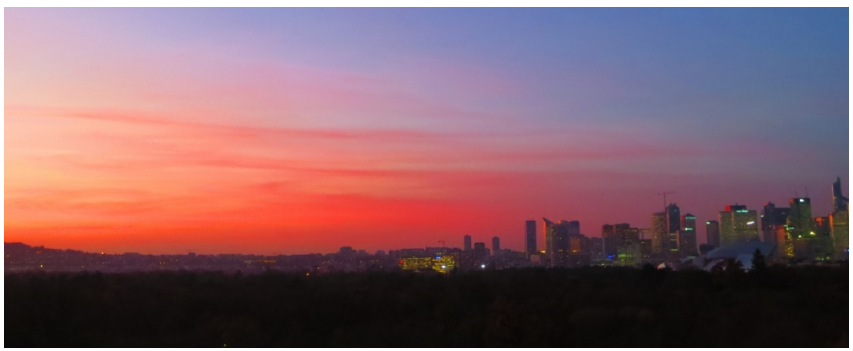
Markov Chain Monte Carlo Methods (introduction)

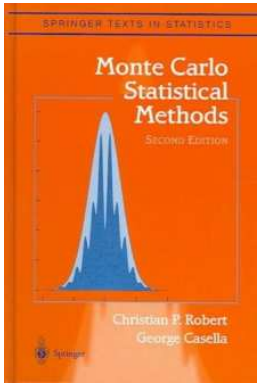
Christian P. Robert

Université Paris-Dauphine, University of Warwick, & CREST



École d'Hiver, Les Diablerets, CH, Feb 5-7 2023

Textbook: *Monte Carlo Statistical Methods* by Christian. P. Robert and George Casella

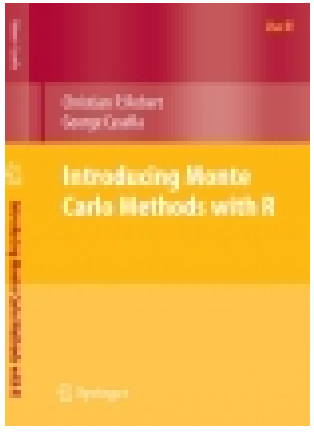


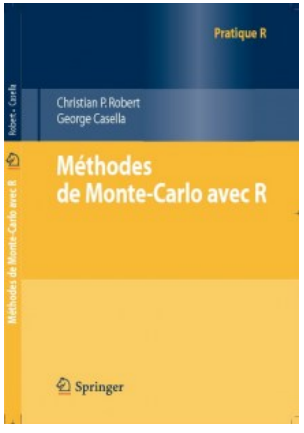
Slides: older slides on

`http://www.ceremade.dauphine.fr/~xian/coursBC.pdf`

Suggested reading

Introducing Monte Carlo Methods with R by Christian. P. Robert and George Casella [trad. française 2010; japonaise 2011]





Outline

The Metropolis-Hastings Algorithm

Running Monte Carlo via Markov Chains

It is not *necessary* to use a sample from the distribution f to approximate the integral

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

We can obtain $X_1, \ldots, X_n \sim f$ without directly simulating from f, **using an ergodic Markov chain with stationary distribution** f

[thanks to the Ergodic Theorem!]

Refresher

Theorem (Ergodic Theorem)

If a Markov chain (X_n) *is Harris positive recurrent, with stationary measure* π*, then for any function* h *with* $E|h| < \infty$*,*

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

Running Monte Carlo via Markov Chains (2)

Idea

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

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

Problem: How can one build a Markov chain with a given stationary distribution?

The Metropolis–Hastings algorithm

Basics

The algorithm uses the **objective (target) density**

$$f \qquad \text{up to a constant}$$

and a conditional density

$$q(y|x)$$

called the **instrumental (or proposal) distribution**

The Metropolis–Hastings algorithm

Algorithm (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

- Independent of normalizing constants for both f and $q(\cdot|x)$ (i.e., *constants* that do not depend on 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 when f is a density wrt Lebesgue measure
- The sequence $(y_t)_t$ is usually **not** a Markov chain

Convergence properties

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

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

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

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

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

Convergence properties (2)

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)
   (i) For h, with $\mathbb{E}_f|h(X)| < \infty$,
$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) = \int h(x)df(x) \qquad \text{a.e. } f.$$
   (ii) 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.

Metropolis algorithm version α

- Simulation method proposed by Metropolis *et al.* (1953)
- Starting from θ_0, ζ is generated from

$$\zeta \sim \text{Uniform in a neighborhood of } \theta_0.$$

- The new value of θ is generated as

$$\theta_1 = \begin{cases} \zeta & \text{with probability } \rho = \exp(\Delta h/T) \wedge 1 \\ \theta_0 & \text{with probability } 1-\rho, \end{cases}$$

  - $\Delta h = h(\zeta) - h(\theta_0)$
  - If $h(\zeta) \geq h(\theta_0)$, ζ is accepted
  - If $h(\zeta) < h(\theta_0)$, ζ may still be accepted
  - which allows escape from local maxima

Temperature decrease (simulated annealing)

Modify temperature T at each iteration, as in

1. Simulate ζ from an instrumental distribution with density $g(|\zeta - \theta_i|)$;
2. Accept $\theta_{i+1} = \zeta$ with probability

$$\rho_i = \exp\{\Delta h_i / T_i\} \wedge 1;$$

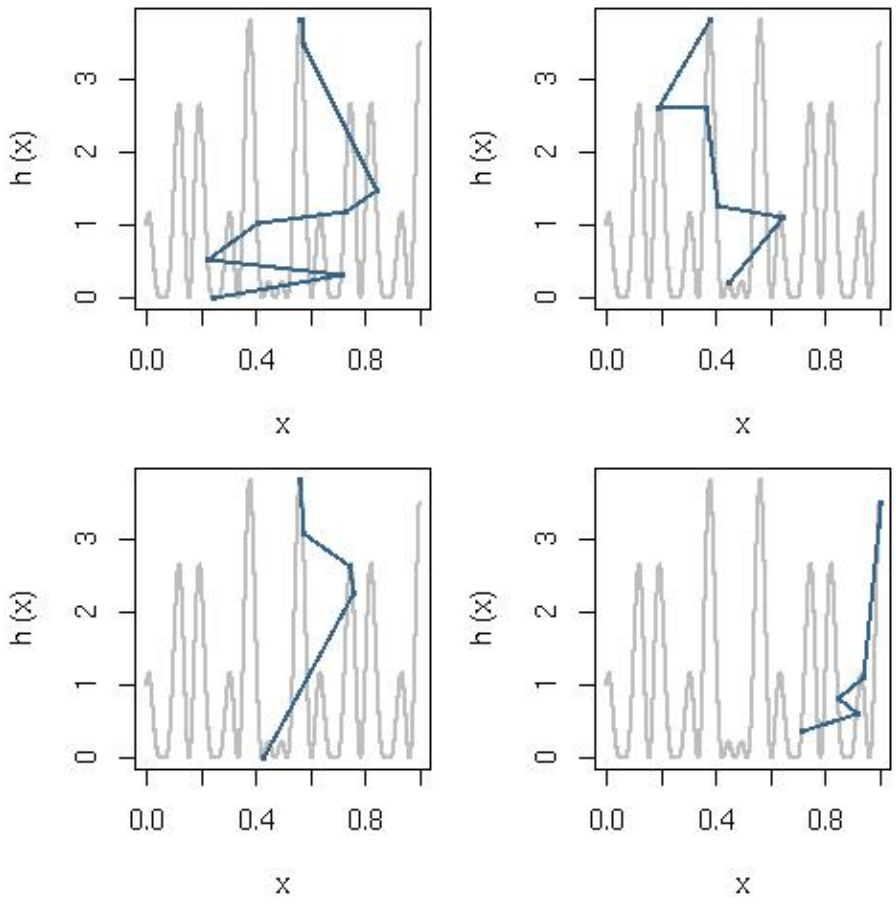
take $\theta_{i+1} = \theta_i$ otherwise.

3. Update T_i to $T_{i+1} \leq T_i$.

- All positive moves accepted
- As $T \downarrow 0$
  - Harder to accept downward moves
  - No big downward moves

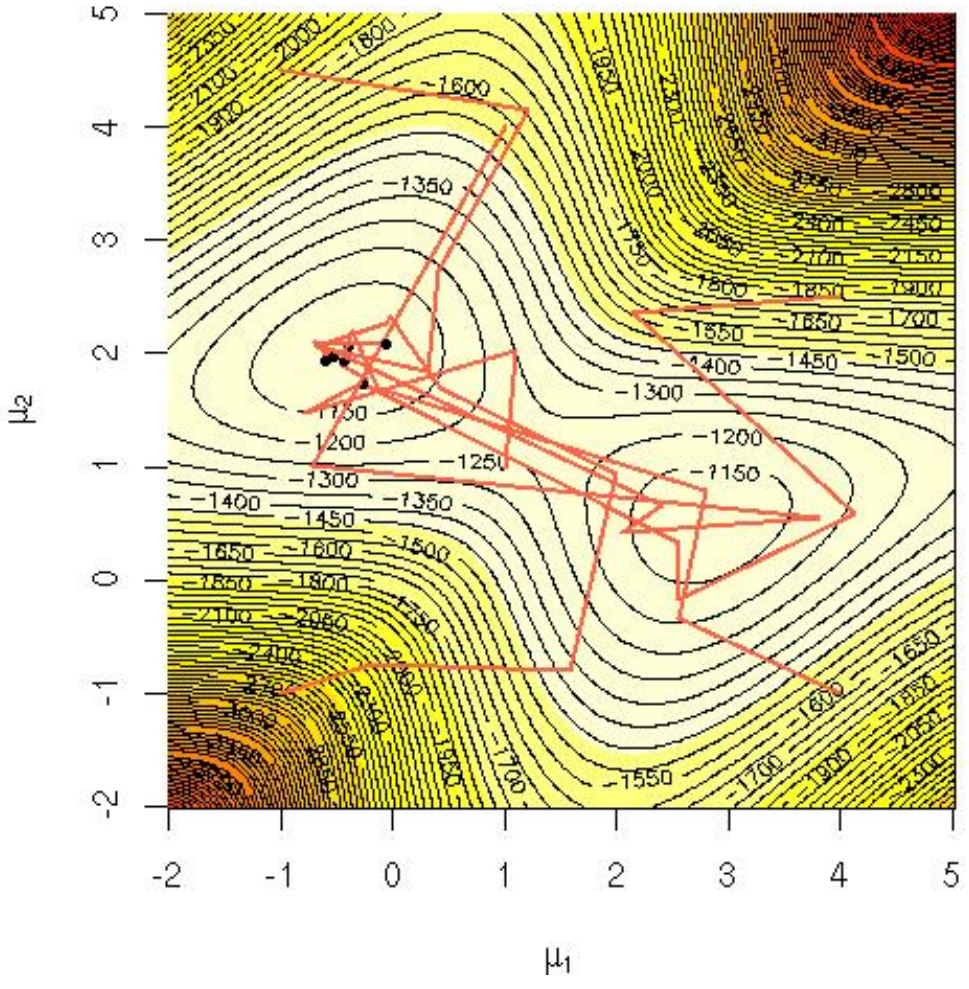
Not a time-homogeneous Markov Chain – more complex to analyze

Illustration



- Trajectory: $T_i = \frac{1}{(1+i)^2}$
- Log trajectory also works
- Can Guarantee Finding Global Max

Normal mixture



- Normal mixture with both means unknown
- Most sequences find max
- They visit both modes

1. The Independent Case

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

Algorithm (Independent Metropolis-Hastings)

Given $x^{(t)}$,

a Generate $Y_t \sim g(y)$

b Take

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

Properties

The resulting sample is **not** iid but there exist strong convergence properties:

Theorem (Ergodicity)

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

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

In this case,

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

[Mengersen & Tweedie, 1996]

Example (Noisy AR(1))

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

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

and observables

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

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

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

Example (Noisy AR(1) too)

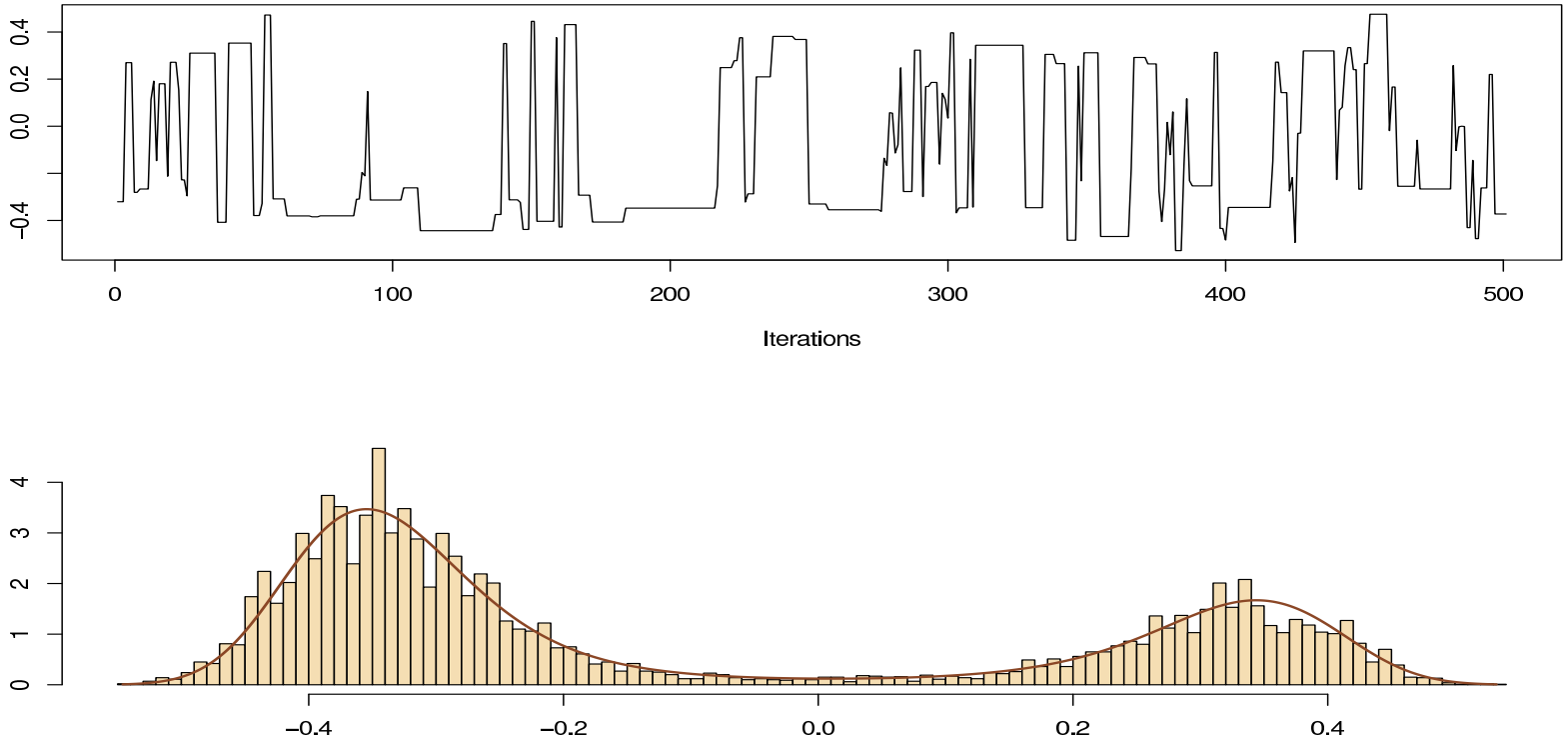
Use for proposal the $\mathcal{N}(\mu_t, \omega_t^2)$ distribution, with

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

Ratio

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

is bounded



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

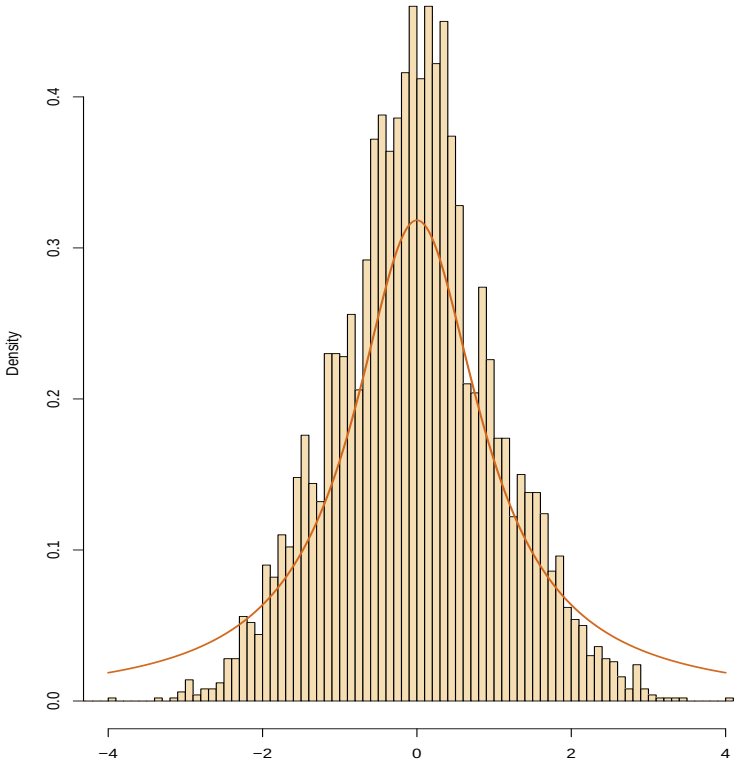
Example (Cauchy by normal)

go random W Given a Cauchy $\mathscr{C}(0,1)$ distribution, consider a normal $\mathscr{N}(0,1)$ proposal
The Metropolis–Hastings acceptance ratio is

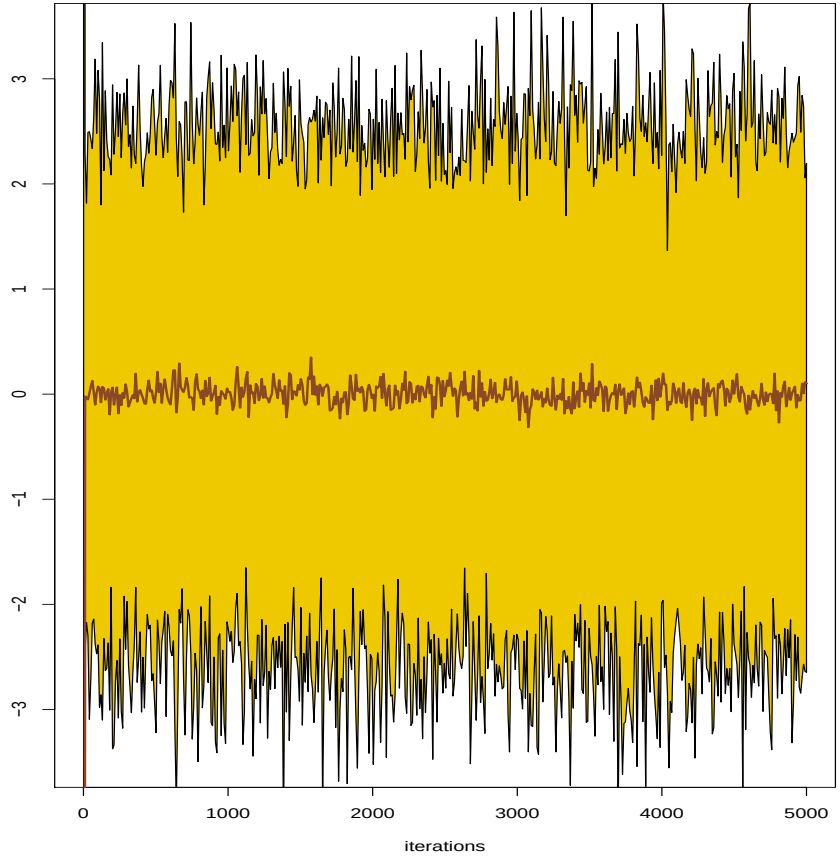
$$\frac{\pi(\xi')/\nu(\xi')}{\pi(\xi)/\nu(\xi))} = \exp\left[\left\{\xi^2 - (\xi')^2\right\}/2\right] \frac{1+(\xi')^2}{(1+\xi^2)}\,.$$

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

[Mengersen & Tweedie, 1996]



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



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

2. Random walk Metropolis–Hastings

Use of a local perturbation as 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* $g(x) = g(-x)$

Algorithm (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 (Random walk and normal target)

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

[Hastings (1970)]

The probability of acceptance is then

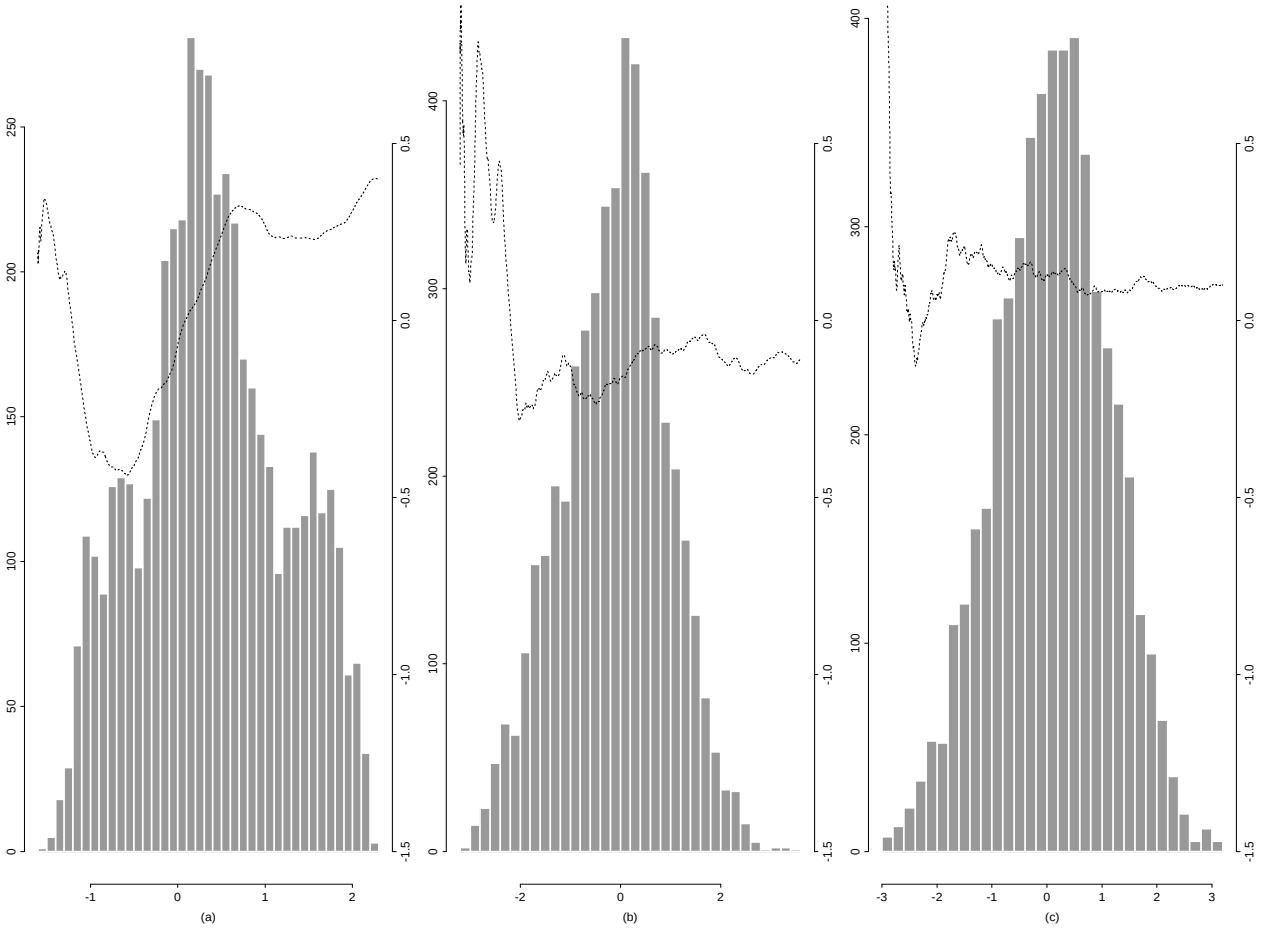
$$\rho(x^{(t)}, y_t) = \exp\{(x^{(t)^2} - y_t^2)/2\} \wedge 1.$$

Example (Random walk & normal (2))

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.



