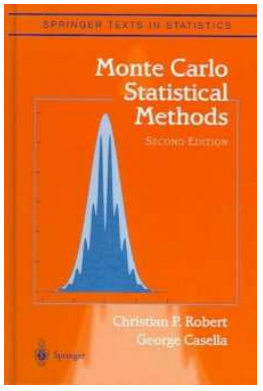
Markov Chain Monte Carlo Methods

Christian P. Robert

Université Paris-Dauphine, & University of Warwick
http://www.ceremade.dauphine.fr/~xian

February 29, 2016

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

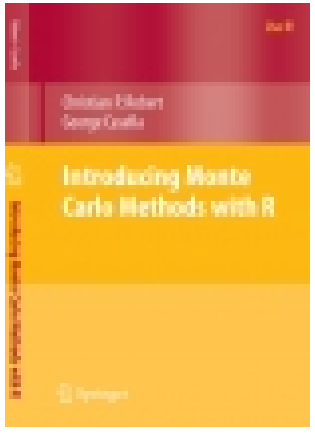


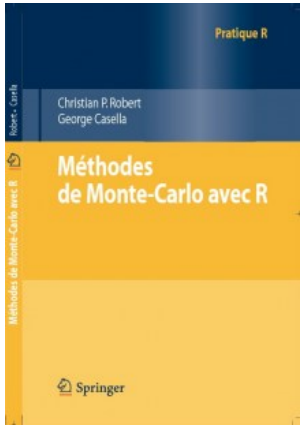
Slides: older slides on

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

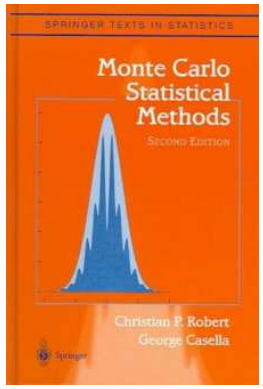
Suggested reading

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





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

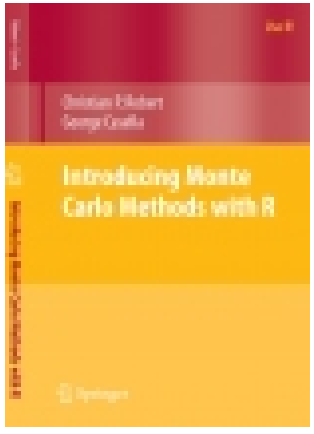


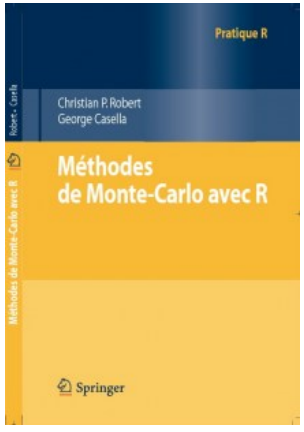
Slides: older slides on

`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

MCMC #1: The Metropolis-Hastings Algorithm

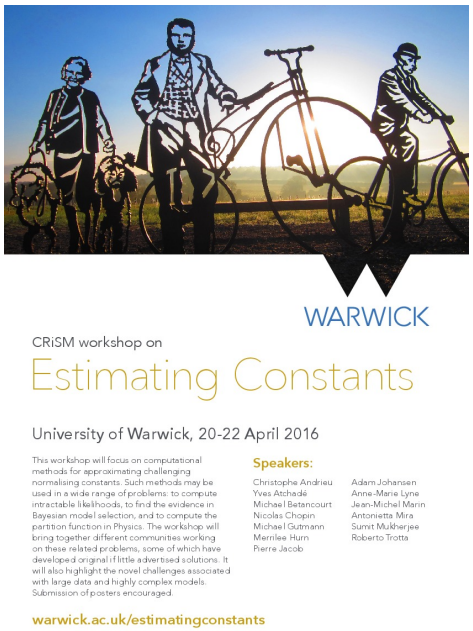
MCMC # 2: Gibbs Sampling

MCMC # 3: Sequential Monte Carlo

MCMC # 4: New directions

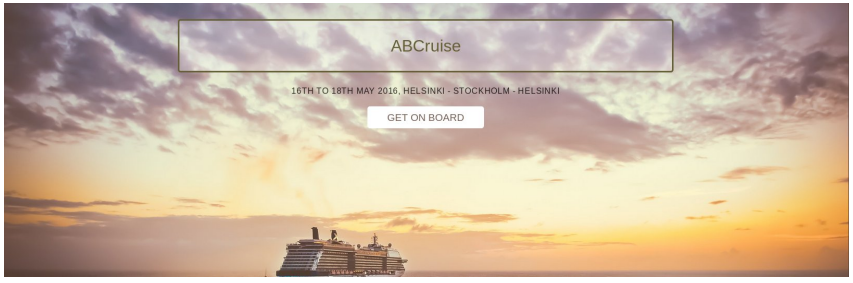
Estimating Constants

- CRiSM Workshop on Estimating Constants



- 20 - 22 April 2016, University of Warwick
- Monte Carlo methods for approximating normalising constants
- 12 plenary speakers
- Call for posters
- cheap registration and loding
- "The Midlands in April, where else...?!"

ABCruise: ABC in Helsinki



- New "ABC in..." workshop between Helsinki and Stockholm
- 16–18 May 2016, on Silja Symphony of Tallink Silja Line
- recent advances in ABC theory and methodology
- 10 plenary speakers
- call for posters
- very cheap registration and loding and meals for 200EUR
- "The Baltic in May, where else...?!"

