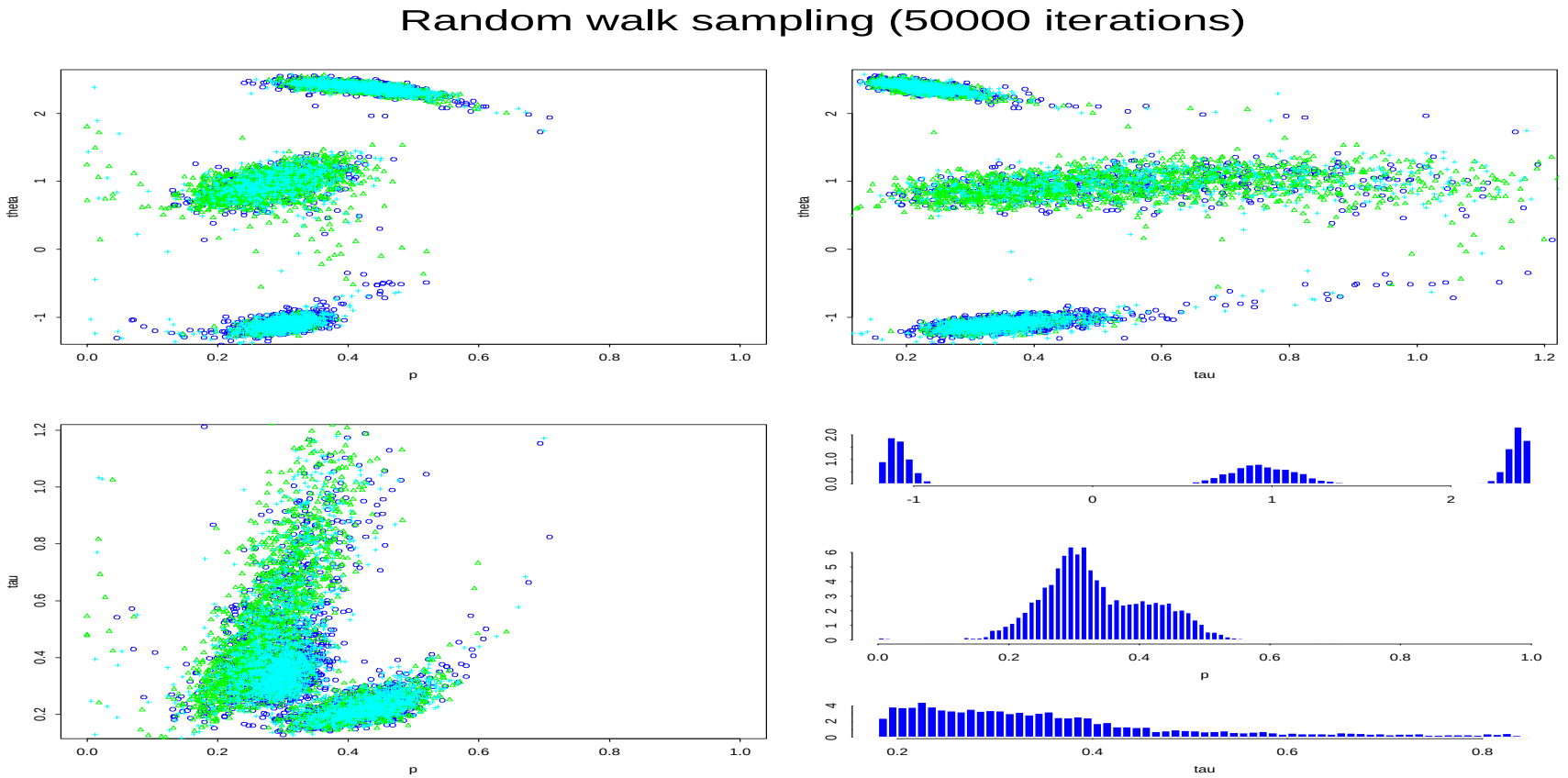
Metropolis-Hastings proposal:

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

where

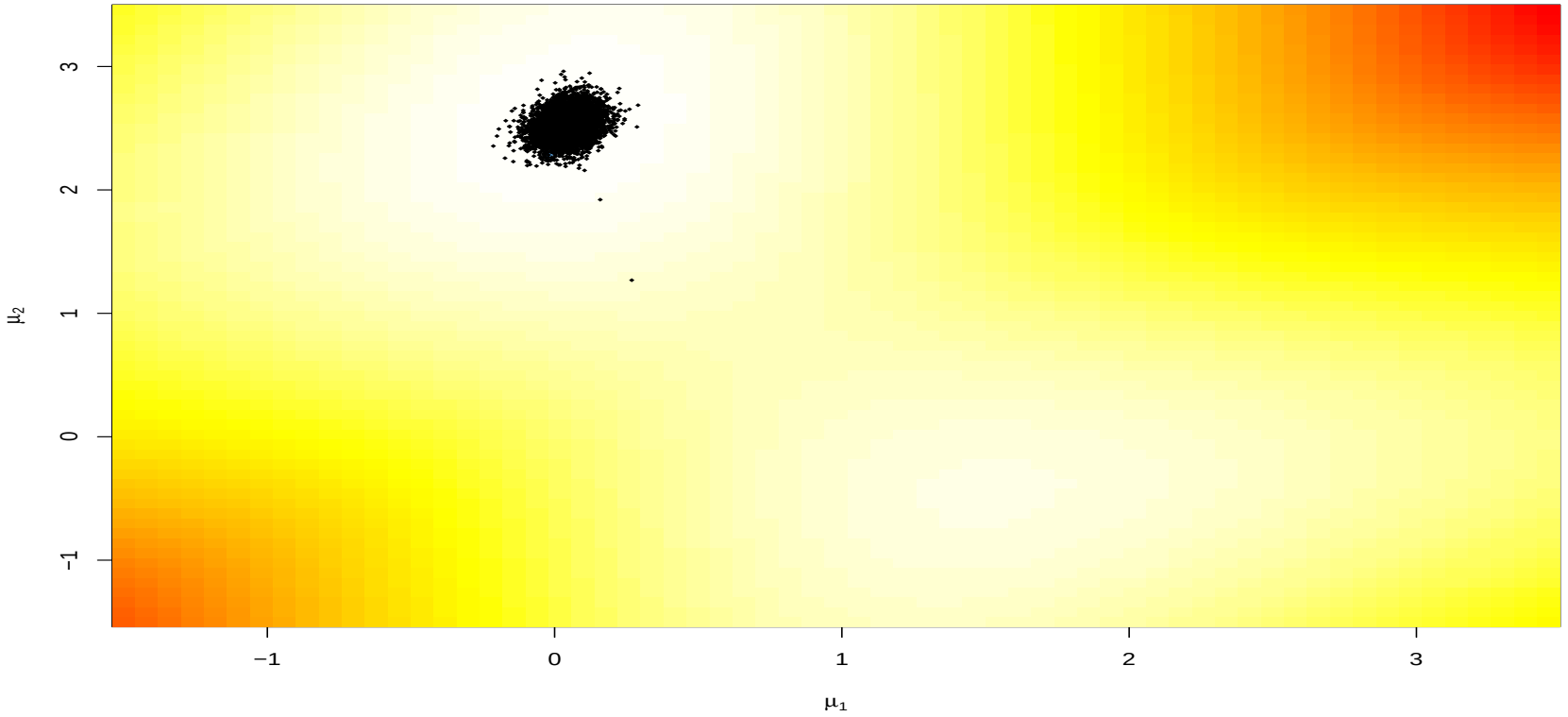
$$\rho^{(t)} = \frac{\pi(\theta^{(t)} + \omega\varepsilon^{(t)}|x)}{\pi(\theta^{(t)}|x)} \wedge 1$$

and ω scaled for good acceptance rate



[Celeux & al., 2000]

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



Convergence properties

Uniform ergodicity prohibited by random walk structure

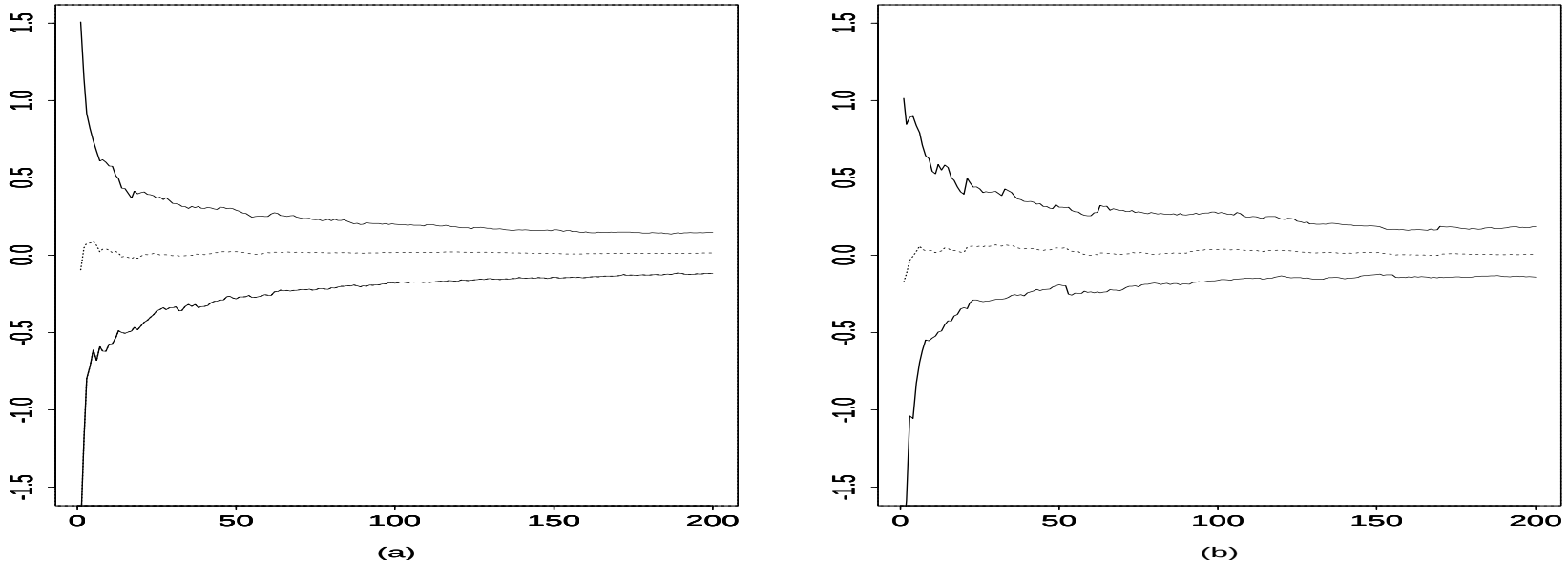
At best, geometric ergodicity:

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 13 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

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

3.4 Extensions

There are many other algorithms

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

to name a few...

3.4.1 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]

3.4.2 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 14 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]

4 The Gibbs Sampler

4.1 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 15 –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 16 –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.

4.1.1 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 17 –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. $Y_2^{(t+1)} \sim g_2(y_2 | y_1^{(t+1)}, y_3^{(t)}, \ldots, y_p^{(t)})$,

   $\ldots$

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

Example 18 —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

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

Corresponding Gibbs sampler

1. Simulate

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

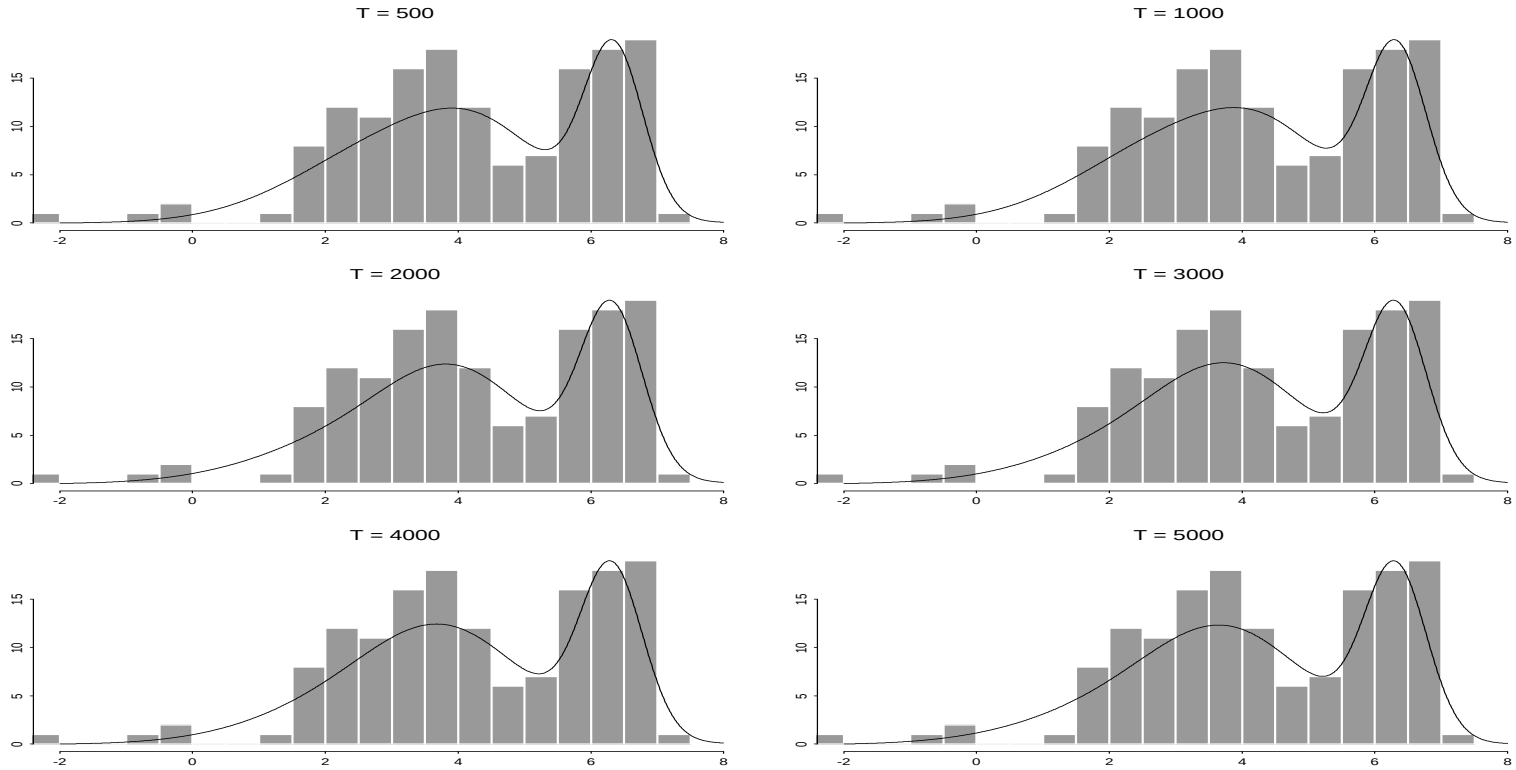
2. Simulate $(j = 1, \ldots, n)$

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

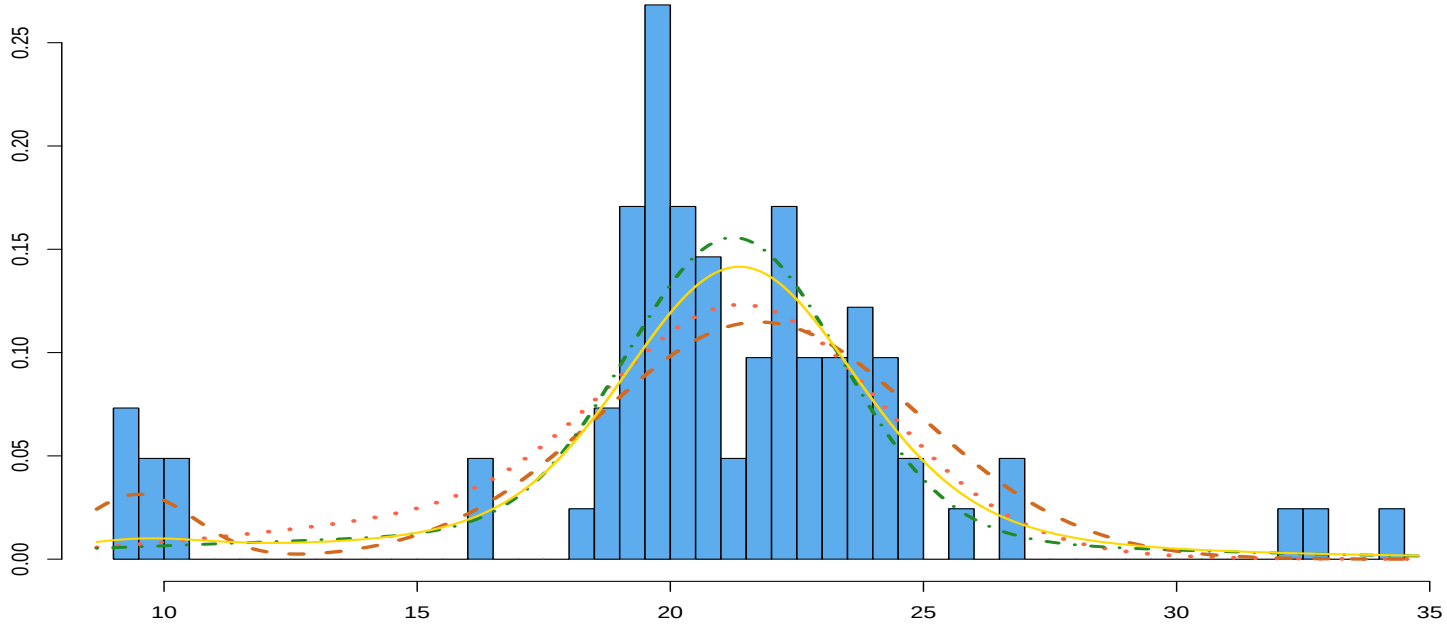
with $(i = 1, \ldots, k)$

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

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)

4.1.2 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.

