Markov Chain Monte Carlo Methods

Christian P. Robert
Université Paris Dauphine

Many thanks to **Peter Green**, **Jim Hobert**, **Eric Moulines**, for their slides

1 Introduction

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

Example 1 –Mixture models–

Models of *mixtures of distributions*:

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

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

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

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

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

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

1.1 Likelihood Methods

Maximum Likelihood Methods

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

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

- Global justifications from asymptotics

Example 2 –Mixtures again–

For a mixture of two normal distributions,

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

likelihood proportional to

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

containing 2^n terms.

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

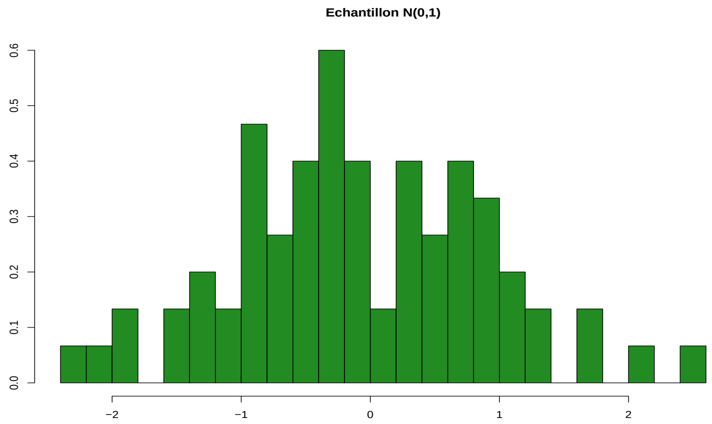
In the special case

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

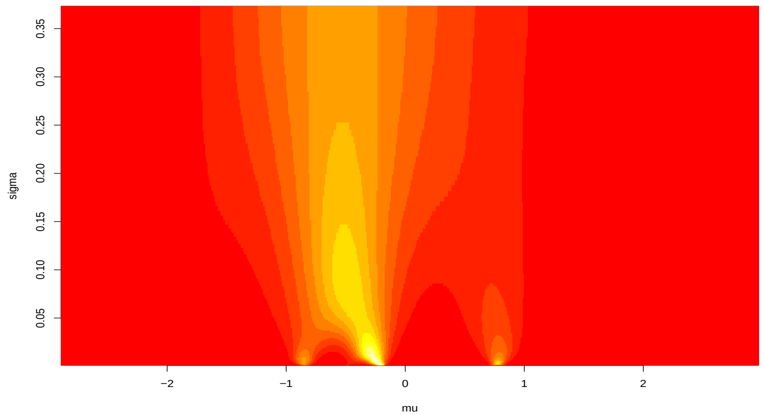
with $\epsilon>0$ known

Then, whatever n, the likelihood is unbounded:

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



Sample from (1)



Likelihood of (1)

1.2 Bayesian Methods

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

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

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

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

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

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

[Bayes Theorem]

where

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

is the *marginal density* of X

Example 3 –Binomial–

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

The classical Bayes estimator δ^π is the posterior mean

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

The curse of conjugate priors

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

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

Example 4 —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

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

[O(2^n)]

For a given permutation (k_t), conditional posterior distribution

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

where

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

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 5 –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 $\Im$ 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 6 –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

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

The Law of Large Numbers implies

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

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

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

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

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

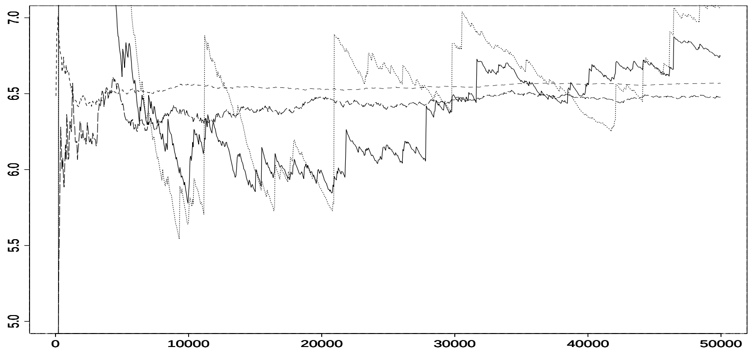
Calculate the integral

$$\int_{2.1}^{\infty} x^5 f(x) dx.$$

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

Simulation results:

- ○ Uniform is best
- ○ Cauchy is OK
- ○ f and Normal are rotten



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

3 Notions on Markov Chains

3.1 Basics

A *Markov chain* is a sequence of random variables that can be thought of as evolving over time.

Probability of a transition depends on the particular set that the chain is in

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

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

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

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

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

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

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

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

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

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

Then, for any bounded ϕ, we may define

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

Note that

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

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

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

Markov chains

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

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

Note that the entire structure of the chain only depends on

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

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

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

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

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

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

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

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

Composition of kernels

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

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

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

Iterated kernel and Chapman-Kolmogorov equations

Set $K^0(x, A) = \delta_x(A)$ the Dirac measure and, for $n \geq 1$, define inductively

$$K^n(x, A) = \int_{\mathcal{X}} \mathfrak{K}(x, dy)K^{n-1}(y, A),$$

We write K^n for the n-th step transition probability kernel

$$\{K^n(x, A), x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$$

Then, for any $0 \leq m \leq n$,

$$K^n(x, A) = \int_{\mathcal{X}} \mathfrak{K}^m(x, dy)K^{n-m}(y, A), \quad x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$$

[Chapman-Kolmogorov equations]

Ionescu-Tulcea Theorem

For any initial measure μ on $\mathcal{B}(\mathcal{X})$ and a family of transition probability kernels $K = \{K_k(x, A), x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$, define, for any $n > 0$ and any $A_0, \ldots, A_n \in \mathcal{B}(\mathcal{X})$,

$$K_\mu^{(n)}[A_0 \times A_1 \times \cdots \times A_n] = \int_{A_0} \int_{A_1} \cdots \int_{A_n} \mu(dx_0) \mathfrak{K}_1(x_0, dx_1) \cdots \mathfrak{K}_n(x_{n-1}, dx_n).$$

Set $K_\mu^{(0)}(A) = \mu(A)$.

[Convention] $K_x^{(n)} = K_\mu^{(n)}$ when μ is the Dirac mass at x.

Then

- $K_\mu^{(n)}$ define a probability measure on $\mathcal{X}^n = (\mathcal{X}^n, \bigvee_{i=1}^n \mathcal{B}(\mathcal{X}))$.
- $K_\mu^{(n)} = \int K_x^{(n)} \mu(dx)$.
- For $m < n$, the projection of $K_\mu^{(n)}$ on $\mathcal{X}^n$ is equal to $K_\mu^{(m)}$,

$$K_\mu^{(m)}(A_1 \times \cdots \times A_m) = K_\mu^{(n)}(A_1 \times \cdots \times A_m \times \mathcal{X} \times \cdots \times \mathcal{X})$$

- **For any initial distribution μ and any family of Markov kernels $(K_k, k \geq 0)$ there exists a unique probability measure on $\mathcal{X}^\infty = (\prod_{i=0}^\infty \mathcal{X}, \bigvee_{i=0}^\infty \mathcal{B}(\mathcal{X}))$ whose projections on $\mathcal{X}^n$ coincide with $K_\mu^{(n)}$**

[Ionescu-Tulcea theorem]

For a discrete state-space, let μ a probability distribution. The Ionescu-Tulcea theorem shows that there exists a sequence of r.v.'s $\{X_n, n \geq 0\}$ such that, for any n, and any $x_0, x_1, \ldots, x_n$ we have

$$K_\mu(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mu(x_0) P_{x_0,x_1} \cdots P_{x_{n-1},x_n} = K_\mu^{(n)}(x_1, \ldots, x_n).$$

Markov property

$X_n, n \geq 0$ sequence of random variables on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, $Q = (Q_k, k \geq 0)$ family of transition kernels on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and μ probability on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Then $X = (X_n, n \geq 0)$ is a Markov chain on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with transition kernel Q and initial distribution μ if

1. **The law of X_0 is μ,**
2. **For any bounded (or positive) function ϕ,**

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] = Q_n\phi(X_n)$$

where $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$, the σ-algebra generated by the r.v.'s $X_0, X_1, \ldots, X_n$.

This later property is often referred to as the **Markov property**

- The Markov chain is time-homogeneous if $Q_n = Q$ for any $n \geq 0$.
- The Markov property implies that, for any bounded function ϕ, any n and any p,

$$\begin{aligned}\mathbb{E}[\phi(X_{n+p})|\mathcal{F}_n] &= Q_n \cdots Q_{n+p-1}\phi(X_n),\\ \mathbb{E}[\phi(X_{n+p})] &= \mu Q_1 \cdots Q_{n+p-1}[\phi].\end{aligned}$$

When the Markov chain is time-homogeneous,

$$\mathbb{E}[\phi(X_{n+p})|\mathcal{F}_n] = Q^p\phi(X_n).$$

- The Markov property also implies that, for any n and bounded functions ϕ_j,

$$\mathbb{E}\left[\phi_{n+1}(X_{n+1})\prod_{k=0}^{n}\phi_k(X_k)\right] = \mathbb{E}\left[Q_n\phi_{n+1}(X_n)\prod_{k=0}^{n}\phi_k(X_k)\right].$$

3.2 Irreducibility

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

It leads to a guarantee of convergence for MCMC algorithms

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

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

τ_y being the first (positive) time y is visited

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

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

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

Minoration condition

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

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

This is called a **minoration condition**.

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

Small sets

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

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

C is called a **small set**

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

3.3 Transience and Recurrence

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

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

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

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

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

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

Similar definitions for the continuous case.

Stronger form of recurrence: **Harris recurrence**

A set A is *Harris recurrent* if

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

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

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

Note that

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

3.4 Invariant Measures

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

Requires the existence of a probability distribution π such that

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

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

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

- The chain is **positive recurrent** if π is a probability measure.
- Otherwise it is **null recurrent** or **transient**
- If π probability measure, π also called *stationary distribution* since

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

- The stationary distribution is unique

Insights

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

To understand why this is so, consider $P_\mu(X_n \in \cdot)$ for any starting distribution μ. If a limiting measure γ_μ exists in a suitable topology on the space of probability measures, such as

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

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

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

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

3.5 Ergodicity and convergence

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

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

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

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

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

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

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

and we want

$$\begin{aligned}&\left\| \int K^n(x, \cdot)\mu(dx) - \pi \right\|_{\mathrm{TV}} \\ &= \sup_A \left| \int K^n(x, A)\mu(dx) - \pi(A) \right|\end{aligned}$$

to be small.

Total variation distance and minoration

Let μ and μ' be two probability measures. Then,

$$1 - \inf \left\{ \sum_i \mu(A_i) \wedge \mu'(A_i) \right\} = \|\mu - \mu'\|_{\mathrm{TV}}.$$

where the infimum is taken over all finite partitions $(A_i)_i$ of $\mathcal{X}$.

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

$$\mu(A) \wedge \mu'(A) \geq \epsilon\nu(A).$$

Then, for all I and all partitions $A_1, A_2, \ldots, A_I$,

$$\sum_{i=1} \mu(A_i) \wedge \mu'(A_i) \geq \epsilon$$

and the previous result thus implies that

$$\|\mu - \mu'\|_{\mathrm{TV}} \leq (1-\epsilon).$$

Dobrushin coefficient

K, a Markov transition kernel on $\mathcal{X}$ has **Dobrushin's coefficient** $\alpha(K)$

$$\begin{aligned}\alpha(K) &= \inf\left\{\sum_{i\in I} K(x,A_i)\wedge K(x',A_i), \quad \forall x,x' \in \mathcal{X}, \right.\\ &\qquad\qquad \left. \forall A_1,\cdots,A_I \text{ partition of } \mathcal{X}\right\}\\ &= 1 - \sup_{x,x'\in\mathcal{X}} \|K(x,\cdot)-K(x',\cdot)\|_{\mathrm{TV}} < 1\end{aligned}$$

Then, for all measures μ, μ' on $\mathcal{B}(\mathcal{X})$

$$\|\mu K - \mu' K\|_{\mathrm{TV}} \leq \tau_1(K)\|\mu - \mu'\|_{\mathrm{TV}}$$

where $\tau_1(K) = 1 - \alpha(K)$

[Dobrushin, 1956]

Total variation distance and minoration

If, for all $x, x' \in \mathcal{X}$ and all $A \in \mathcal{B}(\mathcal{X})$,

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

then

$$\alpha(K) \leq \epsilon,$$

and thus, for any initial measures μ and μ' we have

$$\|\mu K - \mu' K\|_{\mathrm{TV}} \leq (1-\epsilon)\|\mu - \mu'\|_{\mathrm{TV}}.$$

Iterating the previous relation shows that, for all $n > 0$,

$$\|\mu K^n - \mu' K^n\|_{\mathrm{TV}} \leq (1-\epsilon)^n\|\mu - \mu'\|_{\mathrm{TV}}.$$

Harris recurrence and ergodicity

If (X_n) Harris positive recurrent and aperiodic, then

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

for every initial distribution μ.

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

[Meyn & Tweedie, 1993]

Convergence in total variation implies

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

for every *bounded function* h.