The Metropolis-Hastings Algorithm

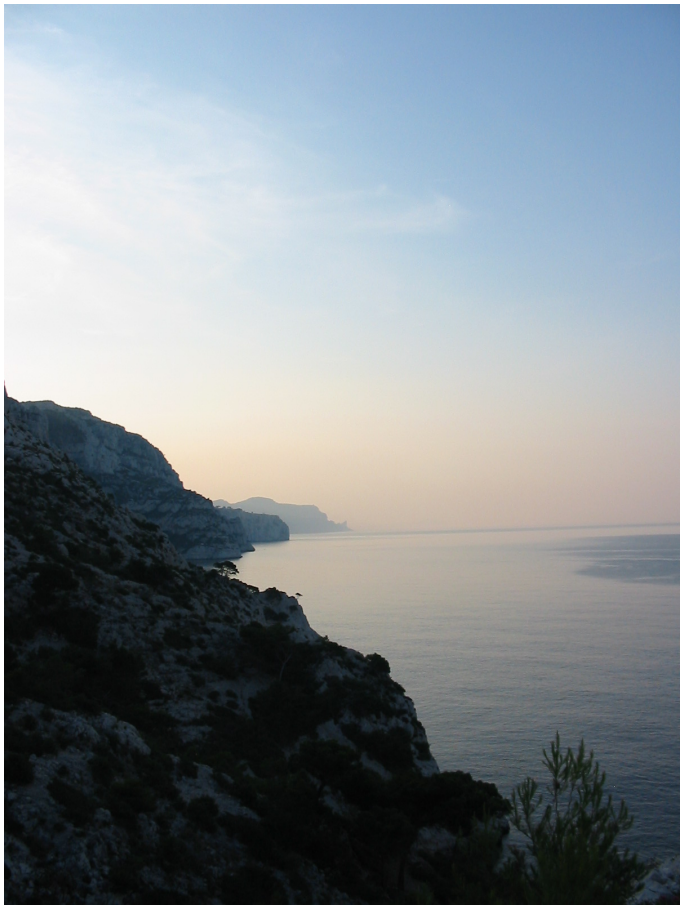
MCMC #1: The Metropolis-Hastings Algorithm

- Monte Carlo Methods based on Markov Chains
- The Metropolis–Hastings algorithm
- A collection of Metropolis-Hastings algorithms
- Extensions

MCMC # 2: Gibbs Sampling

MCMC # 3: Sequential Monte Carlo

MCMC # 4: New directions



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

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

Running Monte Carlo via Markov Chains (2)

Idea

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

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

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

Running Monte Carlo via Markov Chains (2)

Idea

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

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

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

Running Monte Carlo via Markov Chains (2)

Idea

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

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

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**

$$f$$

and a conditional density

$$q(y|x)$$

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

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

Convergence properties

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

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

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

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

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

Convergence properties

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

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

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

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

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

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.

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.

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.

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

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

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

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

In this case,

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

[Mengersen & Tweedie, 1996]

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

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

In this case,

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

[Mengersen & Tweedie, 1996]

Independent MH: illustration

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

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

Independent MH: illustration

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

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

Independent MH: illustration

Example (Noisy AR(1))

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

Independent MH: illustration

Example (Noisy AR(1))

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

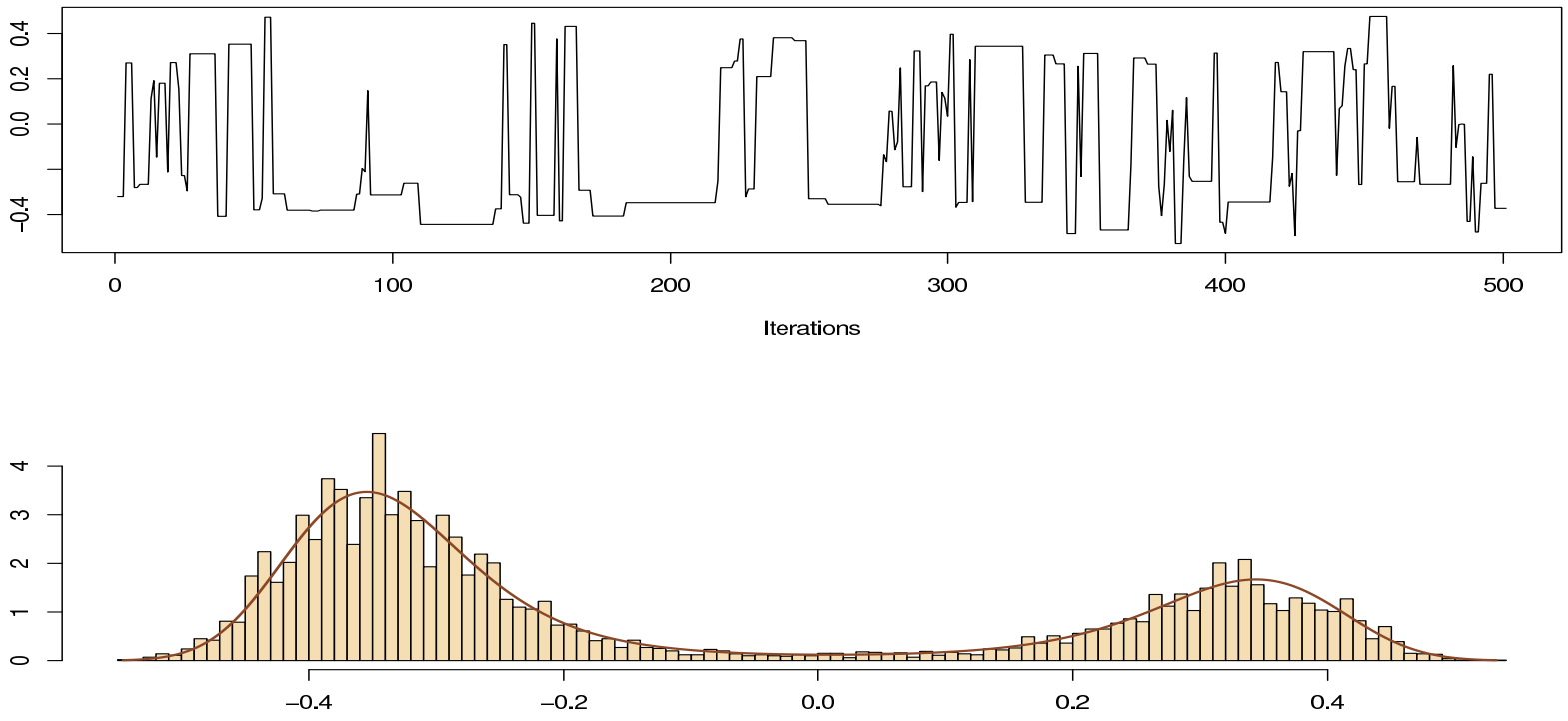
$$\mu_t = \varphi \frac{x_{t-1} + x_{t+1}}{1 + \varphi^2} \quad \text{and} \quad \omega_t^2 = \frac{\tau^2}{1 + \varphi^2}\,.$$