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).

Convergence properties

Uniform ergodicity prohibited by random walk structure
At best, geometric ergodicity:

Theorem (Sufficient ergodicity)

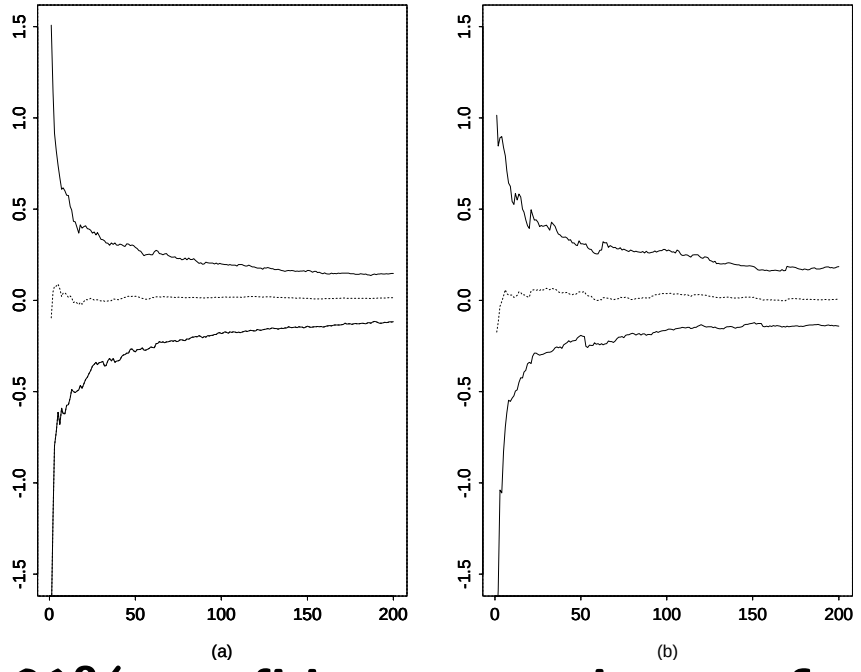
For a symmetric density f*, log-concave in the tails, and a positive and symmetric density* g*, the chain* $(X^{(t)})$ *is geometrically ergodic.*

[Mengersen & Tweedie, 1996]

▸ no tail effect

Example (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

Extensions

There are many other families of HM algorithms

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

to name just a few...

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*

Theorem

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

Discretization

Instead, consider the sequence

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

where σ^2 corresponds to the discretization step

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

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

[Cf unscented Langevin]

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

[Roberts & Rosenthal, 1998]

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 MH 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

Case of the independent MH algorithm (2)

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 (Inverse Gaussian distribution)

▸ no inverse

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

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

Example (Inverse Gaussian distribution (2))

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.

Case of the random walk (2)

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%.

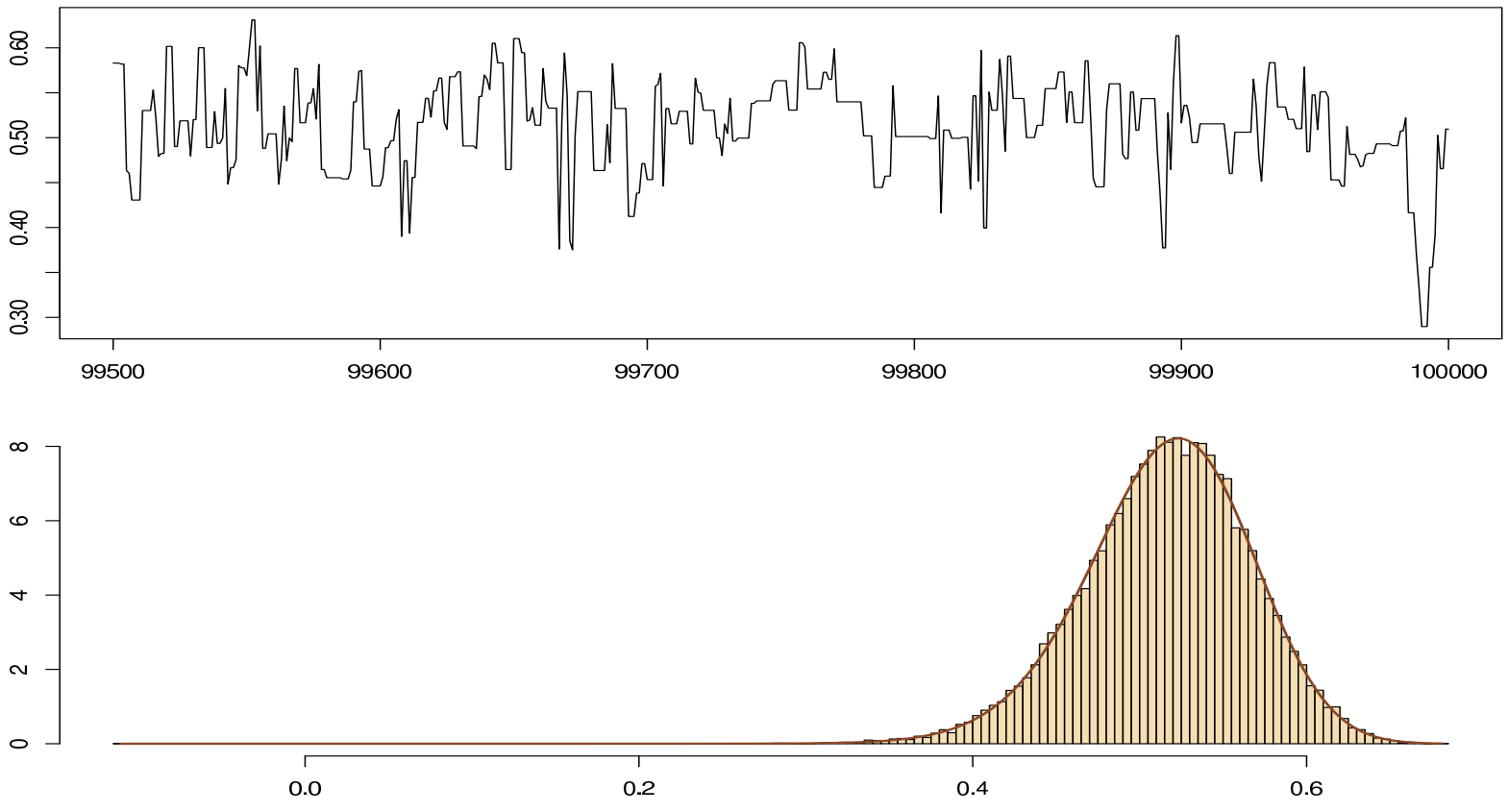
[Gelman,Gilks and Roberts, 1995]

Not highly constrictive since relies on formalised setting of limiting diffusion

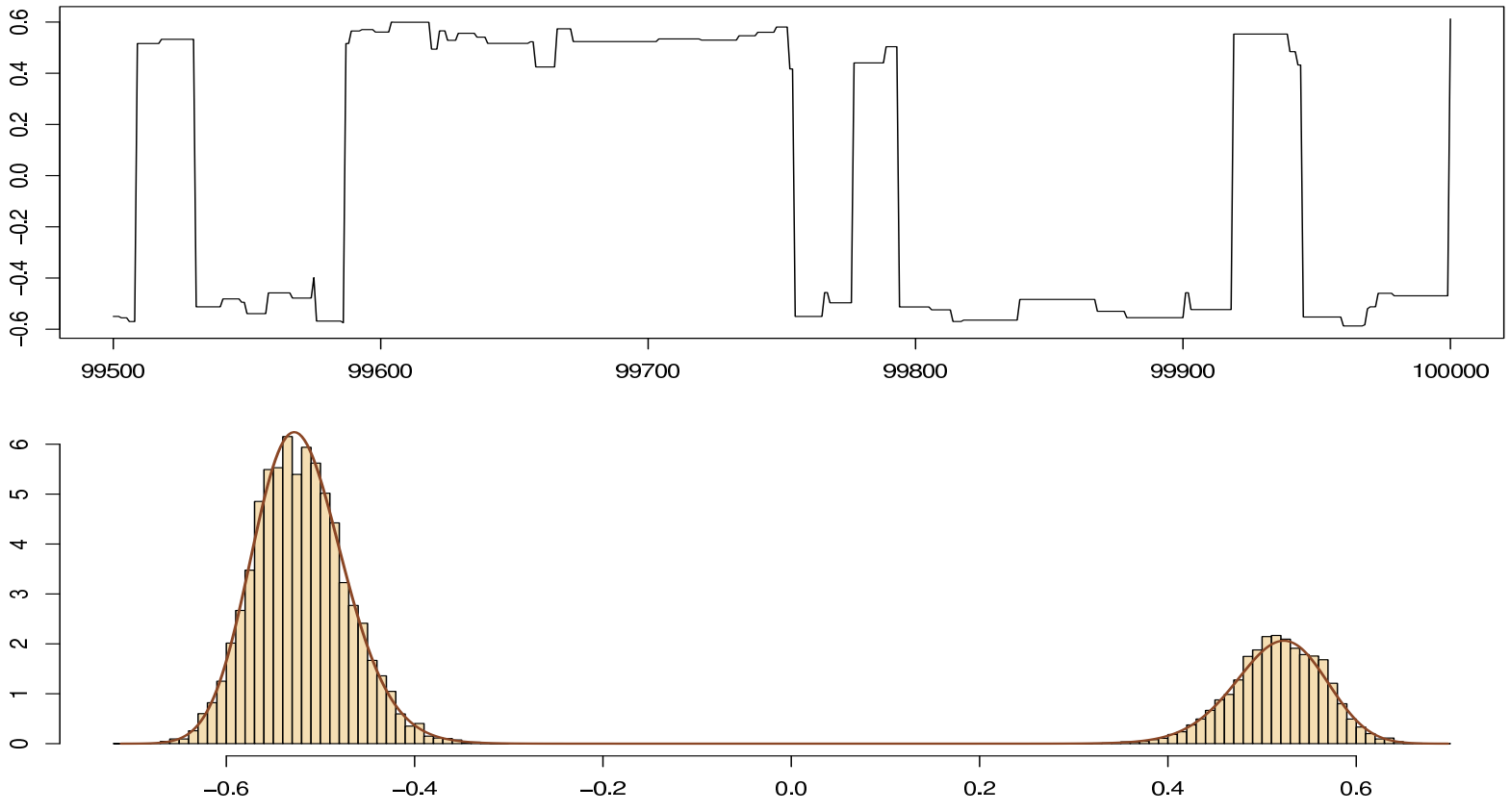
Impact of scale

Example (Noisy AR(1) continued)

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



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



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

1. Rao-Blackwellisation

Given a density $f(\cdot)$ to simulate take $g(\cdot)$ density such that

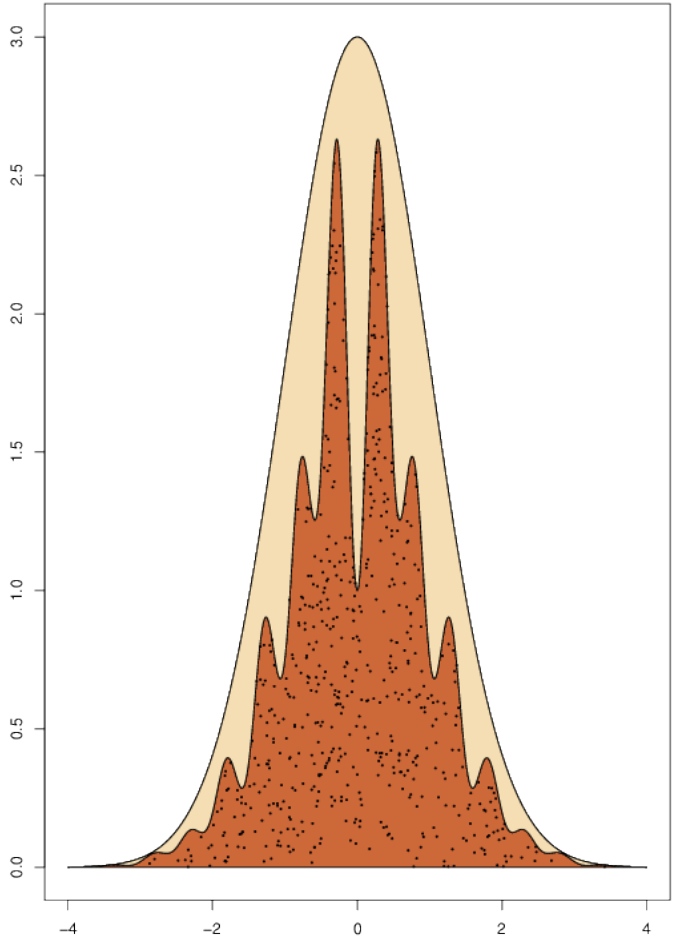
$$f(x) \leq Mg(x)$$

for $M \geq 1$
To simulate $X \sim f$, it is sufficient to generate

$$Y \sim g \; U|Y = y \sim \mathcal{U}(0, Mg(y))$$

until

$$0 < u < f(y)$$



Much ado about...

Raw outcome: id sequences $Y_1, Y_2, \ldots, Y_t \sim g$ and $U_1, U_2, \ldots, U_t \sim \mathcal{U}(0,1)$
Random number of accepted Y_i's

$$\mathbb{P}(N = n) = \binom{n-1}{t-1} (1/M)^t (1 - 1/M)^{n-t} ,$$

Joint density of $(N, \mathbf{Y}, \mathbf{U})$

$$\begin{aligned}
&\mathbb{P}(N = n, Y_1 \leq y_1, \ldots, Y_n \leq y_n, U_1 \leq u_1, \ldots, U_n \leq u_n) \\
&= \int_{-\infty}^{y_n} g(t_n)(u_n \wedge w_n) dt_n \int_{-\infty}^{y_1} \ldots \int_{-\infty}^{y_{n-1}} g(t_1) \ldots g(t_{n-1}) \\
&\quad \times \sum_{(i_1, \cdots, i_{t-1})} \prod_{j=1}^{t-1} (w_{i_j} \wedge u_{i_j}) \prod_{j=t}^{n-1} (u_{i_j} - w_{i_j})^+ dt_1 \cdots dt_{n-1},
\end{aligned}$$

where $w_i = f(y_i)/Mg(y_i)$ and sum over all subsets of

Much ado about noise

Accept-reject sample $(X_1, \ldots, X_m)$ associated with $(U_1, \ldots, U_N)$ and $(Y_1, \ldots, Y_N)$

N is stopping time for acceptance of m variables among Y_j's

Rewrite estimator of $\mathbb{E}[h]$ as

$$\frac{1}{m} \sum_{i=1}^{m} h(X_i) = \frac{1}{m} \sum_{j=1}^{N} h(Y_j)\, \mathbb{I}_{U_j \leq w_j}\,,$$

with $w_j = f(Y_j)/Mg(Y_j)$

Rao-Blackwellisation: smaller term-wise variance when integrating out the U_i's,

$$\frac{1}{m} \sum_{j=1}^{N} \mathbb{E}[\mathbb{I}_{U_j \leq w_j} | N, Y_1, \ldots, Y_N]\, h(Y_j) = \frac{1}{m} \sum_{i=1}^{N} \rho_i h(Y_i),$$

where $(i < n)$

$$\rho_i = \mathbb{P}(U_i < w_i | N = n, Y_1, \ldots, Y_n)$$

extension to Metropolis–Hastings case

Sample produced by Metropolis–Hastings algorithm

$$x^{(1)}, \ldots, x^{(T)}$$

based on two samples,

$$y_1, \ldots, y_T \quad \text{and} \quad u_1, \ldots, u_T$$

Ergodic mean rewritten as

$$\delta^{MH} = \frac{1}{T} \sum_{t=1}^{T} h(x^{(t)}) = \frac{1}{T} \sum_{t=1}^{T} h(y_t) \sum_{i=t}^{T} \mathbb{I}_{x^{(i)}=y_t}$$

Conditional expectation

$$\delta^{RB} = \frac{1}{T} \sum_{t=1}^{T} h(y_t)\, \mathbb{E}\left[\sum_{i=t}^{T} \mathbb{I} X^{(i)} = y_t \middle| y_1, \ldots, y_T \right]$$

$$= \frac{1}{T} \sum_{t=1}^{T} h(y_t) \left(\sum_{i=t}^{T} \mathbb{P}(X^{(i)} = y_t | y_1, \ldots, y_T) \right)$$

weight derivation

Take

$$\rho_{ij} = \frac{f(y_j)/q(y_j|y_i)}{f(y_i)/q(y_i|y_j)} \wedge 1 \qquad (j > i),$$

$$\overline{\rho}_{ij} = \rho_{ij} q(y_{j+1}|y_j), \quad \underline{\rho}_{ij} = (1 - \rho_{ij}) q(y_{j+1}|y_i) \qquad (i < j < T),$$

$$\zeta_{jj} = 1, \quad \zeta_{jt} = \prod_{l=j+1}^{t} \underline{\rho}_{jl} \qquad (i < j < T),$$

$$\tau_0 = 1, \quad \tau_j = \sum_{t=0}^{j-1} \tau_t \zeta_{t(j-1)}\, \overline{\rho}_{tj}, \quad \tau_T = \sum_{t=0}^{T-1} \tau_t \zeta_{t(T-1)} \rho_{tT} \qquad (i < T),$$

$$\omega_T^i = 1, \quad \omega_i^j = \overline{\rho}_{ji} \omega_{i+1}^i + \underline{\rho}_{ji} \omega_{i+1}^j \qquad (0 \le j < i < T).$$

Theorem

The estimator δ^{RB} *satisfies*

2. Another Rao–Blackwellisation

Alternative representation of Metropolis–Hastings estimator δ as

$$\delta = \frac{1}{n}\sum_{t=1}^{n} h(x^{(t)}) = \frac{1}{n}\sum_{i=1}^{M_n} n_i h(z_i)\,,$$

where rv's defined as

- z_i's are the accepted y_j's,
- M_n is the number of accepted y_j's till time n,
- n_i is the number of times z_i appears in the sequence $(x^{(t)})_t$.

The "accepted candidates"

Define

$$\tilde{q}(\cdot|z_i) = \frac{\alpha(z_i,\cdot)\, q(\cdot|z_i)}{p(z_i)} \leq \frac{q(\cdot|z_i)}{p(z_i)}$$

where $p(z_i) = \int \alpha(z_i, y)\, q(y|z_i) dy$

To simulate from $\tilde{q}(\cdot|z_i)$

1. Propose a candidate $y \sim q(\cdot|z_i)$
2. Accept with probability

$$\tilde{q}(y|z_i) \Big/ \left(\frac{q(y|z_i)}{p(z_i)}\right) = \alpha(z_i, y)$$

Otherwise, reject it and starts again.