There are difference speeds of convergence

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

Geometric ergodicity

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

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

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

Geometric ergodicity implies a lot of important results

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

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

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

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

Forster-Lyapunov conditions

A central instrument to prove geometric ergodicity is made of **Forster-Lyapunov drift** conditions: There exist a small set C, a function $V \geq 1$, $\{V < \infty\} \neq \emptyset$, constants $\lambda < 1$, $b < \infty$, such that

$$PV \leq \lambda V + b\mathbb{I}_C$$

Main achievement of Markov chain theory: **Forster-Lyapunov conditions + proper irreducibility and aperiodicity conditions** are **necessary and sufficient** for geometric ergodicity.

[Meyn & Tweedie, 1993]

Minoration condition and uniform ergodicity

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

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

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

The following conditions are equivalent:

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

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

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

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

[Meyn and Tweedie, 1993]

3.6 Limit theorems

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

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

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

$$X_n \sim P_x^n$$

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

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

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

Ergodic Theorem

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

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

Central Limit Theorem

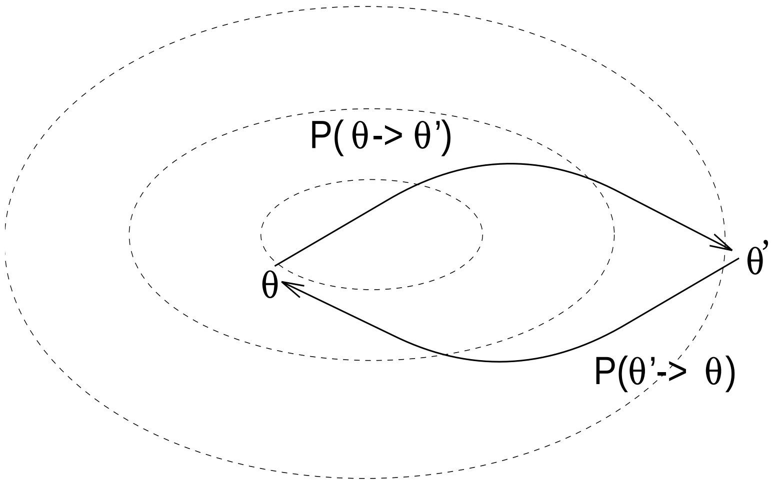
To get a CLT, we need more assumptions.

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

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

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

The direction of time does not matter



[Green, 1995]

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

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

where

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

[Kipnis & Varadhan, 1986]

3.7 Quantitative convergence rates

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

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

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

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

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

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

[Douc, Moulines and Rosenthal, 2001]

[Jones and Hobert, 2001]

For MCMC algorithms, kernels are "explicitly" known.

Type of quantities (more or less directly) available:

- Minoration constants

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

- Foster-Lyapunov Drift conditions,

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

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

Coupling

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

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

The **basic coupling construction**

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

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

respectively.

[Thorisson, 2000]

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

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

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

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

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

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

Small set and coupling sets

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

$$K(x,A) \geq \epsilon \nu(A), \qquad \forall x \in C. \tag{2}$$

Small sets always exist when the MC is φ-irreducible

[Jain and Jamieson, 1967]

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

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

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

Coupling sets can be larger than product of small sets

Assume

$$X_{n+1} = \phi(X_n) + \xi_{n+1}$$

where ϕ is a uniformly continuous function on $\mathbb{R}^d$, and $(\xi_n, n \geq 0)$ i.i.d. $\xi_n \sim g$

$$\|K(x,.) - K(x',.)\|_{\mathrm{TV}} = \int [g(y) - g(y - \{\phi(x) - \phi(x')\})]\, dy$$

Then, under regularity conditions for g and for M small enough,

$$\bar{C} = \left\{(x,x') \in \mathbb{R}^d, |x - x'| \leq M\right\}$$

is an ϵ-coupling set.

Coupling for Markov chains

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

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

For example,

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

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

Coupling algorithm

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

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

Coupling inequality

Define the coupling time T as

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

Coupling inequality

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

[Pitman, 1976; Lindvall, 1992]

Drift conditions

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

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

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

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

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

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

[Meyn and Tweedie, 1993]

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

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

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

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

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

DMR'01 result

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

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

where

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

Proof of the main result

Define N_k the number of visits to $\bar{C}$ before k:

$$N_k \ = \ \#\{m \,:\, 0 \le m \le k,\ (X_m, X_m') \in \bar{C}\}\,,$$

For any $0 \le j \le k$,

$$\|\xi P^k(\cdot) - \xi' P^k(\cdot)\|_{TV}$$
$$\le P_{\xi,\xi',0}[T > k,\ N_{k-1} \ge j] \ + \ P_{\xi,\xi',0}[T > k,\ N_{k-1} < j]$$

[Coupling inequality]

$\{T > k,\ N_{k-1} \ge j\}$ is contained in the event that the first j coin flips all came up tails. Hence, for $j < k$,

$$P_{\xi,\xi',0}[T > k,\ N_{k-1} \ge j] \le (1-\epsilon)^j.$$

By construction $P_{\xi,\xi',0}[T > k,\ N_{k-1} \ge k] = 0$.

Let

$$M_k \ = \lambda^{-k} B^{-N_{k-1}} V(X_k, X_k') \mathbb{I}(d_k = 0).$$

M_k is a **supermartingale**. Hence,

$$\mathbb{E}_{\xi,\xi',0}[M_k] \le \mathbb{E}_{\xi,\xi',0}[M_0]$$

Lindvall's inequality thus implies:

$$P_{\xi,\xi',0}[T > k,\ N_{k-1} \le j-1] \le \lambda^k B^{j-1} \mathbb{E}_{\xi,\xi',0}[V(X_0, X_0')]$$

Possible optimization of the bound

The drift condition also implies that, for any $0 < \gamma < 1$ and $(x, x') \notin C$

$$\bar{P} V^\gamma(x, x') \le \lambda^\gamma V^\gamma(x, x').$$

Using V^γ instead of V, the bound can be rewritten as, for $j \le k$,

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

where

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

Setting $j = k$, the bound above yields

$$\|\xi P^k(\cdot) - \xi' P^k(\cdot)\|_{\mathrm{TV}} \leq \left\{\lambda^\gamma \vee (1-\epsilon) \sup_{\bar{C}} \overline{R} V^\gamma \right\}^k \mathbb{E}_{\xi,\xi',0}[V^\gamma(X_0, X_0')],$$

which is similar to the bound found in Roberts and Tweedie (1999)

Optimal rate of convergence

Since

$$\gamma \to \lambda^\gamma$$

is a decreasing function of γ while

$$\gamma \to (1-\epsilon) \sup_{\bar{C}} \overline{R} V^\gamma$$

is increasing, the best rate of convergence is then obtained by choosing γ as the solution of the equation

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

Noting that,

$$\sup_{\bar{C}} \overline{R} V^\gamma \leq \sup_{\bar{C}} (\bar{R} V)^\gamma$$

it is possible to obtain a lower bound for the solution in closed form

$$\gamma_* = \log(1-\epsilon) / (\log(\lambda) - \log(\sup_{\bar{C}} \bar{R})).$$

The exponential rate of convergence is thus faster that λ^{γ_*}, bound found in Roberts and Tweedie (1999).

Stochastically monotone chains

Assume that the state-space $\mathcal{X}$ is totally ordered.

For λ and μ two probability measures on $\mathcal{X}$,

$$\lambda \leq \mu$$

if, for all $a \in \mathcal{X}$, $\lambda((-\infty, a]) \geq \mu((-\infty, a])$.

A Markov kernel is **stochastically monotone** if for all $x \leq x'$,

$$K(x, \cdot) \leq K(x', \cdot)$$

[Lund, Meyn and Tweedie, 1996; Roberts and Tweedie, 2000]

For $x \in \mathcal{X}$, define the **quantile function**

$$P^{-1}(x,u) = \inf\left\{y \in \mathcal{X},\ P\left(x,(-\infty,y]\right) \geq u\right\}.$$

and

$$\bar{P}(x,x';A \times A') = \int_0^1 \mathbb{I}_A(P^{-1}(x,u))\, \mathbb{I}_{A'}(P^{-1}(x',u))du.$$

This kernel **preserves the order** through successive iterations.

Assume that for P there exists a drift function V satisfying the Foster-Lyapunov condition $PV \leq \lambda V + b\mathbb{I}_C$ where $C = (-\infty, c]$ is a small set.

Set $\bar{V}(x,x') = V(x \vee x')$. Then

$$\bar{P}\bar{V}(x,x') = \int_0^1 V\left(P^{-1}(x,u) \vee P^{-1}(x',u)\right) du$$

If $x' \geq x$, $P^{-1}(x',u) \geq P^{-1}(x,u)$ for all $0 \leq u \leq 1$. Thus :

$$\bar{P}\bar{V}(x,x') = PV(x') \leq \lambda V(x') + b\mathbb{I}_{(-\infty,c]}((x,x'))$$

Hence, **the double chain satisfies the double drift conditions with the same constants as the single chain**

3.8 Renewal and CLT

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

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

is

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

Standard MC if CLT

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

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

Minoration

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

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

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

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

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

Split chain

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

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

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

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

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

Renewals

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

Then

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

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

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

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

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

Moment conditions

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

If

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

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

[Hobert & al., 2002]

Note that drift + minoration ensures geometric ergodicity

[Rosenthal, 1995; Roberts & Tweedie, 1999]

4 The Metropolis-Hastings Algorithm

4.1 Monte Carlo Methods based on Markov Chains

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

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

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

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

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

4.2 The Metropolis–Hastings algorithm

4.2.1 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 8 –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

4.2.2 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

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

   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.

4.3 A Collection of Metropolis-Hastings Algorithms

4.3.1 The Independent Case

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

Algorithm 9 –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

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

In this case,

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

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

[Mengersen & Tweedie, 1996]

Example 10 –Generating gamma variables–

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

Algorithm 11 –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

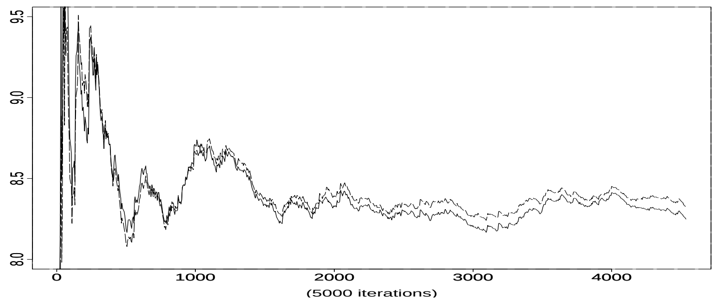
Algorithm 12 –Gamma Metropolis-Hastings–

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 $\mathcal{G}a(2, 2/2.43)$

4.3.2 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 13 –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 14 –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.

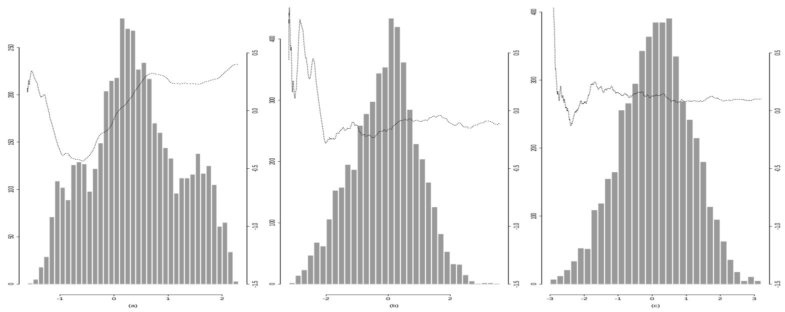


Figure 1: Three samples based on $\mathcal{U}[-\delta, \delta]$ with (a) $\delta = 0.1$, (b) $\delta = 0.5$ and (c) $\delta = 1.0$, superimposed with the convergence of the means ($15,000$ simulations).

Example 15 —Mixture models—

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

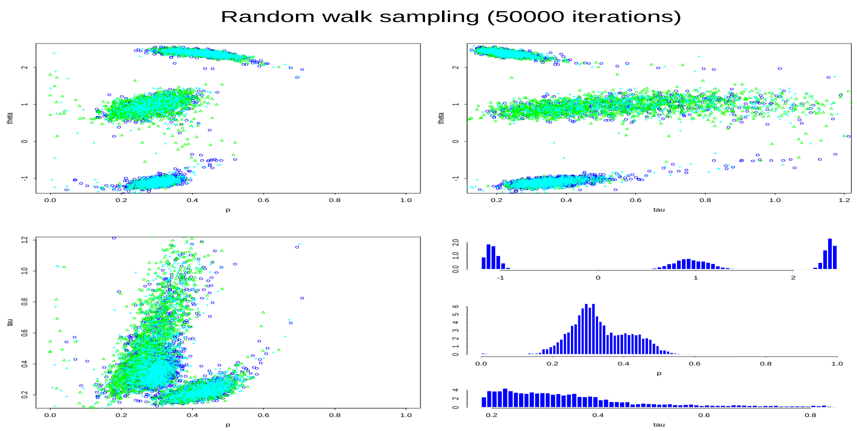
Metropolis-Hastings proposal:

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

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]

Convergence properties

Uniform ergodicity prohibited by random walk structure

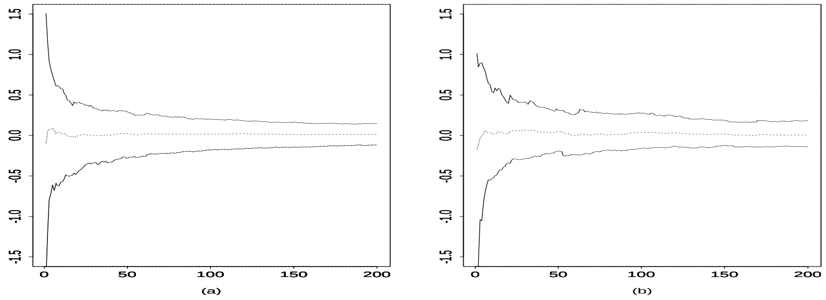
At best, geometric ergodicity:

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 16 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 $n, m \in \mathbb{N}$,

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

and $r(0) = 1$, for C a small set, $\tau_C = \inf\{n \ge 1, X_n \in C\}$, and $h \ge 1$, 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) $PV_{n+1} \leq V_n - r(n) h + b r(n) \mathbb{I}_C$.