Ratio

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

is bounded

Independent MH: illustration



(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.

Independent MH: counterexample

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

[Mengersen & Tweedie, 1996]

Independent MH: counterexample

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

[Mengersen & Tweedie, 1996]

Independent MH: counterexample

Example (Cauchy by normal)

go random W Given a Cauchy $\mathscr{C}(0,1)$ distribution, consider a normal $\mathscr{N}(0,1)$ proposal

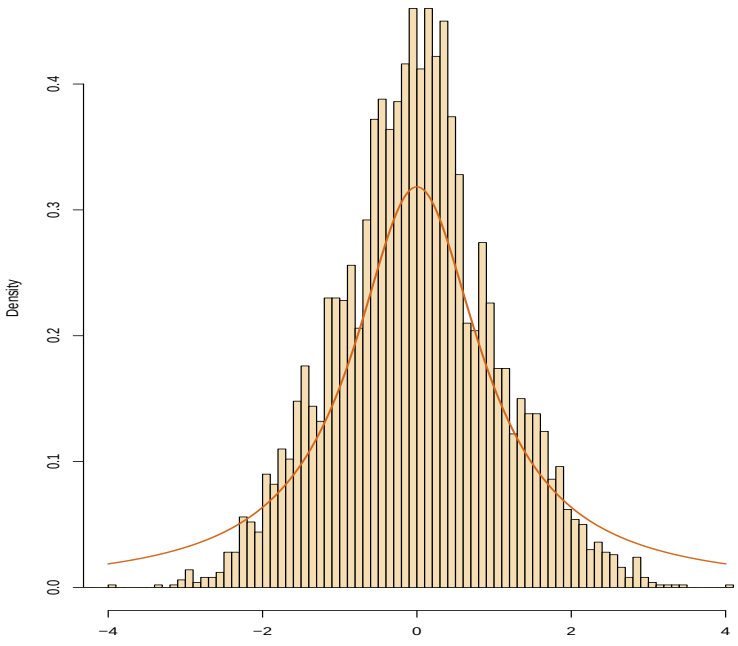
The Metropolis–Hastings acceptance ratio is

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

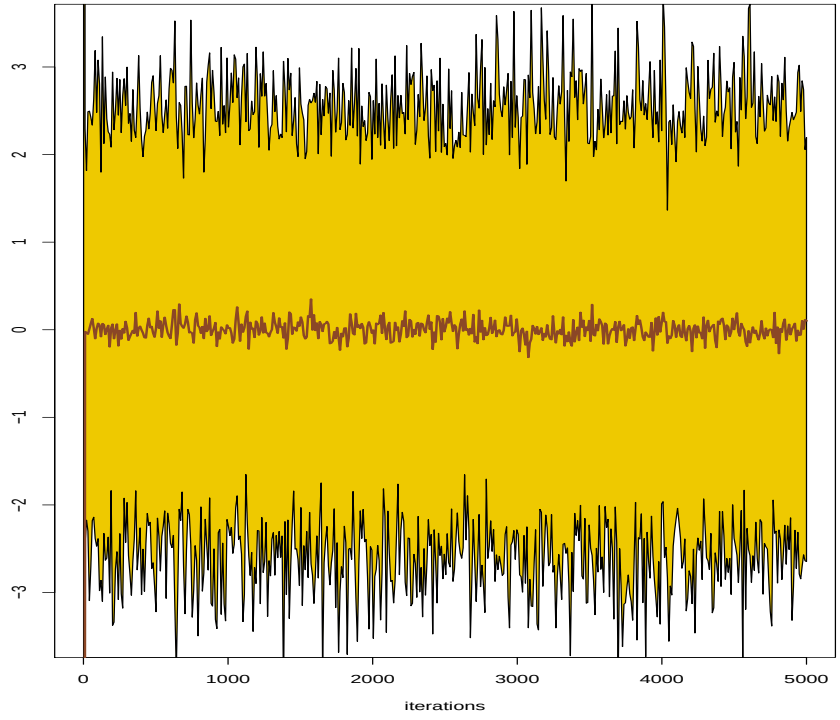
Poor performances: the proposal distribution has lighter tails than the target Cauchy and convergence to the stationary distribution is not even geometric!