4.1.3 Slice sampler

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

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

it can be completed

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

leading to the following Gibbs algorithm:

Algorithm 19 –Slice sampler–

Simulate

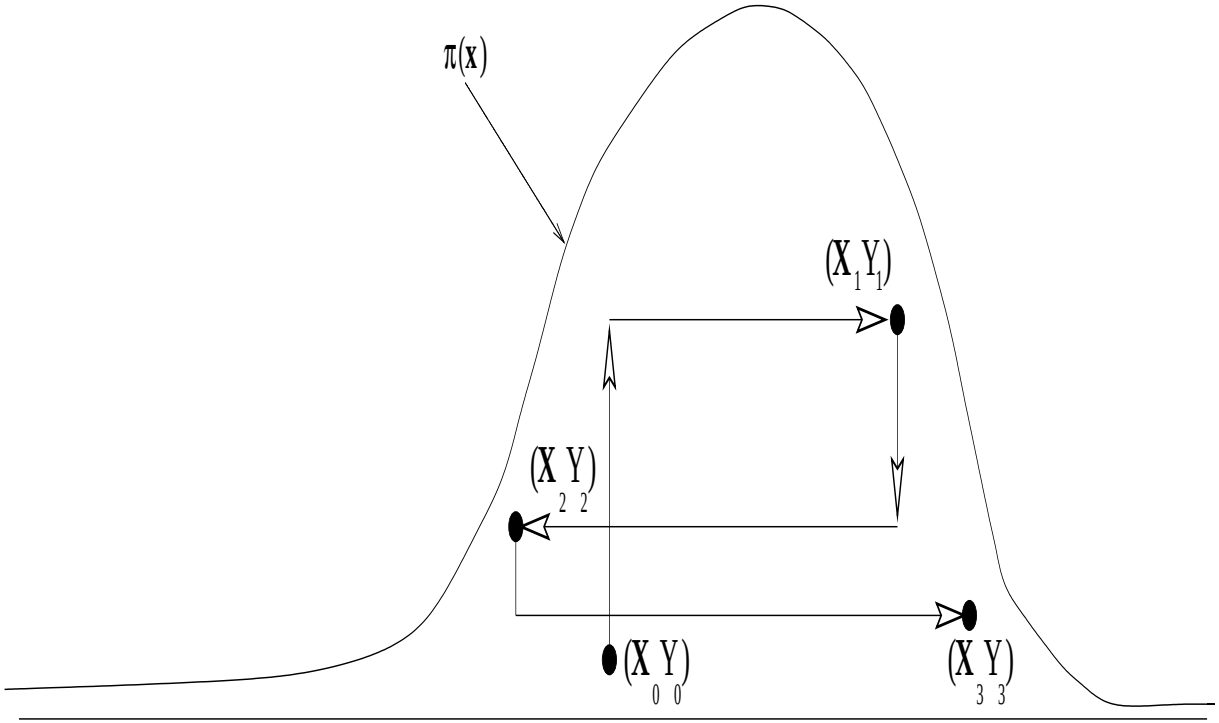
1. $\omega_1^{(t+1)} \sim \mathcal{U}_{[0,f_1(\theta^{(t)})]}$;

$\ldots$

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

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



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.

4.1.4 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, $\left(Y^{(t)}\right)$ 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 20 —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 21 —A poor slice sampler—

Consider

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

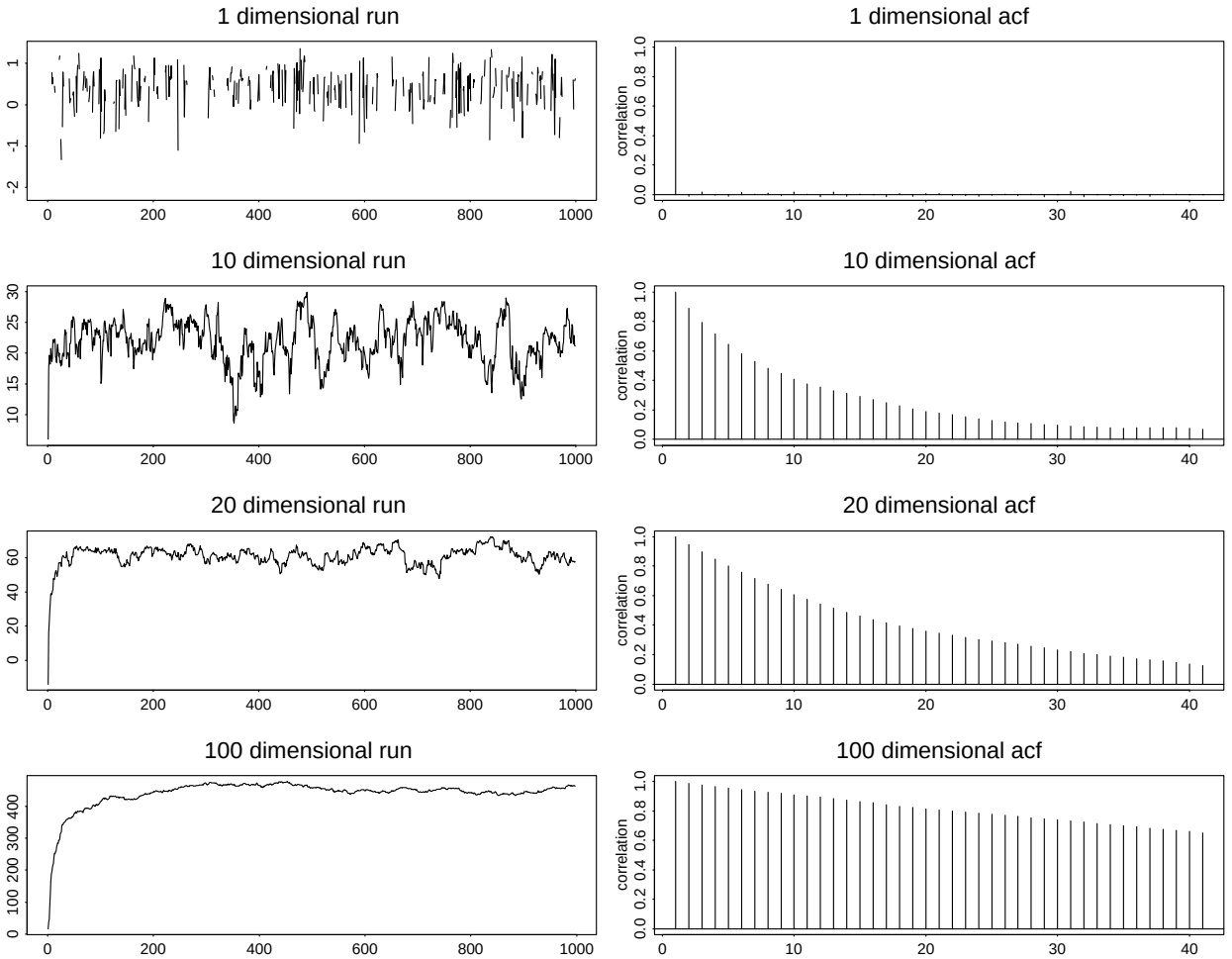
Slice sampler equivalent to one-dimensional slice sampler on

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

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)

4.1.5 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 $y' \in \mathcal{Y}$.

4.1.6 Hierarchical models

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

Example 22 –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

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

4.2 Data Augmentation

The Gibbs sampler with only two steps is particularly useful

Algorithm 23 –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 24 –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

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

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 25 –Poisson-Gamma Gibbs–

1.. Simulate $Y_i^{(t)} \sim \mathcal{P}(\lambda^{(t-1)})\, \mathbb{I}_{y\geq 4} \quad i = 1, \ldots, 13$

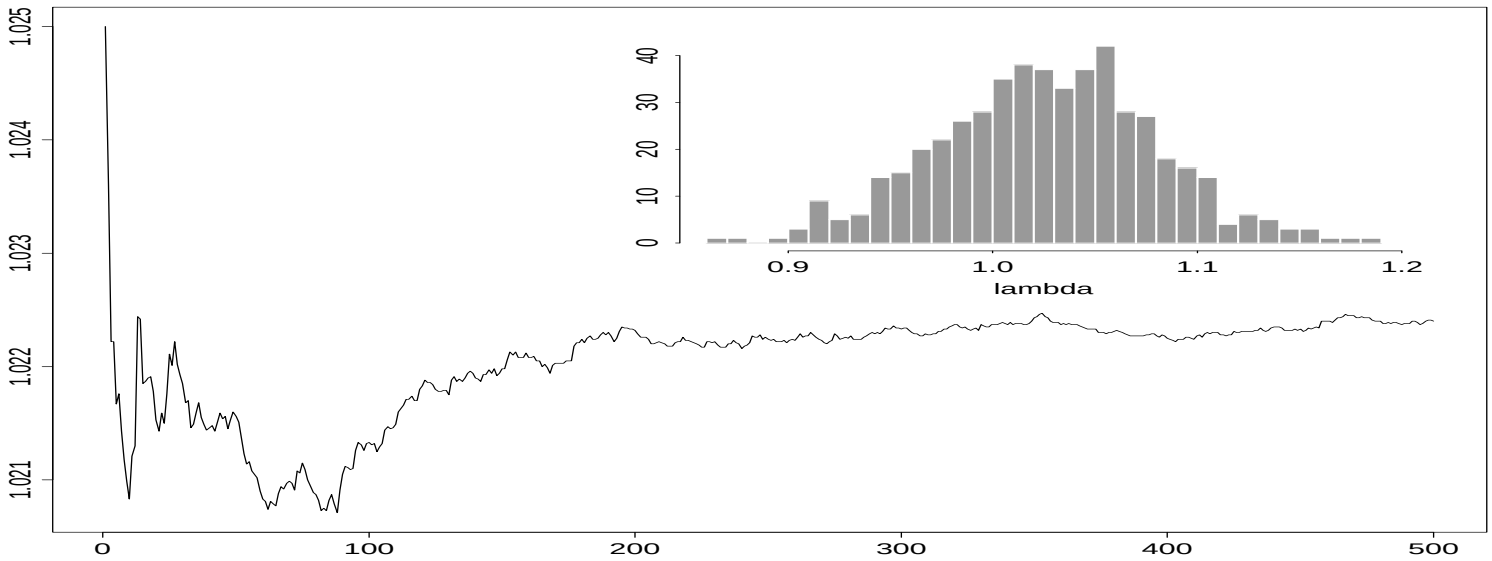
2.. Simulate

$$\lambda^{(t)} \sim \mathcal{G}a\left(313 + \sum_{i=1}^{13} y_i^{(t)},\ 360\right) .$$

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


1.025
1.024
1.023
1.022
1.021
0
100
200
300
400
500
0
10
20
30
40
0.9
1.0
1.1
1.2
lambda

4.2.1 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

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

and is unbiased.

4.2.2 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.

4.3 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 26 –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 27 –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

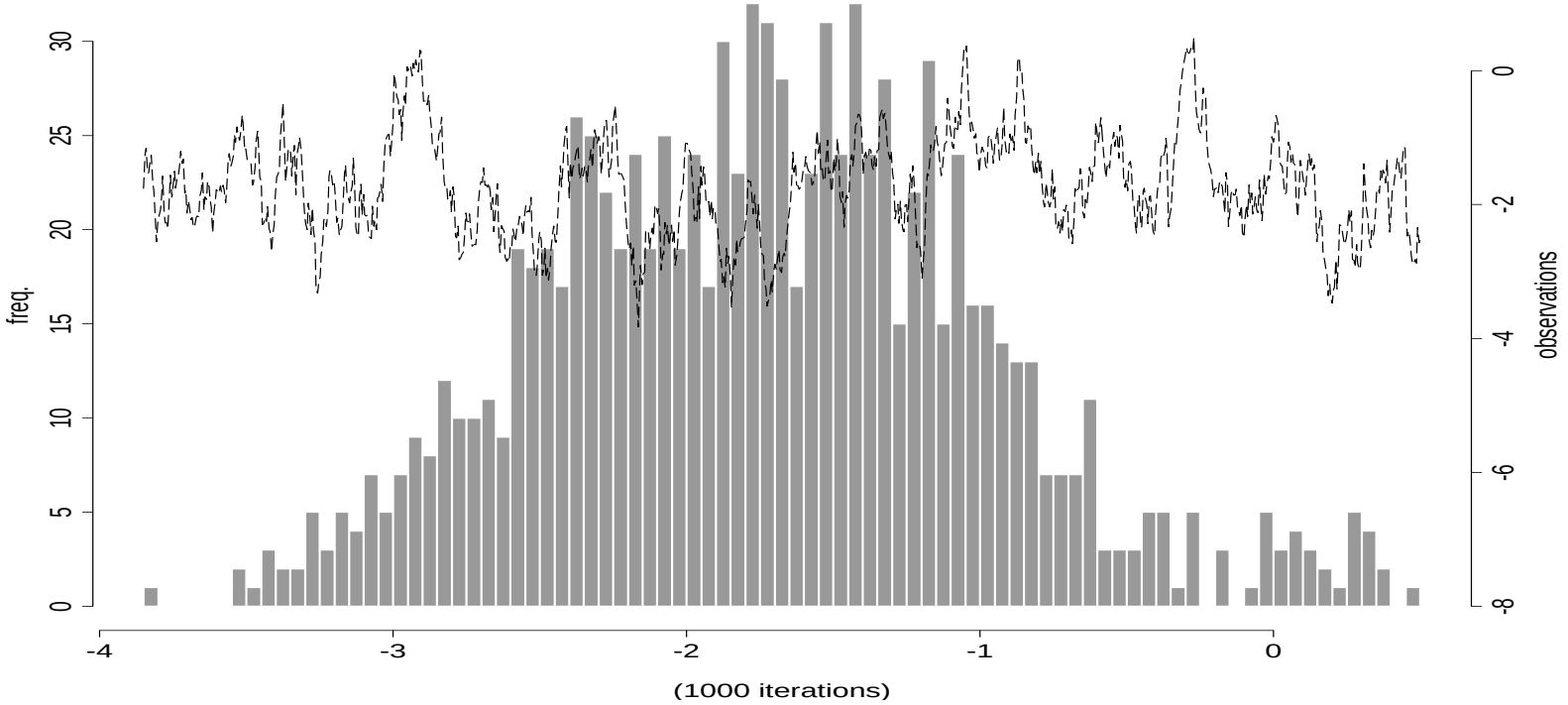
$$\pi(\mu, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2} \ .$$

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 $\left(\mu^{(t)}\right)$ 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

5 MCMC tools for variable dimension problems

5.1 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)**

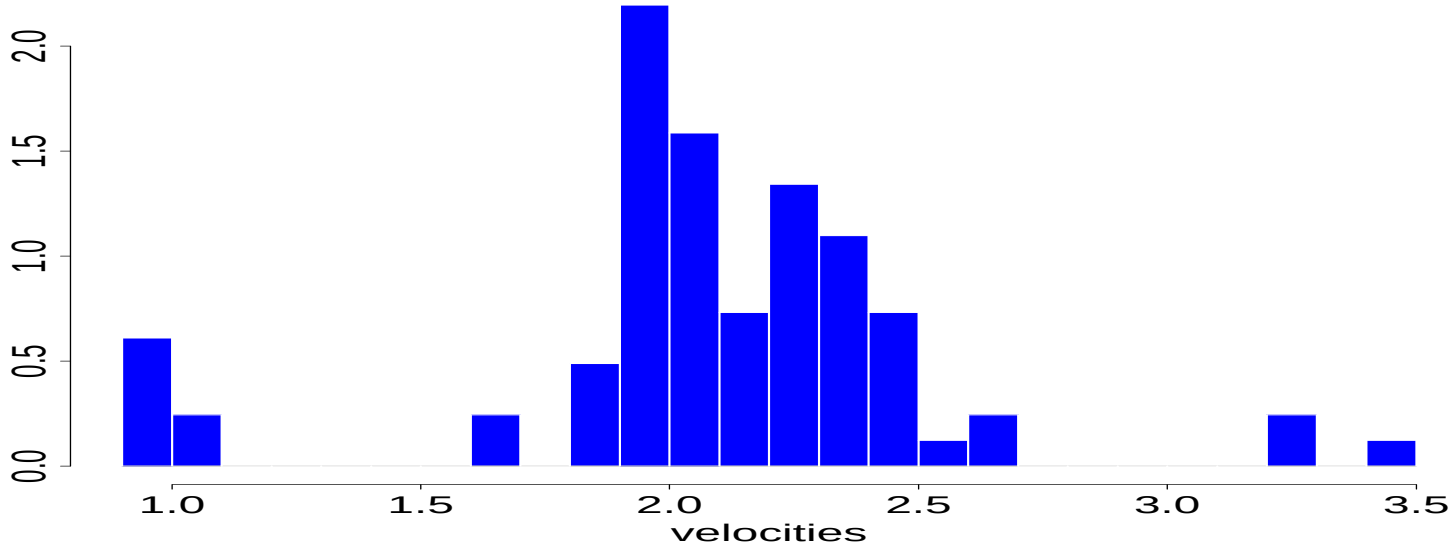
Many areas of application

- *variable* **selection**
- **change point(s) determination**
- **image analysis**
- **graphical models and expert systems**
- *variable* **dimension models**
- **causal inference**

Example 28 —Mixture modelling–

Benchmark dataset: Speed of galaxies

[Roeder, 1990; Richardson & Green, 1997]



Modelling by a mixture model

$$\mathfrak{M}_i : x_j \sim \sum_{\ell=1}^{i} p_{\ell i} \mathcal{N}(\mu_{\ell i}, \sigma^2_{\ell i}) \qquad (j = 1, \ldots, 82)$$

$\mathbf{i}?$

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

$$p(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}$$

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

5.2 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 $\mathfrak{K}_{1\to 2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to 1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

$$\pi(d\theta_1)\, \mathfrak{K}_{1\to 2}(\theta_1, d\theta) = \pi(d\theta_2)\, \mathfrak{K}_{2\to 1}(\theta_2, d\theta)\,,$$

and take, wlog, $\dim(\mathfrak{M}_2) > \dim(\mathfrak{M}_1)$.