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

4.4 Extensions

There are many other algorithms

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

to name a few...

4.4.1 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]

4.4.2 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 17 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]

5 The Gibbs Sampler

5.1 General Principles

A very specific simulation algorithm based on the target f

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

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

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

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 constituents is.

The Gibbs sampler can be turned into a reversible sampler, either using the *Random Scan Gibbs sampler* (see below) or running instead the (double) sequence

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

Example 19 –Bivariate Gibbs sampler–

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

Generate a sequence of observations by

Set $X_0 = x_0$, and for $t = 1, 2, \ldots,$ generate

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

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

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

Example 20 –Auto-exponential model–

On $\mathbb{R}^3_+$, density

$$f(y_1, y_2, y_3) \propto \exp\{-(y_1 + y_2 + y_3 + \theta_{12} y_1 y_2 + \theta_{23} y_2 y_3 + \theta_{31} y_3 y_1)\} ,$$

with known $\theta_{ij} > 0$.

The full conditional densities are exponential

$$Y_3 | y_1, y_2 \sim \mathcal{E}xp \left(1 + \theta_{23} y_2 + \theta_{31} y_1\right) ,$$

In contrast, the other conditionals, and the marginal distributions are difficult.

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.

5.1.1 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 21 –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 22 –Cauchy-normal –

Consider the density

$$f(\theta|\theta_0) \propto \frac{e^{-\theta^2/2}}{[1+(\theta-\theta_0)^2]^\nu}$$

posterior from the model

$$X|\theta \sim \mathcal{N}(\theta,1) \text{ and } \theta \sim \mathcal{C}(\theta_0,1).$$

Then

$$f(\theta|\theta_0) \propto \int_0^\infty e^{-\theta^2/2}\, e^{-[1+(\theta-\theta_0)^2]\,\eta/2}\, \eta^{\nu-1}\, d\eta,$$

and therefore

$$g(\theta,\eta) \propto e^{-\theta^2/2}\, e^{-[1+(\theta-\theta_0)^2]\,\eta/2}\, \eta^{\nu-1},$$

with conditional densities

$$g_1(\eta|\theta) \quad = \quad \mathcal{G}a\left(\nu, \frac{1+(\theta-\theta_0)^2}{2}\right),$$

$$g_2(\theta|\eta) \quad = \quad \mathcal{N}\left(\frac{\theta_0\eta}{1+\eta}, \frac{1}{1+\eta}\right).$$

The parameter η is completely meaningless for the problem at hand but serves to facilitate computations.

Example 23 —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 $(i = 1, \ldots, k)$

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

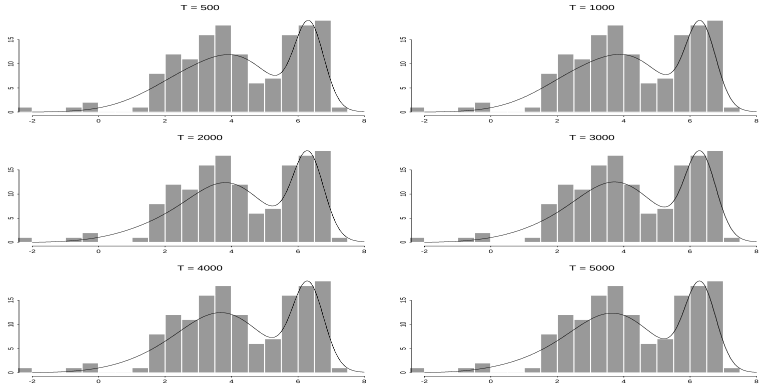
and update n_i and $\bar{x}_i$ $(i = 1, \ldots, k)$.

149 observations of acidity levels in lakes in the American North-East

Mixture model fit with the Gibbs sampler

Lack of evolution of estimated (plug-in) density from the Gibbs sampler when iterations increase

Phenomenon which occurs often in mixture settings, due to weak identifiability of these models.



Estimation of the density for 3 components and T iterations

5.1.2 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.

5.1.3 Slice sampler

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

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

it can be completed

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

leading to the following Gibbs algorithm:

Algorithm 24 –Slice sampler–

Simulate

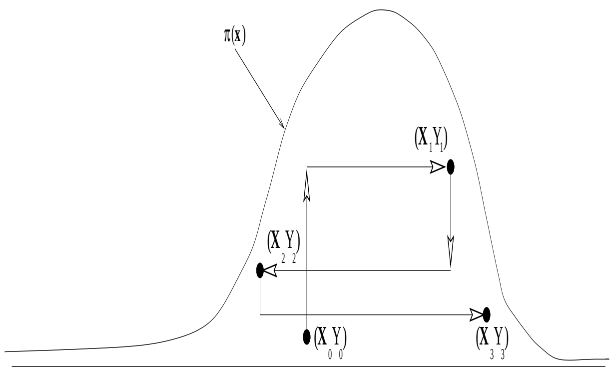
1. $\omega_1^{(t+1)} \sim \mathcal{U}_{[0, f_1(\theta^{(t)})]}$;

   $\dots$

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

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

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



Representation of a few steps of the slice sampler

[Roberts & Rosenthal, 1998]

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

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

Example 25 –Normal simulation–
For the standard normal density,

$$f(x) \propto \exp(-x^2/2),$$

a slice sampler is based on

$$\begin{aligned} \omega|x &\sim \mathcal{U}_{[0,\exp(-x^2/2)]}\ , \\ X|\omega &\sim \mathcal{U}_{[-\sqrt{-2\log(\omega)},\sqrt{-2\log(\omega)}]} \end{aligned}$$

5.1.4 Properties of the Gibbs sampler

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

If either

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

[Positivity condition]

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

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

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

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

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

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

for every initial distribution μ.

Slice sampler

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

If f is bounded and $\operatorname{supp} f$ 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**:

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

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 26 —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 27 —A poor slice sampler—

Consider

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

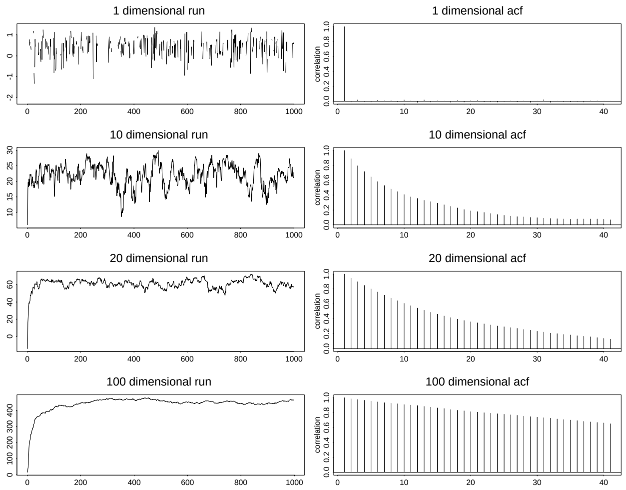
Slice sampler equivalent to one-dimensional slice sampler on

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

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)

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

5.1.6 Hierarchical models

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

Example 28 –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

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

5.2 Data Augmentation

The Gibbs sampler with only two steps is particularly useful

Algorithm 29 –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 30 –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 31 –Poisson-Gamma Gibbs–

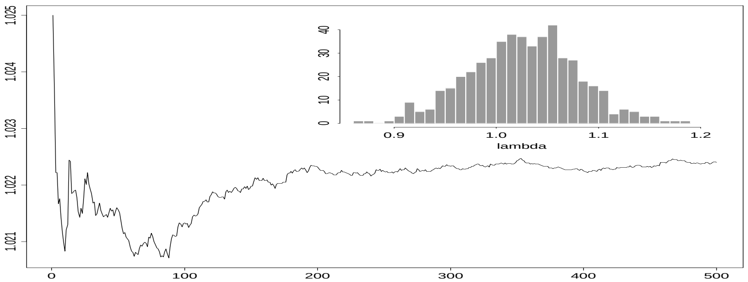
1.. Simulate $Y_i^{(t)}\sim\mathcal{P}(\lambda^{(t-1)})\,\mathbb{I}_{y\geq 4}\quad i=1,\ldots,13$

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



5.2.1 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
$$\mathrm{var}\left(\mathbb{E}\left[h(Y_1)|Y_2^{(t)},\ldots,Y_p^{(t)}\right]\right) \le \mathrm{var}(h(Y_1)),$$
so δ_{rb} is uniformly better (for Data Augmentation)

Some examples of Rao-Blackwellization

- For the bivariate normal
$$(X,Y)' \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$
the Gibbs sampler is based upon
$$\begin{aligned} X \mid y &\sim \mathcal{N}(\rho y,\, 1-\rho^2)\\ Y \mid x &\sim \mathcal{N}(\rho x,\, 1-\rho^2). \end{aligned}$$

To estimate $\mu = \mathbb{E}(X)$ we could use
$$\delta_0 = \frac{1}{T}\sum_{i=1}^{T} X^{(i)}$$
or its Rao-Blackwellized version
$$\delta_1 = \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}[X^{(i)}|Y^{(i)}] = \frac{1}{T}\sum_{i=1}^{T}\varrho Y^{(i)},$$
which satisfies $\sigma^2_{\delta_0}/\sigma^2_{\delta_1} = \frac{1}{\rho^2} > 1.$

- For the Poisson-Gamma Gibbs sampler, we could estimate λ with
$$\delta_0 = \frac{1}{T}\sum_{t=1}^{T}\lambda^{(t)},$$
but we instead used the Rao-Blackwellized version
$$\begin{aligned}\delta^\pi &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\lambda^{(t)}|x_1,x_2,\ldots,x_5,y_1^{(i)},y_2^{(i)},\ldots,y_{13}^{(i)}]\\ &= \frac{1}{360T}\sum_{t=1}^{T}\left(313+\sum_{i=1}^{13}y_i^{(t)}\right),\end{aligned}$$

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

The estimator

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

and is unbiased.

5.2.2 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.

5.2.3 Parameterization

Convergence of both Gibbs sampling and Metropolis–Hastings algorithms may suffer from a poor choice of parameterization

The overall advice is to try to make the components “as independent as possible”.

Example 32 –Random effects model–

In the simple random effects model

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

where

$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$

for a flat prior on μ, the Gibbs sampler implemented for the

$$(\mu, \alpha_1, \ldots, \alpha_I)$$

parameterization exhibits high correlation if $\sigma_y^2/(IJ\sigma_\alpha^2)$ is large and consequent slow convergence

If the model is rewritten as the hierarchy

$$y_{ij} \sim \mathcal{N}(\eta_i, \sigma_y^2), \qquad \eta_i \sim \mathcal{N}(\mu, \sigma_\alpha^2),$$

the correlations between the η_i's and between μ and the η_i's are lower

5.3 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 33 –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 function $f(x_1, x_2)$ that is a candidate for the joint density is

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

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

Thus, these conditional distributions do not correspond to a joint probability distribution.

Example 34 –Improper random effects–

For a random effect model,

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

where

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

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

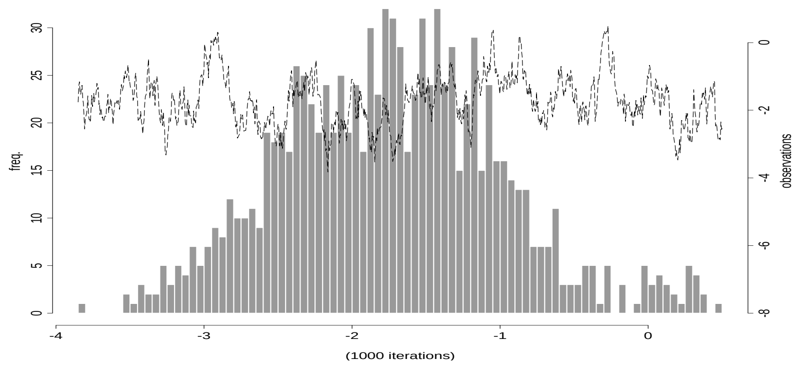
$$\pi(\mu, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2}\ .$$

The conditional distributions

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

are well-defined and a Gibbs sampling can be easily implemented in this setting.

Evolution of $(\mu^{(t)})$ and corresponding histogram



The figure shows the sequence of the $\mu^{(t)}$ and the corresponding histogram for 1000 iterations. The trend of the sequence and the histogram **do not** indicate that the corresponding "joint distribution" **does not exist**

Final notes on impropriety

The improper posterior Markov chain
cannot be positive recurrent

The major task in such settings is to find indicators that flag that something is wrong. However, the output of an "improper" Gibbs sampler may not differ from a positive recurrent Markov chain.