[Mengersen & Tweedie, 1996]

Independent MH: counterexample



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.

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* $g(x) = g(-x)$

Random walk Metropolis–Hastings

Algorithm (Random walk Metropolis)

Given $x^{(t)}$

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

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

Illustration

Example (A toy example)

perturbed version of the normal $\mathcal{N}(0,1)$ density, $\varphi(\cdot)$

$$\tilde{\pi}(x) \propto \sin^2(x) \times \sin^2(2x) \times \varphi(x)\,.$$

And proposal uniform $\mathcal{U}(x-\alpha, x+\alpha)$ kernel

Metropolis-Hastings step:

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

where

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

and α scaled for good acceptance rate

Illustration

Example (A toy example)

perturbed version of the normal $\mathcal{N}(0,1)$ density, $\varphi(\cdot)$

$$\tilde{\pi}(x) \propto \sin^2(x) \times \sin^2(2x) \times \varphi(x)\,.$$

And proposal uniform $\mathcal{U}(x-\alpha, x+\alpha)$ kernel
Metropolis-Hastings step:

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

where

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

and α scaled for good acceptance rate

Illustration

R code

```
target=function(x){
  sin(x)^2*sin(2*x)^2*dnorm(x)}

metropolis=function(x,alpha=1){
  y=runif(1,x-alpha,x+alpha)
  if (runif(1)>target(y)/target(x)) y=x
  return(y)}
```

with arbitrary starting value

```
T=10^4
x=rep(3.14,T)
for (t in 2:T) x[t]=metropolis(x[t-1])
```

Illustration

R code

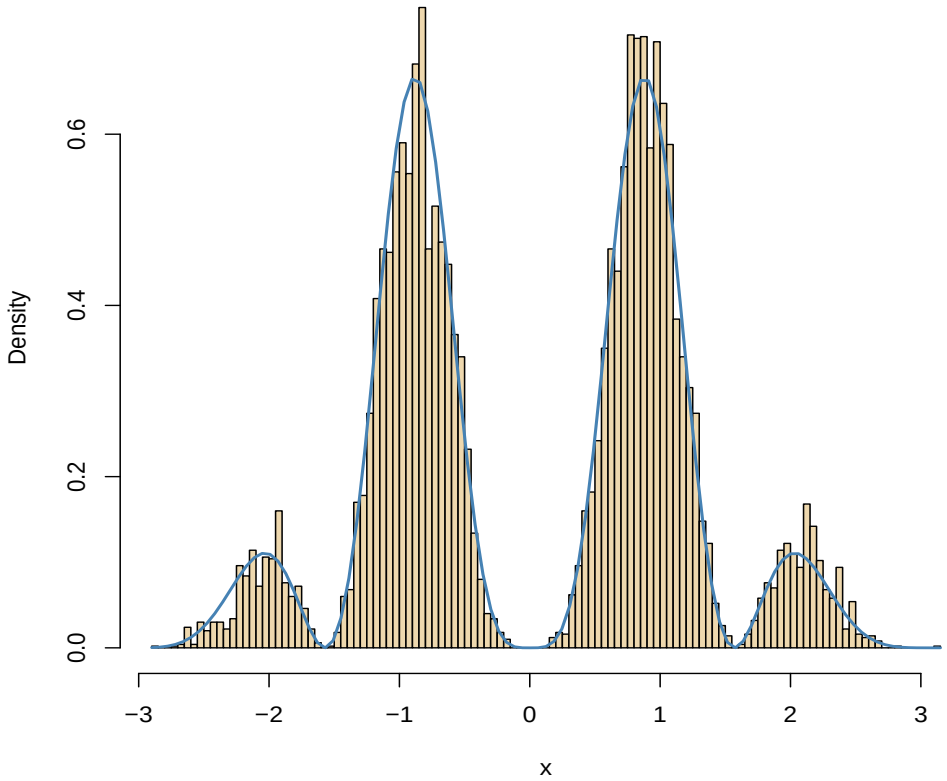
```
target=function(x){
  sin(x)^2*sin(2*x)^2*dnorm(x)}

metropolis=function(x,alpha=1){
  y=runif(1,x-alpha,x+alpha)
  if (runif(1)>target(y)/target(x)) y=x
  return(y)}
```

with arbitrary starting value

```
T=10^4
x=rep(3.14,T)
for (t in 2:T) x[t]=metropolis(x[t-1])
```

Illustration



$T = 10^4,\ \alpha = 1,\ x^{(1)} = \pi$

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

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.