Proposal expressed as

$$\theta_2 = \Psi_{1\to 2}(\theta_1, v_{1\to 2})$$

where $v_{1\to 2}$ is a random variable of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to 2} \sim \varphi_{1\to 2}(v_{1\to 2})\,.$$

In this case, $\mathfrak{q}_{1\to 2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to 2}(v_{1\to 2}) \left| \frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})} \right|^{-1} ,$$

by the Jacobian rule.

If probability $\varpi_{1\to 2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

$$\alpha(\theta_1, v_{1\to 2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to 1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to 2}\, \varphi_{1\to 2}(v_{1\to 2})} \left| \frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})} \right| .$$

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

$$\theta_2 \sim \mathcal{N}(\Psi_{1\to 2}(\theta_1, v_{1\to 2}), \varepsilon) \qquad \text{and} \qquad \Psi_{1\to 2}(\theta_1, v_{1\to 2}) \sim \mathcal{N}(\theta_2, \varepsilon)$$

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 29 —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

$$(T) \begin{cases} u_1, u_2, u_3 & \sim & \mathcal{U}(0,1) \\ p_{j(k+1)} & = & u_1 p_{jk} \\ \mu_{j(k+1)} & = & u_2 \mu_{jk} \\ \sigma^2_{j(k+1)} & = & u_3 \sigma^2_{jk} \end{cases}$$

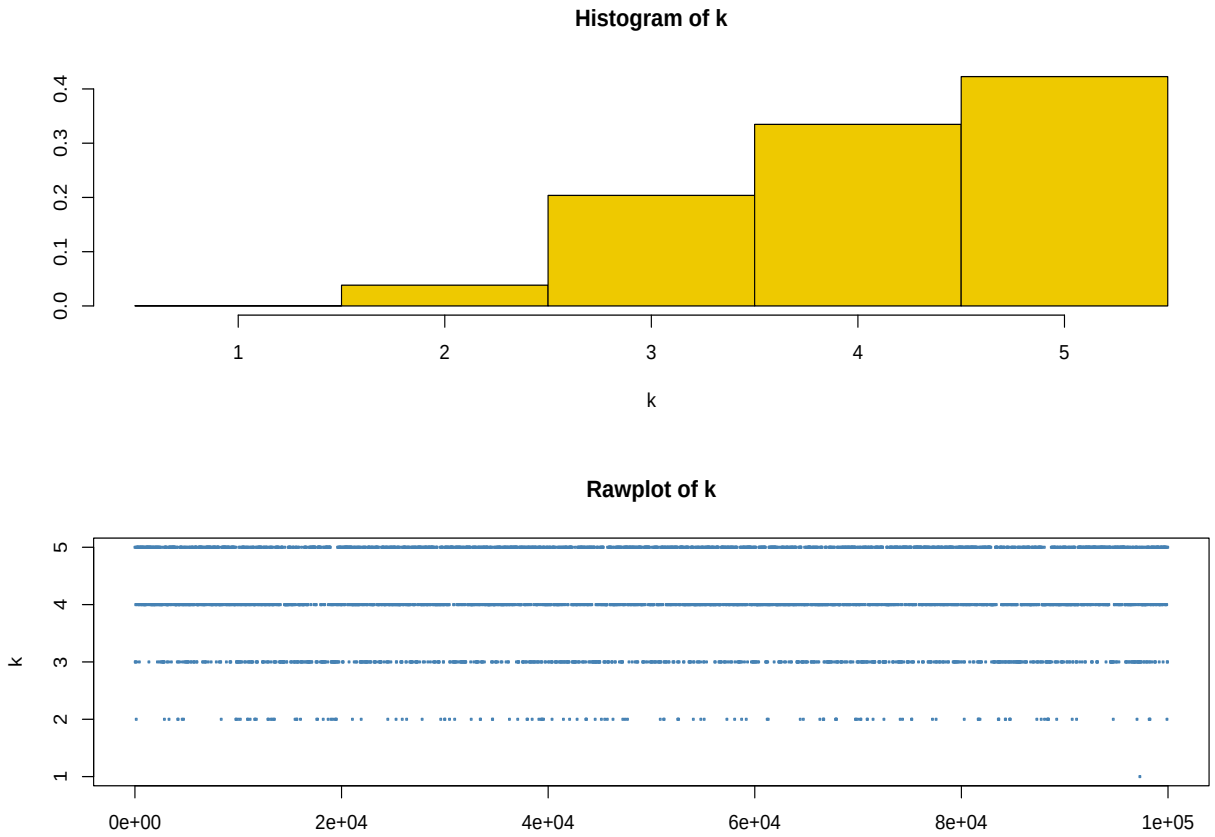
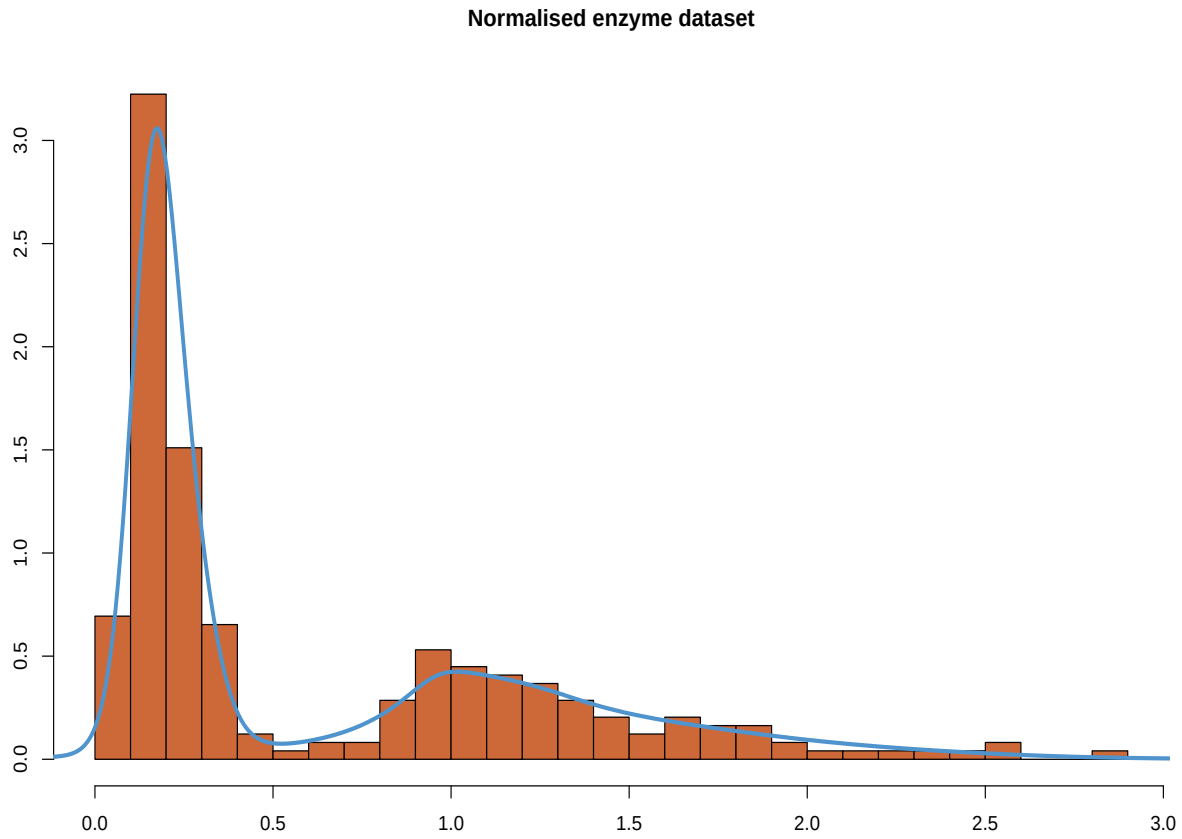


Figure 2: Histogram and rawplot of $100,000$ k's produced by RJMCMC under the imposed constraint $k \leq 5$.


Normalised enzyme dataset
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Example 30 —Hidden Markov model—

$$\begin{aligned} P(X_t + 1 = j | X_t = i) &= w_{ij}, \\ w_{ij} &= \omega_{ij} / \sum_{\ell} \omega_{i\ell}, \\ Y_t | X_t = i &\sim \mathcal{N}(\mu_i, \sigma_i^2). \end{aligned}$$

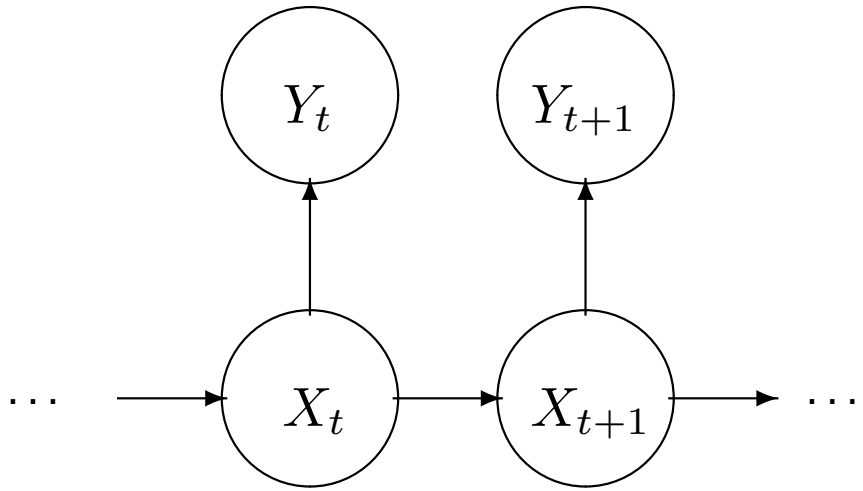


Figure 3: DAG representation of a simple hidden Markov model

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.;

$$\mu_{j_1} = \mu_{j_\star} - 3\sigma_{j_\star}\varepsilon_\mu, \quad \mu_{j_2} = \mu_{j_\star} + 3\sigma_{j_\star}\varepsilon_\mu, \quad \varepsilon_\mu \sim \mathcal{N}(0,1);$$

$$\sigma^2_{j_1} = \sigma^2_{j_*}\xi_\sigma, \quad \sigma^2_{j_2} = \sigma^2_{j_*}/\xi_\sigma, \quad \xi_\sigma \sim \log\mathcal{N}(0,1).$$

[Robert & al., 2000]

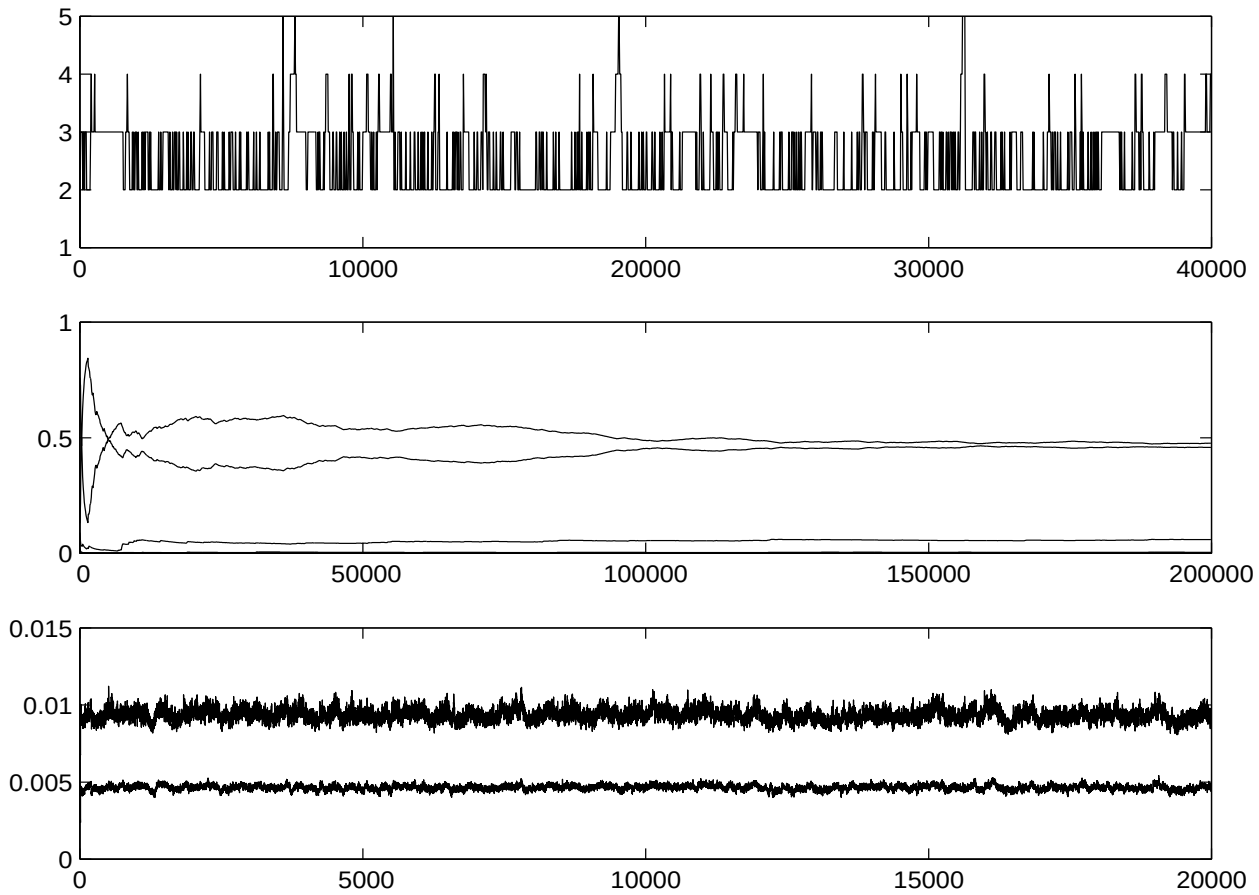


Figure 4: Upper panel: First 40,000 values of k for S&P 500 data, plotted every 20th sweep. Middle panel: estimated posterior distribution of k for S&P 500 data as a function of number of sweeps. Lower panel: σ_1 and σ_2 in first 20,000 sweeps with $k = 2$ for S&P 500 data.

Example 31 —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]

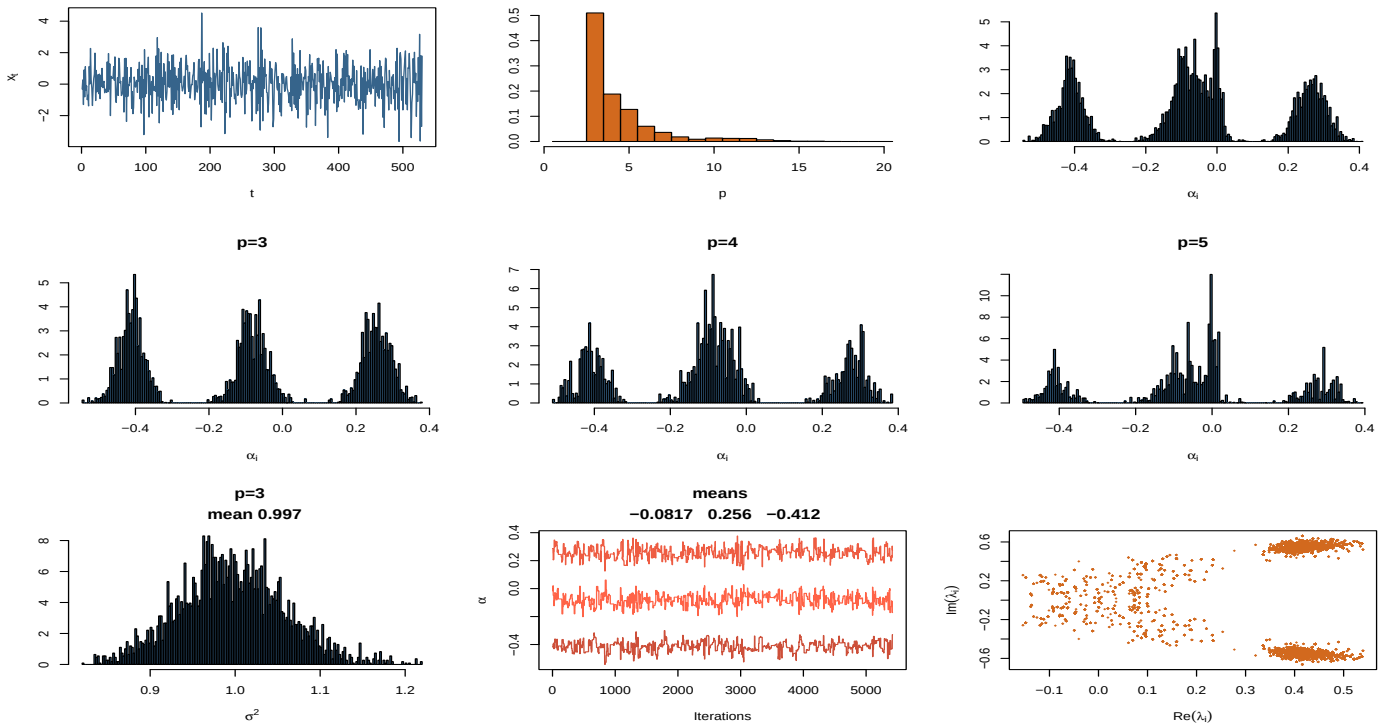


Figure 5: Reversible jump algorithm based on an $AR(3)$ simulated dataset of 530 points *(upper left)* with true parameters α_i $(-0.1, 0.3, -0.4)$ and $\sigma = 1$. First histogram associated with p, the following histograms with the α_i's, for different values of p, and of σ^2. Final graph: scatterplot of the complex roots. One before last: evolution of $\alpha_1, \alpha_2, \alpha_3$.