▶ this is the transition of the HM algorithm The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \underbrace{\frac{\pi(x)p(x)}{\int \pi(u)p(u)du}}\ \underbrace{\frac{\alpha(x,y)q(y|x)}{p(x)}}\ \frac{\pi(x)\alpha(x,y)q(y|x)}{\int \pi(u)p(u)du}\ \frac{\pi(y)\alpha(y,$$

The "accepted chain"

Lemma (Douc & X., *AoS*, 2011)

The sequence (z_i, n_i) *satisfies*

1. $(z_i, n_i)_i$ *is a Markov chain;*
2. z_{i+1} *and* n_i *are independent given* z_i*;*
3. n_i *is distributed as a geometric random variable with probability parameter*

$$p(z_i) := \int \alpha(z_i, y)\, q(y|z_i)\, dy\,; \tag{1}$$

4. $(z_i)_i$ *is a Markov chain with transition kernel* $\tilde{Q}(z, dy) = \tilde{q}(y|z)dy$ *and stationary distribution* $\tilde{\pi}$ *such that*

$$\tilde{q}(\cdot|z) \propto \alpha(z, \cdot)\, q(\cdot|z) \quad \text{and} \quad \tilde{\pi}(\cdot) \propto \pi(\cdot)p(\cdot)\,.$$

Importance sampling perspective

1. A natural idea:
$$\delta^* = \frac{1}{n} \sum_{i=1}^{M_n} \frac{h(z_i)}{p(z_i)},$$
2. A natural idea:
$$\delta^* \simeq \frac{\sum_{i=1}^{M_n} \frac{h(z_i)}{p(z_i)}}{\sum_{i=1}^{M_n} \frac{1}{p(z_i)}} = \frac{\sum_{i=1}^{M_n} \frac{\pi(z_i)}{\tilde{\pi}(z_i)} h(z_i)}{\sum_{i=1}^{M_n} \frac{\pi(z_i)}{\tilde{\pi}(z_i)}}.$$
3. But p not available in closed form.
4. The geometric n_i is the replacement, an obvious solution that is used in the original Metropolis–Hastings estimate since $\mathbb{E}[n_i] = 1/p(z_i)$.

The Bernoulli factory

The crude estimate of $1/p(z_i)$,

$$n_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \mathbb{I}\{u_\ell \ge \alpha(z_i, y_\ell)\}\,,$$

can be improved:

Lemma (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|z_i)$, the quantity

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \{1 - \alpha(z_i, y_\ell)\}$$

is an unbiased estimator of $1/p(z_i)$ which variance, conditional on z_i, is lower than the conditional variance of n_i, $\{1 - p(z_i)\}/p^2(z_i)$.

Rao-Blackwellised, for sure?

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \{1 - \alpha(z_i, y_\ell)\}$$

1. Infinite sum but finite with at least positive probability:

$$\alpha(x^{(t)}, y_t) = \min \left\{ 1, \frac{\pi(y_t)}{\pi(x^{(t)})} \frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})} \right\}$$

   For example: take a symmetric random walk as a proposal.

2. What if we wish to be sure that the sum is finite?

Finite horizon k version:

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \le \ell \le k \wedge j} \{1 - \alpha(z_i, y_j)\} \prod_{k+1 \le \ell \le j} \mathbb{I}\{u_\ell \ge \alpha(z_i, y_\ell)\}$$

Variance improvement

Theorem (Douc & X., *AoS*, 2011)

If $(y_j)_j$ *is an iid sequence with distribution* $q(y|z_i)$ *and* $(u_j)_j$ *is an iid uniform sequence, for any* $k \geq 0$*, the quantity*

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \leq \ell \leq k \wedge j} \{1 - \alpha(z_i, y_j)\} \prod_{k+1 \leq \ell \leq j} \mathbb{I}\{u_\ell \geq \alpha(z_i, y_\ell)\}$$

is an unbiased estimator of $1/p(z_i)$ *with an almost sure finite number of terms. Moreover, for* $k \geq 1$*,*

$$\mathbb{V}\hat{\xi}_i^k z_i = \frac{1 - p(z_i)}{p^2(z_i)} - \frac{1 - (1 - 2p(z_i) + r(z_i))^k}{2p(z_i) - r(z_i)} \left(\frac{2 - p(z_i)}{p^2(z_i)} \right) (p(z_i) - r(z_i)) ,$$

where $p(z_i) := \int \alpha(z_i, y)\, q(y|z_i)\, dy$*. and* $r(z_i) := \int \alpha^2(z_i, y)\, q(y|z_i)\, dy$*. Therefore, we have*

$$\mathbb{V}\hat{\xi}_i z_i \leq \mathbb{V}\hat{\xi}_i^k z_i \leq \mathbb{V}\hat{\xi}_i^0 z_i = \mathbb{V} n_i z_i .$$

3. Delayed acceptance

Motivation: Non-informative inference for mixture models

Standard mixture of distributions model

$$\sum_{i=1}^{k} w_i \, f(x|\theta_i)\,, \quad \text{with} \quad \sum_{i=1}^{k} w_i = 1\,. \tag{1}$$

[Titterington et al., 1985; Fruhwirth, 2006]

Jeffreys' prior for mixture not available due to computational reasons : it has not been tested so far

[Jeffreys, 1939]

Warning: Jeffreys' prior improper in some settings

[Grazian & Robert, 2015]

Grazian & Robert (2015) consider genuine Jeffreys' prior for complete set of parameters in (1), deduced from Fisher's information matrix

Computation of prior density costly, relying on many integrals like

The "Big Data" plague

Simulation from posterior distribution with large sample size n

- Computing time at least of order $O(n)$
- solutions using likelihood decomposition

$$\prod_{i=1}^{n} \ell(\theta|x_i)$$

and handling subsets on different processors (CPU), graphical units (GPU), or computers

[Scott et al., 2013, Korattikara et al., 2013]

- no consensus on method of choice, with instabilities from removing most prior input and uncalibrated approximations

[Neiswanger et al., 2013]

Proposed solution

"There is no problem an absence of decision cannot solve."
Anonymous

Given $\alpha(x, y) := 1 \wedge r(x, y)$, factorise

$$r(x, y) = \prod_{k=1}^{d} \rho_k(x, y)$$

under constraint $\rho_k(x, y) = \rho_k(y, x)^{-1}$
Delayed Acceptance Markov kernel given by

$$\tilde{P}(x, A) := \int_A q(x, y)\tilde{\alpha}(x, y)dy + \left(1 - \int_X q(x, y)\tilde{\alpha}(x, y)dy\right) \mathbf{1}_A(x)$$

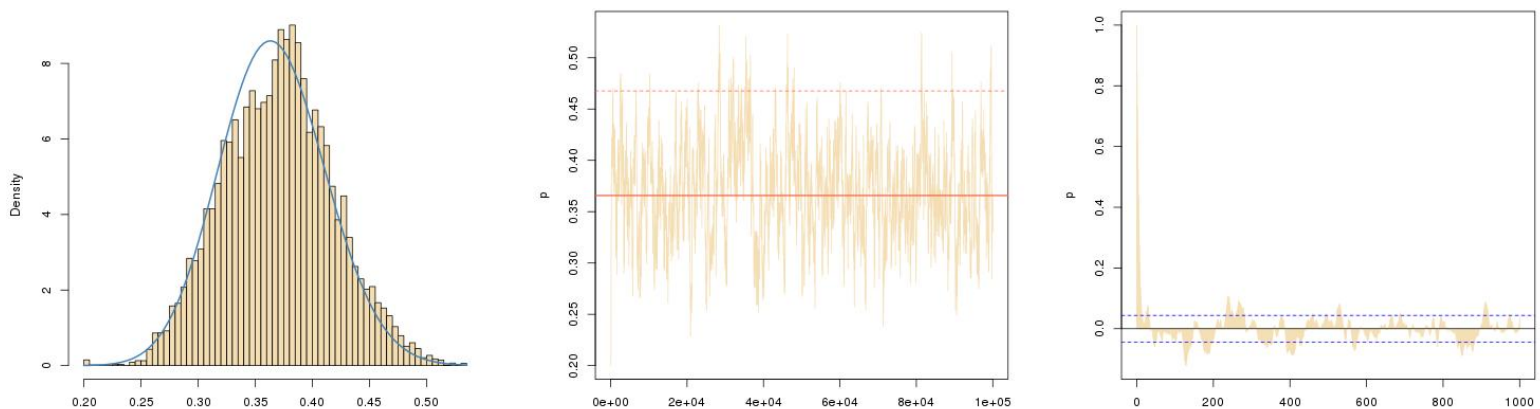
where

$$\tilde{\alpha}(x, y) := \prod_{k=1}^{d}\{1 \wedge \rho_k(x, y)\}.$$

Algorithm 1 Delayed Acceptance

Potential drawbacks

- Delayed Acceptance *efficiently* reduces computing cost only when approximation $\tilde{\pi}$ is "good enough" or "flat enough"
- Probability of acceptance always smaller than in the original Metropolis–Hastings scheme
- Decomposition of original data in likelihood bits may however lead to deterioration of algorithmic properties without impacting computational efficiency...
- ...e.g., case of a term explosive in $x = 0$ and computed by itself: leaving $x = 0$ near impossible



The "Big Data" plague

Delayed Acceptance intended for likelihoods or priors, but not a clear solution for "Big Data" problems

1. all product terms must be computed
2. all terms previously computed either stored for future comparison or recomputed
3. sequential approach limits parallel gains...
4. ...unless prefetching scheme added to delays

[Strid (2010)]

Validation of the method

Lemma (1)

For any Markov chain with transition kernel Π *of the form*

$$\Pi(x, A) = \int_A q(x, y)a(x, y)dy + \left(1 - \int_X q(x, y)a(x, y)dy\right) \mathbf{1}_A(x),$$

and satisfying detailed balance, the function $a(\cdot)$ *satisfies (for* π*-a.e.* x, y*)*

$$\frac{a(x, y)}{a(y, x)} = r(x, y).$$

Lemma

$(\tilde{X}_n)_{n\geq 1}$*, the Markov chain associated with* $\tilde{P}$*, is a* π*-reversible Markov chain.*

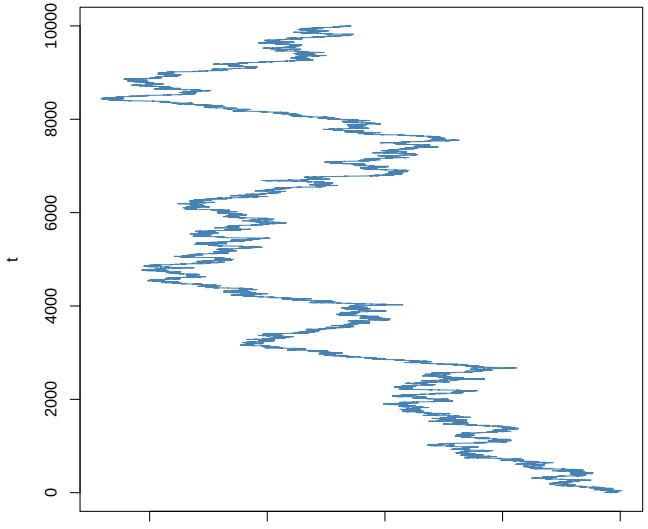
4. Folding MCMC

Motivating example: Consider the target

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

standard Cauchy distribution
Basic Metropolis-Hastings algorithm with uniform proposal
$z_t \sim \mathcal{U}(x_t - \epsilon, x_t + \epsilon)$ cannot be geometrically ergodic

[Mengersen and Tweedie (1996)]



Dynamics of a standard random-walk Metropolis–Hastings algorithm when targeting a Cauchy distribution, based on 10^4 iterations and a uniform scale of $\varepsilon = .1$.

new proposal

Metropolis-Hastings alternative:

1. the current value x_t of the Markov chain is first inverted into $y_t = 1/x_t$ if found outside $(-1, 1)$,
2. then moved by a random walk on $(-1, 1)$ to $z_t \sim \mathcal{U}(y_t - \epsilon, y_t + \epsilon)$, which value is accepted or not according to the standard Metropolis-Hastings ratio,
3. and outcome inverted into $x_{t+1} = 1/y_{t+1}$ with probability $1/2$

simple version of the folding algorithm, with folding set the unit interval $(-1, 1)$

validation

simple version of the folding algorithm, with folding set the unit interval $(-1, 1)$

validation

simple version of the folding algorithm, with folding set the unit interval $(-1, 1)$

- Cauchy target still stationary for this distribution
- probability $1/2$ resulting from Jacobian rather than from $\mathbb{P}(|X| < 1) = 1/2$
- not-so-simple [but still-manageable] probabilty if chosing folding interval $(-2, 2)$ and inversion $y_t = 4/x_t$
- fundamental reason is that Cauchy is invariant by inversion
- resulting Markov chain is uniformly ergodic

simulation outcome

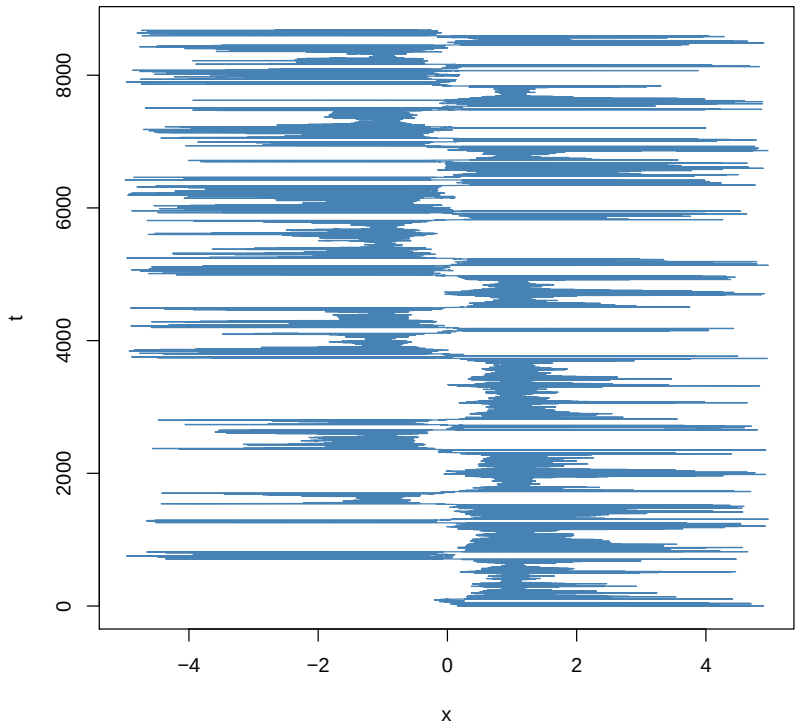


Figure: Folded Markov chain for Cauchy target with same scale of the random walk

folding the Markov chain

Consider target π on state space X
Let $A_0, A_1, \ldots, A_M$ be a finite partition of the state space and create differentiable bijections $g_1, \ldots, g_M$ from A_0 to $A_1, \ldots, A_M$, respectively. Set $X^\star = A_0$ as the folded space
Define the distribution

$$\pi^\star(x^\star) = \pi(x^\star) + \pi(g_1 x^\star)\,|\partial_x g_1\,(x^\star)| + \ldots + \pi(g_M x^\star)\,|\partial_x g_M\,(x^\star)|$$

on $X^\star$

© $\pi^\star(\cdot)$ is a proper density on $X^\star$

unfolding the folded Markov chain

Simulating from $\pi^\star$ is equivalent to simulating from π:

Lemma

If $x^\star \sim \pi^\star$, *then*

$$x = \begin{cases} x^\star & \text{with probability } \pi(x^\star)/\pi^\star(x^\star) \\ g_1 x^\star & \text{with probability } \pi(g_1 x^\star)\,|\partial_x g_1\,(x^\star)|\,/\pi^\star(x^\star) \\ \qquad \ldots & \\ g_M x^\star & \text{with probability } \pi(g_M x^\star)\,|\partial_x g_M\,(x^\star)|\,/\pi^\star(x^\star) \end{cases}$$

is distributed from the target π.

© build MCMC sampler aiming at $\pi^\star$

Cauchy example validated

For the Cauchy example:

- $A_0 = (-1, 1)$, $A_1 = (-1, 1)^c$, $g_1 x^\star = 1/x^\star$
- and

$$\begin{aligned}\pi^\star(x) &= \pi(x^\star) + \pi(g_1 x^\star)|\partial_x g_1(x^\star)| \\ &= \frac{1}{(1+x^2)\pi} + \frac{1}{(1+1/x^2)\pi}\frac{1}{x^2} \\ &= \frac{2}{(1+x^2)\pi}\end{aligned}$$

- unfolding by $x = \begin{cases} x^\star & \text{w.p. } 1/2 \\ 1/x^\star & \text{w.p. } 1/2 \end{cases}$

For the alternative

- $A_0 = (-2, 2)$, $A_1 = (-2, 2)^c$, $g_1 x^\star = 4/x^\star$
- and

$$\pi^\star(x) = \pi(x^\star) + \pi(g_1 x^\star)|\partial_x g_1(x^\star)|$$

folding set

Unless target distribution simple enough for informed choice, natural choice for A_0 is HPD region

$$H_\alpha = \{x \in X;\ \pi(x) \geq \alpha\}$$

as

- $\pi^\star$ [and hence π] lower bounded on H_α
- resulting H_α compact
- some transition kernels produce uniform ergodic chains
- partition of X into A_0, A_0^c with natural stereoscopic projection [provided A_0 star-convex]

$$g_1(x^\star) = \frac{\rho^2}{|x^\star|^2} x^\star$$

practical implementation

While H_α usually unavailable, approximations can be found from preliminary MCMC runs when $\pi(x)$ or unnormalised version of it can be computed

- preliminary run produces simulations with [relative] values of π, $\pi(x^1), \ldots, \pi(x^N)$
- derivation of higher density values [and potential clustering]
- choice of an HPD approximation as ball and g_1 as natural projection
- reevaluation of the folding set after further simulations

Note: black box compatibility with MCMC code

5. Pseudo-marginal version

Many settings where numerically computing target density $\pi(\cdot)$ is impossible, even up a normalising constant
Example of doubly intractable likelihoods, when likelihood function contains intractable non-constant term

$$\ell(\theta|x) \propto g(x|\theta)$$

and intractable normalising constant

$$\mathfrak{Z}(\theta) = \int_{\mathcal{X}} g(x|\theta)\,dx$$

See for instance Ising model

pseudo-marginal extension

Approach based on unbiased estimator of $\pi(\cdot|x)$ and retaining Metropolis–Hastings validity

If $\hat{\pi}(\theta|z)$ is unbiased estimator of $\pi(\theta)$ when $z \sim q(\cdot|\theta)$

$$\int_{\mathcal{Z}} \overbrace{\hat{\pi}(\theta|z)q(\cdot|\theta)}^{\text{same } \theta} \, dz = \pi(\theta)$$

then acceptance ratio

$$\frac{\hat{\pi}(\theta^*|z^*)q(z^*|\theta^*)}{\hat{\pi}(\theta|z)q(z|\theta)} \, \frac{q(\theta^*,\theta)q(z|\theta)}{q(\theta,\theta^*)q(z^*|\theta^*)}$$

© preserves stationarity wrt extended target

Reason: auxiliary variable z makes simulation of joint (θ, z) a regular Metropolis-Hastings move

[Beaumont & al, 2003; Andrieu & Roberts, 2009]

Performances depend on quality of estimators $\hat{\pi}$ but always poorer

Alternative explanation

Take importance weight

$$w = \hat{\pi}(\theta|z)/\pi(\theta)$$

as auxiliary variable with constant conditional expectation c and distribution $p(w|\theta)$
Corresponding joint proposal $q(\theta, \theta^*)p(w^*|\theta^*)$ and associated acceptance proposal

$$\frac{w^*\pi(\theta^*)p(w^*|\theta^*) \times q(\theta^*, \theta)p(w|\theta)}{w\pi(\theta)p(w|\theta) \times q(\theta, \theta^*)p(w^*|\theta^*)}$$

leads to joint target (proportional to)

$$\pi(\theta)\, w\, p(w|x)$$

with marginal $\pi(\theta)$

[Andrieu & Roberts, 2009; Wilkinson, 2010]

Illustration: particle MCMC

Hidden Markov model, where latent Markov chain $x_{0:T}$ with density

$$p_0(x_0|\theta)p_1(x_1|x_0,\theta)\cdots p_T(x_T|x_{T-1},\theta)\,,$$

associated with observed sequence $y_{1:T}$ such that

$$y_{1:T}|x_{1:T},\theta \sim \prod_{i=1}^{T} q_i(y_i|x_i,\theta)\,,$$

pMCMC

At iteration t

- propose value $\theta' \sim \mathfrak{h}(\theta|\theta^{(t)})$
- propose value of latent series $x'_{0:T}$ via particle filter approximation of $p(x_{0:T}|\theta', y_{1:T})$
- derive unbiased estimator of marginal posterior of $y_{1:T}$, $\hat{q}(y_{1:T}|\theta')$

The Gibbs sanoker

General Principle

A very **specific** Markov chain Monte Carlo simulation algorithm based on the target distribution f:

1. Uses the conditional densities $f_1, \ldots, f_p$ from f
2. Start with the random variable $\mathbf{X} = (X_1, \ldots, X_p)$
3. Simulate from the conditional densities,

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

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

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

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

Properties

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

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

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

Example (Bivariate Gibbs sampler)

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

Generate a sequence of observations by
Set $X_0 = x_0$
For $t = 1, 2, \ldots,$ generate

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

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

A Very Simple Example: Independent $\mathcal{N}(\mu, \sigma^2)$ Observations

When $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(y|\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate outside a standard familly

But...

$$\mu|\mathbf{Y}_{0:n}, \sigma^2 \sim \mathcal{N}\left(\mu \left| \frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{\sigma^2}{n} \right.\right)$$

$$\sigma^2|\mathbf{Y}_{1:n}, \mu \sim \mathcal{IG}\left(\sigma^2 \left| \frac{n}{2} - 1, \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^2 \right.\right)$$

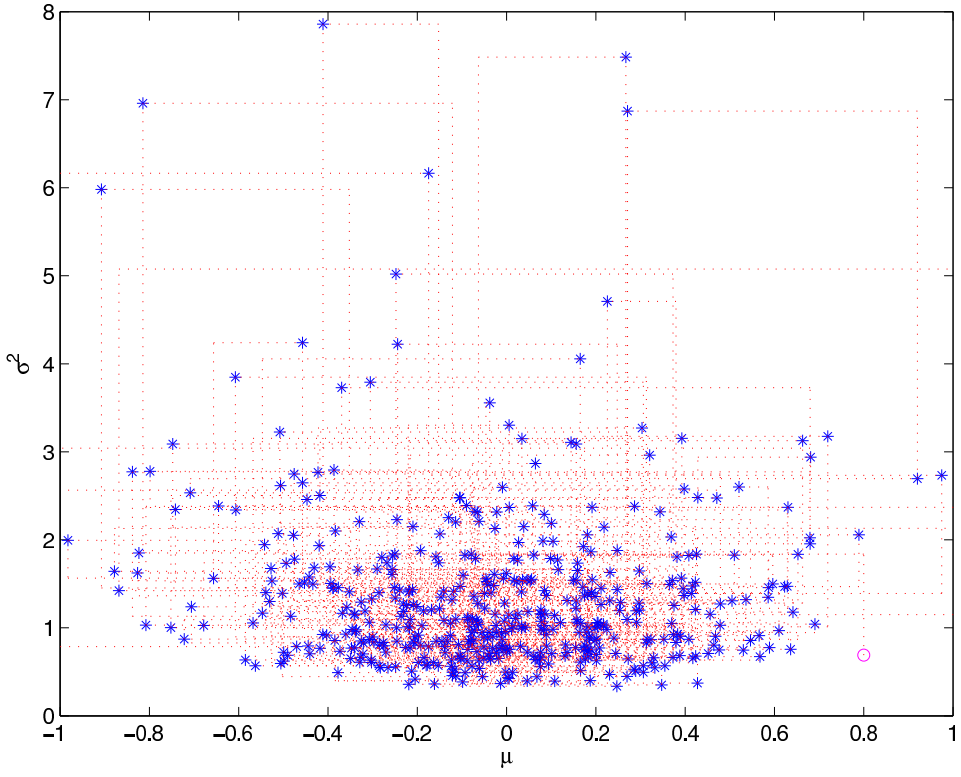
assuming constant (improper) priors on both μ and σ^2

- Hence we may use the Gibbs sampler for simulating from the posterior of (μ, σ^2)

R Gibbs Sampler for Gaussian posterior

```
n = length(Y);
S = sum(Y);
mu = S/n;
for (i in 1:500)
    S2 = sum((Y-mu)^2);
    sigma2 = 1/rgamma(1,n/2-1,S2/2);
    mu = S/n + sqrt(sigma2/n)*rnorm(1);
```

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

Limitations 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.