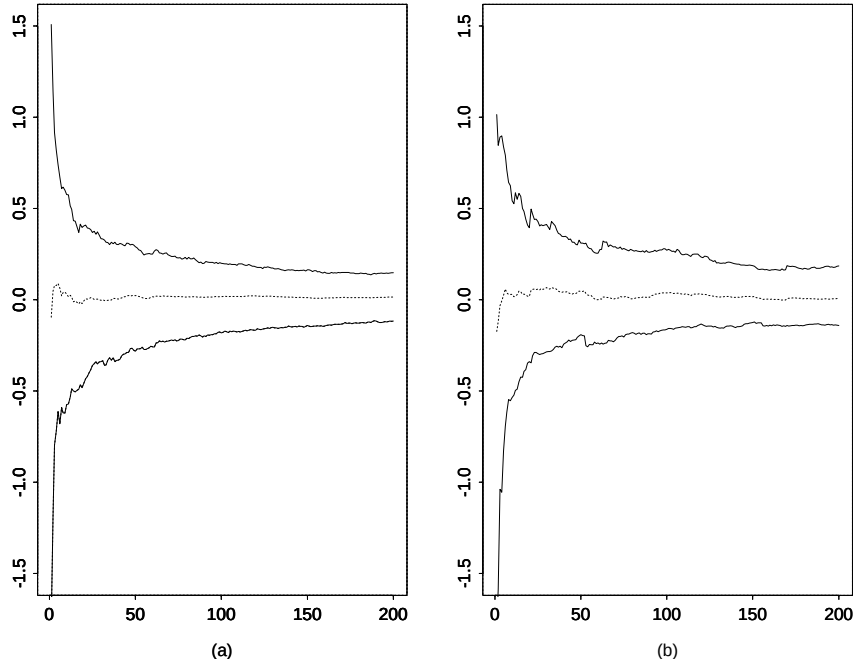
[Mengersen & Tweedie, 1996]

▸ no tail effect

Convergence properties

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

Convergence properties

Example (Cauchy by normal continued)

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

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

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

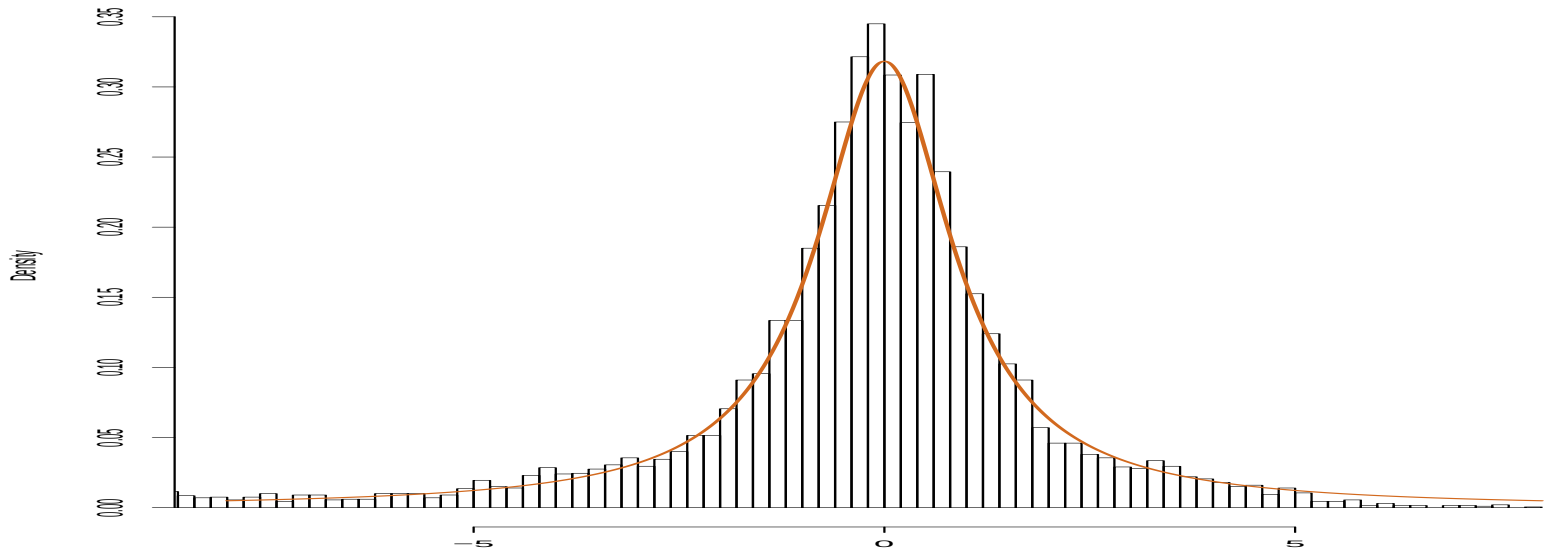
[Roberts & Tweedie, 2004]

Convergence properties

Again, lack of geometric ergodicity!