5.3 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 32 —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. $\delta(\Phi) \leftarrow \sum_{j=1}^{k} \delta_j(\Phi_j), \xi \leftarrow \lambda_0 + \delta(\Phi), u \sim \mathcal{U}([0, 1])$
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

$$\begin{aligned}
&\lim_{N\to\infty} N\delta_{k+1} \times \frac{1}{k+1} \times \min(A^{-1}, 1) \\
&\qquad = \lim_{N\to\infty} N\frac{1}{k+1} \times \text{likelihood ratio}^{-1} \times \frac{\beta_k}{\delta_{k+1}} \times \frac{b(w,\phi)}{(1-w)^{k-1}} \\
&\qquad = \text{likelihood ratio}^{-1} \times \frac{\lambda_k}{k+1} \times \frac{b(w,\phi)}{(1-w)^{k-1}}.
\end{aligned}$$

Hence **“RJMCMC→BDMCMC”**. This holds more generally.

Example 33 **—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]

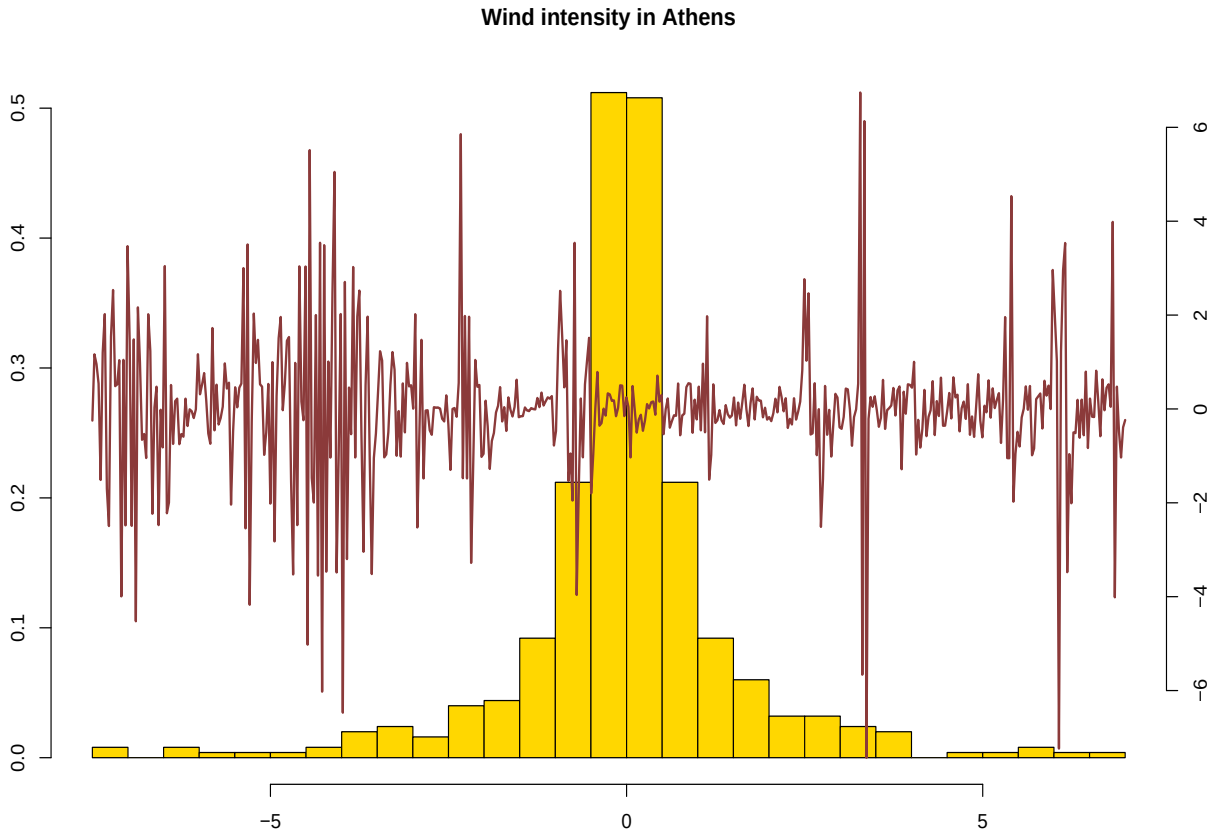


Figure 6: Histogram and rawplot of 500 wind intensities in Athens

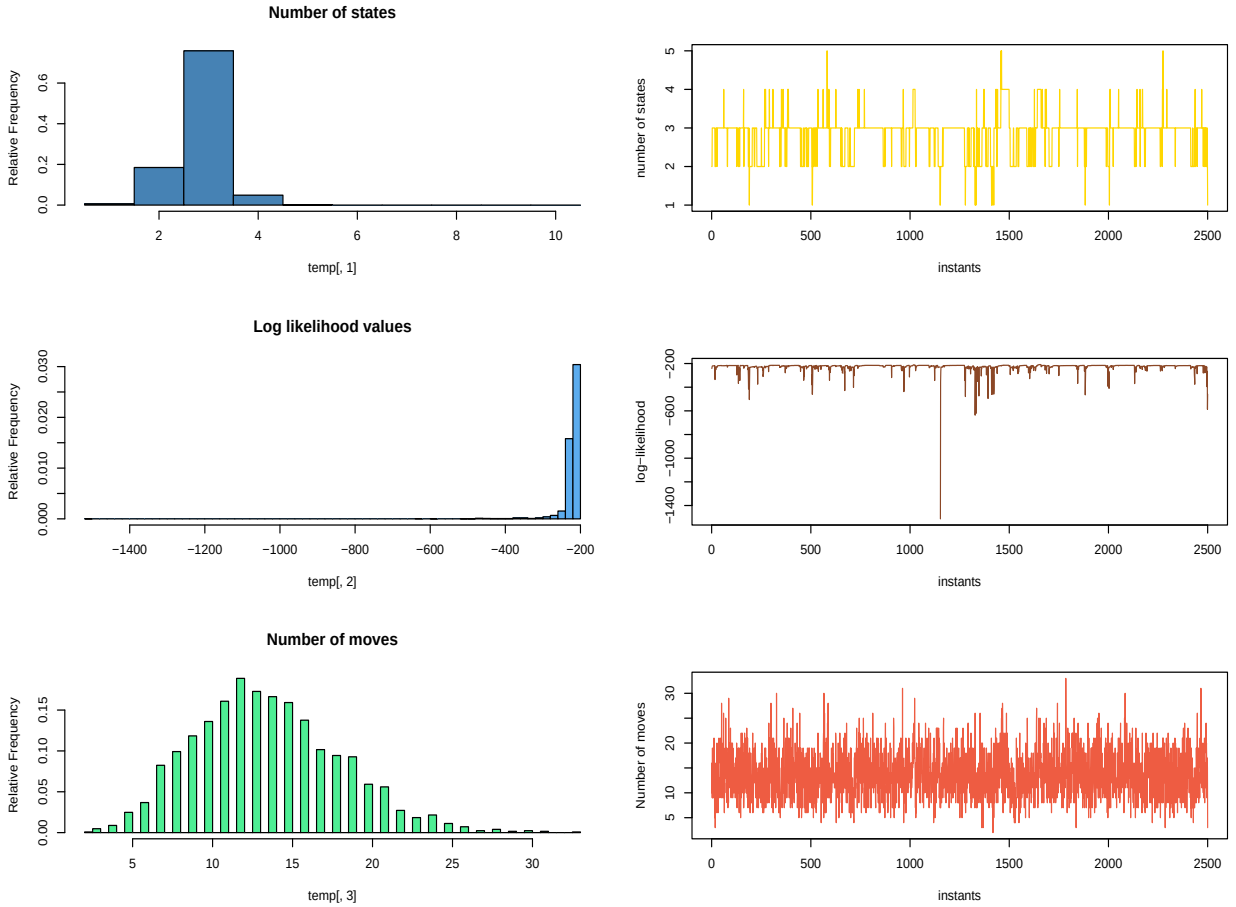


Figure 7: MCMC output on k (histogram and rawplot), corresponding loglikelihood values (histogram and rawplot), and number of moves (histogram and rawplot)

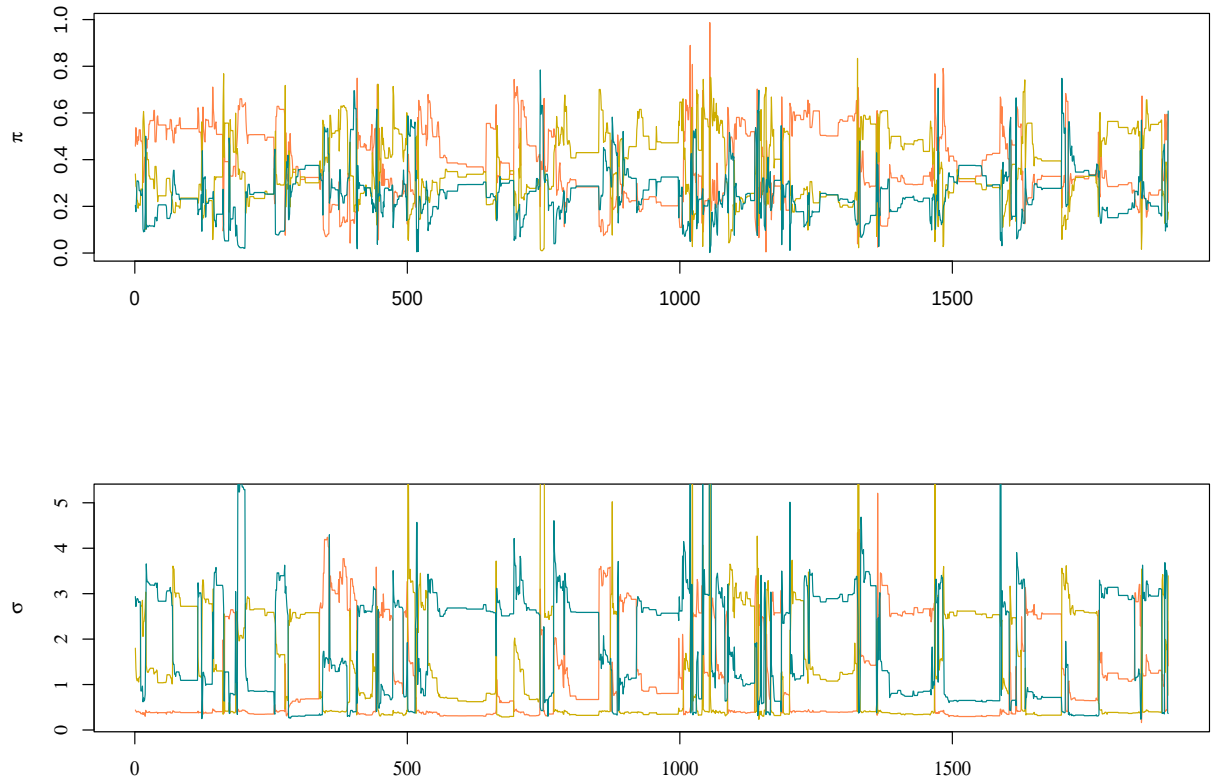


Figure 8: MCMC sequence of the probabilities π_j of the stationary distribution (top) and the parameters σ (bottom) of the three components when conditioning on $k = 3$

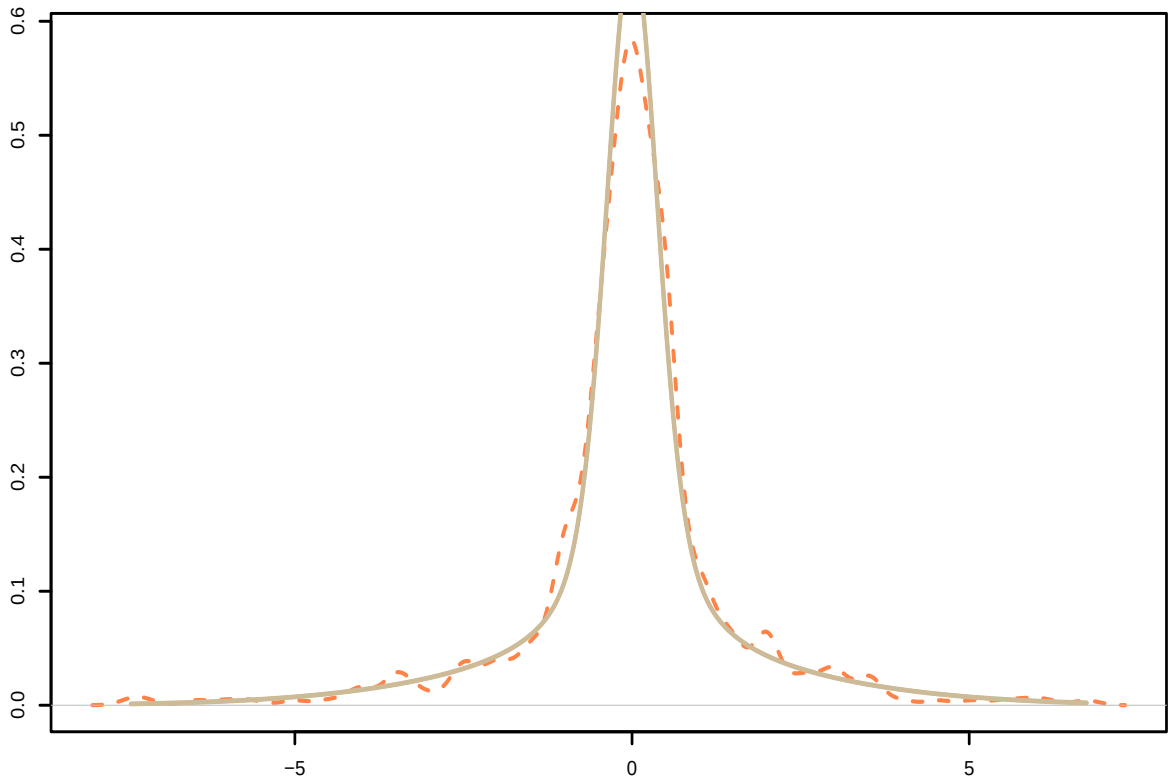


Figure 9: MCMC evaluation of the marginal density of the dataset (dashes), compared with R nonparametric density estimate (solid lines).

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 34 —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.