Latent variables are back

The Gibbs sampler can be generalized in much wider generality
A density g is a completion of f if

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

Note

The variable z may well be meaningless for the problem

Purpose

Demarginalisation 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

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

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

Algorithm (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 (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

Example (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

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

Z is the component indicator associated with observation x

Example (Mixtures (2))

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

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

Algorithm (Mixture Gibbs sampler)

1. Simulate

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

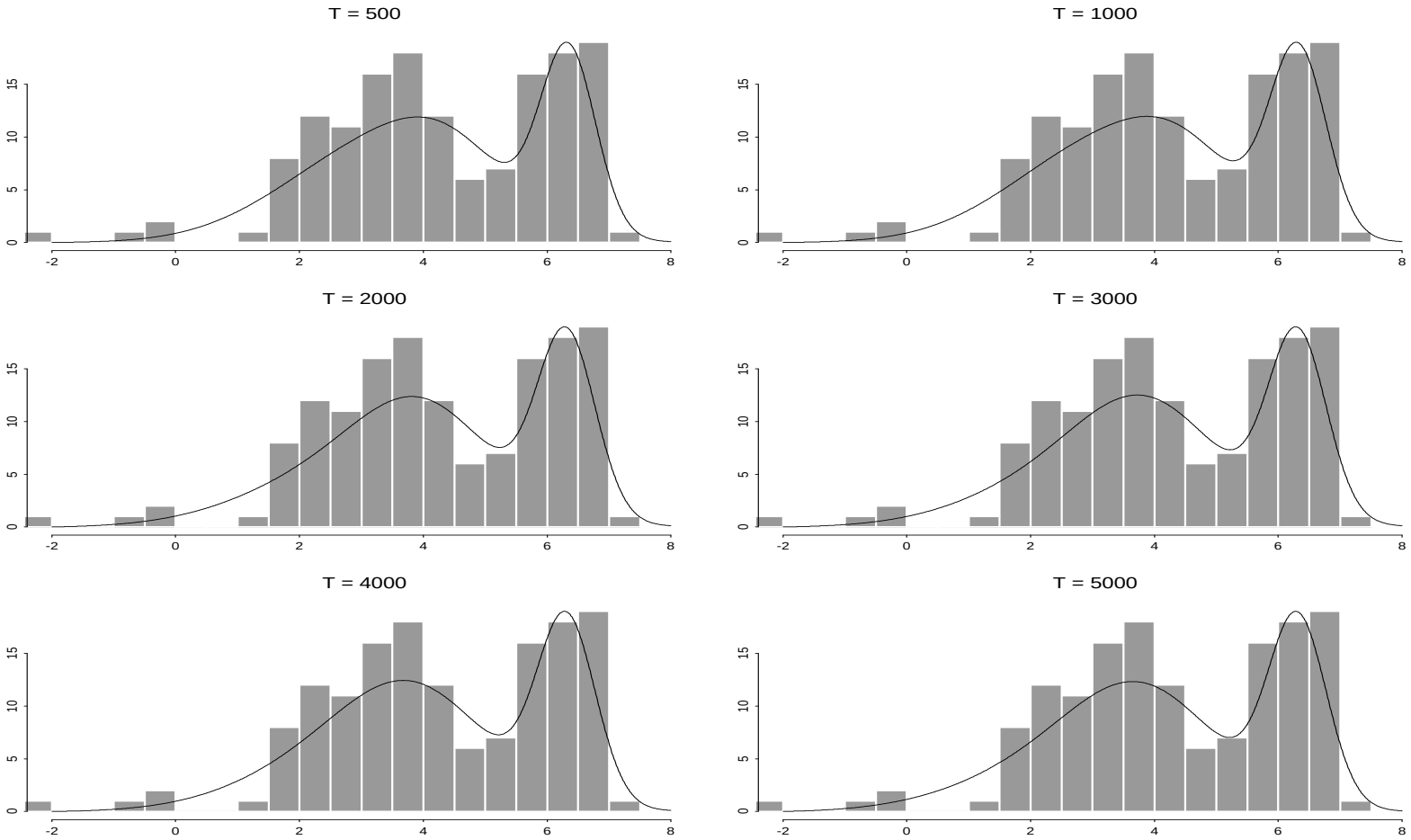
2. Simulate $(j = 1, \ldots, n)$

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

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

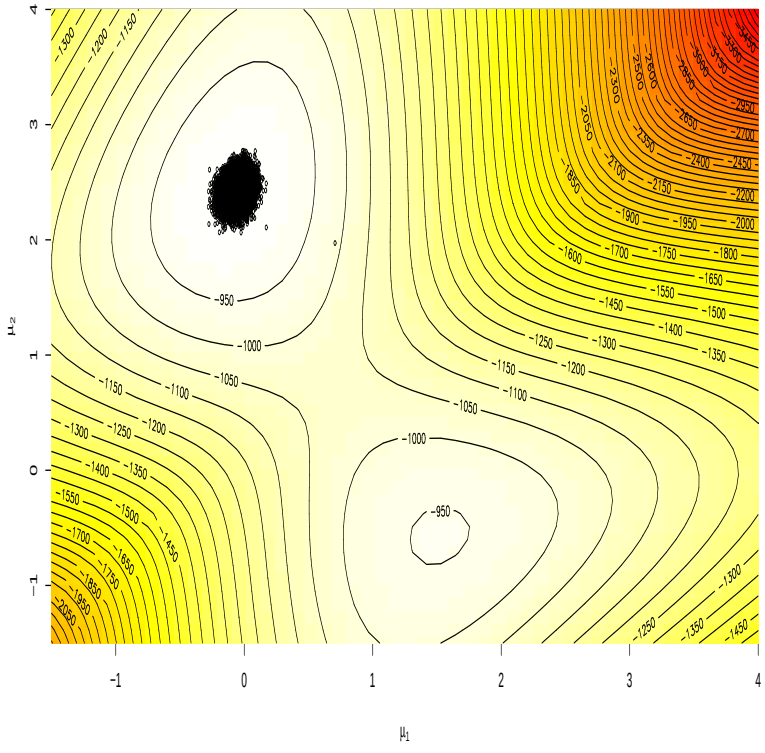
$$p_{ij} \propto p_i f(x_j | \theta_i)$$

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



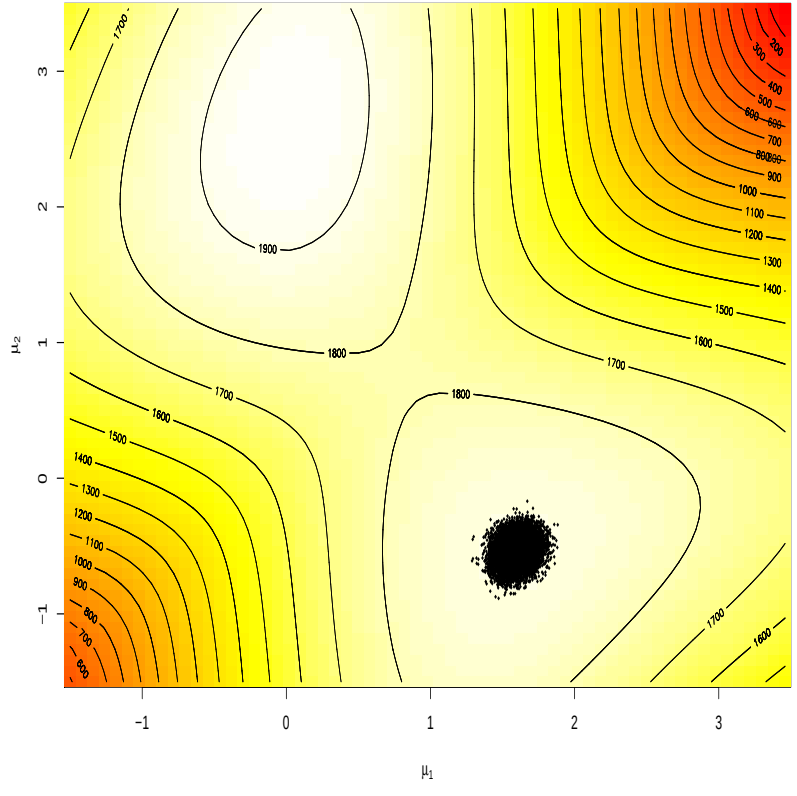
Estimation of the pluggin density for 3 components and T iterations for 149 observations of acidity levels in US lakes

A wee problem



Gibbs started at random

Gibbs stuck at the wrong mode



Random Scan Gibbs sampler

back to basics

don't do random

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

Motivation

The Random Scan Gibbs sampler is reversible.

Slice sampler as generic Gibbs

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

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

it can be completed as

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

leading to the following Gibbs algorithm:

Algorithm (Slice sampler)

Simulate

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

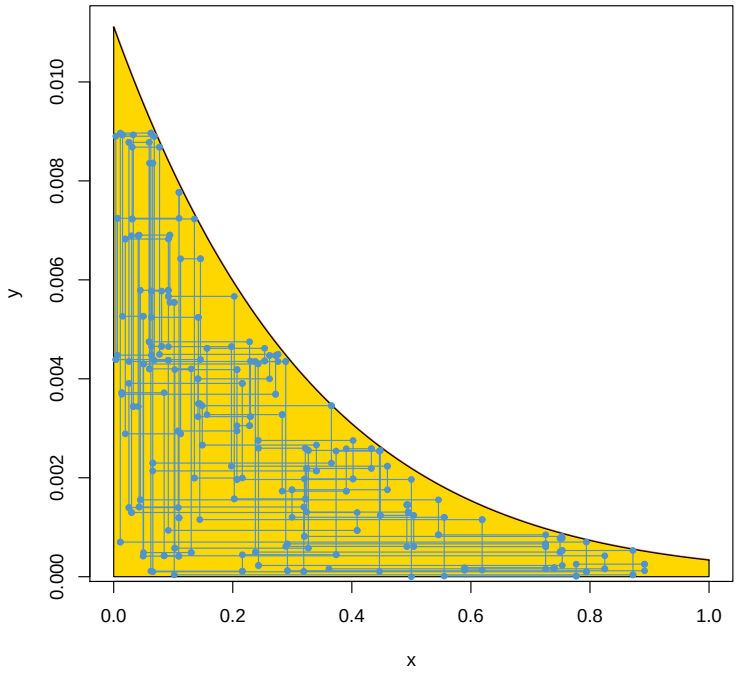
   ...

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

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

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

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



Number of Iterations 2, 3, 4, 5, 10, 50, 100

Good slices

The slice sampler usually enjoys good theoretical properties (like geometric ergodicity and even uniform ergodicity under bounded f and bounded $\mathscr{X}$).
As k increases, the determination of the set $A^{(t+1)}$ may get increasingly complex.

Properties of the Gibbs sampler

Theorem (Convergence)

For

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

if either

[Positivity condition]

(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*

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

then the chain is irreducible and positive Harris recurrent.

Properties of the Gibbs sampler (2)

Consequences

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

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

(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

▸ fast on that slice

For convergence, the properties of X_t and of $f(X_t)$ are identical

Theorem (Uniform ergodicity)

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

[Mira & Tierney, 1997]

A small set for a slice sampler

no slice detail

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

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

is a **small set**:

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

where

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

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

[Roberts & Rosenthal, 1998]

Slice sampler: drift

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

[Roberts & Rosenthal, 1998]

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

For $n > 23$,

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

Slice sampler: convergence

▸ no more slice detail

Theorem

For any density *such that*

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

then

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

[Roberts & Rosenthal, 1998]

A poor slice sampler

Example

Consider

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

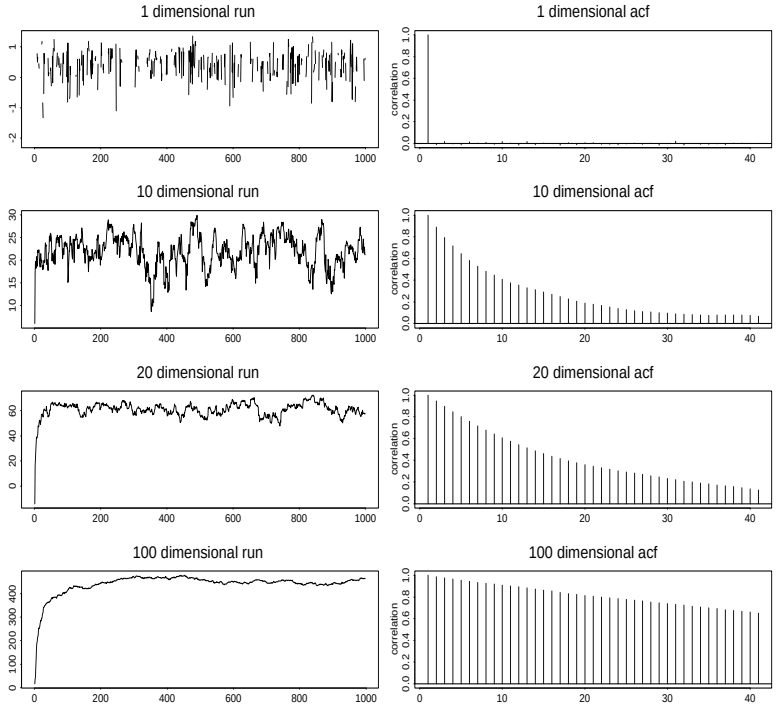
Slice sampler equivalent to one-dimensional slice sampler on

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

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)

Hammersley-Clifford theorem

An illustration that conditionals determine the joint distribution

Theorem

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*

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

[Hammersley & Clifford, circa 1970]

General HC decomposition

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

Hierarchical models

▸ no hierarchy

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

Example (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

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

where λ_i is the underlying rate of infection in herd i
Lack of independence might manifest itself as overdispersion.

Example (Animal epidemiology (2))

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

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

▸ if you hate rats

Example (Rats)

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

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

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

Example (Rats (2))

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

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

and

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

if ith rat treated with a placebo (P) or a drug (D)
Hyperparameters of the model,

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

associated with Jeffreys' noninformative priors.
Alternative prior with two possible levels of intoxication

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

Conditional decompositions

Easy decomposition of the posterior distribution
For instance, if

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

then

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

Conditional decompositions (2)

where

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

Conditional decompositions (3)

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

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

where

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

Example (Rats inc., continued ▸ if you still hate rats)

Posterior complete distribution given by

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

Local conditioning property

For the hierarchical model

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

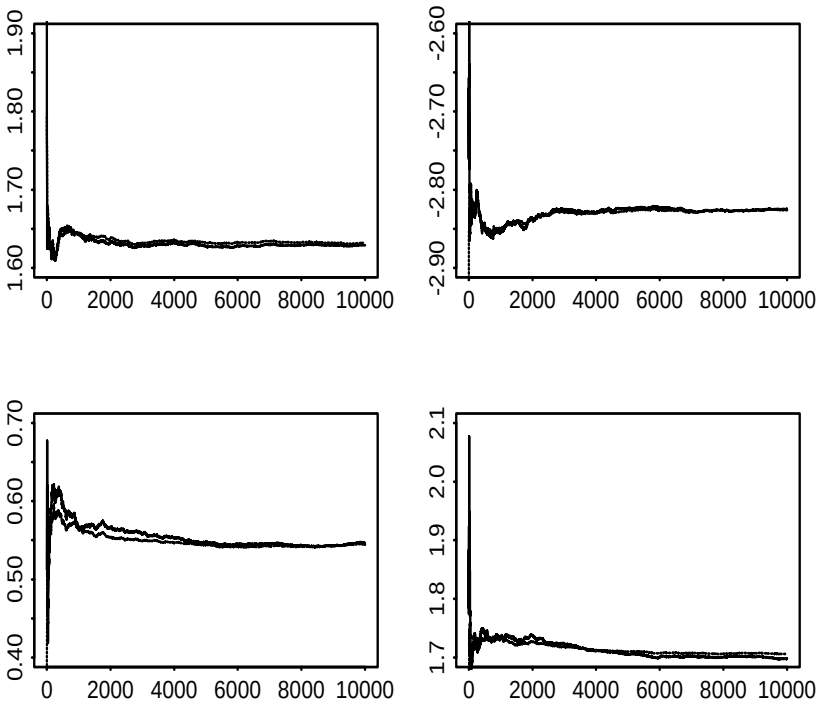
we have

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

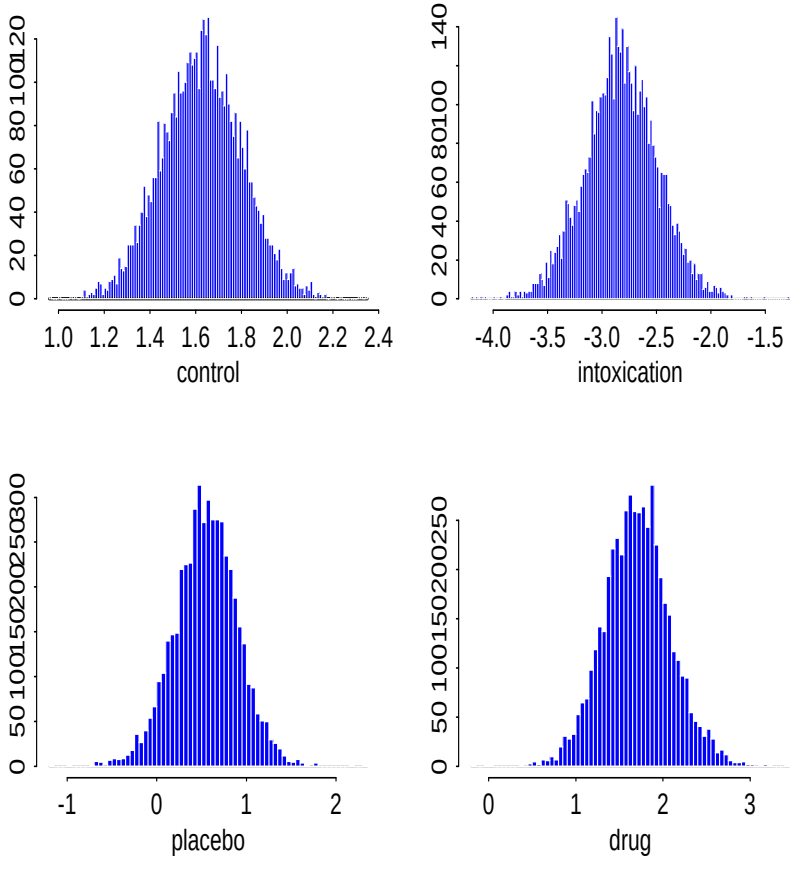
with the convention $\theta_0 = \theta$ and $\theta_{n+1} = 0$.