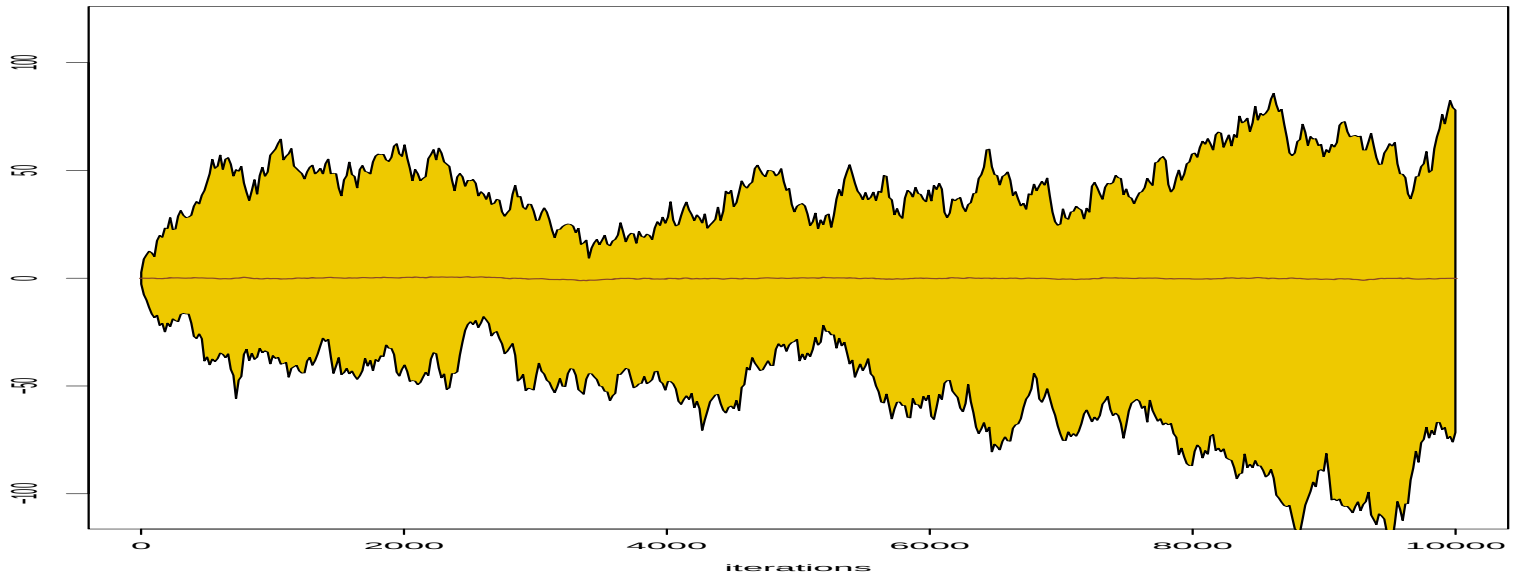
[Mengersen & Tweedie, 1996]

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



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

Convergence properties



Range of 500 parallel runs for the same setup

Further convergence properties

Under assumptions

skip detailed convergence

- **(A1)** f is super-exponential, *i.e.* it is positive with positive continuous first derivative such that $\lim_{|x|\to\infty} n(x)'\nabla \log f(x) = -\infty$ where $n(x) := x/|x|$.
  In words : exponential decay of f in every direction with rate tending to ∞
- **(A2)** $\limsup_{|x|\to\infty} n(x)'m(x) < 0$, where $m(x) = \nabla f(x)/|\nabla f(x)|$.
  In words: non degeneracy of the countour manifold $\mathcal{C}_{f(y)} = \{y : f(y) = f(x)\}$

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

[Jarner & Hansen, 2000]

Further [further] convergence properties

skip hyperdetailed convergence

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

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

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

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

Further [further] convergence properties

skip hyperdetailed convergence

then,

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

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

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

[Tuominen & Tweedie, 1994]

Comments

- **[CLT, Rosenthal's inequality...]** h-ergodicity implies CLT for additive (possibly unbounded) functionals of the chain, Rosenthal's inequality and so on...
- **[Control of the moments of the return-time]** The condition implies (because $h \geq 1$) that

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

where $r_0(n) = \sum_{l=0}^{n} r(l)$ Can be used to derive bounds for the coupling time, an essential step to determine computable bounds, using coupling inequalities

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

Alternative conditions

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

[Possible alternative conditions]

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

- (i) $\sup_C V_0 < \infty$,
- **(ii)** $\{V_0 = \infty\} \subset \{V_1 = \infty\}$ and
- **(iii)** $PV_{n+1} \leq V_n - r(n)h + br(n)\mathbb{I}_C$.

Alternative conditions

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

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

where σ_C is the hitting time on C and

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

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

Extensions

There are many other families of HM algorithms

- *Adaptive Rejection Metropolis Sampling*
- *Reversible Jump*
- *Langevin algorithms*
- *Hamiltonian MC*

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

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

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

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

MH correction (MALA)

Accept the new value Y_t with probability

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

Choice of the scaling factor σ

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

[Roberts & Rosenthal, 1998]

Optimizing the Acceptance Rate

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

Most common alternatives:

(a) a fully automated algorithm like ARMS;

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

(c) a random walk

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

Case of the independent Metropolis–Hastings algorithm

Choice of g that maximizes the average acceptance rate

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

Related to the speed of convergence of

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

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

Case of the independent Metropolis–Hastings algorithm

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.

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

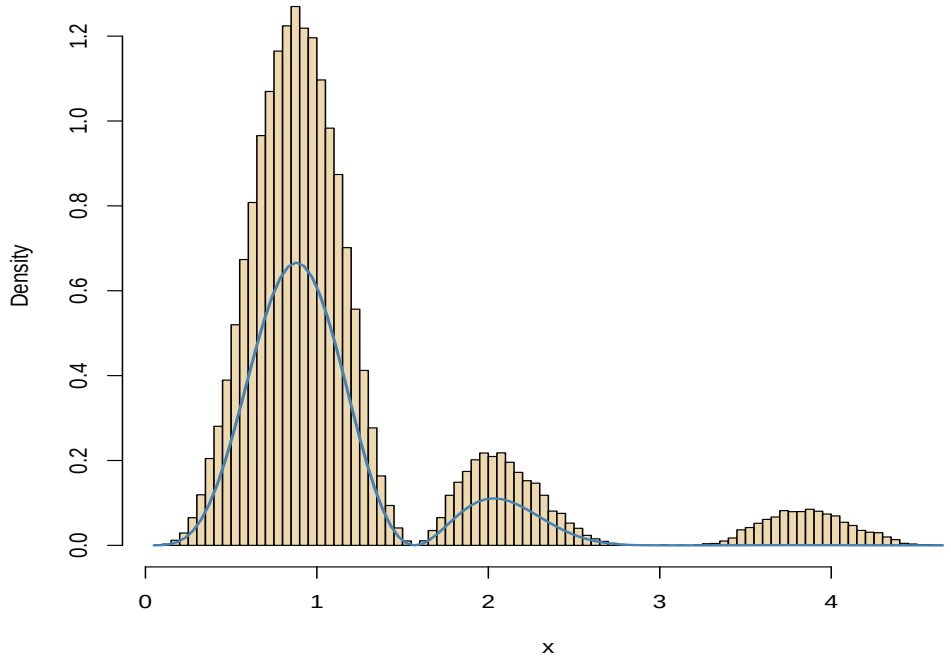
Different approach to acceptance rates
If the average acceptance rate is low, the successive values of $f(y_t)$ tend to be small compared with $f(x^{(t)})$, which means that the random walk moves quickly on the surface of f since it often reaches the "borders" of the support of f