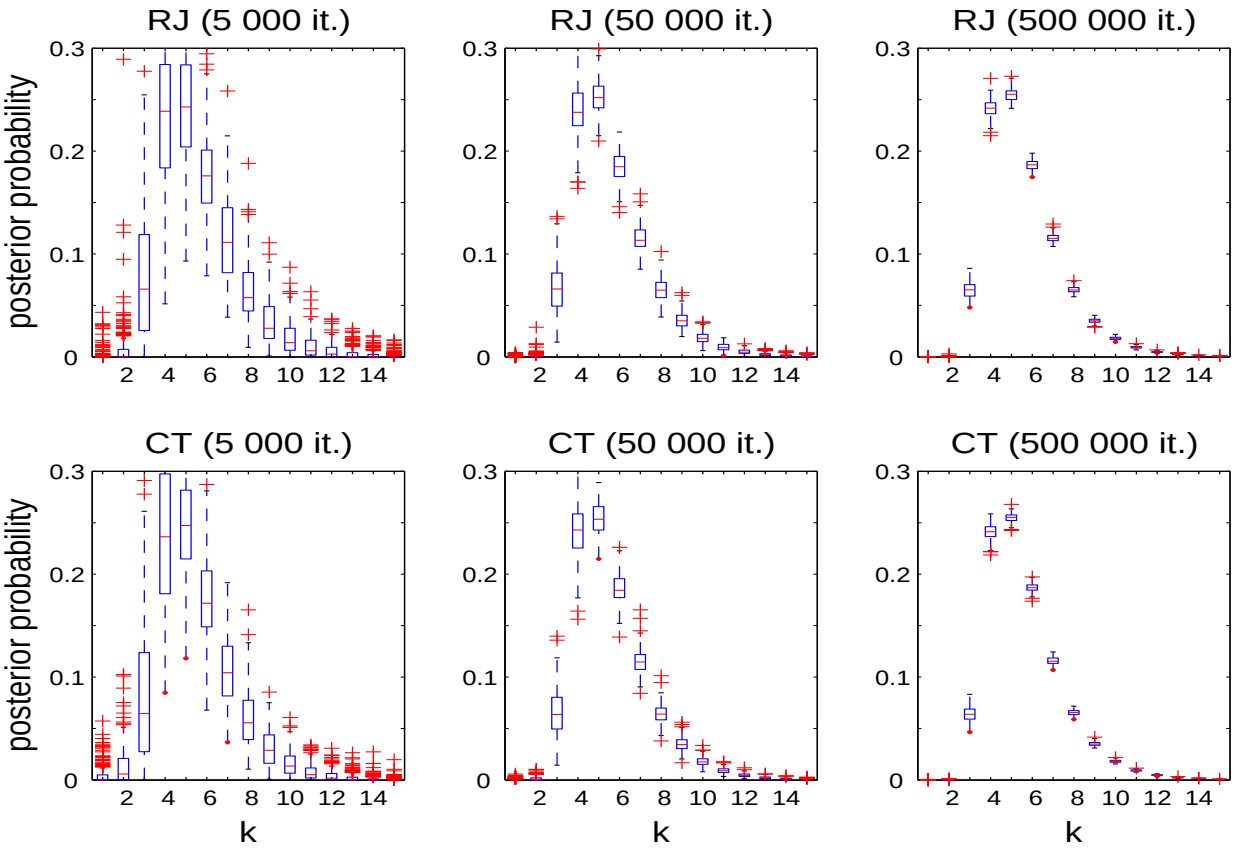


Figure 10: Galaxy dataset, box plot for the estimated posterior on k obtained from 200 independent runs: RJMCMC (top) and BDMCMC (bottom). The number of iterations varies from 5 000 (left), to 50 000 (middle) and 500 000 (right).

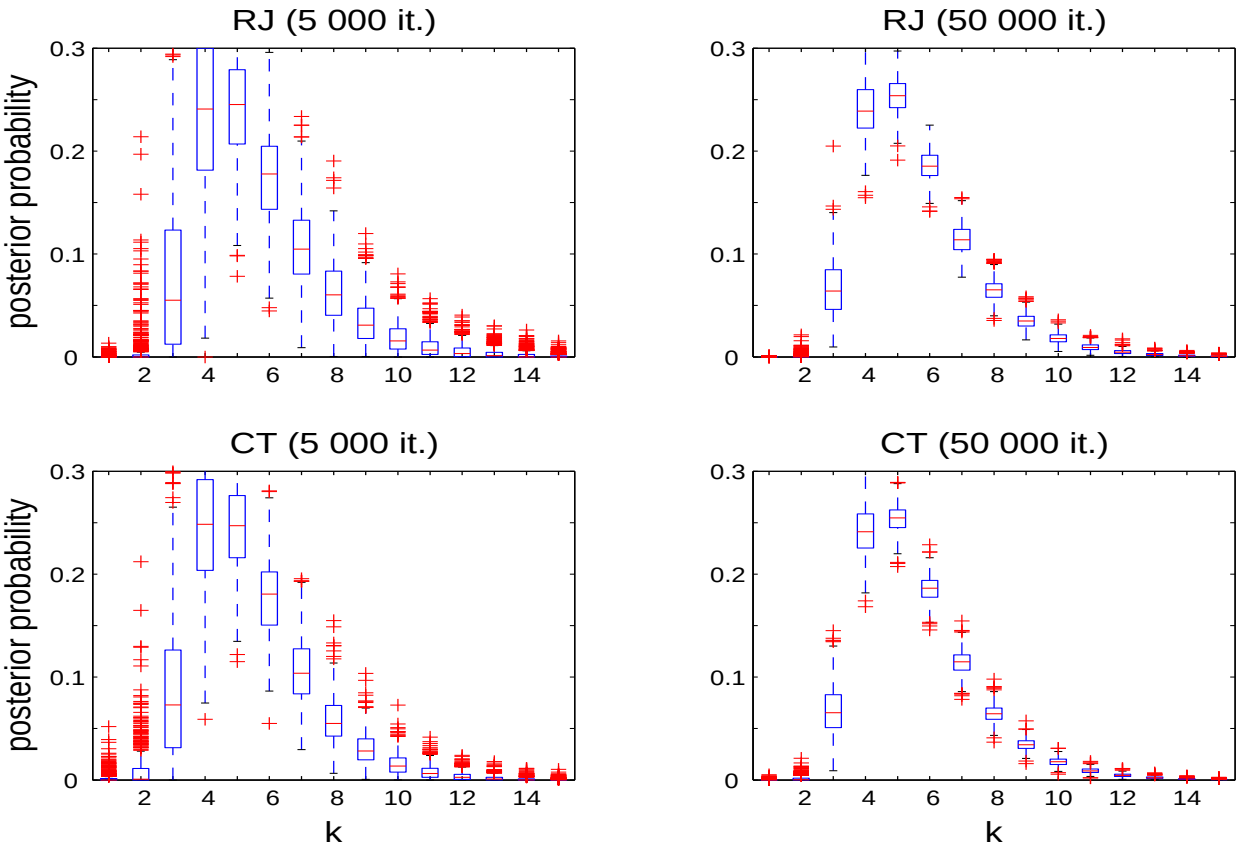


Figure 11: Galaxy dataset, box plot for the estimated posterior on k obtained from 500 independent runs: Top RJMCMC and bottom, CTMCMC. The number of iterations varies from 5 000 (left plots) to 50 000 (right plots).

6 Population Monte Carlo

6.1 Importance sampling

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:

$$\mathbb{E}\left[h(X^{(t)})\ \frac{\pi(X^{(t)})}{q_t(X^{(t)}|X^{(t-1)})}\right]$$
$$= \int h(x)\,\frac{\pi(x)}{q_t(x|y)}\,q_t(x|y)\,g(y)\,dx\,dy$$
$$= \int h(x)\,\pi(x)\,dx$$

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 $\varrho_i^{(t)}$

6.2 Dynamic sampling

More global dynamic scheme:

6.3 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,

$$\text{var}\left(\hat{\mathfrak{I}}_t\right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}\left(\varrho_i^{(t)} h(x_i^{(t)})\right) \ ,$$

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

6.4 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.

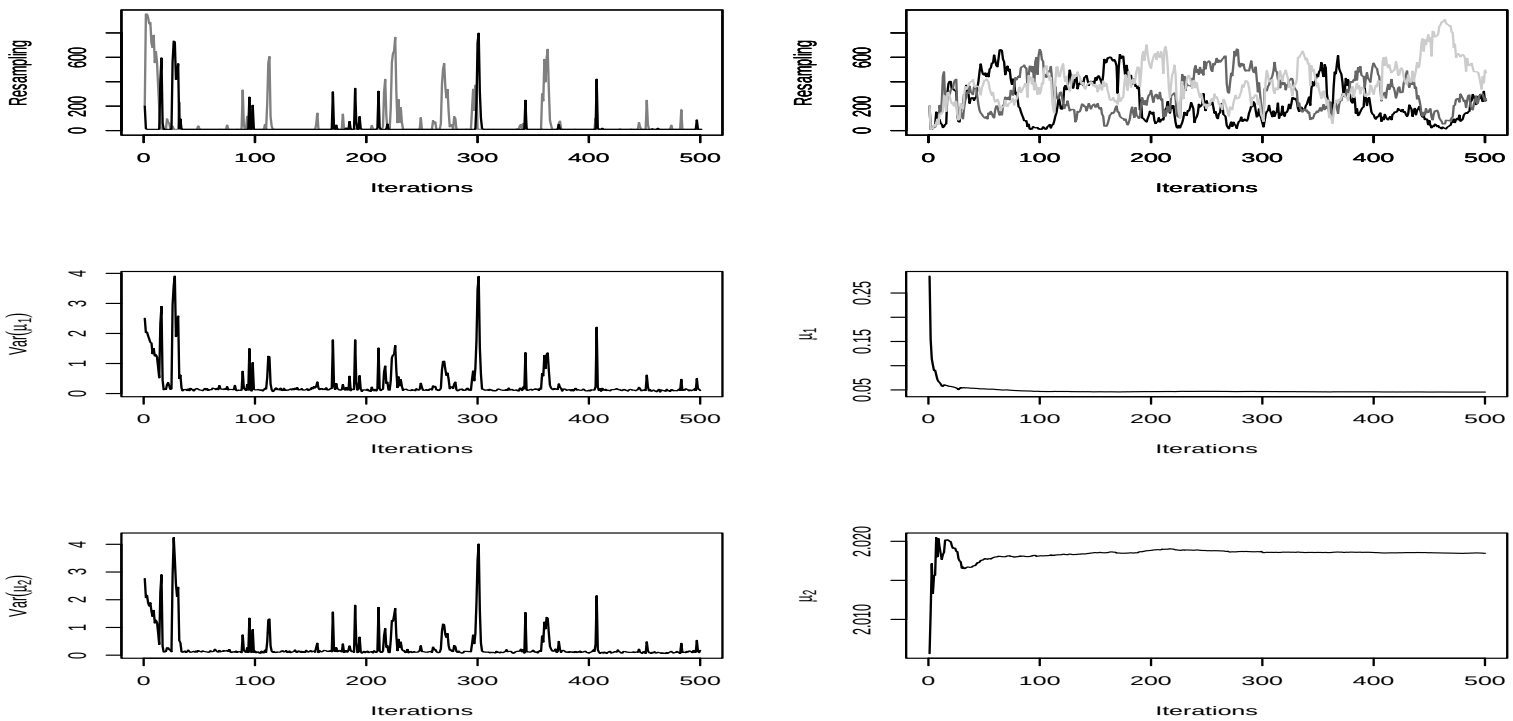


Figure 12: *(u.left)* Number of resampled points for $v_1 = 5$ (darker) and $v_2 = 2$; *(u.right)* Number of resampled points for the other variances; *(m.left)* Variance of the μ_1's along iterations; *(m.right)* Average of the μ_1's over iterations; *(l.left)* Variance of the μ_2's along iterations; *(l.right)* Average of the simulated μ_2's over iterations.

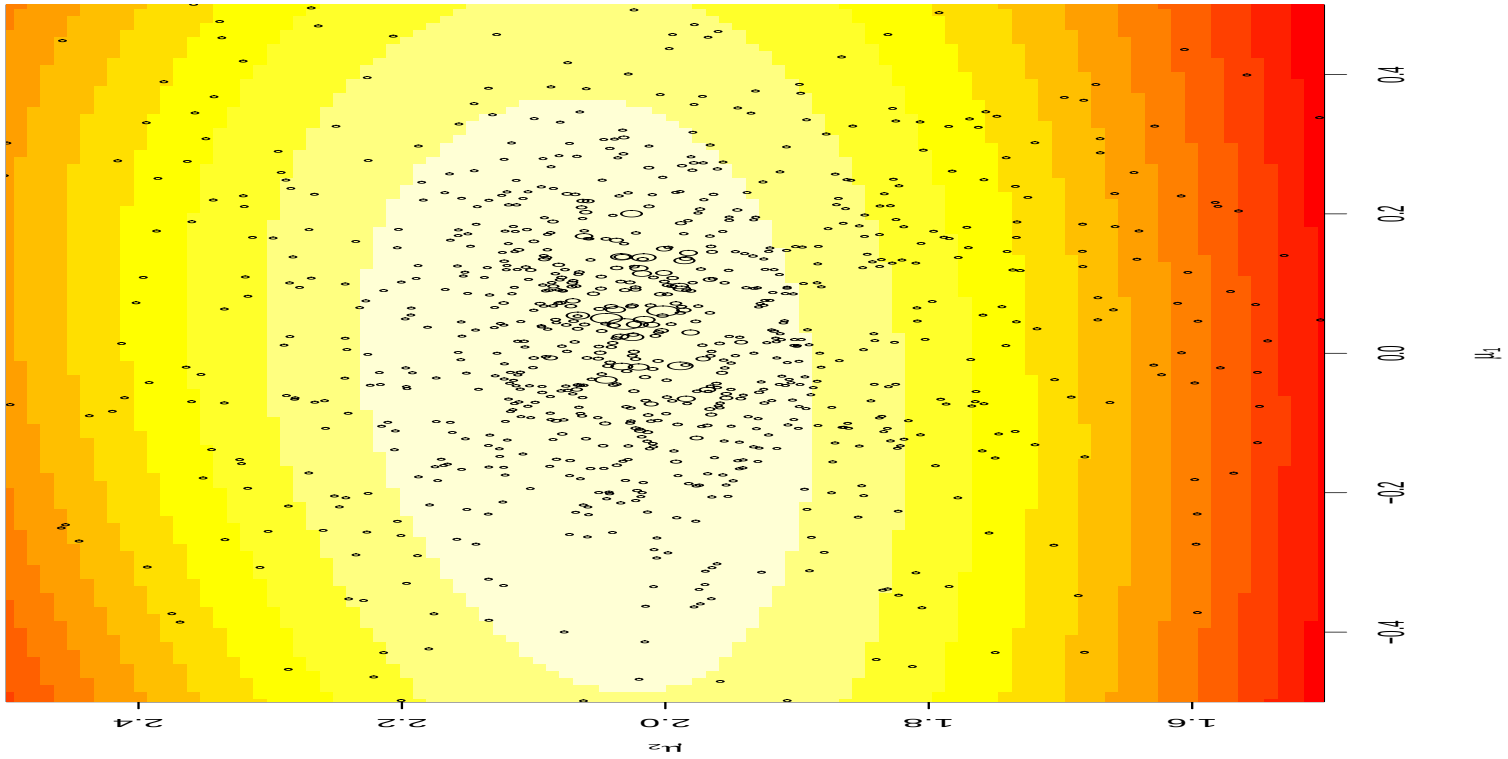


Figure 13: Log-posterior distribution and sample of means

6.5 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]

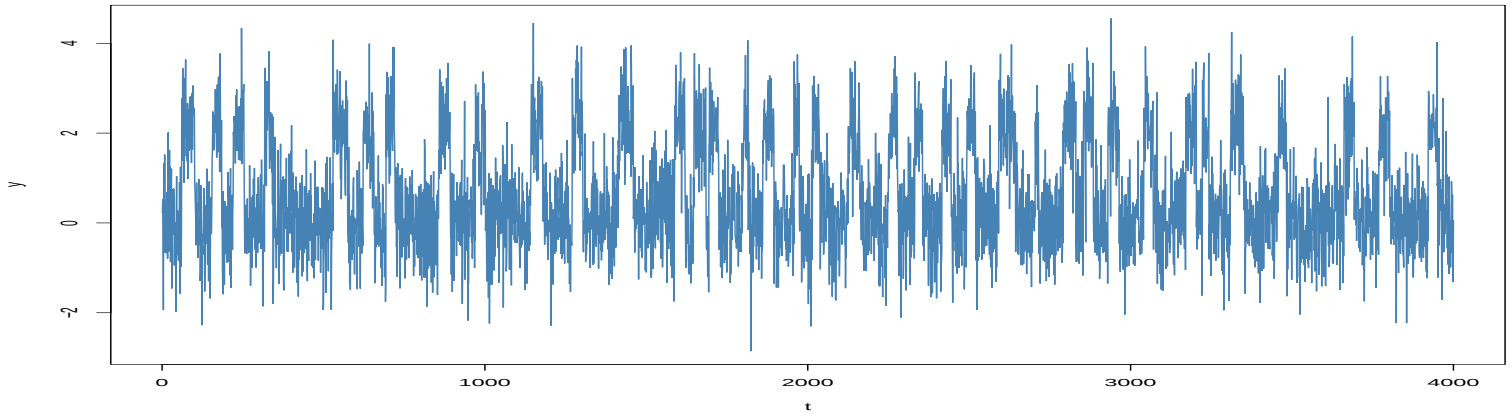


Figure 14: Simulated sample of size 4000

Our assumptions

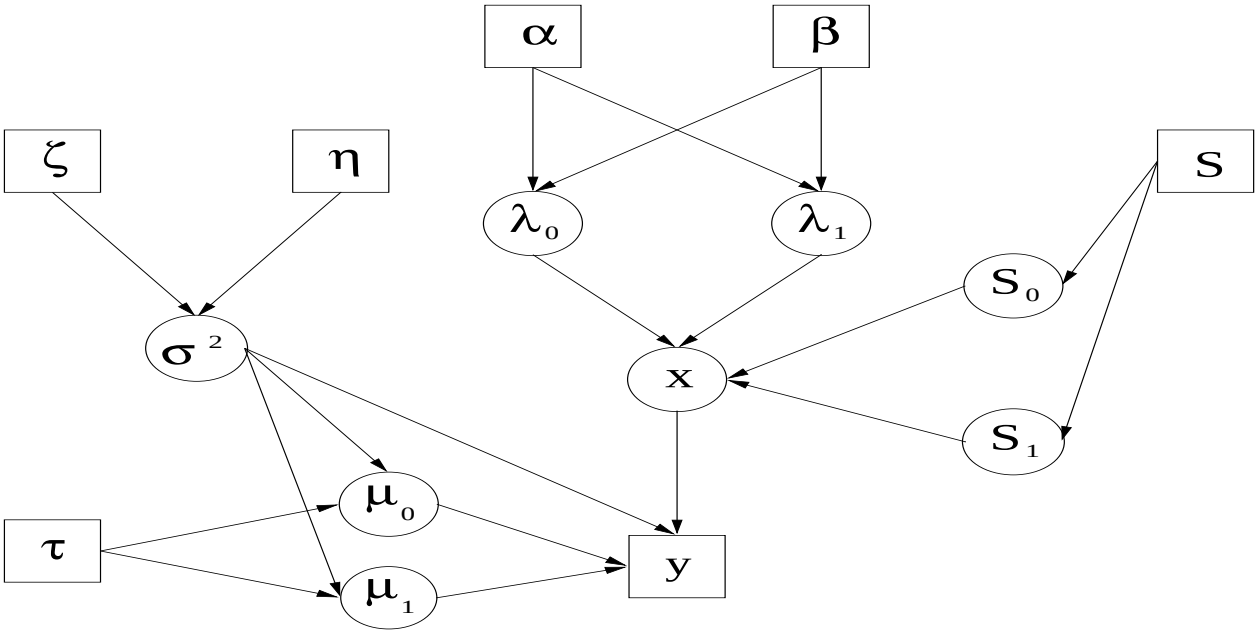
- The durations d_j are **integers**
  - generalisation of HMM: geometric *vs.* negative binomial
  - identifiability issue
- s_0 and s_1 are **integers** and uniform on $\{1, \ldots, S\}$
  - generalisation of HMM: exponential *vs.* sum of exponentials
  - alternative to duplicate states