Example (Rats inc., terminated ▸ still this zemmiphobia?!)

The full conditional distributions correspond to standard distributions and Gibbs sampling applies.



Convergence of the posterior means



Posteriors of the effects

Posterior Gibbs inference

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

Posterior probabilities of significant effects

Improper Priors

↯ Unsuspected danger resulting from careless use of MCMC algorithms:
It may 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 (Conditional exponential distributions)

For the model

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

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

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

but

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

© **These conditionals do not correspond to a joint probability distribution**

Example (Improper random effects)

Consider

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

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

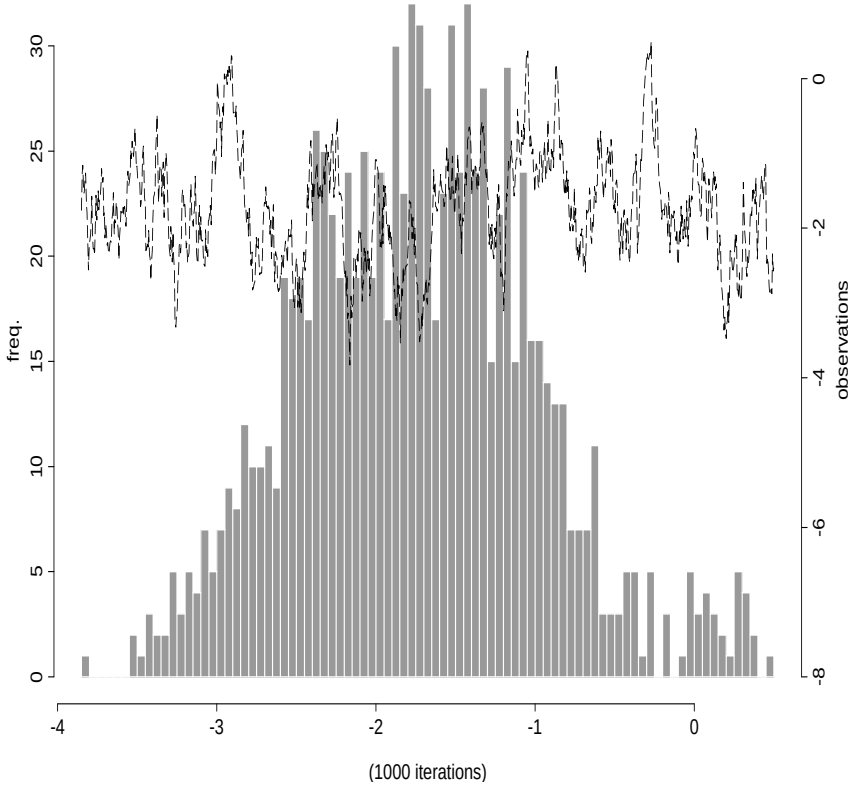
$$\pi(\mu, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2} \; .$$

Example (Improper random effects 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 sampler can be easily implemented in this setting.



Example (Improper random effects 2)

The figure shows the sequence of $\mu^{(t)}$'s and its histogram over $1,000$ iterations. They both **fail to** 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

Hamiltonian Monte Carlo and other PDMPs

Continuous time Markov process

Hamiltonian (or hybrid) Monte Carlo (HMC) auxiliary variable technique that takes advantage of a continuous time Markov process to sample from target $\pi(\theta)$

Auxiliary variable $\vartheta \in \mathbb{R}^{\mathrm{d}}$ introduced along with a density $\varpi(\vartheta|\theta)$ so that the joint distribution of (θ, ϑ) enjoys $\pi(\theta)$ as its marginal

$$\pi(\theta) = \int \pi(\theta)\varpi(\vartheta|\theta)\mathrm{d}\vartheta$$

Based on representation of joint distribution

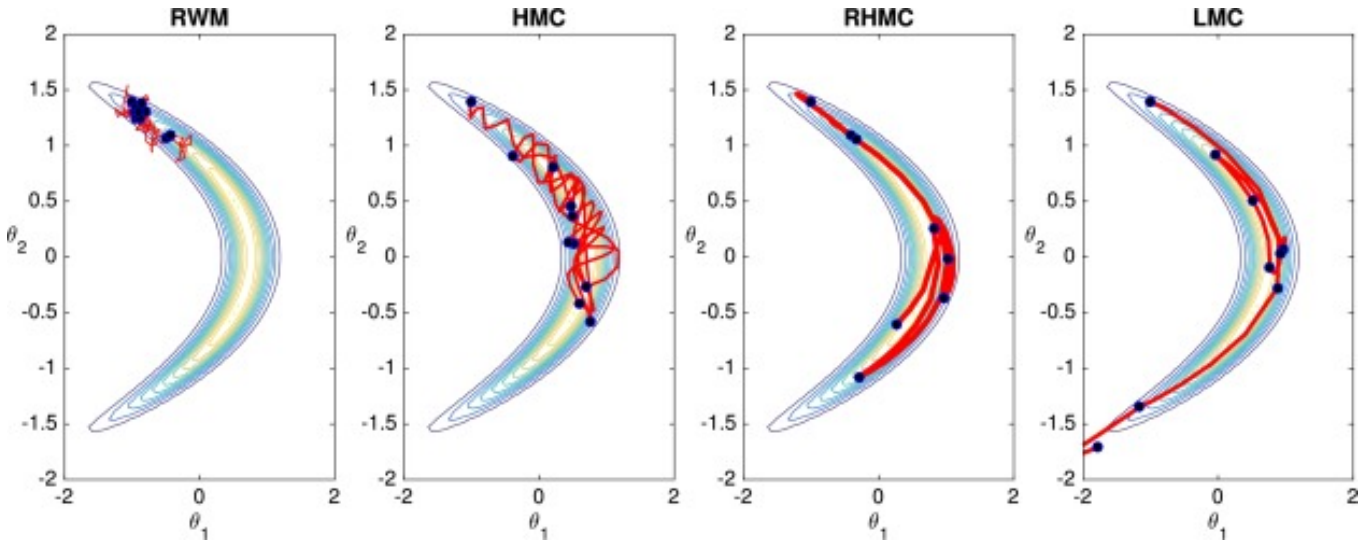
$$\omega(\theta, \vartheta) = \pi(\theta)\varpi(\vartheta|\theta) \propto \exp\{-\mathrm{H}(\theta, \vartheta)\},$$

where $\mathrm{H}(\cdot)$ called *Hamiltonian*

Hamiltonian Monte Carlo (HMC) associated with the continuous time process (θ_t, ϑ_t) generated by the so-called *Hamiltonian equations*

Background

Approach from physics (Duane et al., 1987) popularised in statistics by Neal (1996, 2002)



[Lan et al., 2016]

- Above continuous time Markov process is deterministic
- Only explores single given level set

$$\{(\theta, \vartheta) : H(\theta, \vartheta) = H(\theta_0, \vartheta_0)\},$$

Practical implementation

Free conditional density $\varpi(\vartheta|\theta)$, usually chosen as Gaussian with either a constant covariance matrix M corresponding to target covariance or as local curvature depending on θ in Riemannian and Lagrangian HMC (Girolami and Calderhead, 2011; Lan et al., 2016)

For fixed covariance matrix M, Hamiltonian equations

$$\frac{d\theta_t}{dt} = M^{-1}\vartheta_t \qquad \frac{d\vartheta_t}{dt} = \nabla\mathcal{L}(\theta_t)\,,$$

equal to the score function

Leapfrog integrator

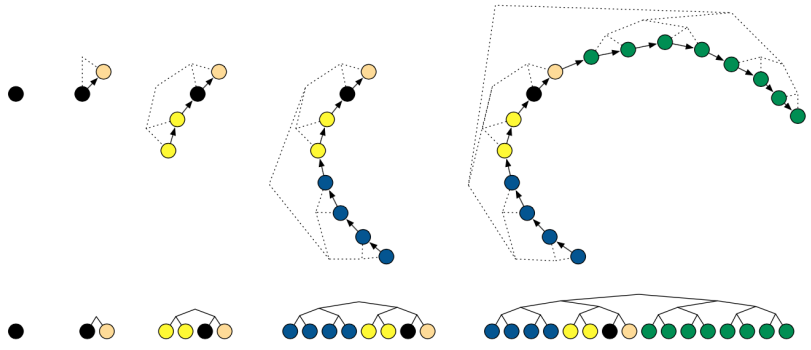
Discretisation simulation technique: symplectic integrator
One version in the independent case with constant covariance M made of leapfrog steps

$$\begin{aligned}\vartheta_{t+\epsilon/2} &= \vartheta_t + \epsilon\nabla\mathcal{L}(\theta_t)/2,\\ \theta_{t+\epsilon} &= \theta_t + \epsilon M^{-1}\vartheta_{t+\epsilon/2},\\ \vartheta_{t+\epsilon} &= \vartheta_{t+\epsilon/2} + \epsilon\nabla\mathcal{L}(\theta_{t+\epsilon})/2,\end{aligned}$$

where ϵ is time-discretisation step
Using proposal on ϑ_0 drawn from Gaussian auxiliary target and deciding on acceptance of the value of $(\theta_{T\epsilon}, \vartheta_{T\epsilon})$ by a Metropolis–Hastings step
Note that first two leapfrog steps induce a Langevin move on θ_t:

$$\theta_{t+\epsilon} = \theta_t + \epsilon^2 M^{-1}\nabla\mathcal{L}(\theta_t)/2 + \epsilon M^{-1}\vartheta_t\,,$$

no U turns



- empirically successful and popular version of HMC: "no-U-turn sampler" (NUTS) adapts value of ϵ based on primal-dual averaging

[workhorse of Stan]

- and eliminates need to choose trajectory length T via a recursive algorithm that builds a set of candidate proposals for a number of forward and backward leapfrog steps, stopping automatically when simulated path traces back

[Hoffman and Gelman, 2014]

Piecewise Deterministic Versions

Goal: sample from a target known up to a constant, defined over $\mathbb{R}^d$,

$$\pi(\mathbf{x}) \propto \gamma(\mathbf{x})$$

with energy $U(\mathbf{x}) = -\log \pi(\mathbf{x})$, $U \in \mathcal{C}^1$.

Marketing arguments

Current default workhorse: reversible MCMC methods

Non-reversible MCMC algorithms based on piecewise deterministic Markov processes perform well empirically
Quantitative convergence rates and variance now available

- Physics (Peters & De With, 2012; Krauth et al., 2009, 2015, 2016) roots
- Mesquita and Hespanha (2010) show geometric ergodicity for exponentially decaying tail targets
- Monmarché (2016) gives sharp results for compact state-spaces
- Bierkens et al. (2016a,b) show ergodicity targets on the real line

Motivation: piecewise deterministic Markov process

PDMP sampler is a (new?) continuous-time, non-reversible MCMC method based on auxiliary variables

1. particle physics simulation

[Peters et al., 2012]

2. empirically state-of-the-art performances

[Bouchard et al., 2017]

3. exact subsampled in big data

[Bierkens et al., 2017]

4. geometric ergodicity for a large class of distribution

[Deligiannidis et al., 2017, Bierkens et al., 2017]

5. Ability to deal with intractable potential $U(x) = \int U_\omega(x)\mu(d\omega)$

[Pakman et al., 2016]

Setup

- All MCMC schemes presented here target an extended distribution on $\mathfrak{Z} = \mathbb{R}^d \times \mathbb{R}^d$

$$\rho(z) = \pi(x) \times \psi(v) = \exp(-H(z))$$

  where $z = (x, v)$ extended state and $\Psi(v)$ [by default] multivariate standard Normal
- Physics takes v as velocity or momentum variables allowing for a deterministic dynamics on $\mathbb{R}^d$
- Obviously sampling from ρ provides samples from π

Piecewise deterministic Markov process

Piecewise deterministic Markov process $\{\boldsymbol{z}_t \in \mathcal{Z}\}_{t\in[0,\infty)}$, with three ingredients,

1. **Deterministic dynamics**: between events, deterministic evolution based on ODE

$$\mathrm{d}\boldsymbol{z}_t/\mathrm{d}t = \Phi(\boldsymbol{z}_t)$$

2. **Event occurrence rate**: $\lambda(t) = \lambda(\boldsymbol{z}_t)$
3. **Transition dynamics**: At event time, τ, state prior to τ denoted by $\boldsymbol{z}_{\tau-}$, and new state generated by $\boldsymbol{z}_\tau \sim Q(\cdot|\boldsymbol{z}_{\tau-})$.

[Davis, 1984, 1993]

Implementation

Algorithm 1 Simulation of PDMP

Input: Starting point z_0, $\tau_0 \leftarrow 0$.

for $k = 1, 2, 3, \cdots$

Sample inter-event time η_k from distribution

$$\mathbb{P}(\eta_k > t) = \exp\left\{-\int_0^t \lambda(z_{\tau_{k-1}+s})ds\right\}.$$

$\tau_k \leftarrow \tau_{k-1} + \eta_k$, $z_{\tau_{k-1}+s} \leftarrow \Psi_s(z_{\tau_{k-1}})$, for $s \in (0, \eta_k)$, where Ψ ODE flow of Φ.

$z_{\tau_k-} \leftarrow \Psi_{\eta_k}(z_{\tau_{k-1}})$, $z_{\tau_k} \sim Q(\cdot|z_{\tau_k-})$.

Simulation of PDMP: constraints

Requires being able to

- compute exactly flow $z_t = \Phi_t(z_0)$

  existing algorithms use $\Phi(z) = (v; 0_d)$ so thatt $\Phi(z_0) = (x_0 + v_0 t; v_0)$
  except for Hamiltonian BPS that uses the Hamiltonian dynamics for a proxy Gaussian Hamiltonian (Vanetti et al., 2017).
- simulate event times (Inversion, thinning, superposition, Devroye, 1986)
- simulate from Q

Basic bouncy particle sampler

Simulation of continuous-time piecewise linear trajectory $(x_t)_t$ with each segment in trajectory specified by

- initial position x
- length τ
- velocity v

[Bouchard et al., 2017]

length specified by inhomogeneous Poisson point process with intensity function

$$\lambda(x, v) = \max\{0, < \nabla U(x), v >\}$$

[Bouchard et al., 2017]

new velocity after bouncing given by Newtonian elastic collision

$$R(x)v = v - 2\frac{< \nabla U(x), v >}{\|\nabla U(x)\|^2}\nabla U(x)$$

[Bouchard et al., 2017]

Implementation hardships

Generally speaking, the main difficulties of implementing PDMP come from

1. Computing the ODE flow Ψ: linear dynamic, quadratic dynamic
2. Simulating the inter-event time η_k: many techniques of superposition and thinning for Poisson processes

[Devroye, 1986]

Poisson process on $\mathbb{R}_+$

Definition (Poisson process)

Poisson process with rate λ on $\mathbb{R}_+$ is sequence

$$\tau_1, \tau_2, \cdots$$

of rv's when intervals

$$\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \cdots$$

are iid with

$$\mathbb{P}(\tau_i - \tau_{i-1} > T) = \exp\left\{-\int_{\tau_{i-1}}^{\tau_{i-1}+T} \lambda(t)dt\right\}, \qquad \tau_0 = 0$$

a rarely available cdf

Simulation by thinning

Theorem (Lewis et al., 1979)

Let

$$\lambda, \Lambda : \mathbb{R}^+ \to \mathbb{R}^+$$

be continuous functions such that $\lambda(\cdot) \leq \Lambda(\cdot)$. *Let*

$$\tau_1, \tau_2, \cdots,$$

be the increasing sequence of a Poisson process with rate $\Lambda(\cdot)$. *For all* i, *if* τ_i *is removed from the sequence with probability*

$$1 - \lambda(t)/\Lambda(t)$$

then the remaining $\tilde{\tau}_1, \tilde{\tau}_2, \cdots$ *form a non-homogeneous Poisson process with rate*$\lambda(\cdot)$

Simulation from upper bound

Simulation by superposition theorem

Theorem (Kingman,1992)

Let $\Pi_1, \Pi_2, \cdots$, *be countable collection of independent Poisson processes on* $\mathbb{R}^+$ *with resp. rates* $\lambda_n(\cdot)$. *If* $\sum_{n=1}^{\infty} \lambda_n(t) < \infty$ *for all* t*'s, then superposition process*

$$\Pi = \bigcup_{n=1}^{\infty} \Pi_n$$

is Poisson process with rate $\mathrm{lambda}(t)$

$$\lambda(t) = \sum_{n=1}^{\infty} \lambda_n(t)$$

Decomposition of $U = \sum_j U_j$ **plus thinning**

Simulation by superposition plus thinning

Almost all implementations of discrete-time schemes consist of sampling a Bernoulli of parameter $\alpha(z)$
For

$$\Phi(z) = (x + v\epsilon, v) \quad \text{and} \quad \alpha(z) = 1 \wedge \pi(x + v\epsilon)/\pi(x)$$

sampling inter-event time for strictly convex $U(\cdot)$ can be obtained by solving $t^\star = \arg\min U(x + vt)$ and additional randomization

- thinning: if there exists $\bar{\alpha}$ such that $\alpha(\Phi^k(z)) \geq \bar{\alpha}(x, k)$, accept-reject
- superposition and thinning: when $\alpha(z) = 1 \wedge \rho(\Phi(z))/\rho(z)$ and $\rho(\cdot) = \prod_i \rho_i(\cdot)$ then $\bar{\alpha}(z, k) = \prod_i \bar{\alpha}_i(z, k)$

Extended generator

Definition

For $\mathcal{D}(\mathcal{L})$ set of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ such that there exists a measurable function $h : \mathcal{Z} \to \mathbb{R}$ with $t \to h(z_t)$ P_z-a.s. for each $z \in \mathcal{Z}$ and the process

$$C_t^f = f(z_t) - f(z_0) - \int_0^t h(z_s)ds$$

a local martingale. Then we write $h = \mathcal{L}f$ and call $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$ the extended generator of the process $\{z_t\}_{t\geq 0}$.

Extended Generator of PDMP

Theorem (Davis, 1993)

The generator, $\mathcal{L}$, of above PDMP is, for $\mathrm{f} \in \mathcal{D}(\mathcal{L})$

$$\mathcal{L}\mathrm{f}(z) = \nabla \mathrm{f}(z) \cdot \Phi(z) + \lambda(z) \int_{z'} \left[\mathrm{f}(z') - \mathrm{f}(z)\right] \mathrm{Q}(\mathrm{d}z'|z)$$

Furthermore, $\mu(\mathrm{d}z)$ is an invariant distribution of above PDMP, if

$$\int \mathcal{L}\mathrm{f}(z)\mu(\mathrm{d}z) = 0, \quad \text{for all } \mathrm{f} \in \mathcal{D}(\mathcal{L})$$

PDMP-based sampler

PDMP-based sampler is an auxiliary variable technique

Given target $\pi(\boldsymbol{x})$,

1. introduce auxiliary variable $\boldsymbol{V} \in \mathcal{V}$ along with a density $\pi(\boldsymbol{v}|\boldsymbol{x})$,
2. choose appropriate Φ, λ and Q

for $\pi(\boldsymbol{x})\pi(\boldsymbol{v}|\boldsymbol{x})$ to be unique invariant distribution of Markov process

Bouncy Particle Sampler (Bouchard et al., 2017)

$\mathcal{V} = \mathbb{R}^d$, and $\pi(v|x) = \varphi(v)$ for $\mathcal{N}(0, I_d)$

1. **Deterministic dynamics**:

$$dx_t/dt = v_t, \, dv_t/dt = 0$$

2. **Event occurrence rate**: $\lambda(x, v) = \langle v, \nabla U(x) \rangle_+ + \lambda^{ref}$
3. **Transition dynamics**:

$$\begin{aligned} &Q((dx', dv')|(x, v)) \\ &= \frac{\langle v, \nabla U(x) \rangle_+}{\lambda(x, v)} \delta_x(dx') \delta_{R_{\nabla U(x)} v}(dv') + \frac{\lambda^{ref}}{\lambda(x, v)} \delta_x(dx') \varphi(dv') \end{aligned}$$

where $R_{\nabla U(x)} v = v - 2 \frac{\langle \nabla U(x), v \rangle}{\langle \nabla U(x), \nabla U(x) \rangle} \nabla U(x)$

Zig-Zag Sampler (Bierkens et al., 2016)

$\mathcal{V} = \{+1, -1\}^d$, and $\pi(v|x) \sim \text{Uniform}(\{+1, -1\}^d)$

1. **Deterministic dynamics**:

$$dx_t/dt = v_t, \, dv_t/dt = 0$$

2. **Event occurrence rate**:
$\lambda(x, v) = \sum_{i=1}^d \lambda_i(x, v) = \sum_{i=1}^d \left[\{v_i \nabla_i U(x)\}_+ + \lambda_i^{\text{ref}}\right]$
3. **Transition dynamics**:

Continuous-time Hamiltonian Monte Carlo (Neal, 1999)

$\mathcal{V} = \mathbb{R}^d$, and $\pi(v|x) = \varphi(v) \sim \mathcal{N}(0, I_d)$

1. **Deterministic dynamics**:

$$dx_t/dt = v_t, \, dv_t/dt = -\nabla U(x_t)$$

2. **Event occurrence rate**: $\lambda(x, v) = \lambda_0(x)$
3. **Transition dynamics**:

$$Q((dx', dv')|(x, v)) = \delta_x(dx')\varphi(dv')$$

Continuous-time Riemannian Manifold HMC (Girolami & Calderhead, 2011)

$\mathcal{V} = \mathbb{R}^d$, and $\pi(v|x) = \mathcal{N}(0, G(x))$, the Hamiltonian is

$$H(x, v) = U(x) + 1/2 v^T G(x)^{-1} v + 1/2 \log(|G(x)|)$$

1. **Deterministic dynamics**:

$$dx_t/dt = \partial H/\partial v(x_t, v_t), dv_t/dt = -\partial H/\partial x(x_t, v_t)$$

2. **Event occurrence rate**: $\lambda(x, v) = \lambda_0(x)$
3. **Transition dynamics**:

$$Q((dx', dv')|(x, v)) = \delta_x(dx')\varphi(dv'|x')$$

Randomized BPS

Define

$$\mathbf{a} = \frac{\langle \boldsymbol{v}, \nabla U(\boldsymbol{x}) \rangle}{\langle \nabla U(\boldsymbol{x}), \nabla U(\boldsymbol{x}) \rangle} \nabla U(\boldsymbol{x}), \quad \mathbf{b} = \boldsymbol{v} - \mathbf{a}$$

Regular BPS, move $\boldsymbol{v}' = -\mathbf{a} + \mathbf{b}$

Alternatives

1. (Fearnhead et al., 2016):
   $\boldsymbol{v}' \sim Q_{\boldsymbol{x}}(d\boldsymbol{v}'|\boldsymbol{v}) = \max\{0, \langle -\boldsymbol{v}', \nabla U(\boldsymbol{x}) \rangle\} \, d\boldsymbol{v}'$
2. (Wu & X, 2017): $\boldsymbol{v}' = -\mathbf{a} + \mathbf{b}'$, where $\mathbf{b}'$ Gaussian variate over the space orthogonal to $\nabla U(\boldsymbol{x})$ in $\mathbb{R}^d$.