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]

Remark: Equivalent goal for MALA



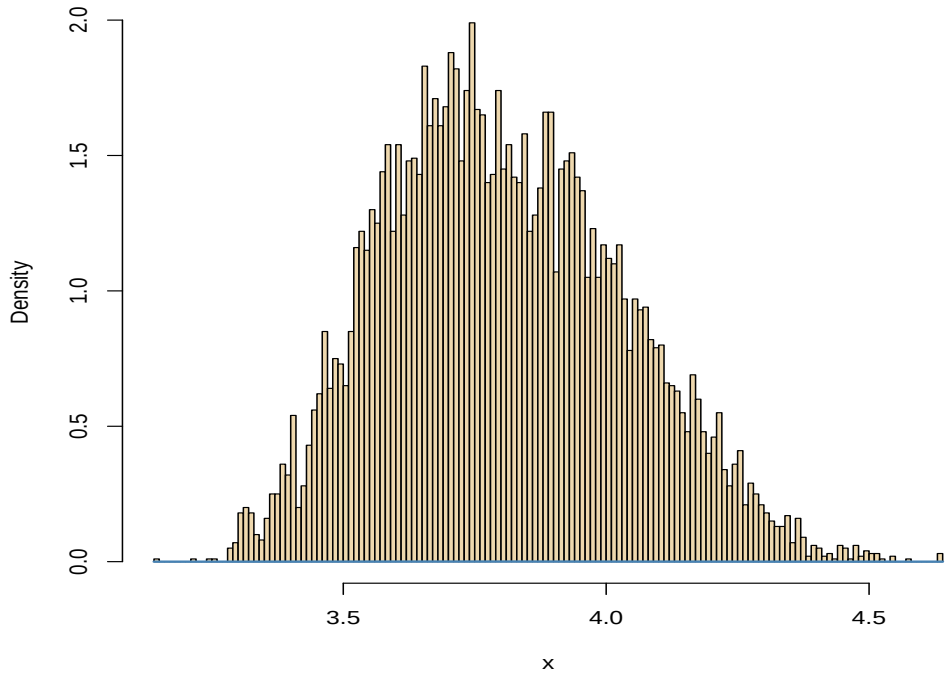
$\alpha = 0.25$

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]