    [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

$\pi\left(\omega^{(j)}|\mathbf{y}, \mathbf{x}_{-}^{(j)}\right)$ Gibbs like:

$$\sigma^{-2}\,|y, x \sim \mathcal{G}\left(\frac{T}{2}+\eta, \left(\frac{1}{2}\right)\left[2\nu + 1_T' y^2 - \frac{\left((1_T - x)'\, y\right)^2}{v_0} - \frac{(x'y)^2}{v_1}\right.\right.$$
$$\left.\left. + \frac{\tau v_0 \left(\theta_0 - \frac{(1_T - x)' y}{v_0}\right)^2}{v_0 + \tau} + \frac{\tau v_1 \left(\theta_0 - \frac{x' y}{v_1}\right)^2}{v_1 + \tau}\right]\right)$$

$$\mu_0 \left| y, x, \sigma^2 \right. \sim \mathcal{N}\left(\frac{\left(1_T - x\right)' y + \theta_0 \tau}{v_0 + \tau}, \frac{\sigma^2}{v_0 + \tau} \right)$$

$$\mu_1 \left| y, x, \sigma^2 \right. \sim \mathcal{N}\left(\frac{x' y + \theta_0 \tau}{v_1 + \tau}, \frac{\sigma^2}{v_1 + \tau} \right)$$

if $v_0 = \left(1_T - x\right)' 1_T$ 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 $(x_t)_{1\leq t\leq T}$, using a pseudo-HMM, and avoids recourse to variable dimension models
2. Provides unrestricted moves between configurations of the process $(x_t)_{1\leq t\leq T}$.
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

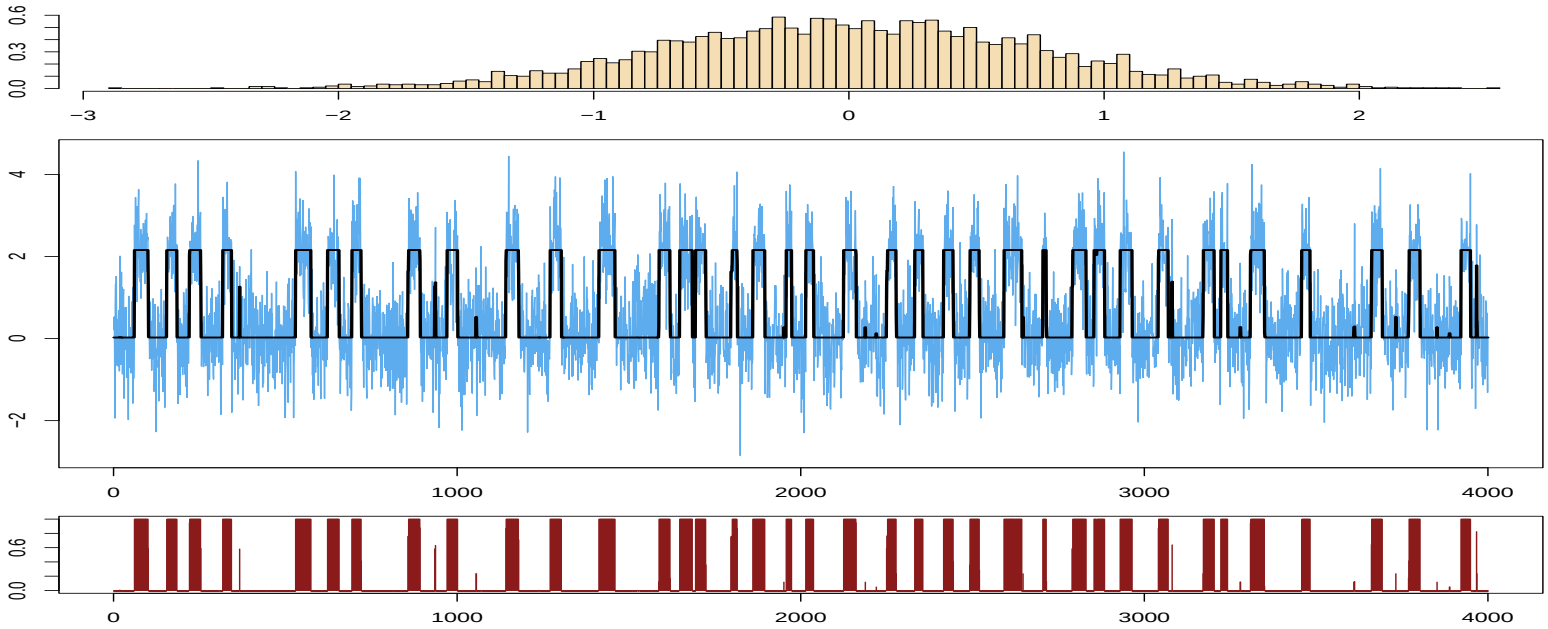


Figure 15: (top) Histograms of residuals after fit by averaged μ_{x_t}; (middle) Simulated sample of size 4000 against fitted averaged μ_{x_t}; (bottom) Probability of allocation to first state for each observation

Population Monte Carlo

Adaptive algorithm: self-improvement of the importance sampler

Long-term behaviour of the algorithm?

stopping rule?

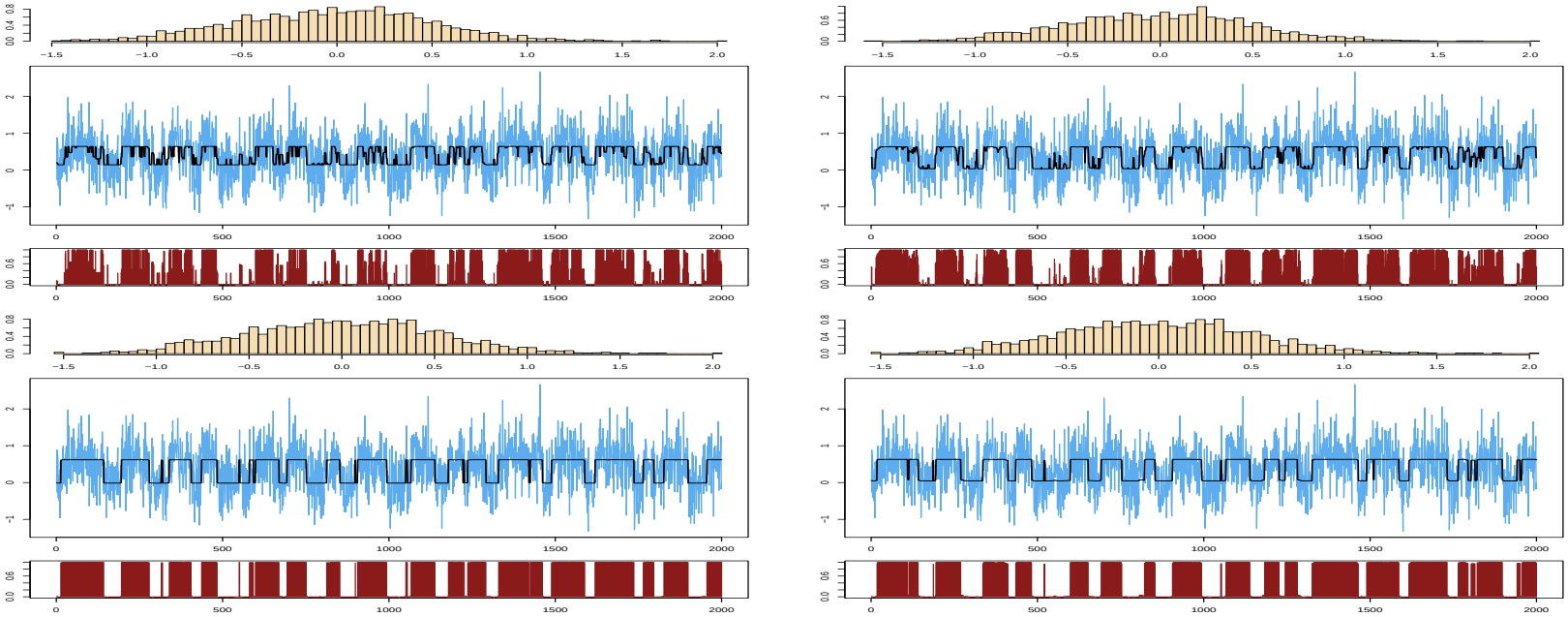


Figure 16: Successive fits for 2000 observations 2000 particles and 1, 2, 5 and 10 iterations.

Degeneracy

Percentage of relevant particles less than 10% on average

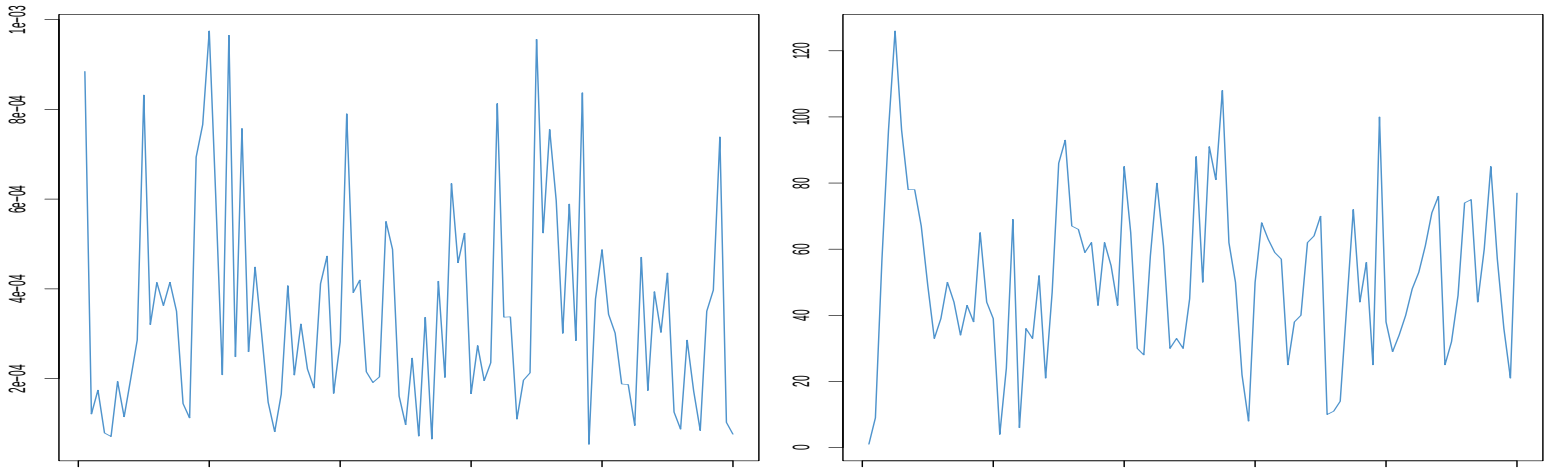


Figure 17: (left) Variance of the weights ϱ_j and (right) Number of particles with descendants along 100 iterations, for 4000 observations and 1000 particles.

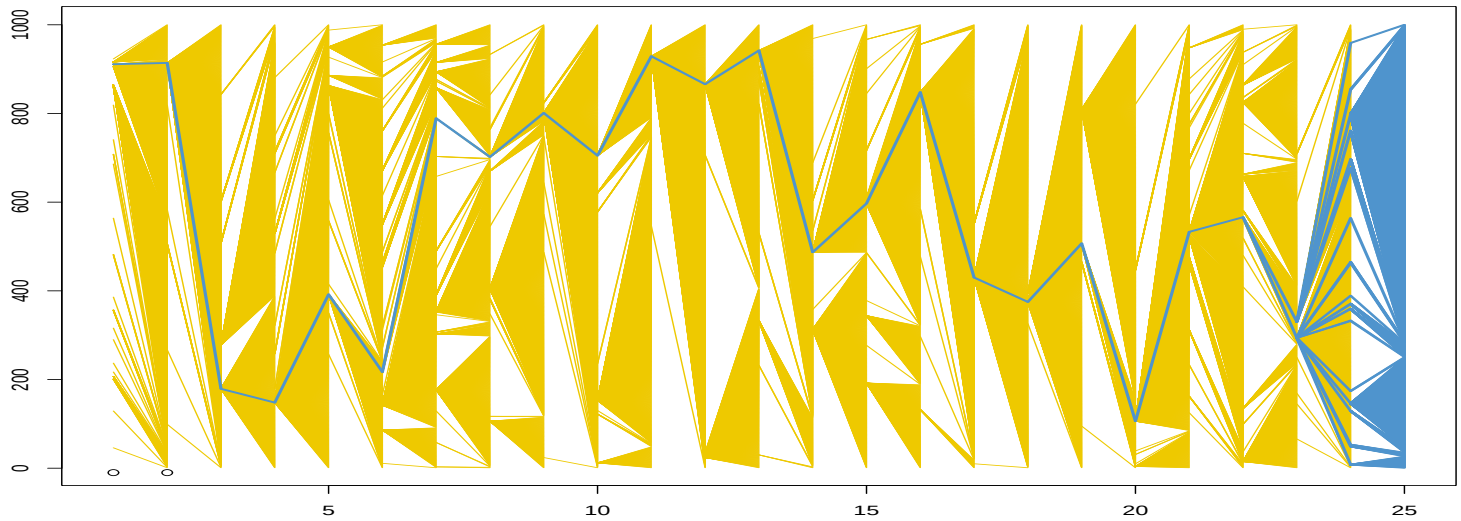


Figure 18: Representation of the sequence of descendents (yellow) and ancestors (blue) for 4000 observations and 1000 particles.