HMC-BPS (Vanetti et al., 2017)

$\rho(\mathbf{x}) \propto \exp\{-V(\mathbf{x})\}$ is a Gaussian approximation of the target $\pi(\mathbf{x})$.

$$\hat{H}(\mathbf{x}, \boldsymbol{v}) = V(\mathbf{x}) + 1/2 \boldsymbol{v}^T \boldsymbol{v}, \quad \tilde{U}(\mathbf{x}) = U(\mathbf{x}) - V(\mathbf{x})$$

1. **Deterministic dynamics**:

$$dx_t/dt = \boldsymbol{v}_t, \, dv_t/dt = -\nabla V(\mathbf{x}_t)$$

2. **Event occurrence rate**: $\lambda(\mathbf{x}, \boldsymbol{v}) = \langle \boldsymbol{v}, \nabla \tilde{U}(\mathbf{x}) \rangle_+ + \lambda^{\text{ref}}$
3. **Transition dynamics**:

$$Q((dx', dv')|(\mathbf{x}, \boldsymbol{v}))$$
$$= \frac{\langle \boldsymbol{v}, \nabla \tilde{U}(\mathbf{x}) \rangle_+}{\lambda(\mathbf{x}, \boldsymbol{v})} \delta_{\mathbf{x}}(dx') \delta_{R_{\nabla \tilde{U}(\mathbf{x})} \boldsymbol{v}}(dv') + \frac{\lambda^{\text{ref}}}{\lambda(\mathbf{x}, \boldsymbol{v})} \delta_{\mathbf{x}}(dx') \varphi(dv')$$

Discretisation

1. Sherlock & Thiery (2017) considers delayed rejection approach with only point-wise evaluations of target, by making speed flip move once proposal involving flip in speed and drift in variable of interest rejected. Also add random perturbation for eergodicity, plus another perturbation based on a Brownian argument. Requires calibration
2. Vanetti et al. (2017) unifies many threads and relates PDMP, HMC, and discrete versions, with convergence results. Main idea improves upon existing deterministic methods by accounting for target. Borrows from earlier slice sampler idea of Murray et al. (AISTATS, 2010), exploiting exact Hamiltonian dynamics for approximation to true target. Except that bouncing avoids the slice step. Eight discrete BPS both correct against target and do not simulating event times.

Benefit: bypassing the generation of inter-event time of

Bayesian importance sampling

Bayesian model choice

▸ directly Markovian

Probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

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

- take largest $\pi(\mathfrak{M}_i|x)$ to determine "best" model,

Bayes factor

Definition (Bayes factors)

For testing hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_a : \theta \notin \Theta_0$, under prior

$$\pi(\Theta_0)\pi_0(\theta) + \pi(\Theta_0^c)\pi_1(\theta)\,,$$

central quantity

$$B_{01} = \frac{\pi(\Theta_0|x)}{\pi(\Theta_0^c|x)} \Big/ \frac{\pi(\Theta_0)}{\pi(\Theta_0^c)} = \frac{\displaystyle\int_{\Theta_0} f(x|\theta)\pi_0(\theta)d\theta}{\displaystyle\int_{\Theta_0^c} f(x|\theta)\pi_1(\theta)d\theta}$$

[Jeffreys, 1939]

Evidence

Problems using a similar quantity, the *evidence*

$$\mathfrak{E}_k = \int_{\Theta_k} \pi_k(\theta_k) L_k(\theta_k)\, d\theta_k,$$

aka the marginal likelihood.

[Jeffreys, 1939]

Bayes factor approximation

When approximating the Bayes factor

$$B_{01} = \frac{\int_{\Theta_0} f_0(x|\theta_0)\pi_0(\theta_0)d\theta_0}{\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}$$

use of importance functions ϖ_0 and ϖ_1 and

$$\widehat{B}_{01} = \frac{n_0^{-1}\sum_{i=1}^{n_0} f_0(x|\theta_0^i)\pi_0(\theta_0^i)/\varpi_0(\theta_0^i)}{n_1^{-1}\sum_{i=1}^{n_1} f_1(x|\theta_1^i)\pi_1(\theta_1^i)/\varpi_1(\theta_1^i)}$$

Diabetes in Pima Indian women

Example (R benchmark)

"A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix (AZ), was tested for diabetes according to WHO criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases."
200 Pima Indian women with observed variables

- plasma glucose concentration in oral glucose tolerance test
- diastolic blood pressure
- diabetes pedigree function
- presence/absence of diabetes

Probit modelling on Pima Indian women

Probability of diabetes function of above variables

$$\mathbb{P}(y = 1|x) = \Phi(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)\,,$$

Test of $H_0 : \beta_3 = 0$ for 200 observations of Pima.tr based on a g-prior modelling:

$$\beta \sim \mathcal{N}_3(0, n\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1})$$

MCMC 101 for probit models

Use of either a random walk proposal

$$\beta' = \beta + \epsilon$$

in a Metropolis-Hastings algorithm (since the likelihood is available)
or of a Gibbs sampler that takes advantage of the missing/latent variable

$$z|y, x, \beta \sim \mathcal{N}(x^{\mathsf{T}}\beta, 1) \left\{ \mathbb{I}_{z\geq 0}^{y} \times \mathbb{I}_{z\leq 0}^{1-y} \right\}$$

(since $\beta|y, X, z$ is distributed as a standard normal)

[Gibbs three times faster]

Importance sampling for the Pima Indian dataset

Use of the importance function inspired from the MLE estimate distribution

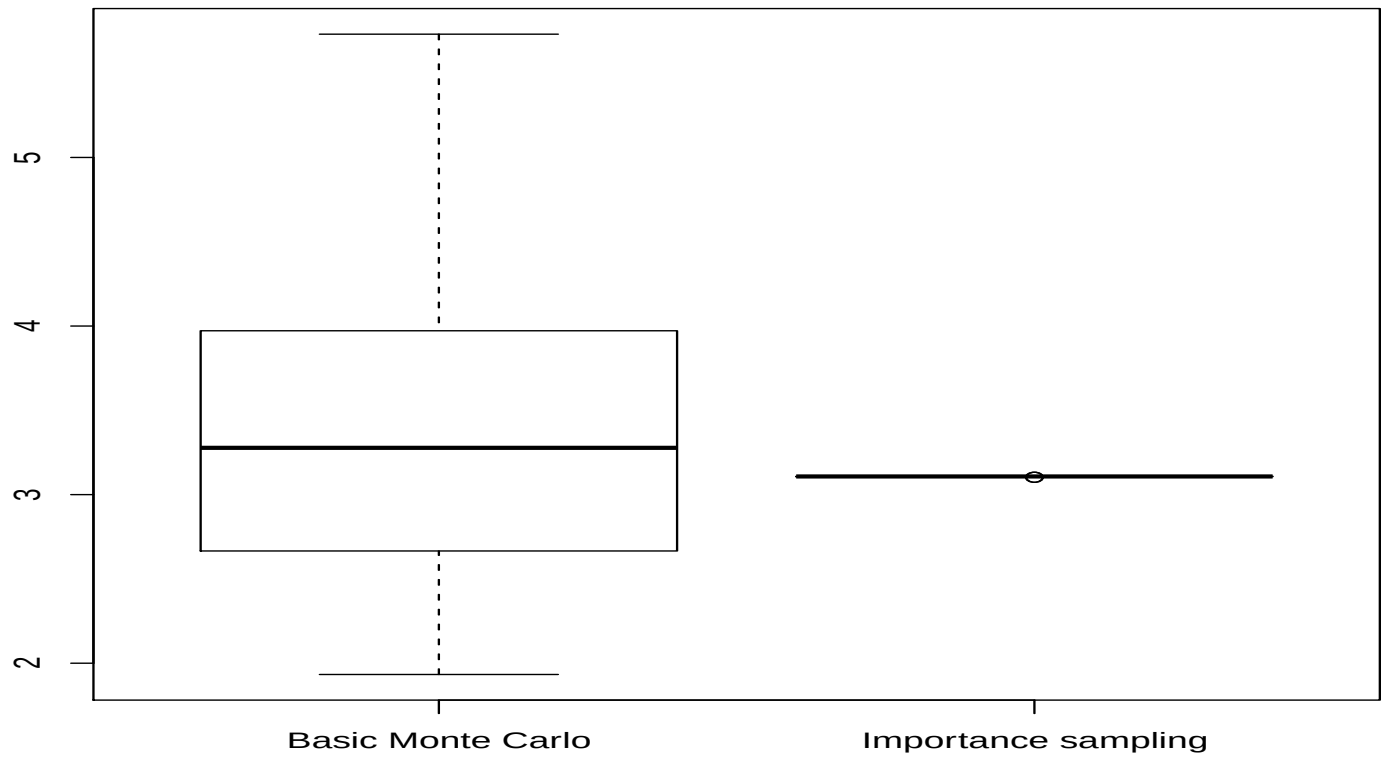
$$\beta \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$$

R Importance sampling code

```
model1=summary(glm(y~-1+X1,family=binomial(link="probit")))
is1=rmvnorm(Niter,mean=model1$coeff[,1],sigma=2*model1$cov.unscaled)
is2=rmvnorm(Niter,mean=model2$coeff[,1],sigma=2*model2$cov.unscaled)
bfis=mean(exp(probitlpost(is1,y,X1)-dmvlnorm(is1,mean=model1$coeff[,1],
     sigma=2*model1$cov.unscaled))) / mean(exp(probitlpost(is2,y,X2)-
     dmvlnorm(is2,mean=model2$coeff[,1],sigma=2*model2$cov.unscaled)))
```

Diabetes in Pima Indian women

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations from the prior and the above MLE importance sampler



Bridge sampling

Special case:

If

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

live on the same space ($\Theta_1 = \Theta_2$), then

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

[Gelman & Meng, 1998; Chen, Shao & Ibrahim, 2000]

Bridge sampling variance

The bridge sampling estimator does poorly if

$$\frac{\mathrm{var}(\widehat{B}_{12})}{B_{12}^2} \approx \frac{1}{n}\,\mathbb{E}\left[\left(\frac{\pi_1(\theta)-\pi_2(\theta)}{\pi_2(\theta)}\right)^2\right]$$

is large, i.e. if π_1 *and* π_2 *have little overlap...*

(Further) bridge sampling

General identity:

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

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

Optimal bridge sampling

The *optimal choice* of auxiliary function is

$$\alpha^\star = \frac{n_1 + n_2}{n_1\pi_1(\theta|x) + n_2\pi_2(\theta|x)}$$

leading to

$$\widehat{B}_{12} \approx \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{\tilde{\pi}_2(\theta_{1i}|x)}{n_1\pi_1(\theta_{1i}|x) + n_2\pi_2(\theta_{1i}|x)}}{\frac{1}{n_2}\sum_{i=1}^{n_2} \frac{\tilde{\pi}_1(\theta_{2i}|x)}{n_1\pi_1(\theta_{2i}|x) + n_2\pi_2(\theta_{2i}|x)}}$$

Back later!

Optimal bridge sampling (2)

Reason:

$$\frac{\mathrm{Var}(\widehat{\mathrm{B}}_{12})}{\mathrm{B}_{12}^2} \approx \frac{1}{n_1 n_2} \left\{ \frac{\int \pi_1(\theta)\pi_2(\theta)[n_1\pi_1(\theta) + n_2\pi_2(\theta)]\alpha(\theta)^2 \, \mathrm{d}\theta}{\left(\int \pi_1(\theta)\pi_2(\theta)\alpha(\theta) \, \mathrm{d}\theta\right)^2} - 1 \right\}$$

(by the δ method)

Drawback: Dependence on the unknown normalising constants solved iteratively

Extension to varying dimensions

When $\dim(\Theta_1) \neq \dim(\Theta_2)$, e.g. $\theta_2 = (\theta_1, \psi)$, introduction of a *pseudo-posterior density,* $\omega(\psi|\theta_1, x)$, augmenting $\pi_1(\theta_1|x)$ into joint distribution

$$\pi_1(\theta_1|x) \times \omega(\psi|\theta_1, x)$$

on Θ_2 so that

$$B_{12} = \frac{\int \tilde{\pi}_1(\theta_1|x)\alpha(\theta_1,\psi)\pi_2(\theta_1,\psi|x)d\theta_1\,\omega(\psi|\theta_1,x)\,d\psi}{\int \tilde{\pi}_2(\theta_1,\psi|x)\alpha(\theta_1,\psi)\pi_1(\theta_1|x)d\theta_1\,\omega(\psi|\theta_1,x)\,d\psi}$$

$$= \mathbb{E}_{\pi_2}\left[\frac{\tilde{\pi}_1(\theta_1)\omega(\psi|\theta_1)}{\tilde{\pi}_2(\theta_1,\psi)}\right] = \frac{\mathbb{E}_\varphi\left[\tilde{\pi}_1(\theta_1)\omega(\psi|\theta_1)/\varphi(\theta_1,\psi)\right]}{\mathbb{E}_\varphi\left[\tilde{\pi}_2(\theta_1,\psi)/\varphi(\theta_1,\psi)\right]}$$

for *any* conditional density $\omega(\psi|\theta_1)$ and *any* joint density φ.

Illustration for the Pima Indian dataset

Use of the MLE induced conditional of β_3 given (β_1, β_2) as a pseudo-posterior and mixture of both MLE approximations on β_3 in bridge sampling estimate

R bridge sampling code

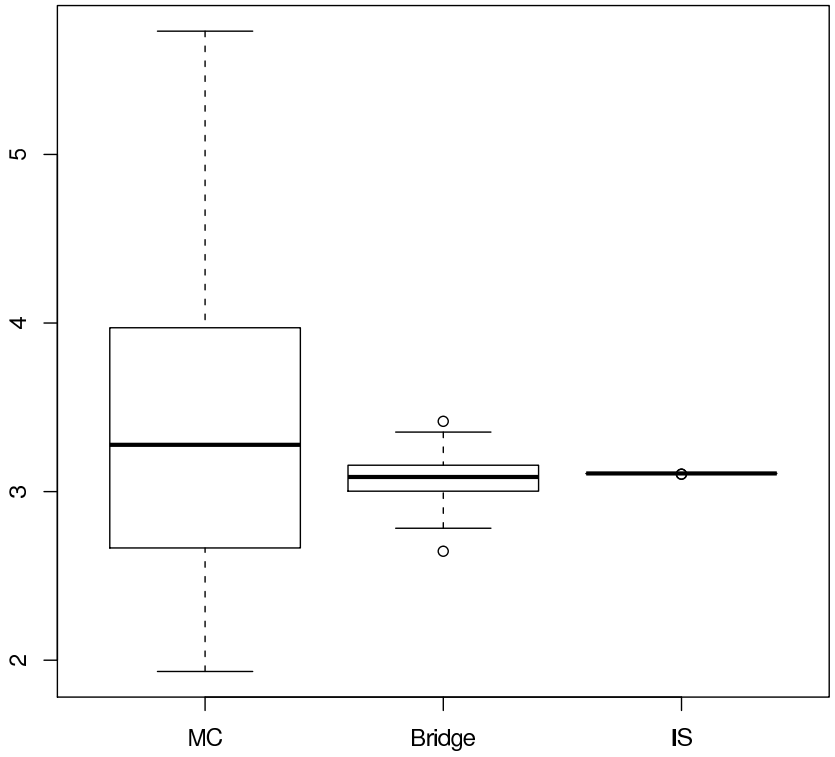
```
cova=model2$cov.unscaled
expecta=model2$coeff[,1]
covw=cova[3,3]-t(cova[1:2,3])%*%ginv(cova[1:2,1:2])%*%cova[1:2,3]

probit1=hmprobit(Niter,y,X1)
probit2=hmprobit(Niter,y,X2)
pseudo=rnorm(Niter,meanw(probit1),sqrt(covw))
probit1p=cbind(probit1,pseudo)

bfbs=mean(exp(probitlpost(probit2[,1:2],y,X1)+dnorm(probit2[,3],meanw(probit2[,1:2]),
     sqrt(covw),log=T))/ (dmvnorm(probit2,expecta,cova)+dnorm(probit2[,3],expecta[3],
     cova[3,3])))/ mean(exp(probitlpost(probit1p,y,X2))/(dmvnorm(probit1p,expecta,cova)+
     dnorm(pseudo,expecta[3],cova[3,3])))
```

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on $100 \times 20,000$ simulations from the prior (MC), the above bridge sampler and the above importance sampler



The original harmonic mean estimator

When $\theta_{ki} \sim \pi_k(\theta|x)$,

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L(\theta_{kt}|x)}$$

is an unbiased estimator of $1/m_k(x)$

[Newton & Raftery, 1994]

Highly dangerous: Most often leads to an infinite variance!!!

"The Worst Monte Carlo Method Ever"

"The good news is that the Law of Large Numbers guarantees that this estimator is consistent ie, it will very likely be very close to the correct answer if you use a sufficiently large number of points from the posterior distribution.

The bad news is that the number of points required for this estimator to get close to the right answer will often be greater than the number of atoms in the observable universe. The even worse news is that it's easy for people to not realize this, and to naïvely accept estimates that are nowhere close to the correct value of the marginal likelihood."

[Radford Neal's blog, Aug. 23, 2008]

Approximating $\mathfrak{Z}_k$ from a posterior sample

Use of the *[harmonic mean]* identity

$$\mathbb{E}^{\pi_k}\left[\frac{\varphi(\theta_k)}{\pi_k(\theta_k)L_k(\theta_k)}\middle|x\right] = \int \frac{\varphi(\theta_k)}{\pi_k(\theta_k)L_k(\theta_k)}\,\frac{\pi_k(\theta_k)L_k(\theta_k)}{\mathfrak{Z}_k}\,d\theta_k = \frac{1}{\mathfrak{Z}_k}$$

no matter what the proposal $\varphi(\cdot)$ is.