Example The random effects model was initially treated in Gelfand *et al.* (1990) as a legitimate model

6 MCMC tools for variable dimension problems

6.1 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)**

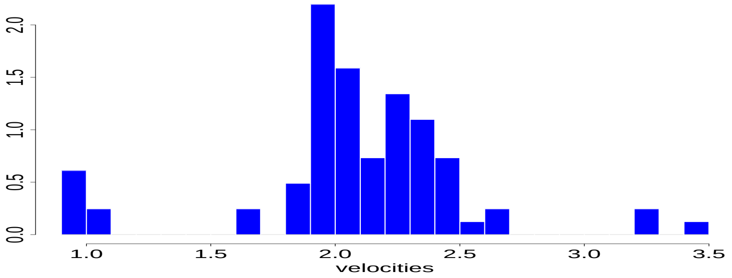
Many areas of application

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

Example 35 —Mixture modelling–

Benchmark dataset: Speed of galaxies

[Roeder, 1990; Richardson & Green, 1997]



Modelling by a mixture model

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

i?

Bayesian variable dimension model

A variable dimension model is defined as a collection of models $(k = 1 \ldots, K)$,

$$\mathfrak{M}_k = \{f(\cdot|\theta_k);\ \theta_k \in \Theta_k\}\ ,$$

associated with a collection of priors on the parameters of these models,

$$\pi_k(\theta_k)\,,$$

and a prior distribution on the indices of these models,

$$\{\varrho(k)\,, k = 1, \ldots, K\}\ .$$

Alternative notation:

$$\pi(\mathfrak{M}_k, \theta_k) = \varrho(k)\,\pi_k(\theta_k)$$

Formally over:

1. Compute

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

2. Take largest $p(\mathfrak{M}_i|x)$ to determine model, or use

$$\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j$$

as predictive

[Different decision theoretic perspectives]

Difficulties

Not at

- (formal) inference level [see above]
- parameter space representation

$$\Theta = \bigoplus_k \Theta_k \,,$$

[even if there are parameters common to several models]

Rather at

- (practical) inference level:
  model separation, interpretation, overfitting, prior modelling, prior coherence
- computational level:
  infinity of models, moves between models, predictive computation

6.2 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

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

Reciprocal proposal has density

$$\frac{\exp\left\{-(\theta_2 - \Psi_{1\to 2}(\theta_1, v_{1\to 2}))^2/2\varepsilon\right\}}{\sqrt{2\pi\varepsilon}} \times \left| \frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})} \right|$$

by the Jacobian rule.

Thus Metropolis–Hastings acceptance probability is

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

Does not depend on ε: **Let ε go to 0**

Interpretation (2): saturation

[Brooks, Giudici, Roberts, 2003]

Consider series of models $\mathfrak{M}_i$ $(i = 1, \ldots, k)$ such that

$$\max_i \dim(\mathfrak{M}_i) = n_{\max} < \infty$$

Parameter of model $\mathfrak{M}_i$ then completed with an auxiliary variable U_i such that

$$\dim(\theta_i, u_i) = n_{\max} \quad \text{and} \quad U_i \sim q_i(u_i)$$

Posit the following joint distribution for [augmented] model $\mathfrak{M}_i$

$$\pi(\mathfrak{M}_i, \theta_i)\, q_i(u_i)$$

Saturation: no varying dimension anymore since (θ_i, u_i) of fixed dimension.

Three stage MCMC update:

1. Update the current value of the parameter, θ_i;
2. Update u_i conditional on θ_i;
3. Update the current model from $\mathfrak{M}_i$ to $\mathfrak{M}_j$ using the bijection

$$(\theta_j, u_j) = \Psi_{i\to j}(\theta_i, u_i)$$

Example 36 —Mixture of normal distributions—

$$\mathfrak{M}_k : \sum_{j=1}^{k} p_{jk}\mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$$

[Richardson & Green, 1997]

Moves:

(i). Split

$$\begin{cases} p_{jk} & = & p_{j(k+1)} + p_{(j+1)(k+1)} \\ p_{jk}\mu_{jk} & = & p_{j(k+1)}\mu_{j(k+1)} + p_{(j+1)(k+1)}\mu_{(j+1)(k+1)} \\ p_{jk}\sigma_{jk}^2 & = & p_{j(k+1)}\sigma_{j(k+1)}^2 + p_{(j+1)(k+1)}\sigma_{(j+1)(k+1)}^2 \end{cases}$$

(ii). Merge (*reverse*)

Additional **Birth and Death** moves for empty components (created from the prior distribution)

Equivalent

(i). Split

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

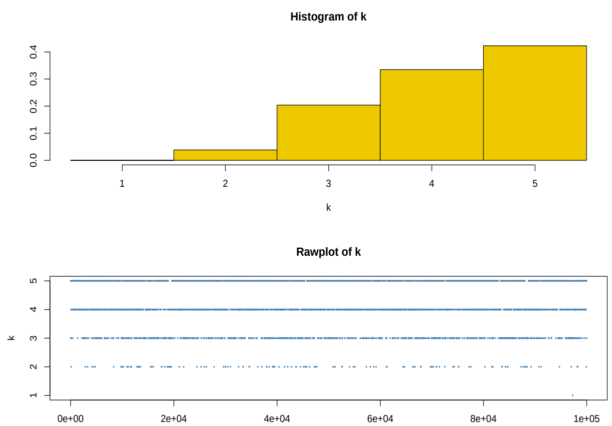
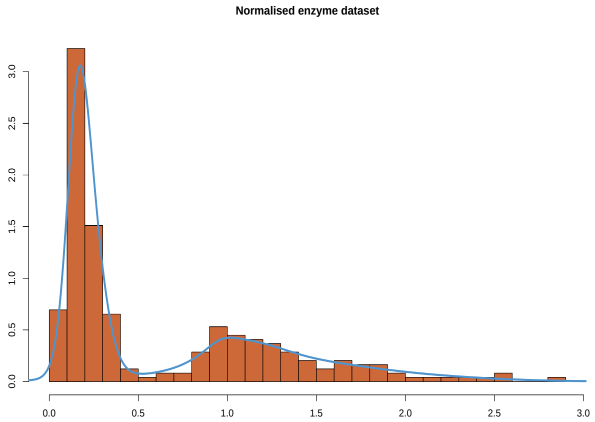


Figure 2: Histogram and rawplot of $100,000$ k's produced by RJMCMC under the imposed constraint $k \leq 5$.



Example 37 —Hidden Markov model—

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

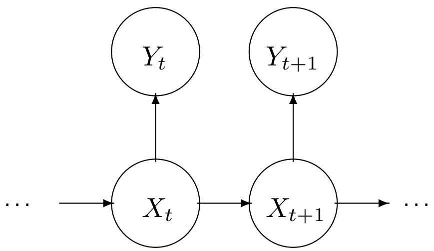


Figure 3: DAG representation of a simple hidden Markov model

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]

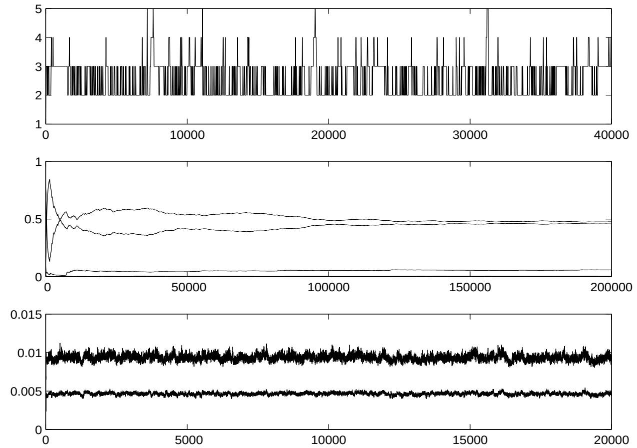


Figure 4: Upper panel: First 40,000 values of k for S&P 500 data, plotted every 20th sweep. Middle panel: estimated posterior distribution of k for S&P 500 data as a function of number of sweeps. Lower panel: σ_1 and σ_2 in first 20,000 sweeps with $k = 2$ for S&P 500 data.

Example 38 —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]

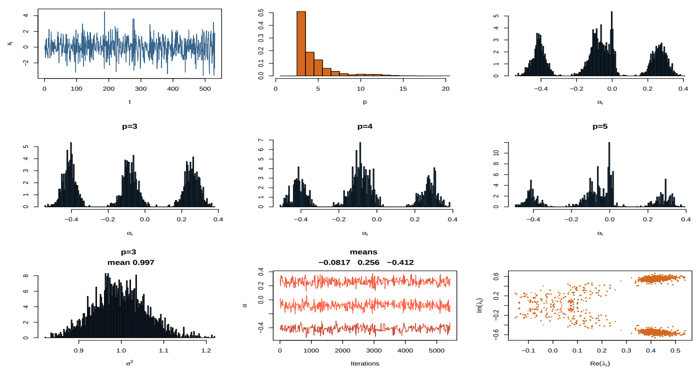


Figure 5: Reversible jump algorithm based on an $AR(3)$ simulated dataset of 530 points *(upper left)* with true parameters α_i $(-0.1, 0.3, -0.4)$ and $\sigma = 1$. First histogram associated with p, the following histograms with the α_i's, for different values of p, and of σ^2. Final graph: scatterplot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

6.3 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**

$$L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\theta}') = L(\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}',\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta},\boldsymbol{\theta}'$$

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 39 —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 $\delta_j(\Phi)/\delta(\Phi)$

   $k \leftarrow k-1$

   $p_\ell \leftarrow p_\ell/(1-p_j) \quad (\ell \neq j)$

   Otherwise,

   Add component j from the prior $\pi(\mu_j,\sigma_j)$

   $p_j \sim \mathcal{B}e(\gamma, k\gamma)$

   $p_\ell \leftarrow p_\ell(1-p_j) \quad (\ell \neq j)$

   $k \leftarrow k+1$

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

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

Hence **"RJMCMC→BDMCMC"**. This holds more generally.

Example 40 **—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]

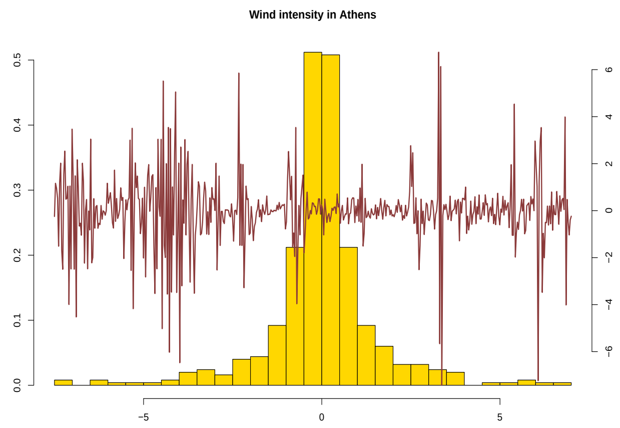


Figure 6: Histogram and rawplot of 500 wind intensities in Athens

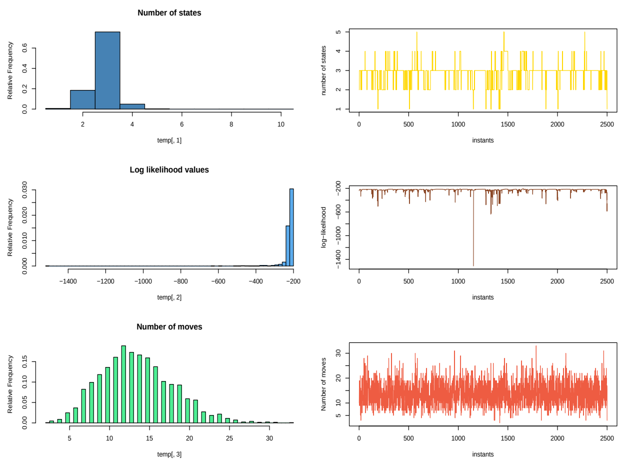


Figure 7: MCMC output on k (histogram and rawplot), corresponding loglikelihood values (histogram and rawplot), and number of moves (histogram and rawplot)

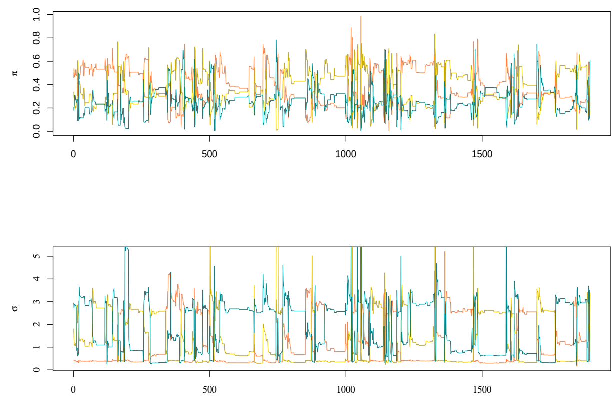


Figure 8: MCMC sequence of the probabilities π_j of the stationary distribution (top) and the parameters σ (bottom) of the three components when conditioning on $k = 3$

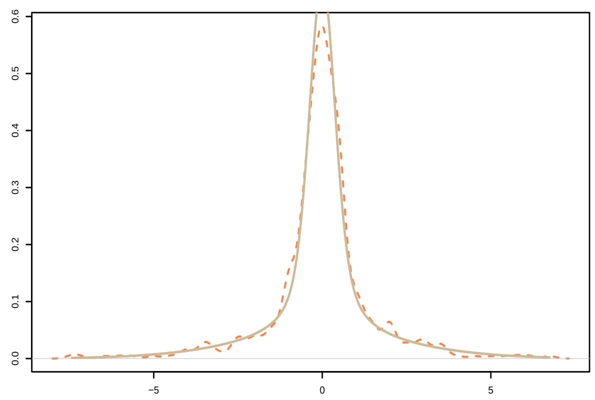


Figure 9: MCMC evaluation of the marginal density of the dataset (dashes), compared with R nonparametric density estimate (solid lines).