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

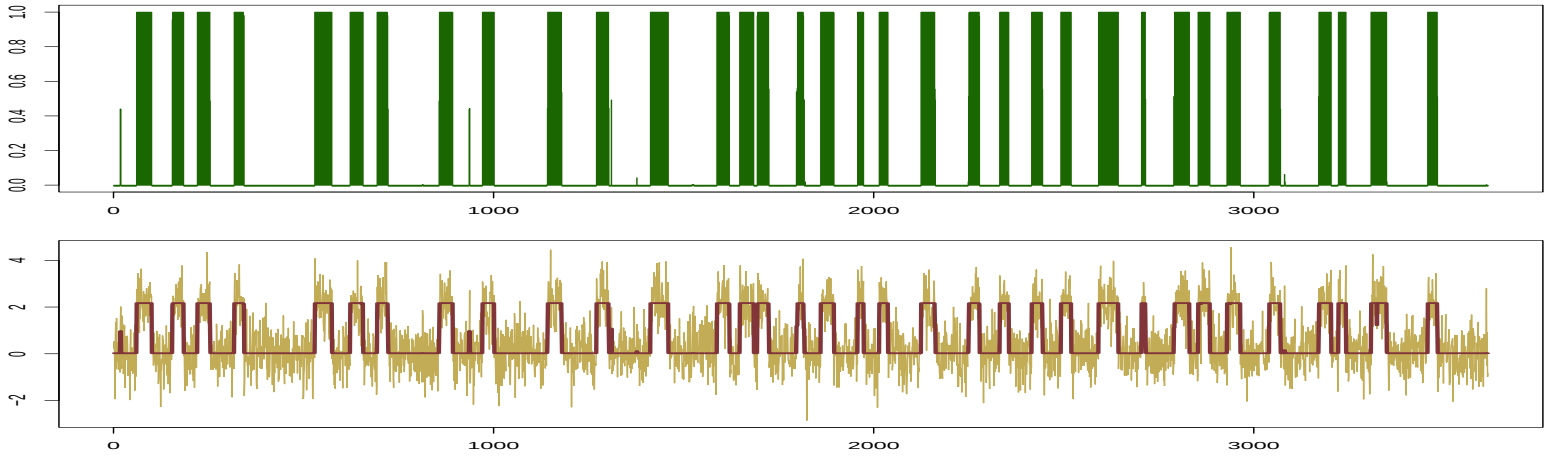


Figure 19: 5000 MCMC iterations

7 Perfect simulation

7.1 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 35 **—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

$$f((x_{t+1},\theta_{t+1})|(x_t,\theta_t)) \quad \propto \quad \binom{n}{x_{t+1}} \theta^{x_{t+1}+\alpha+x_t-1} (1-\theta)^{\beta+2n-x_t-x_{t+1}-1}.$$

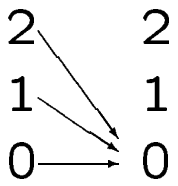
$n = 2$, $\alpha = 2$ **and** $\beta = 4$**.**

State space

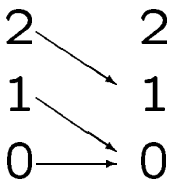
$$\mathcal{X} = \{0, 1, 2\}\,.$$

Transition probabilities

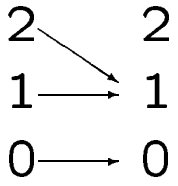
$$\begin{aligned}
\Pr(0 \mapsto 0) &= .583, \quad \Pr(0 \mapsto 1) = .333, \quad \Pr(0 \mapsto 2) = .083, \\
\Pr(1 \mapsto 0) &= .417, \quad \Pr(1 \mapsto 1) = .417, \quad \Pr(1 \mapsto 2) = .167, \\
\Pr(2 \mapsto 0) &= .278, \quad \Pr(2 \mapsto 1) = .444, \quad \Pr(2 \mapsto 2) = .278
\end{aligned}$$



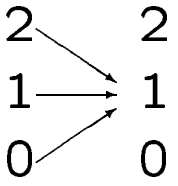
$u_{t+1} < .278$



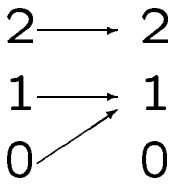
$u_{t+1} \in (.278, .417)$



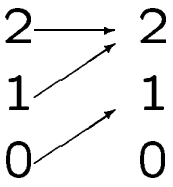
$u_{t+1} \in (.417, .583)$



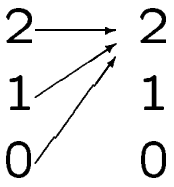
$u_{t+1} \in (.583, .722)$



$u_{t+1} \in (.722, .833)$



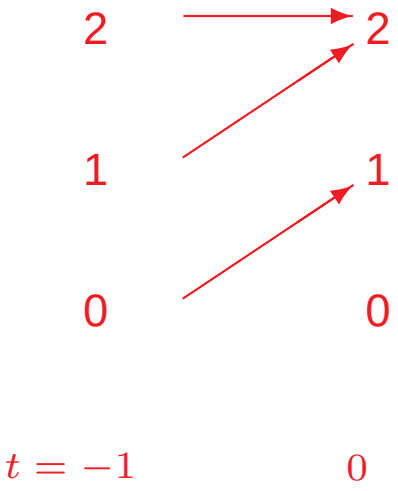
$u_{t+1} \in (.833, .917)$



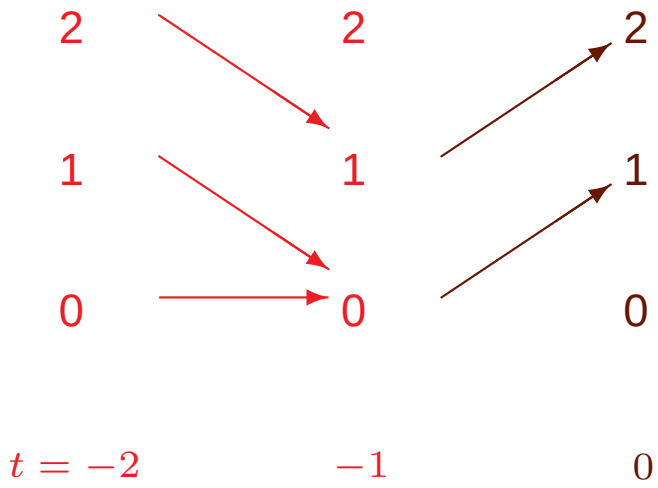
$u_{t+1} > .917$

All possible transitions for the Beta-Binomial(2,2,4) example

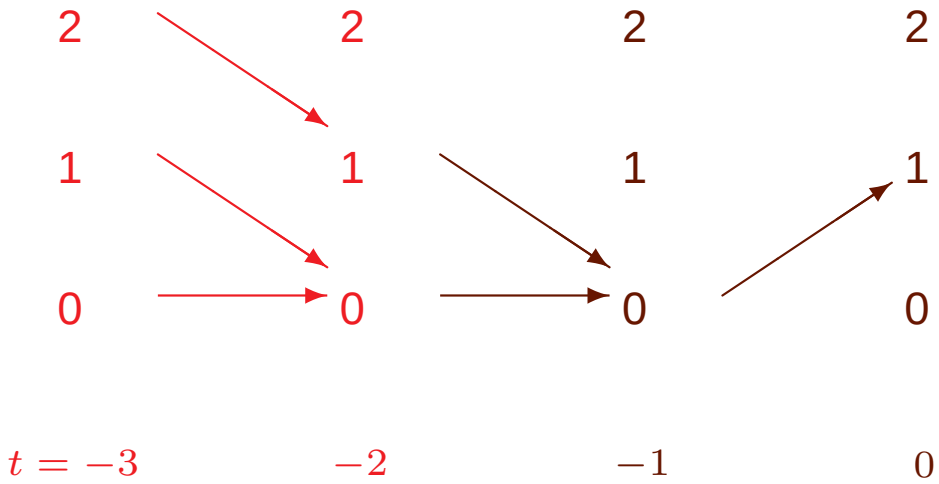
Begin at time $t = -1$ and draw U_0. Suppose $U_0 \in (.833, .917)$.



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 36 **—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 $m = 0, \ldots, n$

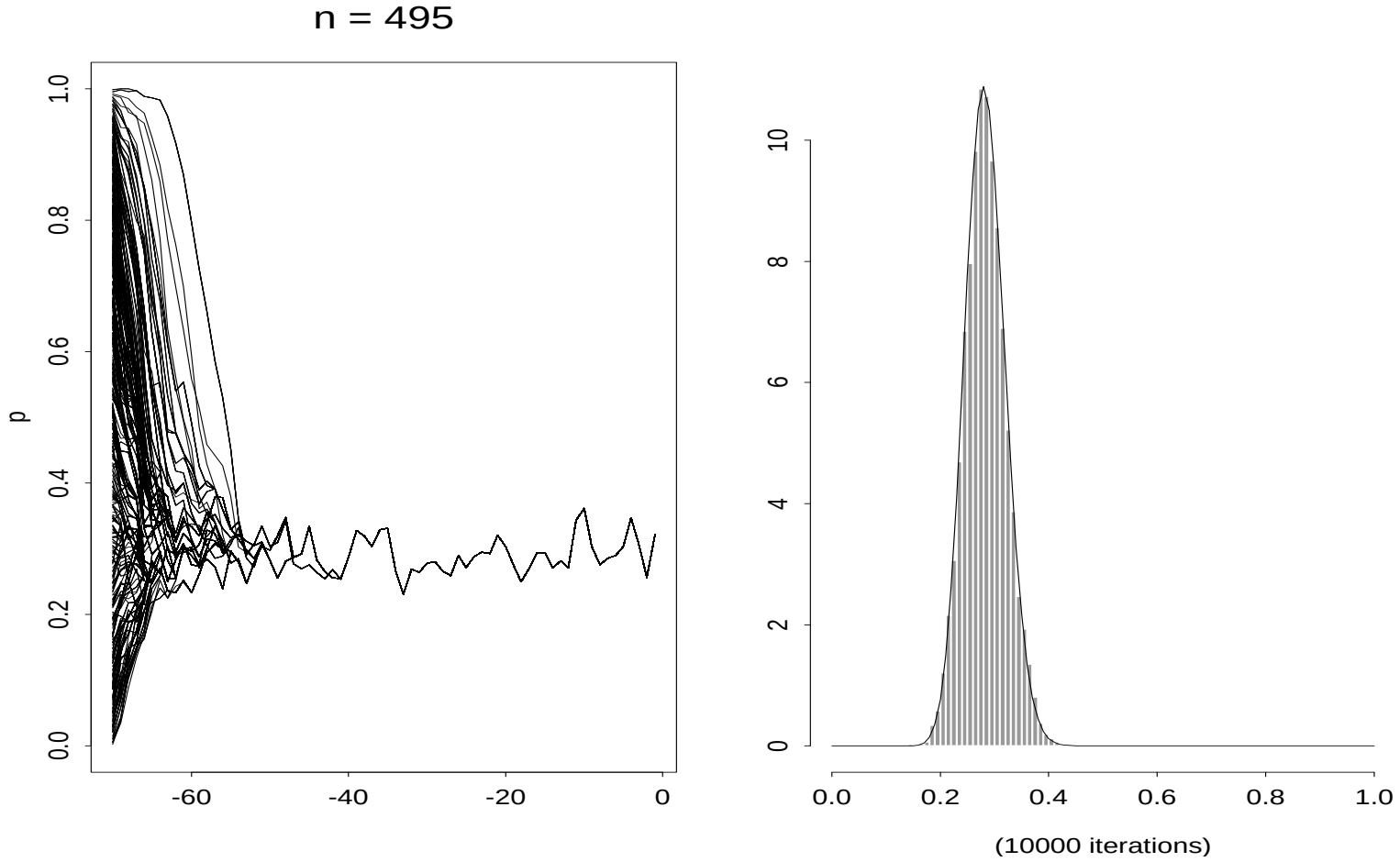


Figure 20: Simulation of $n = 495$ iid rv's from $.33\,\mathcal{N}(3.2, 3.2) + .67\,\mathcal{N}(1.4, 1.4)$ and coalescence at $t = -73$.

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

$$m_1 = \sum_{i=1}^{n} \mathbb{I}\left(u_i \leq \left[1 + \frac{p_2 f_2(x_i) + p_3 f_3(x_i)}{p_1 f_1(x_i)} \right]^{-1} \right) .$$

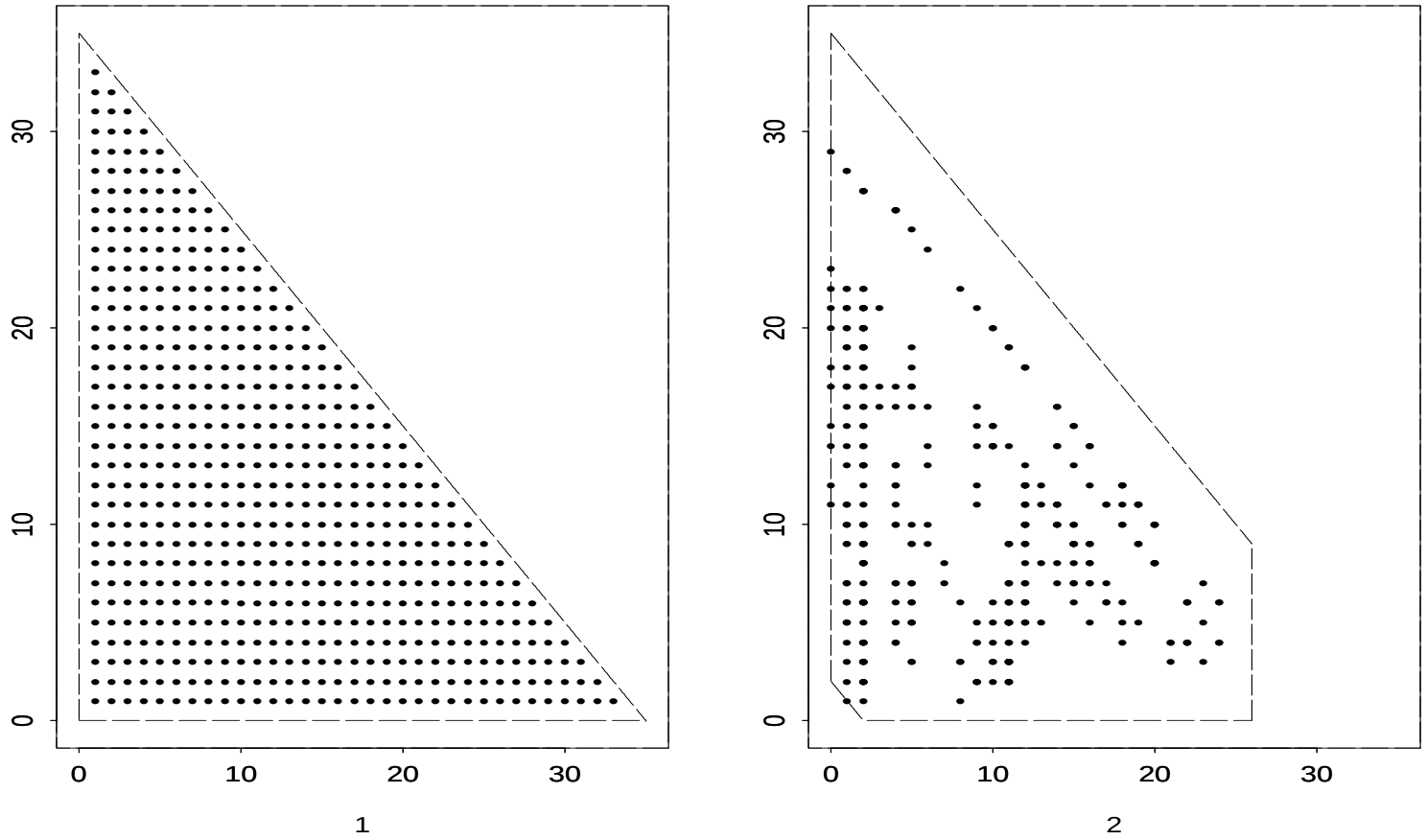


Figure 21: Sample of $n = 35$ observations from $.23\,\mathcal{N}(2.2, 1.44) + .62\,\mathcal{N}(1.4, 0.49) + .15\,\mathcal{N}(0.6, 0.64)$

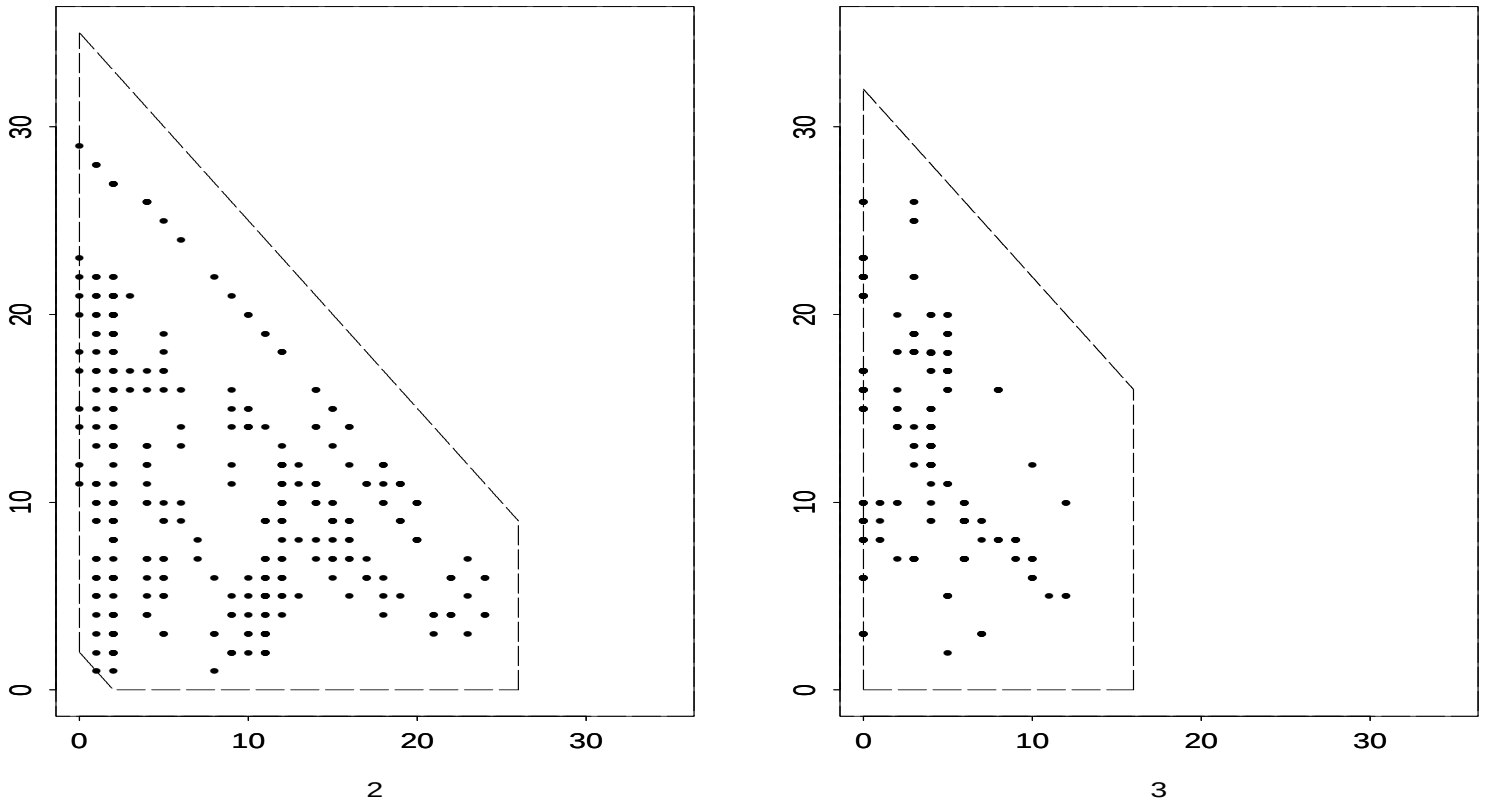
Lozenge monotonicity (preserved)