[Gelfand & Dey, 1994; Bartolucci et al., 2006]

Direct exploitation of the MCMC output

Comparison with regular importance sampling

Harmonic mean: Constraint opposed to usual importance sampling constraints: $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi_k(\theta_k)L_k(\theta_k)$ for the approximation

$$\widehat{\mathfrak{Z}_{1k}} = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta_k^{(t)})}{\pi_k(\theta_k^{(t)})L_k(\theta_k^{(t)})}$$

to have a finite variance.
E.g., use finite support kernels (like Epanechnikov's kernel) for φ

Comparison with regular importance sampling (cont'd)

Compare $\widehat{\mathfrak{Z}_{1k}}$ with a standard importance sampling approximation

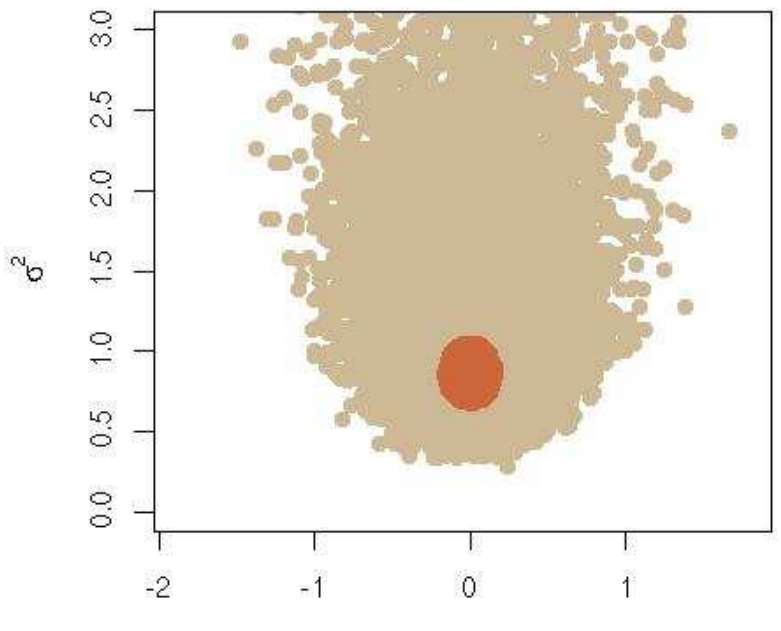
$$\widehat{\mathfrak{Z}_{2k}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)})}{\varphi(\theta_k^{(t)})}$$

where the $\theta_k^{(t)}$'s are generated from the density $\varphi(\cdot)$ (with fatter tails like t's)

HPD indicator as φ

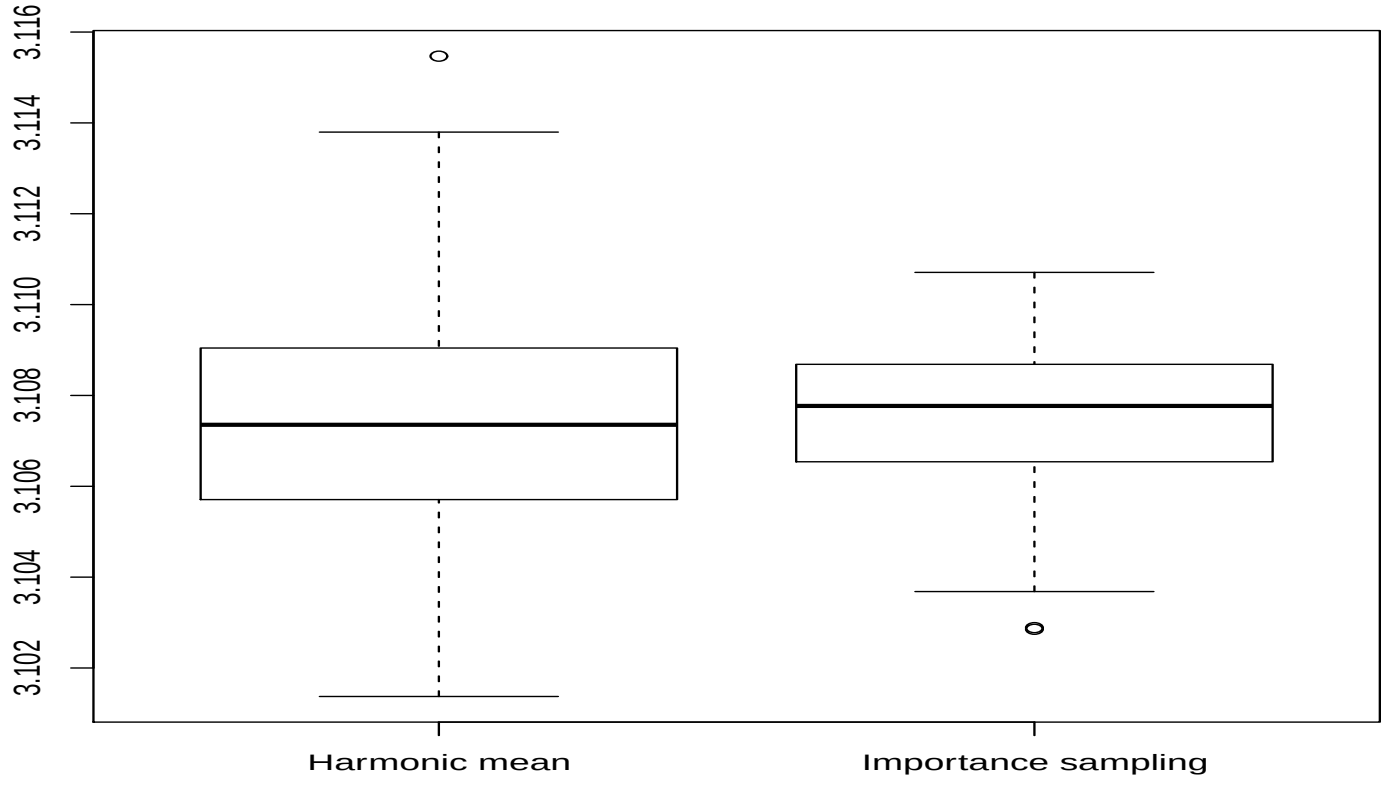
Use the convex hull of MCMC simulations corresponding to the 10% HPD region (easily derived!) and φ as indicator:

$$\varphi(\theta) = \frac{10}{T} \sum_{t \in \text{HPD}} \mathbb{I}_{d(\theta,\theta^{(t)}) \leq \epsilon}$$



Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations for a simulation from the above harmonic mean sampler and importance samplers



Approximating $\mathfrak{Z}_k$ using a mixture representation

◂ Bridge sampling redux

Design a specific mixture for simulation [importance sampling] purposes, with density

$$\widetilde{\varphi}_k(\theta_k) \propto \omega_1 \pi_k(\theta_k) L_k(\theta_k) + \varphi(\theta_k)\,,$$

where $\varphi(\cdot)$ is arbitrary (but normalised)
Note: ω_1 is *not* a probability weight

Approximating $\mathfrak{Z}$ using a mixture representation (cont'd)

Corresponding MCMC (=Gibbs) sampler

At iteration t

1. Take $\delta^{(t)} = 1$ with probability

$$\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) \Big/ \left(\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) + \varphi(\theta_k^{(t-1)})\right)$$

   and $\delta^{(t)} = 2$ otherwise;

2. If $\delta^{(t)} = 1$, generate $\theta_k^{(t)} \sim \text{MCMC}(\theta_k^{(t-1)}, \theta_k)$ where $\text{MCMC}(\theta_k, \theta_k')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi_k(\theta_k|x) \propto \pi_k(\theta_k) L_k(\theta_k)$;

3. If $\delta^{(t)} = 2$, generate $\theta_k^{(t)} \sim \varphi(\theta_k)$ independently

Evidence approximation by mixtures

Rao-Blackwellised estimate

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) \Big/ \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)}) \,,$$

converges to $\omega_1 \mathfrak{Z}_k / \{\omega_1 \mathfrak{Z}_k + 1\}$
Deduce $\hat{\mathfrak{Z}}_{3k}$ from $\omega_1 \hat{\mathfrak{E}}_{3k} / \{\omega_1 \hat{\mathfrak{E}}_{3k} + 1\} = \hat{\xi}$ ie

$$\hat{\mathfrak{E}}_{3k} = \frac{\sum_{t=1}^{T} \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) \Big/ \omega_1 \pi(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)})}{\sum_{t=1}^{T} \varphi(\theta_k^{(t)}) \Big/ \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)})}$$

[Bridge sampler]

Chib's representation

Direct application of Bayes' theorem: given $\mathbf{x} \sim \mathrm{f}_k(\mathbf{x}|\theta_k)$ and $\theta_k \sim \pi_k(\theta_k)$,

$$\mathfrak{E}_k = \mathrm{m}_k(\mathbf{x}) = \frac{\mathrm{f}_k(\mathbf{x}|\theta_k)\,\pi_k(\theta_k)}{\pi_k(\theta_k|\mathbf{x})}$$

Use of an approximation to the posterior

$$\widehat{\mathfrak{E}}_k = \widehat{\mathrm{m}_k}(\mathbf{x}) = \frac{\mathrm{f}_k(\mathbf{x}|\theta_k^*)\,\pi_k(\theta_k^*)}{\widehat{\pi_k}(\theta_k^*|\mathbf{x})}\,.$$

Case of latent variables

For missing variable z as in mixture models, natural Rao-Blackwell estimate

$$\widehat{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \pi_k(\theta_k^*|\mathbf{x}, \mathbf{z}_k^{(t)}) \,,$$

where the $\mathbf{z}_k^{(t)}$'s are Gibbs sampled latent variables

Label switching

A mixture model [special case of missing variable model] is invariant under permutations of the indices of the components. E.g., mixtures

$$0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.3,1)$$

and

$$0.7\mathcal{N}(2.3,1) + 0.3\mathcal{N}(0,1)$$

are **exactly** the same!

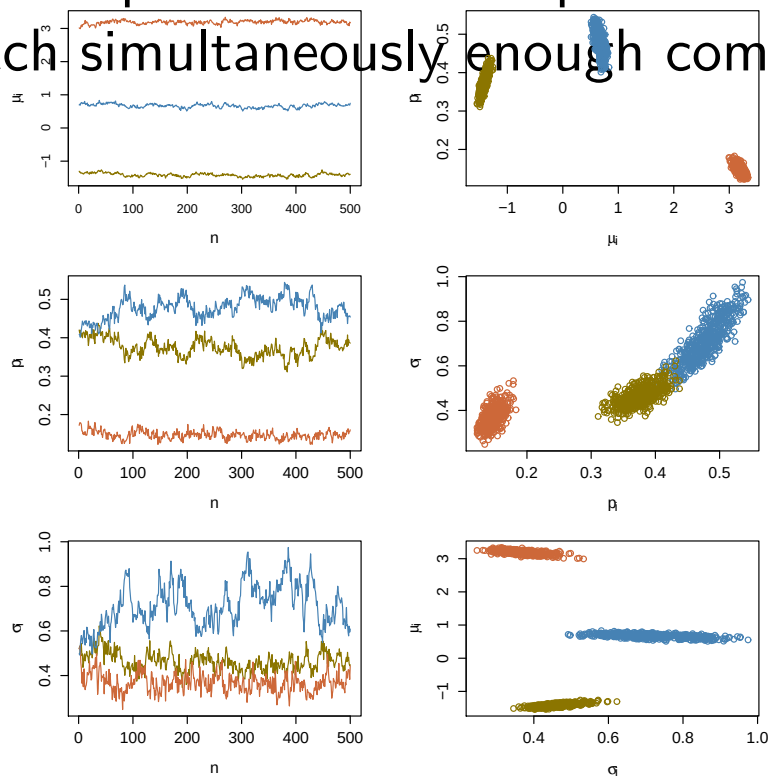
© **The component parameters θ_i are not identifiable *marginally* since they are exchangeable**

Connected difficulties

1. Number of modes of the likelihood of order $\mathrm{O}(\mathrm{k!})$:
   © Maximization and even [MCMC] exploration of the posterior surface harder
2. Under exchangeable priors on $(\theta, \mathbf{p})$ *[prior invariant under permutation of the indices]*, all posterior marginals are identical:
   © Posterior expectation of θ_1 equal to posterior expectation of θ_2

License

Since Gibbs output does not produce exchangeability, the Gibbs sampler has not explored the whole parameter space: it lacks energy to switch simultaneously enough component allocations at once



Label switching paradox

We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler.
If we observe it, then we do not know how to estimate the parameters.
If we do not, then we are uncertain about the convergence!!!

Compensation for label switching

For mixture models, $z_k^{(t)}$ usually fails to visit all configurations in a balanced way, despite the symmetry predicted by the theory

$$\pi_k(\theta_k|\mathbf{x}) = \pi_k(\sigma(\theta_k)|\mathbf{x}) = \frac{1}{k!}\sum_{\sigma\in\mathfrak{S}} \pi_k(\sigma(\theta_k)|\mathbf{x})$$

for all σ's in $\mathfrak{S}_k$, set of all permutations of $\{1,\ldots,k\}$.
Consequences on numerical approximation, biased by an order $k!$
Recover the theoretical symmetry by using

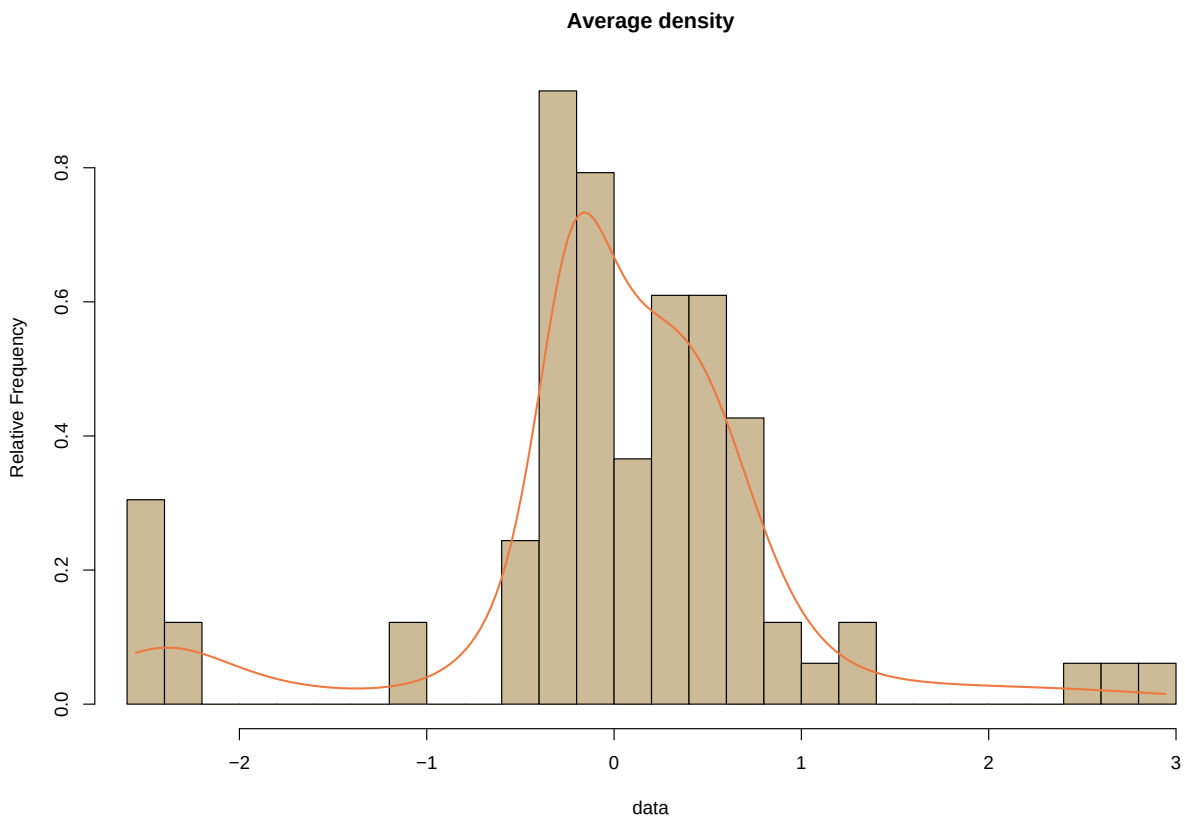
$$\widetilde{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T\,k!}\sum_{\sigma\in\mathfrak{S}_k}\sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|\mathbf{x}, z_k^{(t)})\,.$$

[Berkhof, Mechelen, & Gelman, 2003]

Galaxy dataset

$\mathrm{n} = 82$ galaxies as a mixture of k normal distributions with both mean and variance unknown.

[Roeder, 1992]



Galaxy dataset (k)

Using only the original estimate, with θ_k^* as the MAP estimator,

$$\log(\hat{m}_k(\boldsymbol{x})) = -105.1396$$

for $k = 3$ (based on 10^3 simulations), while introducing the permutations leads to

$$\log(\hat{m}_k(\boldsymbol{x})) = -103.3479$$

Note that

$$-105.1396 + \log(3!) = -103.3479$$

k	2	3	4	5	6	7	8
$m_k(\boldsymbol{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

[Lee, Marin, Mengersen & Robert, 2008]

Case of the probit model

For the completion by z,

$$\hat{\pi}(\theta|x) = \frac{1}{T} \sum_{t} \pi(\theta|x, z^{(t)})$$

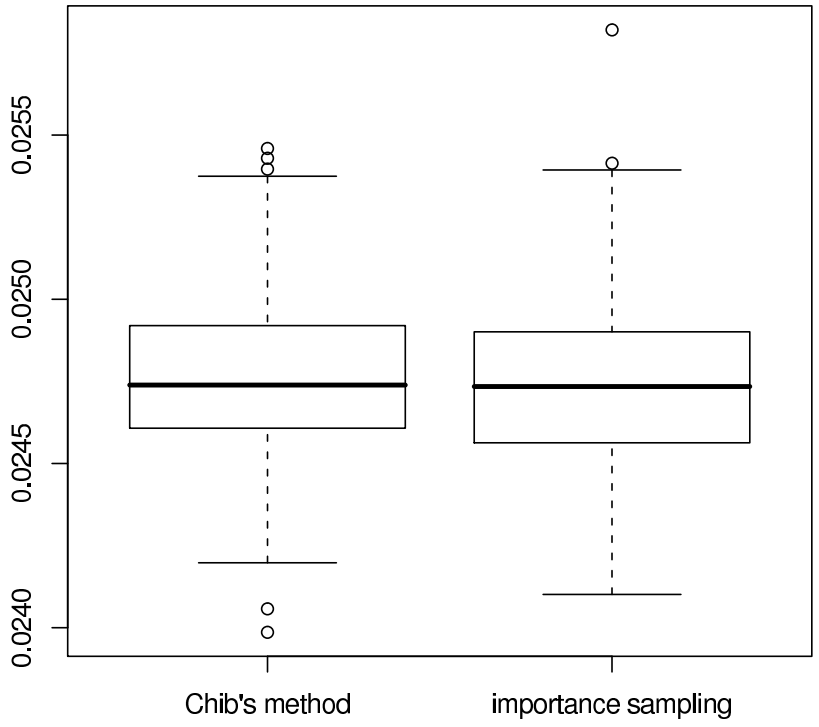
is a simple average of normal densities

R Bridge sampling code

```
gibbs1=gibbsprobit(Niter,y,X1)
gibbs2=gibbsprobit(Niter,y,X2)
bfchi=mean(exp(dmvlnorm(t(t(gibbs2$mu)-model2$coeff[,1]),mean=rep(0,3),
        sigma=gibbs2$Sigma2)-probitlpost(model2$coeff[,1],y,X2)))/
      mean(exp(dmvlnorm(t(t(gibbs1$mu)-model1$coeff[,1]),mean=rep(0,2),
        sigma=gibbs1$Sigma2)-probitlpost(model1$coeff[,1],y,X1)))
```

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations for a simulation from the above Chib's and importance samplers



The Savage–Dickey ratio

Special representation of the Bayes factor used for simulation
Given a test $H_0 : \theta = \theta_0$ in a model $f(x|\theta, \psi)$ with a nuisance parameter ψ, under priors $\pi_0(\psi)$ and $\pi_1(\theta, \psi)$ such that

$$\pi_1(\psi|\theta_0) = \pi_0(\psi)$$

then

$$B_{01} = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)}\,,$$

with the obvious notations

$$\pi_1(\theta) = \int \pi_1(\theta, \psi)d\psi\,, \quad \pi_1(\theta|x) = \int \pi_1(\theta, \psi|x)d\psi\,,$$

[Dickey, 1971; Verdinelli & Wasserman, 1995]

Measure-theoretic difficulty

The representation depends on the choice of versions of conditional densities:

$$
\begin{aligned}
B_{01} &= \frac{\int \pi_0(\psi) f(x|\theta_0,\psi)\, d\psi}{\int \pi_1(\theta,\psi) f(x|\theta,\psi)\, d\psi d\theta} && \text{[by definition]} \\
&= \frac{\int \pi_1(\psi|\theta_0) f(x|\theta_0,\psi)\, d\psi\, \pi_1(\theta_0)}{\int \pi_1(\theta,\psi) f(x|\theta,\psi)\, d\psi d\theta\, \pi_1(\theta_0)} && \text{[specific version of } \pi_1(\psi|\theta_0)\text{]} \\
&= \frac{\int \pi_1(\theta_0,\psi) f(x|\theta_0,\psi)\, d\psi}{m_1(x)\pi_1(\theta_0)} && \text{[specific version of } \pi_1(\theta_0,\psi)\text{]} \\
&= \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)}
\end{aligned}
$$

© Dickey's (1971) condition is not a condition

Similar measure-theoretic difficulty

Verdinelli-Wasserman extension:

$$B_{01} = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \, \mathbb{E}^{\pi_1(\psi|x,\theta_0,x)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)} \right]$$

depends on similar choices of versions

Monte Carlo implementation relies on continuous versions of all densities *without making mention of it*

[Chen, Shao & Ibrahim, 2000]

Computational implementation

Starting from the (new) prior

$$\tilde{\pi}_1(\theta, \psi) = \pi_1(\theta)\pi_0(\psi)$$

define the associated posterior

$$\tilde{\pi}_1(\theta, \psi|x) = \pi_0(\psi)\pi_1(\theta)f(x|\theta, \psi)/\tilde{m}_1(x)$$

and impose

$$\frac{\tilde{\pi}_1(\theta_0|x)}{\pi_0(\theta_0)} = \frac{\int \pi_0(\psi)f(x|\theta_0, \psi)\, d\psi}{\tilde{m}_1(x)}$$

to hold.
Then

$$B_{01} = \frac{\tilde{\pi}_1(\theta_0|x)}{\pi_1(\theta_0)} \frac{\tilde{m}_1(x)}{m_1(x)}$$

First ratio

If $(\theta^{(1)}, \psi^{(1)}), \ldots, (\theta^{(T)}, \psi^{(T)}) \sim \tilde{\pi}(\theta, \psi | x)$, then

$$\frac{1}{T} \sum_{t} \tilde{\pi}_1(\theta_0 | x, \psi^{(t)})$$

converges to $\tilde{\pi}_1(\theta_0|x)$ (if the right version is used in θ_0).
When $\tilde{\pi}_1(\theta_0|x, \psi$ unavailable, replace with

$$\frac{1}{T} \sum_{t=1}^{T} \tilde{\pi}_1(\theta_0 | x, z^{(t)}, \psi^{(t)})$$

Bridge revival (1)

Since $\tilde{m}_1(x)/m_1(x)$ is unknown, apparent failure!
Use of the identity

$$\mathbb{E}^{\tilde{\pi}_1(\theta,\psi|x)}\left[\frac{\pi_1(\theta,\psi)f(x|\theta,\psi)}{\pi_0(\psi)\pi_1(\theta)f(x|\theta,\psi)}\right] = \mathbb{E}^{\tilde{\pi}_1(\theta,\psi|x)}\left[\frac{\pi_1(\psi|\theta)}{\pi_0(\psi)}\right] = \frac{m_1(x)}{\tilde{m}_1(x)}$$

to (biasedly) estimate $\tilde{m}_1(x)/m_1(x)$ by

$$T/\sum_{t=1}^{T}\frac{\pi_1(\psi^{(t)}|\theta^{(t)})}{\pi_0(\psi^{(t)})}$$

based on the same sample from $\tilde{\pi}_1$.