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 $\tau_1, \dots, \tau_N$ 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 41 —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.

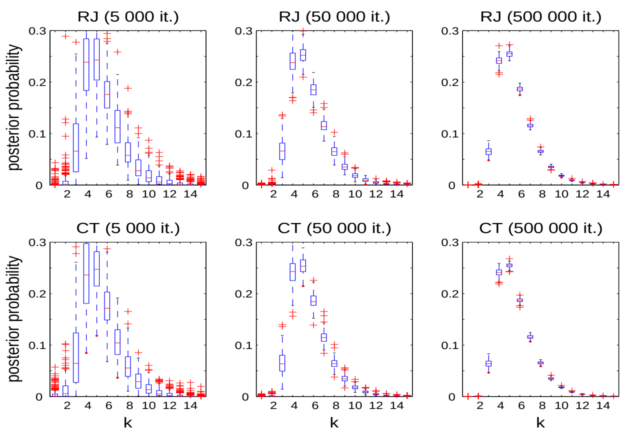


Figure 10: Galaxy dataset, box plot for the estimated posterior on k obtained from 200 independent runs: RJMCMC (top) and BDMCMC (bottom). The number of iterations varies from 5 000 (left), to 50 000 (middle) and 500 000 (right).

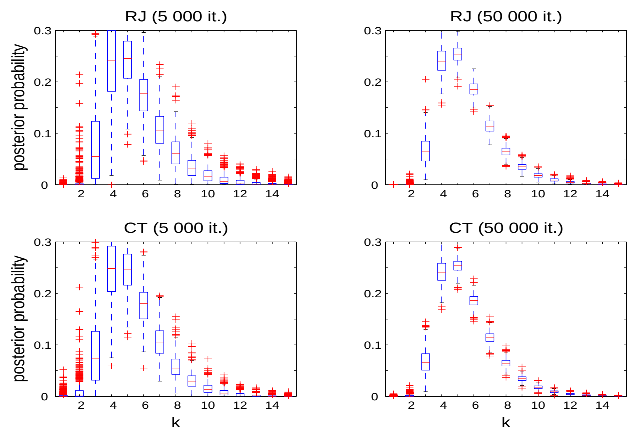


Figure 11: Galaxy dataset, box plot for the estimated posterior on k obtained from 500 independent runs: Top RJMCMC and bottom, CTMCMC. The number of iterations varies from 5 000 (left plots) to 50 000 (right plots).

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 42 **—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

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

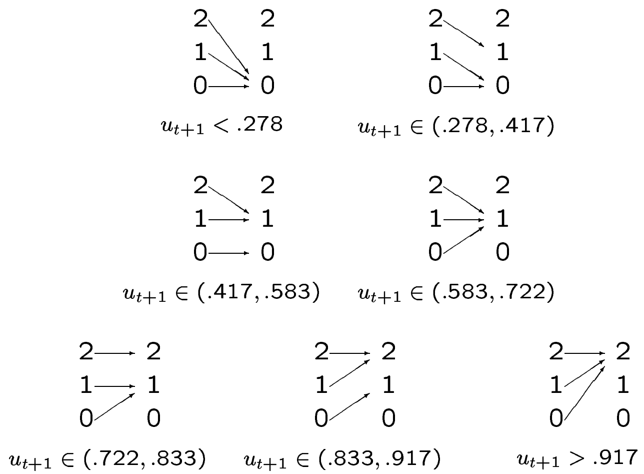
$n = 2$, $\alpha = 2$ **and** $\beta = 4$.

State space

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

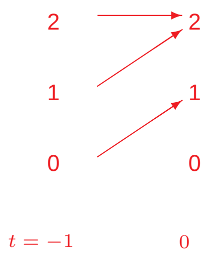
Transition probabilities

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

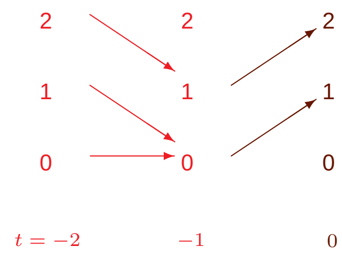


All possible transitions for the Beta-Binomial(2,2,4) example

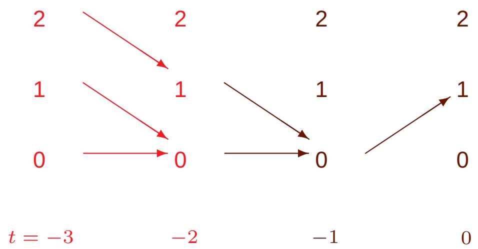
Begin at time $t = -1$ and draw U_0. Suppose $U_0 \in (.833, .917)$.



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 43 **—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 $(0 \leq j \leq n)$ 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 $m = 0, \ldots, n$

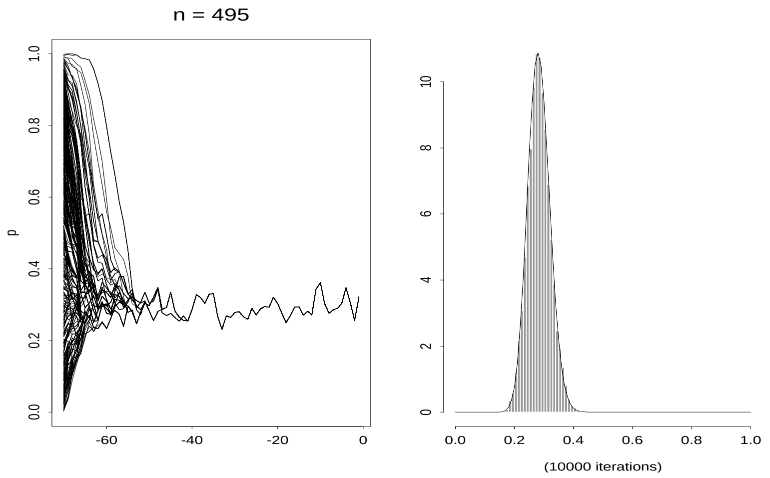


Figure 12: Simulation of $n = 495$ iid rv's from $.33\mathcal{N}(3.2, 3.2) + .67\mathcal{N}(1.4, 1.4)$ and coalescence at $t = -73$.

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

$$\begin{aligned} n_1 &= \sum_{i=1}^{n} \mathbb{I}\left(u_i \le \frac{p_1 f_1(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i) + p_3 f_3(x_i)}\right), \\ n_2 &= \sum_{i=1}^{n} \left\{ \mathbb{I}\left(u_i > \frac{p_1 f_1(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i) + p_3 f_3(x_i)}\right) \right. \\ &\qquad \left. \times \mathbb{I}\left(u_i \le \frac{p_1 f_1(x_i) + p_2 f_2(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i) + p_3 f_3(x_i)}\right) \right\}, \end{aligned}$$

   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

$$\mathcal{T} = \{(n_1, n_2); n_1 + n_2 \le n\}$$

by Gibbs is contained in the lozenge

$$\mathcal{L} = \{(n_1, n_2); \underline{n}_1 \le n_1 \le \overline{n}_1,\ n_2 \ge 0,\ \underline{n}_3 \le n - n_1 - n_2 \le \overline{n}_3\}\,,$$

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,

$$\frac{p_2}{p_1} = \sum_{i=1}^{n_2+1} w_{2i} \Bigg/ \sum_{i=1}^{n_1+1} w_{1i} \quad \text{and} \quad \frac{p_3}{p_1} = \sum_{i=1}^{n-n_1-n_2+1} w_{3i} \Bigg/ \sum_{i=1}^{n_1+1} w_{1i}$$

are both decreasing in n_1.

So is

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

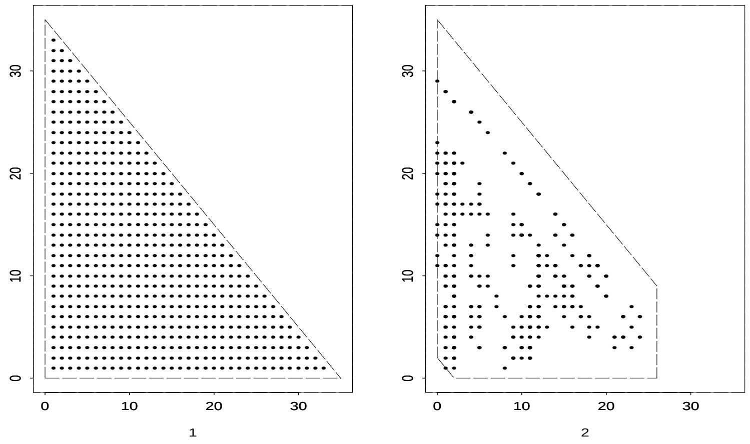


Figure 13: Sample of $n = 35$ observations from $.23\,\mathcal{N}(2.2, 1.44) + .62\,\mathcal{N}(1.4, 0.49) + .15\,\mathcal{N}(0.6, 0.64)$

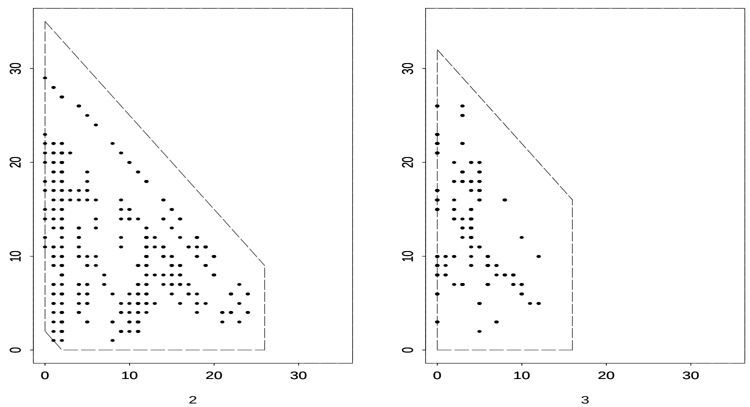
Lozenge monotonicity (preserved)

The image of $\mathcal{L}$ is contained in

$$\mathcal{L}' = \{(m_1, m_2); \underline{m}_1 \le m_1 \le \overline{m}_1\ , m_2 \ge 0,\ \underline{m}_3 \le m_3 \le \overline{m}_3\},$$

where

- $\underline{m}_1$ is $\min n_1$ over the images of the left border $\{n_1 = \underline{n}_1\}$
- $\overline{m}_1$ is $\max n_1$ over the images of the right border $\{n_1 = \overline{n}_1\}$
- $\underline{m}_3$ is $\min n_3$ over the images of the upper border $\{n_3 = \underline{n}_3\}$
- $\overline{m}_3$ is $\max n_3$ of the images of the lower border $\{n_3 = \overline{n}_3\}$



Lozenge monotonicity (completed)

- Envelope result: generation of the images of all points on the borders of $\mathcal{L}$

  [Kendall, 1998]
- $O(n)$ complexity versus $O(n^2)$ for brute force CFTP
- Checking for coalescence of the borders only : *almost perfect* !
- Extension to $k = 4$ underway

  [Machida, 1999]

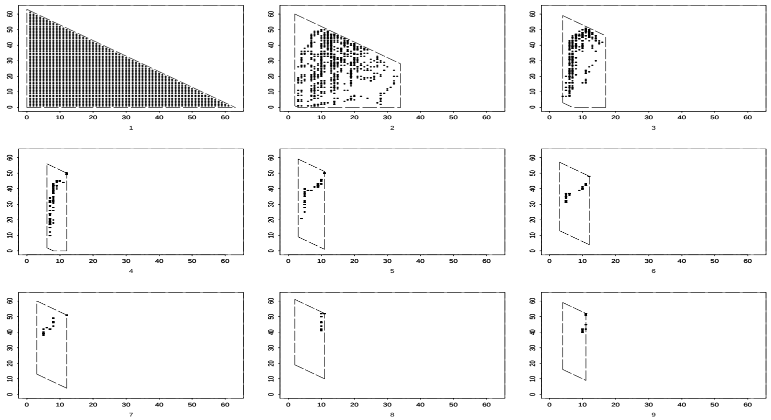


Figure 14: $n = 63$ observations from $.12\,\mathcal{N}(1.1, 0.49) + .76\,\mathcal{N}(3.2, 0.25) + .12\,\mathcal{N}(2.5, 0.09)$

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,

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

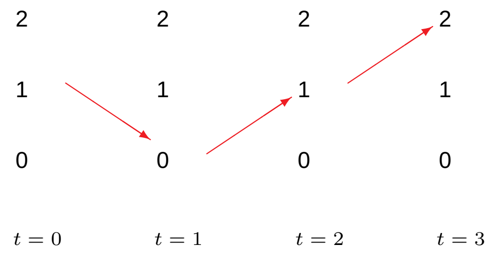
Example 44 **—Beta-Binomial—**

Choose $T = 3$ and $X_T = 2$.