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 $\{n_3 = \overline{n}_3\}$


0
10
20
30
0
10
20
30
2
0
10
20
30
0
10
20
30
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]

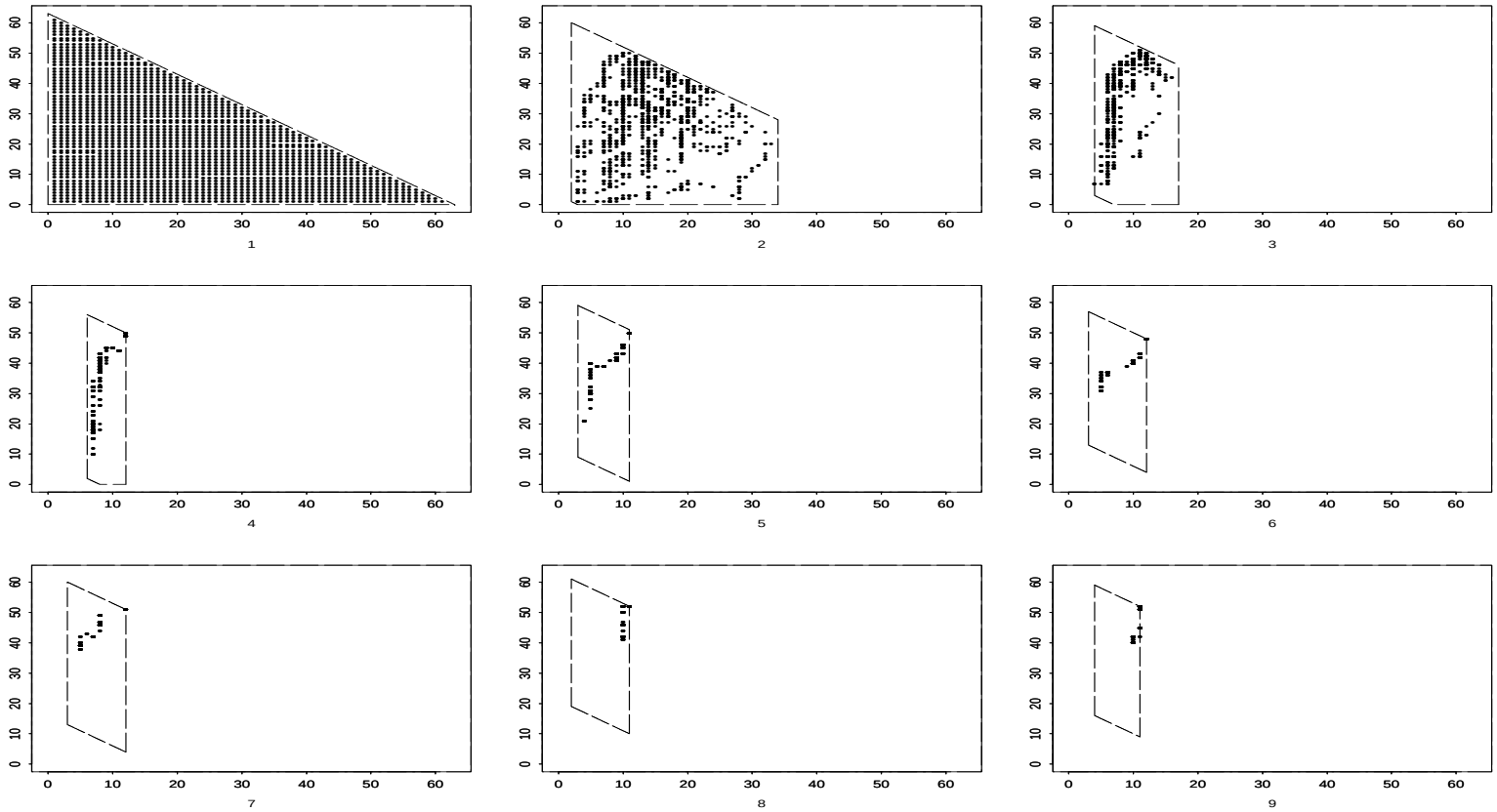


Figure 22: $n = 63$ observations from $.12\mathcal{N}(1.1, 0.49) + .76\mathcal{N}(3.2, 0.25) + .12\mathcal{N}(2.5, 0.09)$

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,

$$\Pr[X_0 = x | C_T(z)] = \frac{\pi(x)/\pi(z)}{\sum_{x'} \pi(x')/\pi(z)} = \pi(x).$$

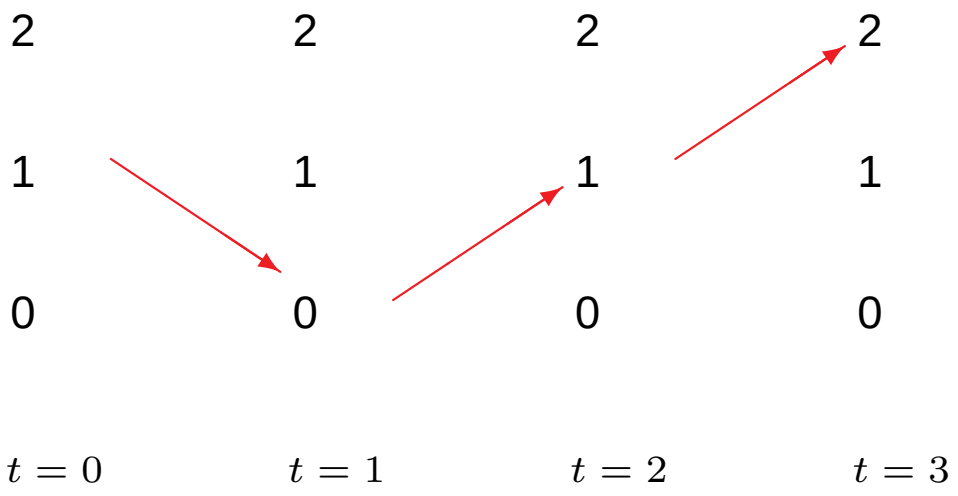
Example 37 **—Beta-Binomial—**

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



$$X_0 = 1, \quad X_1 = 0, \quad X_2 = 1 \quad \text{and} \quad X_3 = 2$$

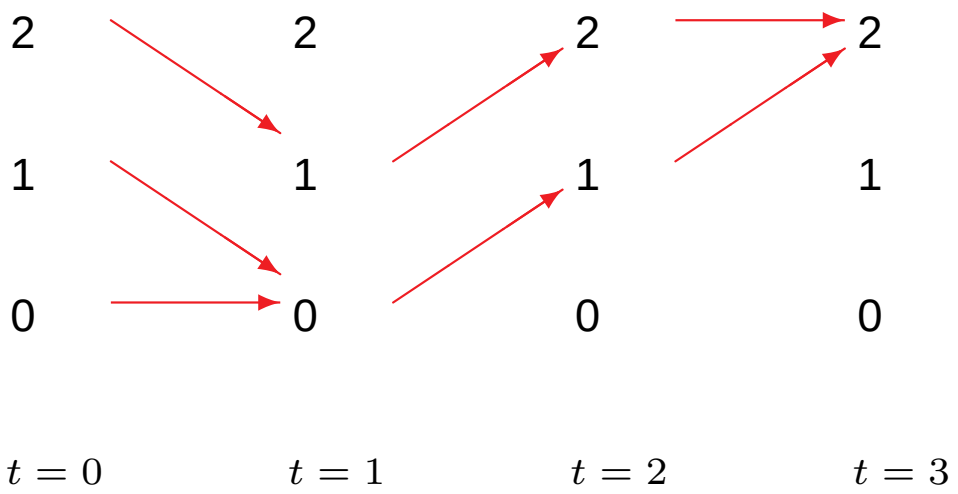
imply

$$U_1 \sim \mathsf{U}(0, .417), \quad U_2 \sim \mathsf{U}(.583, .917), \quad U_3 \sim \mathsf{U}(.833, 1)$$

Suppose

$$U_1 \in (.278, .417) \quad U_2 \in (.833, .917) \quad U_3 > .917$$

Begin chains in states 0, 1 and 2.



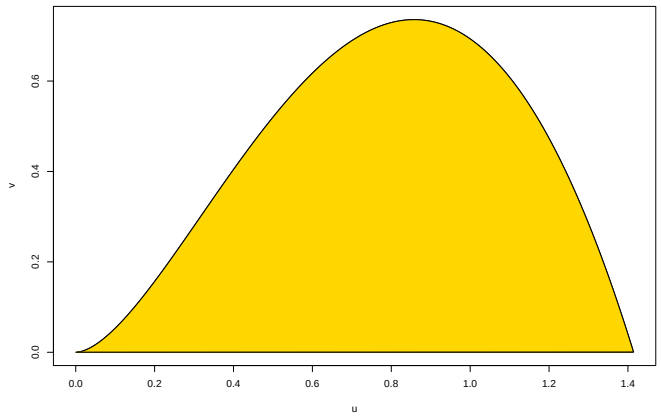
The chains coalesce in $X_3 = 2$; so we accept $X_0 = 1$ as a draw from π.

7.2 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 38 —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

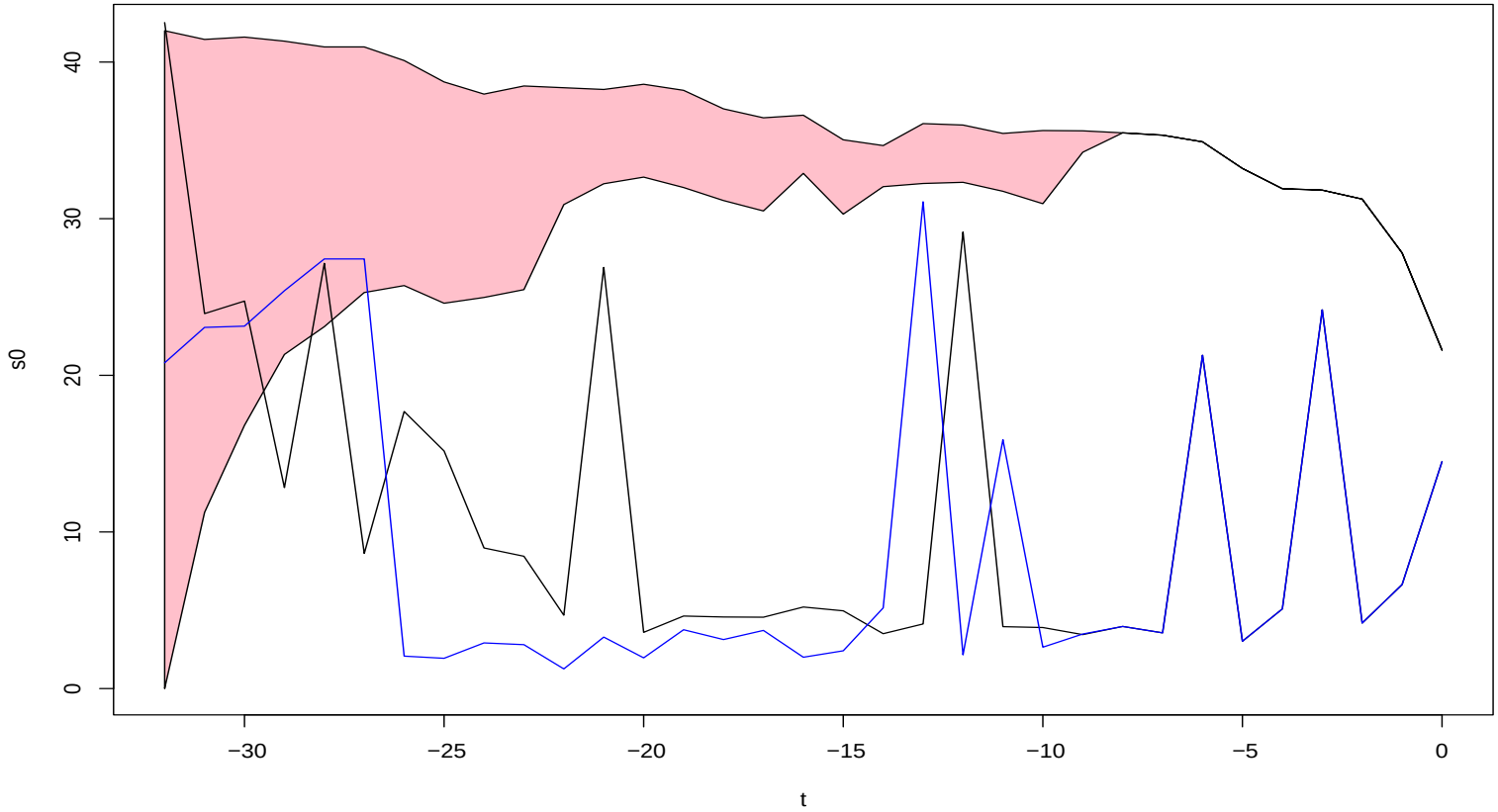
$$\mathbf{z} \mid \mathbf{x} \sim p^{n_0}(1-p)^{n_1}\frac{\Gamma(\alpha_0+n_0-1)\Gamma(\alpha_1+n_1-1)}{(\beta_0+s_0)^{\alpha_0+n_0}(\beta_1+s_1)^{\alpha_1+n_1}}\,.$$

- 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

$$x_1^{(t+1)} = \begin{cases} y_t \sim q(y|x_1^{(t)}) & \text{if } u_t \leq \dfrac{\pi(y_t)\, q(x_1^{(t)}|y_t)}{\pi(x_1^{(t)})\, q(y_t|x_1^{(t)})}, \\ x_1^{(t)} & \text{otherwise.} \end{cases}$$

generate

$$x_2^{(t+1)} = \begin{cases} y_t & \text{if } u_t \leq \dfrac{\pi(y_t)\, q(x_2^{(t)}|x_1^{(t)})}{\pi(x_2^{(t)})\, q(y_t|x_1^{(t)})}, \\ x_2^{(t)} & \text{otherwise.} \end{cases} \qquad (2)$$

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 (2) preserves the ordering.

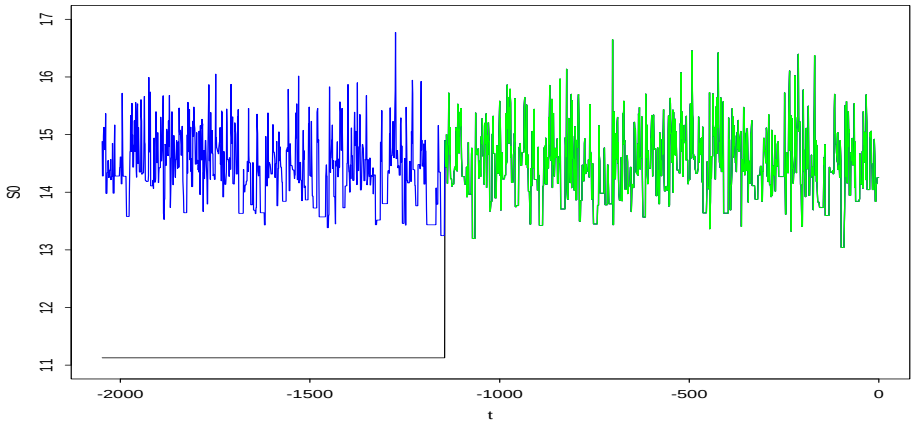
[Now, this is a result from Corcoran and Tweedie!!!]

Example When state space $\mathcal{X}$ compact,

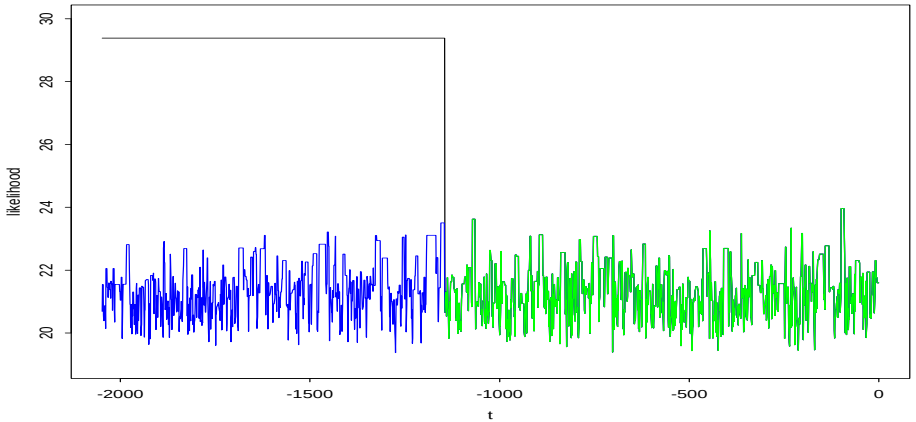
use for h the uniform distribution on $\mathcal{X}$.

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

7.3 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

$$\pi(A) = \frac{1}{\mathbb{E}_\alpha(\tau_\alpha)} \sum_{t=1}^{\infty} \mathsf{Pr}_\alpha(X_t \in A, \tau_\alpha \geq t)$$

Can be extended to stationary measures