Bridge revival (2)

Alternative identity

$$\mathbb{E}^{\pi_1(\theta,\psi|x)}\left[\frac{\pi_0(\psi)\pi_1(\theta)f(x|\theta,\psi)}{\pi_1(\theta,\psi)f(x|\theta,\psi)}\right] = \mathbb{E}^{\pi_1(\theta,\psi|x)}\left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta)}\right] = \frac{\tilde{m}_1(x)}{m_1(x)}$$

suggests using a second sample $(\bar{\theta}^{(1)},\bar{\psi}^{(1)},z^{(1)}),\ldots,$ $(\bar{\theta}^{(T)},\bar{\psi}^{(T)},z^{(T)}) \sim \pi_1(\theta,\psi|x)$ and

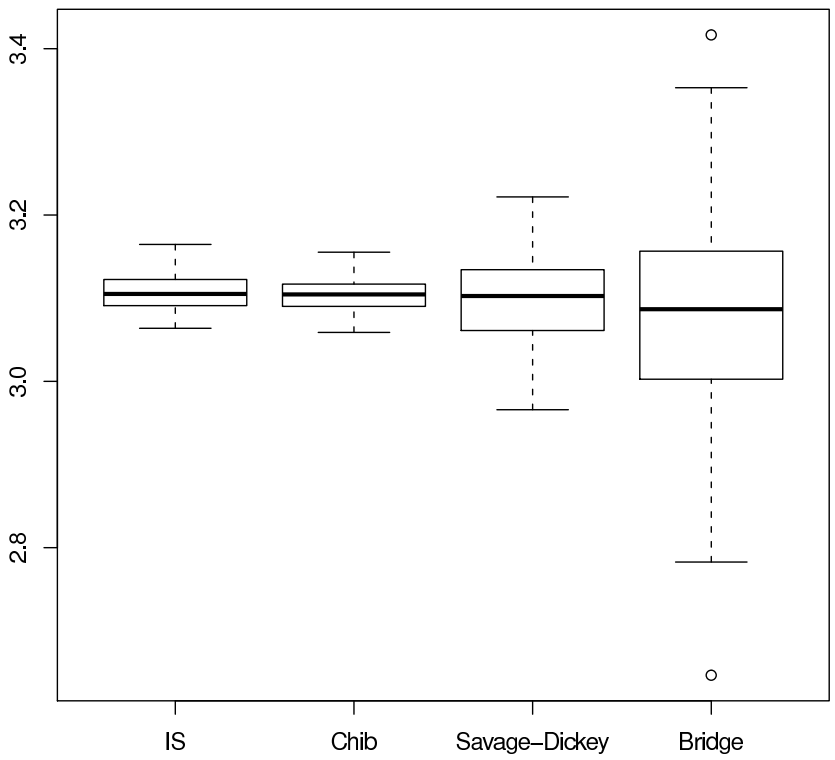
$$\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\bar{\psi}^{(t)})}{\pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})}$$

Resulting estimate:

$$\widehat{B_{01}} = \frac{1}{T}\frac{\sum_t \tilde{\pi}_1(\theta_0|x,z^{(t)},\psi^{(t)})}{\pi_1(\theta_0)}\,\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\bar{\psi}^{(t)})}{\pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})}$$

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations for a simulation from the above importance, Chib's, Savage–Dickey's and bridge samplers



Nested sampling: Goal

Skilling's (2007) technique using the one-dimensional representation:

$$\mathfrak{E} = \mathbb{E}^{\pi}[L(\theta)] = \int_0^1 \varphi(x)\,dx$$

with

$$\varphi^{-1}(l) = P^{\pi}(L(\theta) > l).$$

Note; $\varphi(\cdot)$ is intractable in most cases.

Nested sampling: First approximation

Approximate $\mathfrak{E}$ by a Riemann sum:

$$\widehat{\mathfrak{E}} = \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi(x_i)$$

where the x_i's are either:

- deterministic: $x_i = e^{-i/N}$
- or random:

$$x_0 = 1, \quad x_{i+1} = t_i x_i, \quad t_i \sim \mathcal{B}e(N, 1)$$

so that $\mathbb{E}[\log x_i] = -i/N$.

Extraneous white noise

Take

$$\mathfrak{E} = \int e^{-\theta}\, d\theta = \int \frac{1}{\delta}\, e^{-(1-\delta)\theta}\, e^{-\delta\theta} = \mathbb{E}_\delta \left[\frac{1}{\delta}\, e^{-(1-\delta)\theta} \right]$$

$$\hat{\mathfrak{E}} = \frac{1}{N} \sum_{i=1}^{N} \delta^{-1}\, e^{-(1-\delta)\theta_i} (x_{i-1} - x_i)\,, \quad \theta_i \sim \mathcal{E}(\delta)\, \mathbb{I}(\theta_i \leq \theta_{i-1})$$

N	deterministic	random
50	4.64	10.5
	4.65	10.5
100	2.47	4.9
	2.48	5.02
500	.549	1.01
	.550	1.14

Comparison of variances and MSEs

Nested sampling: Second approximation

Replace (intractable) $\varphi(x_i)$ by φ_i, obtained by

Nested sampling

Start with N values $\theta_1, \ldots, \theta_N$ sampled from π
At iteration i,

1. Take $\varphi_i = L(\theta_k)$, where θ_k is the point with smallest likelihood in the pool of θ_i's
2. Replace θ_k with a sample from the prior *constrained to* $L(\theta) > \varphi_i$: the current N points are sampled from *prior constrained to* $L(\theta) > \varphi_i$.

Nested sampling: Third approximation

Iterate the above steps until a given stopping iteration j is reached: e.g.,

- observe very small changes in the approximation $\widehat{\mathfrak{Z}}$;
- reach the maximal value of $L(\theta)$ when the likelihood is bounded and its maximum is known;
- truncate the integral $\mathfrak{E}$ at level ϵ, i.e. replace

$$\int_0^1 \varphi(x)\,dx \qquad \text{with} \qquad \int_\epsilon^1 \varphi(x)\,dx$$

Approximation error

$$
\begin{aligned}
\text{Error} &= \widehat{\mathfrak{E}} - \mathfrak{E} \\
&= \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i - \int_0^1 \varphi(x)\,dx = -\int_0^{\epsilon} \varphi(x)\,dx \\
&+ \left[\sum_{i=1}^{j} (x_{i-1} - x_i)\varphi(x_i) - \int_{\epsilon}^1 \varphi(x)\,dx \right] \qquad \text{(Quadrature Error)} \\
&+ \left[\sum_{i=1}^{j} (x_{i-1} - x_i)\{\varphi_i - \varphi(x_i)\} \right] \qquad \text{(Stochastic Error)}
\end{aligned}
$$

[Dominated by Monte Carlo!]

A CLT for the Stochastic Error

The (dominating) stochastic error is $O_P(N^{-1/2})$:

$$N^{1/2}\{\text{Stochastic Error}\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$$

with

$$V = -\int_{s,t\in[\epsilon,1]} s\varphi'(s)t\varphi'(t)\log(s \vee t)\,ds\,dt.$$

[Proof based on Donsker's theorem]

The number of simulated points equals the number of iterations j, and is a *multiple* of N: if one stops at first iteration j such that $e^{-j/N} < \epsilon$, then: $j = N\lceil -\log \epsilon \rceil$.

Curse of dimension

For a simple Gaussian-Gaussian model of dimension $\dim(\theta) = \mathrm{d}$, the following 3 quantities are $\mathrm{O}(\mathrm{d})$:

1. asymptotic variance of the NS estimator;
2. number of iterations (necessary to reach a given truncation error);
3. cost of one simulated sample.

Therefore, CPU time necessary for achieving error level e is

$$\mathrm{O}(\mathrm{d}^3/e^2)$$

Sampling from constr'd priors

Exact simulation from the constrained prior is *intractable* in most cases!

Skilling (2007) proposes to use MCMC, but:

- this introduces a bias (stopping rule).
- if MCMC stationary distribution is unconst'd prior, more and more difficult to sample points such that $\mathrm{L}(\theta) > \mathrm{l}$ as l increases.

If implementable, then *slice sampler* can be devised at the same cost!

A IS variant of nested sampling

Consider *instrumental* prior $\widetilde{\pi}$ and likelihood $\tilde{L}$, weight function

$$w(\theta) = \frac{\pi(\theta)L(\theta)}{\widetilde{\pi}(\theta)\widetilde{L}(\theta)}$$

and weighted NS estimator

$$\widehat{\mathfrak{E}} = \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i w(\theta_i).$$

Then choose $(\widetilde{\pi}, \widetilde{L})$ so that sampling from $\widetilde{\pi}$ constrained to $\widetilde{L}(\theta) > l$ is easy; e.g. $\mathcal{N}(c, I_d)$ constrained to $\|c - \theta\| < r$.

Benchmark: Target distribution

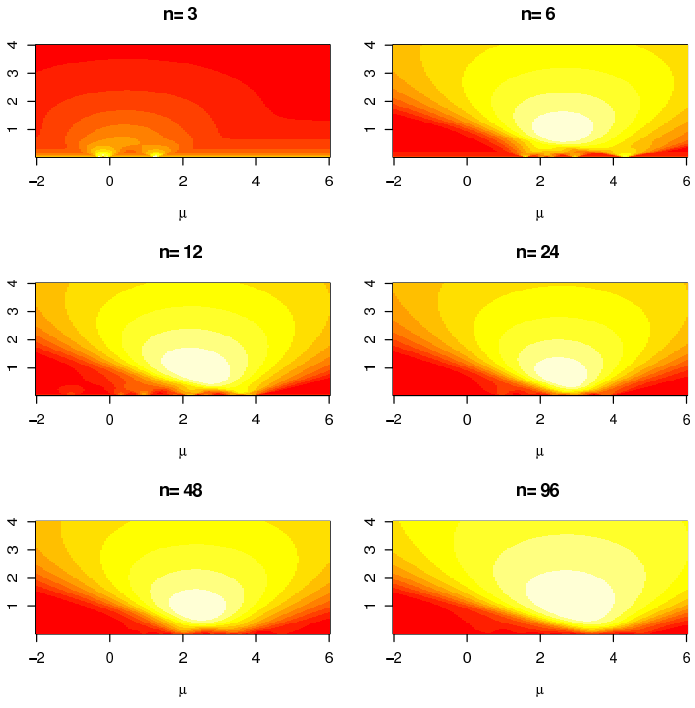
Posterior distribution on (μ, σ) associated with the mixture

$$p\mathcal{N}(0,1) + (1-p)\mathcal{N}(\mu, \sigma)\,,$$

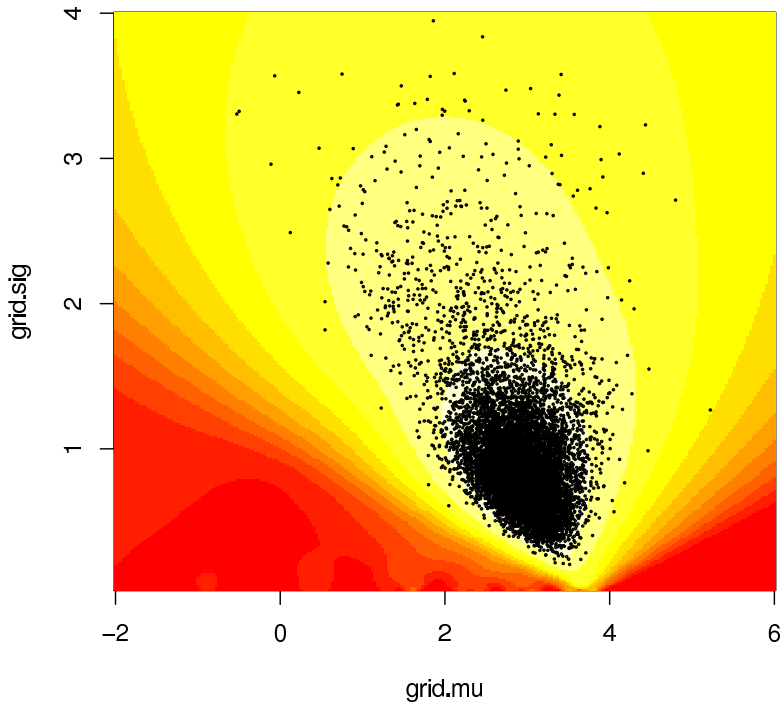
when p is known

Experiment

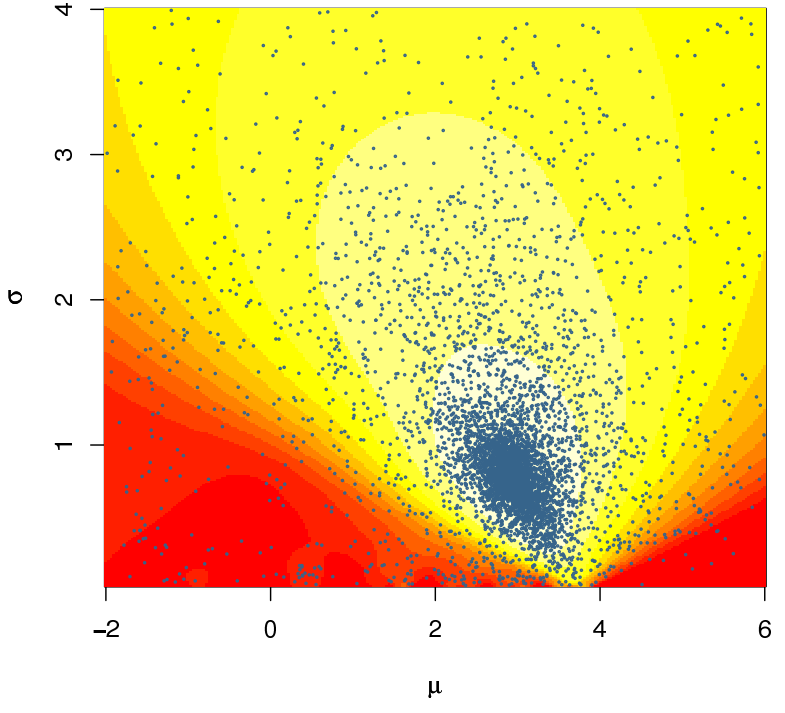
- n observations with $\mu = 2$ and $\sigma = 3/2$,
- Use of a uniform prior both on $(-2, 6)$ for μ and on $(.001, 16)$ for $\log \sigma^2$.
- occurrences of posterior bursts for $\mu = x_i$
- computation of the various estimates of $\mathfrak{E}$



Experiment (cont'd)

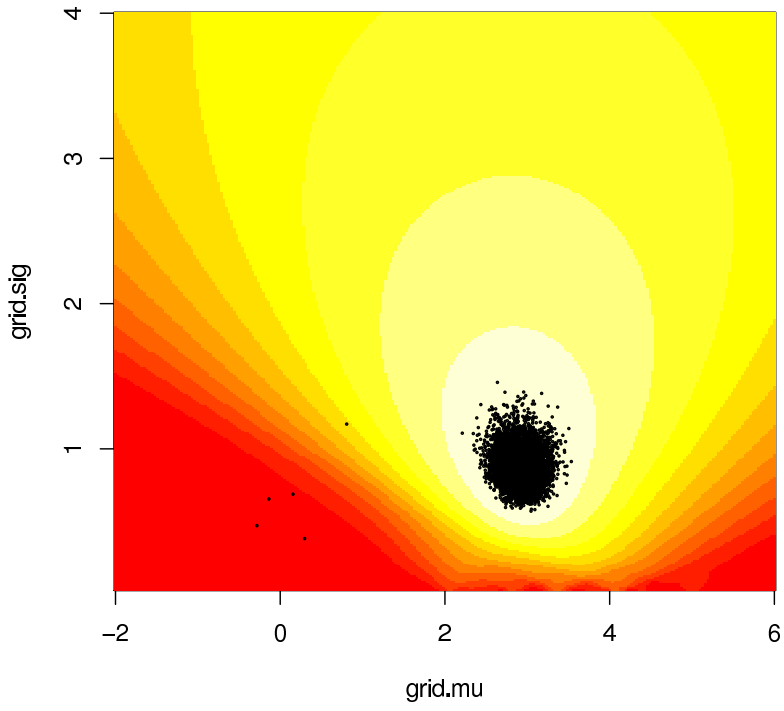


MCMC sample for $n = 16$ observations from the mixture.

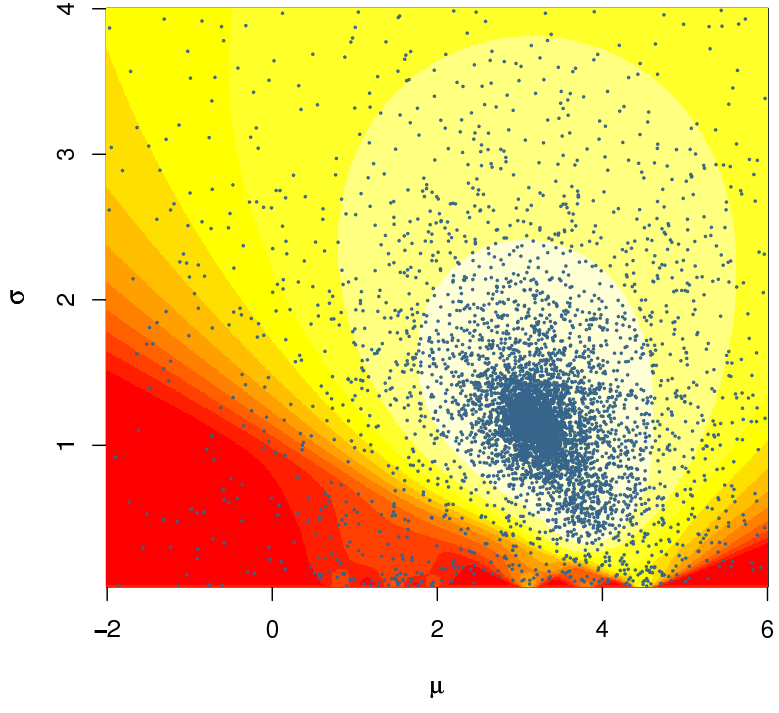


Nested sampling sequence with $M = 1000$ starting points.

Experiment (cont'd)



MCMC sample for $\mathrm{n} = 50$ observations from the mixture.

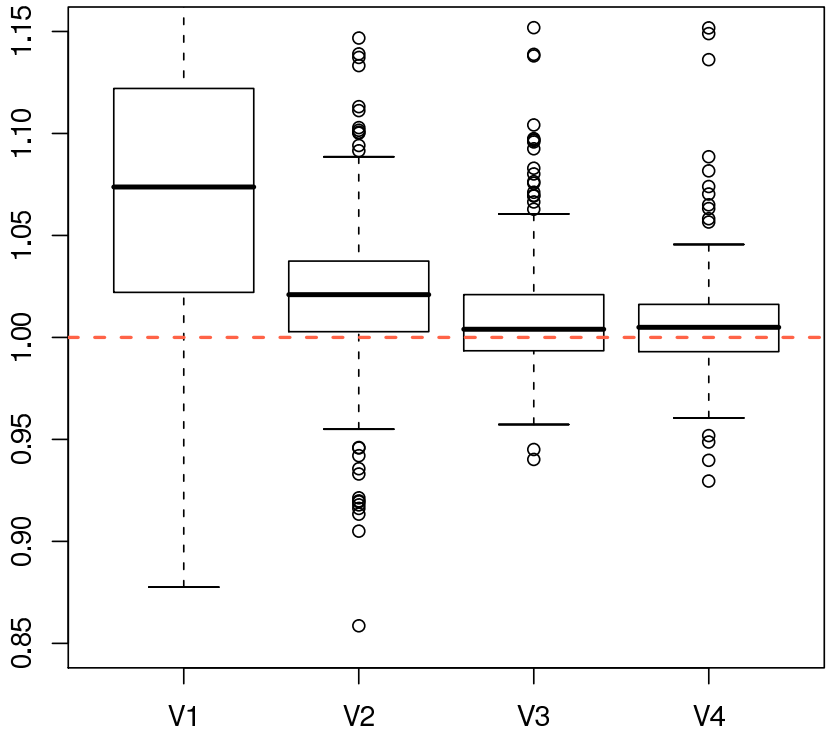


Nested sampling sequence with $\mathrm{M} = 1000$ starting points.

Comparison

Monte Carlo and MCMC (=Gibbs) outputs based on $T = 10^4$ simulations and numerical integration based on a 850×950 grid in the (μ, σ) parameter space.
Nested sampling approximation based on a starting sample of $M = 1000$ points followed by at least 103 further simulations from the constr'd prior and a stopping rule at 95% of the observed maximum likelihood.
Constr'd prior simulation based on 50 values simulated by random walk accepting only steps leading to a lik'hood higher than the bound

Comparison (cont'd)



Graph based on a sample of 10 observations for $\mu = 2$ and $\sigma = 3/2$ (150 replicas).

Comparison (cont'd)

Nested sampling gets less reliable as sample size increases
Most reliable approach is mixture $\widehat{\mathfrak{E}}_3$ although harmonic solution $\widehat{\mathfrak{E}}_1$ close to Chib's solution [taken as golden standard]
Monte Carlo method $\widehat{\mathfrak{E}}_2$ also producing poor approximations to $\mathfrak{E}$
(Kernel ϕ used in $\widehat{\mathfrak{E}}_2$ is a t non-parametric kernel estimate with standard bandwidth estimation.)