Reversible chain, so

$$X_2|X_3 = 2 \sim \text{BetaBin}(2,4,4)$$
$$X_1|X_2 = 1 \sim \text{BetaBin}(2,3,5)$$
$$X_0|X_1 = 2 \sim \text{BetaBin}(2,4,4)$$

Suppose



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

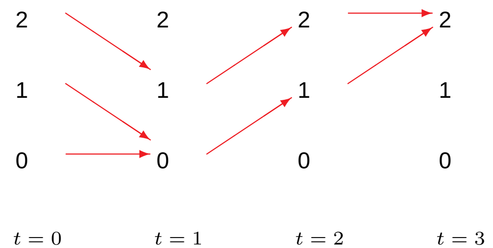
imply

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

Suppose

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

Begin chains in states 0, 1 and 2.



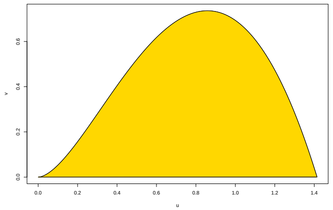
The chains coalesce in $X_3 = 2$; so we accept $X_0 = 1$ as a draw from π.

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 $\quad \pi(\omega_1) \leq \pi(\omega_2)$

$$\mathcal{A}_2 = \{\omega\,;\,\pi(\omega) \geq u\pi(\omega_2)\} \subset \mathcal{A}_1 = \{\omega\,;\,\pi(\omega) \geq u\pi(\omega_1)\}$$

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 45 —Exponential example ($k = 2, p$ known)

Joint distribution

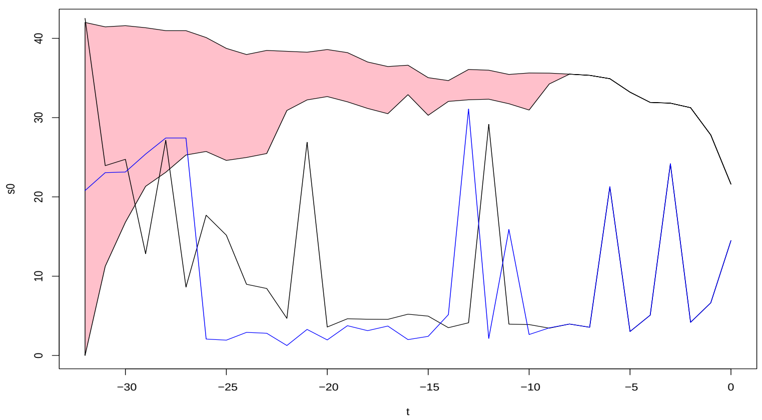
$$\prod_{i=1}^{n} p^{(1-z_i)}(1-p_i)^{z_i}\lambda_{z_i}\exp(-\lambda_{z_i}x_i)\prod_{j=1}^{k}\lambda_j^{\alpha_j-1}\exp(-\lambda_j\beta_j)$$

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

$$x_2^{(t+1)} = \begin{cases} y_t & \text{if } u_t \leq \dfrac{\pi(y_t)\, q(x_2^{(t)}|x_1^{(t)})}{\pi(x_2^{(t)})\, q(y_t|x_1^{(t)})}, \\ x_2^{(t)} & \text{otherwise.} \end{cases} \quad (3)$$

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

$$\tilde{1} = \arg\max \pi/h,$$

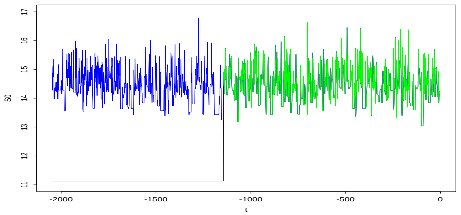
the coupling (3) preserves the ordering.

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

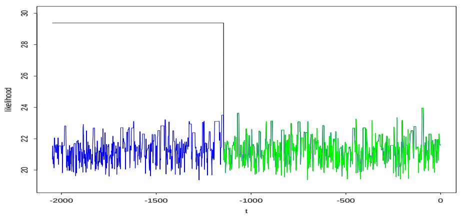
Example When state space $\mathcal{X}$ compact, use for h the uniform distribution on $\mathcal{X}$.

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

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

Can be extended to stationary measures

8 Controlled MCMC algorithms

8.1 Adaptive MCMC algorithms

How to efficiently estimate

$$\mathfrak{I}(h) = \int_{\mathcal{X}} h(x) f(dx)\,?$$

with an MCMC based estimator

$$\widehat{\mathfrak{I}}_N(h) = \frac{1}{N+1}\sum_{i=0}^{N} h(x_i)\,?$$

[Andrieu & Robert, 2002]

Metropolis-Hastings algorithm

Given that the Markov chain is at x, **proposal distribution**

$$y|x \sim q(x, y)$$

Then the Markov chain

1. goes to y with probability

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

2. Otherwise stays at x.

The choice of q

Key to the success of the MCMC approach.

Typically q depends on a parameter θ

$$q = q_\theta$$

Example 46 —Symmetric Gaussian random walk—

$$q_\theta(x,y) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(\frac{-1}{2\theta^2}(y-x)^2\right)$$

Variance of $\widehat{\mathfrak{I}}_N(h)$ large for values of θ^2 either too small or too large.

Example 47 —Mixture of kernels

$$\sum_{i=1}^{k} \omega_i \mathfrak{K}_i(\,\cdot\,; x)$$

depends on the weight vector $\theta = (\omega_1, \ldots, \omega_k)$ for its efficiency (model choice, blocking, etc.)

Example 48 Choice of a auxilliary parameter in a completion scheme

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

Goals

1. We want to choose θ in an "**optimal**" manner (to be defined below!)
2. We want this choice to be **automatic** **[minimise human intervention and time waste]**
3. We want to use a **single** run of the algorithm **[adaptive algorithm]**

Potential problem

If at iteration i we adjust θ in the light of the **whole past of the chain**,

$$x_0, x_1, \ldots, x_{i-1} \,,$$

then it is not a Markov chain anymore.

What about ergodicity then???

Updating scheme preserving ergodicity proposed in Gelman *et al.* (1995).

- Relies on the notion of **regeneration**
- Theoretically valid, but practically difficult to apply

Adaptive scale of Haario *et al.* (2000)

- Variance update $\theta^{(t+1)} = \frac{t}{t+1}\theta^{(t)} + \frac{1}{t+1}(X_t - \mu)(X_t - \mu)^{\mathrm{T}}$
- Complex (local) proof of ergodicity

8.2 Performances of MCMC algorithms

Define a **loss** criterion/function for the evaluation of the performances of an MCMC algorithm

$$\eta\,(\theta)$$

in such a way that optimum value θ_* is root of

$$\eta\,(\theta) = 0$$

Reformulated as a **minimisation problem**

$$\theta_* = \arg\min \Psi(\eta(\theta))$$

where

$$\eta(\theta) = \int_{\mathcal{X}} \mathfrak{H}(\theta, x)\mu_\theta(dx) \qquad (4)$$

Example 49 —Coerced acceptance—

Define an optimal acceptance rate $\alpha_\star$ and set

[Gelman & al., 1995]

$$\begin{aligned}\Psi(\eta) &= \eta^2 \\ \eta(\theta) &= \overline{\alpha}_\theta - \alpha_\star \\ &= \int_{\mathcal{X}^2} \left[1 \wedge \frac{f(y)\, q_\theta(y,x)}{f(x)\, q_\theta(x,y)} - \alpha_\star\right] q_\theta(x,y)\, f(x)\, dxdy.\end{aligned}$$

and

$$\mathfrak{H}(\theta,x,y) = \left[\min\left\{1, \frac{f(dy) q(y,dx;\theta)}{f(dx)\, q(x,dy;\theta)}\right\} - \alpha_*\right]$$

Example 50 —Autocorrelations—

Asymptotic variance of $\sqrt{N}\,\widehat{\mathfrak{I}}_N(h)$, approximated by its truncated version

[Geyer, 1992]

$$\eta(\theta) = \Sigma_{h,\tau}(\theta) = \mathrm{var}_f(h(x_0)) + 2\sum_{i=1}^{\tau} \mathrm{cov}\left(h(x_0), h(x_i); \theta\right),$$

and

$$\mathfrak{H}(\theta,x) = h(x_0)h(x_0)' + 2\sum_{i=1}^{\tau} h(x_0)h(x_i)'$$

Example 51 —Moment matching—

Force the proposal to match some moments of the target

$$\int_{\mathcal{X}} (\phi(x) - \theta)\, f(dx) = 0$$

[Haario & al., 2000]

Th

$$\eta(\theta) = \int_{\mathcal{X}} (\phi(x) - \theta)\, f(dx)$$

and

$$\Psi(\eta) = |\eta|^2$$

8.3 Adaptation towards efficiency

How can one find the roots of

$$h(\theta) = 0$$

when

$$h(\theta) = \int_{\mathcal{X}} \mathfrak{H}(\theta,x)\, \mu_\theta(dx)\ ?$$

In most cases of interest, it is not possible to evaluate the integral for a fixed θ, and one needs to resort to numerical methods

The Robbins-Monro algorithm

Iterative techniques called **stochastic approximation** follow from the Robbins-Monro algorithm

[Robbins & Monro, 1954]

Noisy gradient optimisation algorithm that takes advantage of the missing data representation of $h(\theta)$:

$$\theta_{i+1} = \theta_i + \gamma_{i+1} \mathfrak{H}\left(\theta_i, x_{i+1}\right),$$

where

$$x_{i+1} | \, \theta_i \sim \mu_{\theta_i}\left(dx\right).$$

and γ_i slowly drifts to 0

MCMC approximation

If sampling from μ_θ difficult, introduce a family of transition probabilities P_θ such that

$$\mu_\theta P_\theta = \mu_\theta$$

in which case

$$\theta_{i+1} = \theta_i + \gamma_{i+1} \mathfrak{H}\left(\theta_i; x_{i+1}\right),$$

where

$$x_{i+1} | \left(\theta_i, x_i\right) \sim P_{\theta_i}\left(x_i; dx_{i+1}\right).$$

Asymptotic behaviour of the algorithm

Since

$$\begin{aligned} \theta_{i+1} &= \theta_i + \gamma_{i+1} h\left(\theta_i\right) + \gamma_{i+1}\left\{\mathfrak{H}\left(\theta_i; x_{i+1}\right) - h\left(\theta_i\right)\right\} \\ &= \theta_i + \gamma_{i+1} h\left(\theta_i\right) + \gamma_{i+1}\, \mathfrak{e}_i \end{aligned}$$

if the effect of the noise series $\{\mathfrak{e}_i\}$ “cancels out”, then the “mean trajectory” of the algorithm is precisely that of the deterministic gradient algorithm.

Intuition

The trajectories $\theta_0, \theta_1, \ldots$ behave asymptotically more or less like the solutions $\theta\left(t\right)$ of the ODE

$$\dot{\theta}\left(t\right) = h\left(\theta\left(t\right)\right).$$

and the solutions of the ODE should converge to a **stationary point**

Many acceleration techniques and variations of the algorithm exist in literature.

Slightly different context here: to solve

$$\min \Psi(\eta(\theta))$$

requires solving the first order equation

$$\begin{aligned}\nabla_\theta \{\Psi(\eta(\theta))\} &= 0 \\ &= \nabla_\theta(\eta(\theta))\,\Psi'(\eta(\theta))\end{aligned}$$

Two-time scale stochastic approximation

Consider jointly

- the Markov chain of interest $\{x_i\}$,
- the proposal parameter θ,

and

- the transforms

$$\xi(\theta, x) = (\mathfrak{H}(\theta, x), \nabla_\theta \mathfrak{H}(\theta, x))$$

and apply "twice" Robbins–Monro

Recursion i

Set $\xi_i = (\eta_i, \dot{\eta}_i)$

Corresponding recursive system

$$\begin{aligned}x_{i+1} &\sim \mathfrak{K}(x_i, dx_{i+1}; \theta_i) \\ \xi_{i+1} &= (1-\gamma_{i+1})\xi_i + \gamma_{i+1}\xi(\theta_i, x_{i+1}) \\ \theta_{i+1} &= \theta_i - \gamma_{i+1}\varepsilon_{i+1}\dot{\eta}_i\Psi'(\eta_i)\end{aligned}$$

where $\{\gamma_i\}$ and $\{\varepsilon_i\}$ go to 0 at infinity

Intuition

Second **time scale** ε_i is there to **slow down** the evolution of the θ_i's

- if $\{\varepsilon_i\}$ goes fast enough to 0, the overall convergence behavior [for the x_i's and ξ_i's] similar to when θ **is fixed**,
- on the time scale $\{\gamma_i\varepsilon_i\}$, θ_i still converges to the solution of

$$\dot{\eta}(\theta)\nabla_\theta\Psi(\eta(\theta)) = 0$$

Convergence conditions on $\{\gamma_i\}$ and $\{\varepsilon_i\}$

- $\{\gamma_i\}$ and $\{\varepsilon_i\}$ go to 0 at infinity
- slow decrease to 0:

$$\sum_i \gamma_i \varepsilon_i = \infty \qquad \sum_i \gamma_i^2 < \infty$$

[Andrieu & Moulines, 2002]

An interesting bound

If $\{\theta_i\}$ remains bounded, there exist constants A and B such that

$$\sqrt{\mathbb{E}\left[\left|\Im(h) - \widehat{\Im}_N(h)\right|^2\right]} \leq \frac{A}{\sqrt{N}} + \frac{B}{N}\sum_{i=1}^{N} \gamma_i \varepsilon_i,$$

Thus if $\gamma_i \varepsilon_i = n^{-\alpha}$ for $\alpha \in [0,1]$, then, by Cesaro's,

$$\frac{\sum_{i=1}^{N} \gamma_i \varepsilon_i}{N} \overset{N \to +\infty}{\sim} N^{-\alpha},$$

the second term will asymptotically be negligible compared to the first term when $\alpha \in (1/2, 1]$

Convergence control for controlled algorithm

Since $\{\theta_i\}$ converges to the solution of

$$\nabla_\theta \left\{\Psi(\eta(\theta))\right\} = 0$$

there must be convergence of

$$\dot{\eta}_i \, \nabla \Psi(\eta_i)$$

to 0

[Convergence monitoring]

8.4 Illustrations

Example 52 —Coerced acceptance—

Imposed an expected acceptance probability $\alpha_\star = 0.4$ for a random walk MH with target

$$0.21\,\mathcal{N}(-5,1) + 0.79\,\mathcal{N}(5,2)$$

Results for $200,000$ iterations.

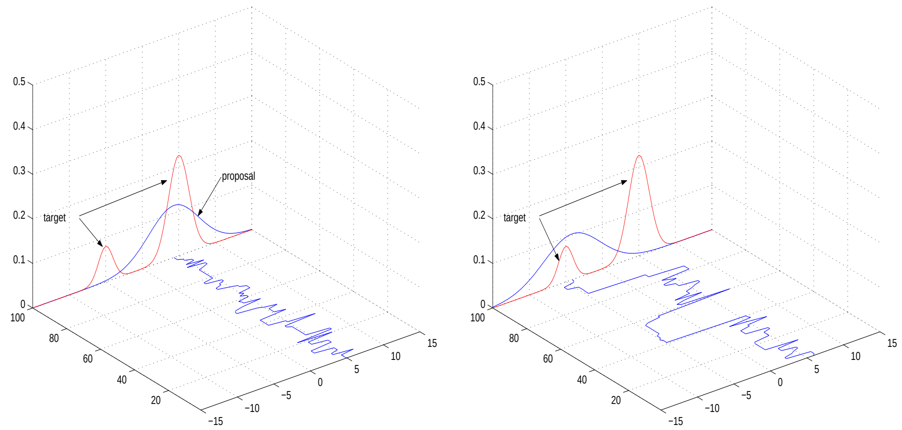


Figure 15: 3D rendering of the mixture target distribution and the proposal distribution for the random walk example.

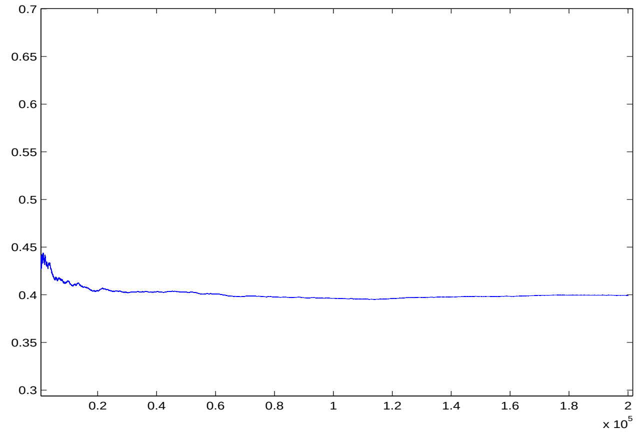


Figure 16: Convergence of the empirical acceptance probability for the bimodal distribution and the random walk proposal.

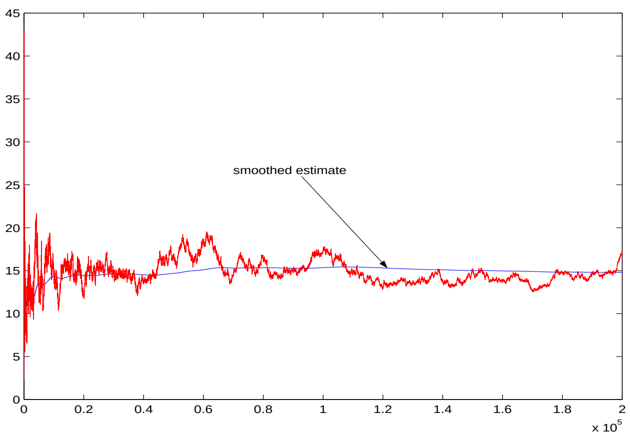


Figure 17: Convergence of the variance of the proposal distribution for the bimodal target distribution.

Example 53 **—Autocorrelation minimisation—**

Back to

$$\Sigma_{h,\tau}(\theta) = \mathrm{var}_f\left(h\left(x_0\right)\right) + 2\sum_{i=1}^{\tau} cov\left(h\left(x_0\right), h\left(x_i\right); \theta\right)$$

Need of a real transform, like

$$\begin{aligned} \Psi\left(\eta(\theta)\right) &= \mathrm{tr}\left(\Sigma_{h,\tau}(\theta)\,\Sigma_{h,\tau}^{\mathsf{T}}(\theta)\right) \\ &= \left\|\Sigma_{h,\tau}(\theta)\right\|^2 \end{aligned}$$

Then

$$h(\theta) = -\nabla_\theta \left\|\Sigma_{h,\tau}(\theta)\right\|^2$$

Recursion

$$\theta_{i+1} = \theta_i - \gamma_{i+1} \nabla_\theta \left\|\widehat{\Sigma_{h,\tau}(\theta)}\right\|^2,$$

where $\nabla_\theta \left\|\widehat{\Sigma_{h,\tau}(\theta)}\right\|^2$ 'unbiased' estimate of $\nabla_\theta \left\|\Sigma_{h,\tau}(\theta)\right\|^2$.

Example Optimize the covariance matrix Σ of a Gaussian random walk

$$q_\theta(x,y) = \left|\frac{\theta}{\sqrt{2\pi}}\right| \exp\left(\frac{-1}{2}(y-x)^{\mathrm{T}}\theta\theta^{\mathrm{T}}(y-x)\right)$$

Reparameterize as $\theta = \Sigma^{-1/2}$, lower triangular matrix such that

$$\Sigma^{-1} = \theta\theta^{\mathrm{T}}$$

Acceptance probability independent from θ,

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

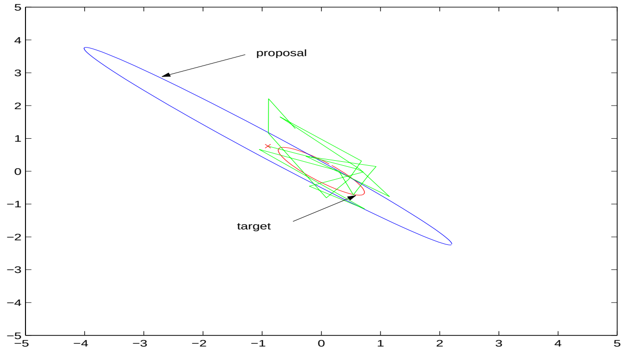


Figure 18: The target Gaussian distribution (red ellipse with center $(0,0)$). The Gaussian proposal distribution after $200,000$ iterations (blue).

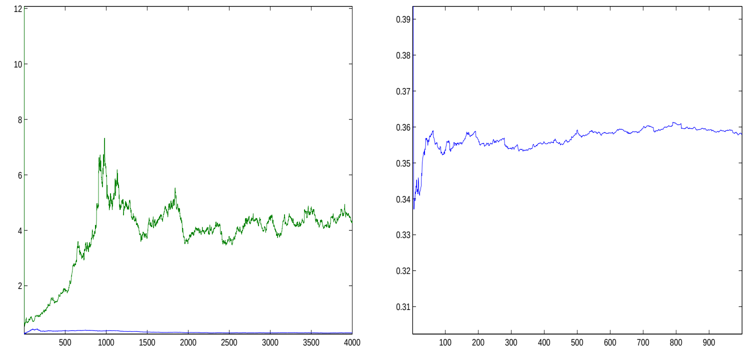


Figure 19: Convergence of parameters a and b of the bivariate Gaussian proposal distribution, subsampled ($1/50$).

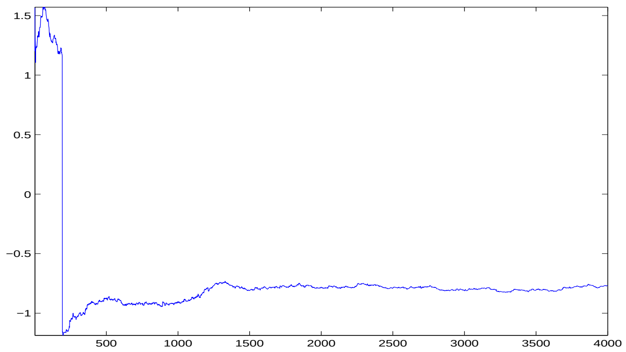


Figure 20: Convergence of parameter α of the bivariate Gaussian proposal distribution, subsampled $(1/50)$.

Example Optimize the weights of a mixture kernel

$$\begin{aligned}\mathfrak{K}_\theta(x,dy) &= \frac{1}{1+\sum_{j=2}^p\left(\theta_j^2+\varepsilon\right)}\mathfrak{K}_1(x,dy) \\ &\quad + \sum_{i=2}^p \frac{\theta_i^2+\varepsilon}{1+\sum_{j=2}^p\left(\theta_j^2+\varepsilon\right)}\mathfrak{K}_i(x,dy)\,.\end{aligned}$$

when the main direction of the Gaussian target is $\pi/4$.

Proposals are normal with orientations $\alpha = 0, \pm\pi/4, \pi/2$ and the same scale.

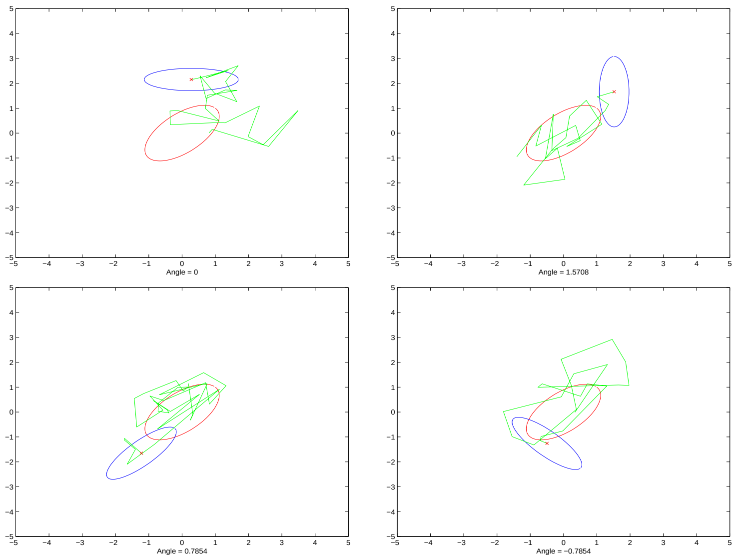


Figure 21: The target distribution and the four possible proposal densities for the mixture of strategies example, along with 50 steps of the corresponding Markov chain.

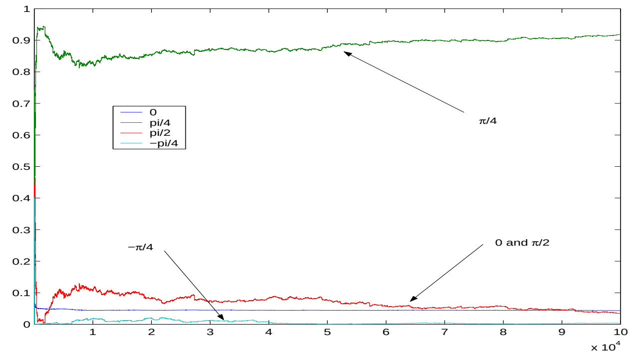


Figure 22: Evolution of the proportions of the mixture of strategies.

Example 54 **—Optimal blocking for SVM's—** Stochastic volatility model

$$\begin{aligned} y_t &= \beta \exp(x_t/2)\, \epsilon_t, \\ x_{t+1} &= \phi x_t + \eta_t \end{aligned}$$

[Shephard & Pitt, 1997]

Data $\{y_t\}$ and (unobserved) volatility $\{x_t\}$

MCMC

Hybrid Gibbs sampler

1. $\xi \triangleq (\phi, \beta, \sigma_\eta^2)|\{x_t\}$
2. $\{x_t\}|\xi$

First stage obvious **[Conjugacy]**
but problem with $\{x_t\}|\xi$

Block updating

Simulation based on a Gaussian approximation of the conditional distribution $f(\{x_t\}|\xi)$

Update of $\{x_t\}$ by blocks $\{x_t\}_{t_1 \le t \le t_2}$

Influence of the size $\mathfrak{s} = t_2 - t_1$ [of the blocks] on convergence performances

Proposal q_θ

1. Select center $c \sim \mathcal{U}(1, \ldots, N)$
2. Generate half-length ℓ from $(-L \le \ell \le L)$

$$p_\ell(\theta) = \frac{\exp(\frac{-1}{2\sigma^2}(\ell - \theta)^2)}{\sum_{m=-L}^{L} \exp(\frac{-1}{2\sigma^2}(m - \theta)^2)}$$

3. Define block as

$$B(c, \ell) \triangleq \{(c - |\ell|)\text{mod}N, \ldots, (c + |\ell|)\text{mod}N\}$$

4. Update $\{x_t : t \in B(c, \ell)\}$ conditional upon ξ and block $B^c(c, \ell)$ based on a Metropolis-Hastings transition $\mathfrak{K}_{c,\ell}$, with a normal proposal $q_{c,\ell}$

Possible repetition of updates before acceptance, leading to kernel

$$\mathfrak{K}(x, dx^*; \theta) = \int_{\mathcal{X}^{M-1}} \prod_{m=1}^{M} \sum_{c,\ell} \omega_{c,\ell}(\theta)\, \mathfrak{K}_{c,\ell}(z_{m-1}, dz_m),$$

where

$$\omega_{c,\ell}(\theta) = \frac{1}{N} p_\ell(\theta)$$

Choice of θ?

Criterion 1 : **Coerced acceptance**

Expected acceptance probability for updating block $B(c,\ell)$:

$$\alpha_{c,\ell} = \int_{\mathcal{X}^{2L+1}} 1 \wedge \frac{f(x^*|\xi) q_{c,\ell}(z_{c,\ell}|x^*)}{f(x|\xi) q_{c,\ell}(z^*_{c,\ell}|x)} q_{c,\ell}(z^*_{c,\ell}|x) f(x|\xi)\, dx\, dz^*_{c,\ell}.$$

where

$$\begin{aligned} x \cap x^* &= \{x_t : t \notin B(c,\ell)\} \\ z_{c,\ell} &= \{x_t : t \in B(c,\ell)\} \end{aligned}$$

Expected acceptance probability for updating one block:

$$\alpha(\theta) = \sum_{c=1}^{N} \sum_{\ell=-L}^{L} \omega_{c,\ell}(\theta)\, \alpha_{c,\ell}$$

Loss function

$$\Psi(\theta) = \frac{1}{2}(\alpha(\theta) - \alpha_*)^2$$

and

$$\begin{aligned} \frac{\partial \Psi(\theta)}{\partial \theta} &= (\alpha(\theta) - \alpha_*) \frac{\partial \alpha(\theta)}{\partial \theta} \\ \frac{\partial \alpha(\theta)}{\partial \theta} &= \sum_{c=1}^{N} \sum_{\ell=-L,}^{L} \frac{\partial \log \omega_{c,\ell}(\theta)}{\partial \theta} \omega_{c,\ell}(\theta)\, \alpha_{c,\ell} \\ \frac{\partial \log \omega_{c,\ell}(\theta)}{\partial \theta} &= \frac{\ell - \theta}{\sigma^2} - \sum_{m=-L}^{L} \frac{m - \theta}{\sigma^2} p_m(\theta) \end{aligned}$$

Stochastic approximation algorithm at iteration i

1. Sample $c \sim \mathcal{U}(1, \dots, N)$, $\ell \sim p_\ell(\theta_i)$ and $z^*_{c,\ell} \sim q_{c,\ell}(z^*_{c,\ell}|x)$
2. Compute

$$\varpi_{c,\ell}(x, x^*) = 1 \wedge \frac{f(x^*|\xi) q_{c,\ell}(z_{c,\ell}|x^*)}{f(x|\xi) q_{c,\ell}(z^*_{c,\ell}|x)}$$

3. Update

$$\begin{aligned} \eta_{i+1} &= (1-\gamma_{i+1})\eta_i + \gamma_{i+1}\,\varpi_{c,\ell}(x,x^*) \\ \dot\eta_{i+1} &= (1-\gamma_{i+1})\dot\eta_i + \gamma_{i+1}\frac{\partial \log \omega_{c,\ell}(\theta_i)}{\partial\theta}\,\varpi_{c,\ell}(x,x^*) \\ \theta_{i+1} &= \theta_i - \gamma_{i+1}\varepsilon_{i+1}\dot\eta_i(\eta_i - \alpha_*) \end{aligned}$$

4. Set x to x^* with probability $\varpi_{c,\ell}(x, x^*)$

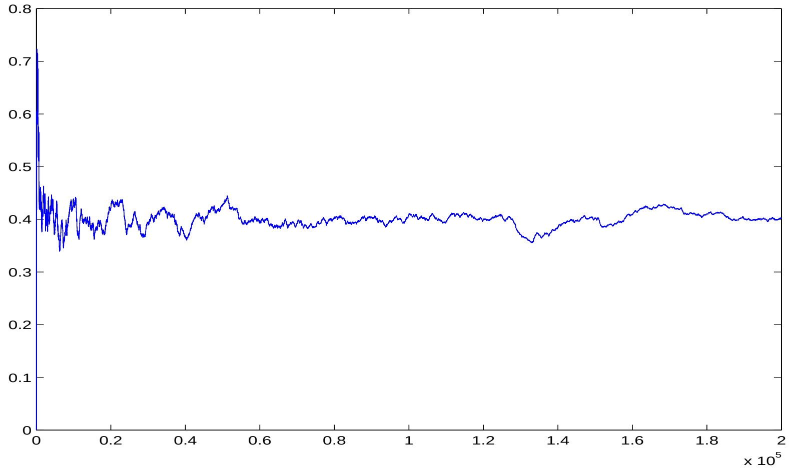


Figure 23: Convergence of the empirical acceptance probability to the coerced value $\alpha^\star = 0.4$.

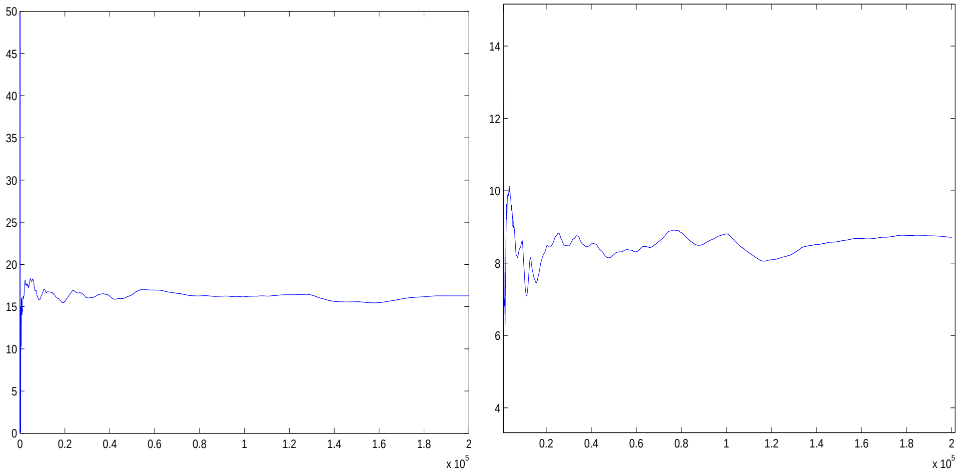


Figure 24: Evolution of θ under coerced value $\alpha^\star = 0.4$ (left) and $\alpha^\star = 0.6$ (right)

Criterion 2 : **Cumulative autocovariance**

Remember that

$$\sigma^2_\tau(\theta) = \mathrm{var}_f(h_0) + 2\sum_{k=1}^{\tau} \mathrm{cov}(h_0, h_k; \theta)$$

where $h_\ell \triangleq h(x_\ell, \xi_\ell)$

Estimation of the roots of the equation

$$\frac{\partial \sigma^2_\tau(\theta)}{\partial\theta} = 0$$

Autocovariance $cov\left(h_0, h_k; \theta\right)$ at lag k involves k iterations:

$$\sum_{\substack{c_1, c_2, \ldots, c_{kM} \\ \ell_1, \ell_2, \ldots, \ell_{kM}}} \int f\left(dx_0, d\xi_0\right) \prod_{p=1}^{k} \Bigg\{ f\left(d\xi_p | z_p\right) \times \prod_{q=1}^{M} \omega_{c_m, l_m}\left(\theta\right) \mathfrak{K}_{c_m, \ell_m}\left(z_{m-1}, dz_m\right) \Bigg\} h_0 h_k'$$

where $m = (p-1)M + q$, $z_0 = x_0$ and $x_k = z_{kM}$

Gradient

$$\sum_{\substack{c_1, \ldots, c_{kM} \\ \ell_1, \ldots, \ell_{kM}}} \int \sum_{n=1}^{kM} \frac{\partial \log \omega_{c_n, \ell_n}\left(\theta\right)}{\partial \theta} f\left(dx_0, d\xi_0\right) \prod_{p=1}^{k} \Bigg\{ f\left(d\xi_p | z_p\right) \times \prod_{q=1}^{M} \omega_{c_m, \ell_m}\left(\theta\right) \mathfrak{K}_{c_m, \ell_m}\left(z_{m-1}, dz_m\right) \Bigg\} h_0 h_k'$$

No need for the two-time scale, since gradient is directly available in integral form

Stochastic approximation algorithm at iteration i

1. Set $\omega = 0$
2. For $m = 1, \ldots, \tau$

   Update ξ

   For $n = (m-1)M + 1, \ldots, mM$

      Update $B\left(c_n, \ell_n\right)$ where $c_n \sim \mathcal{U}(1, N)$, $\ell_n \sim p_\ell\left(\theta_i\right)$

      $\omega \leftarrow \omega + \frac{\partial \log \omega_{c_n, \ell_n}(\theta)}{\partial \theta}|_{\theta_i}$

   $g_m \leftarrow \omega \times h_0 h_m'$
3. $\theta_{i+1} = \theta_i - \gamma_{i+1} \sum_{m=1}^{\tau} g_m$

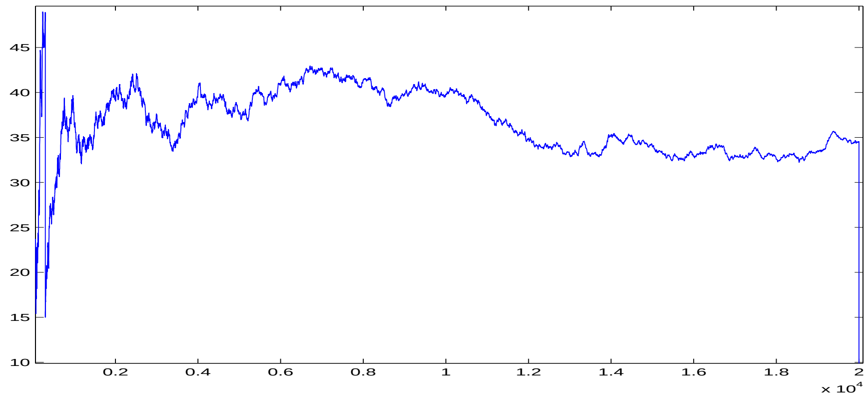


Figure 25: Convergence of θ for the autocovariance criterion for $\tau^2 = 25$ and $\mu = 5.$

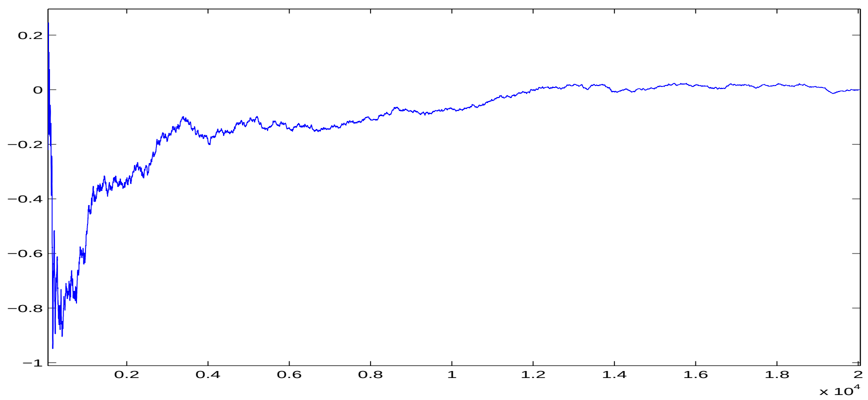


Figure 26: Convergence of the smoothed estimated gradient

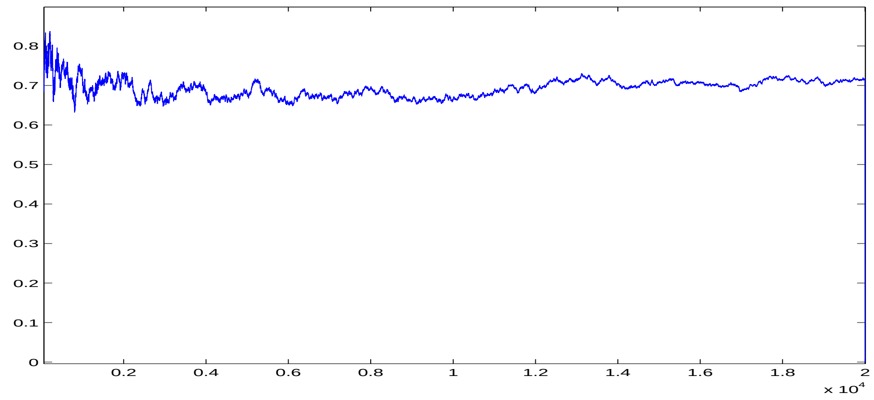


Figure 27: Convergence of the empirical acceptance probability for the autocovariance criterion for $\tau^2 = 25$ and $\mu = 5$.