Remark: Equivalent goal for MALA

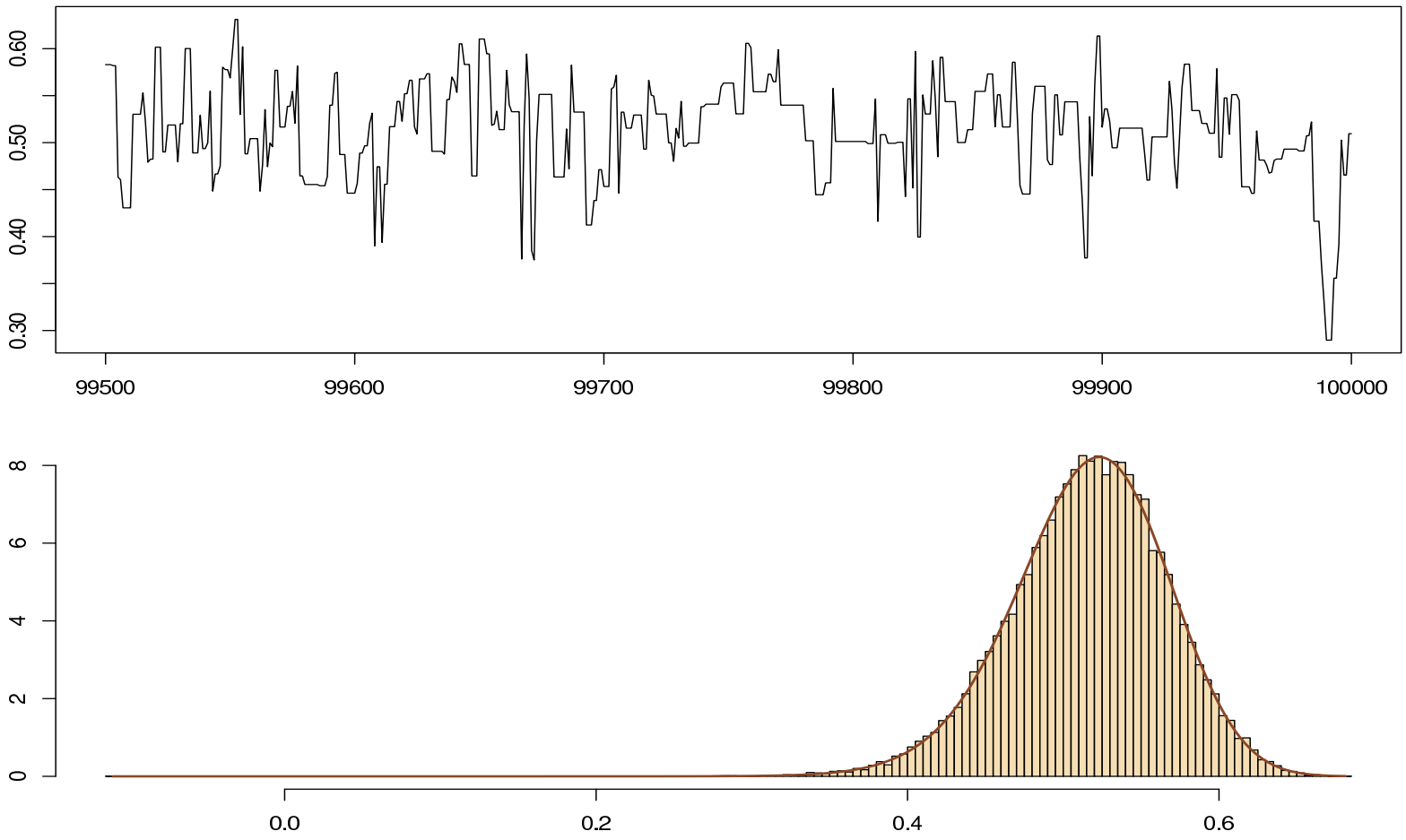


$\alpha = 0.1$

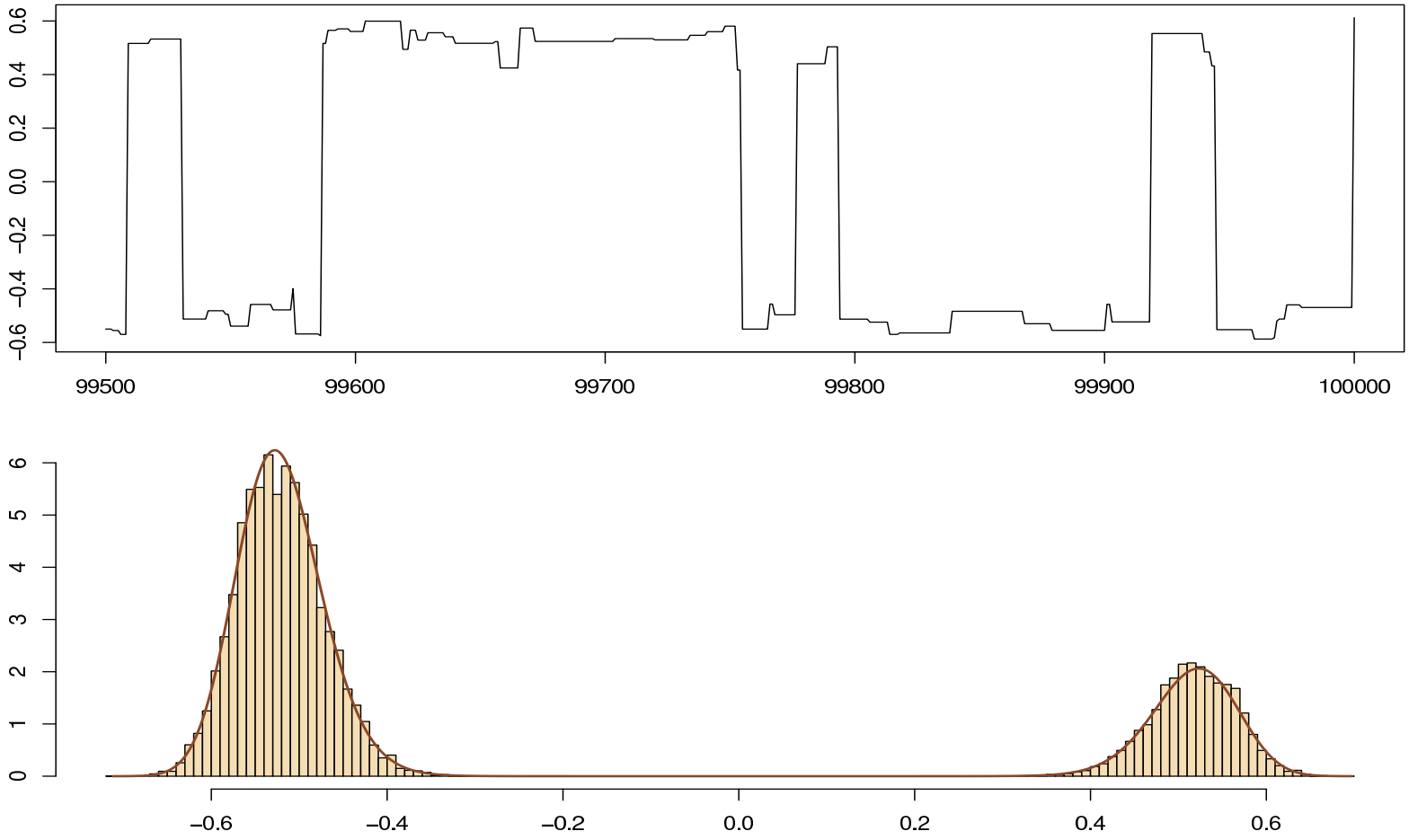
Illustration

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 $\omega = .1$.

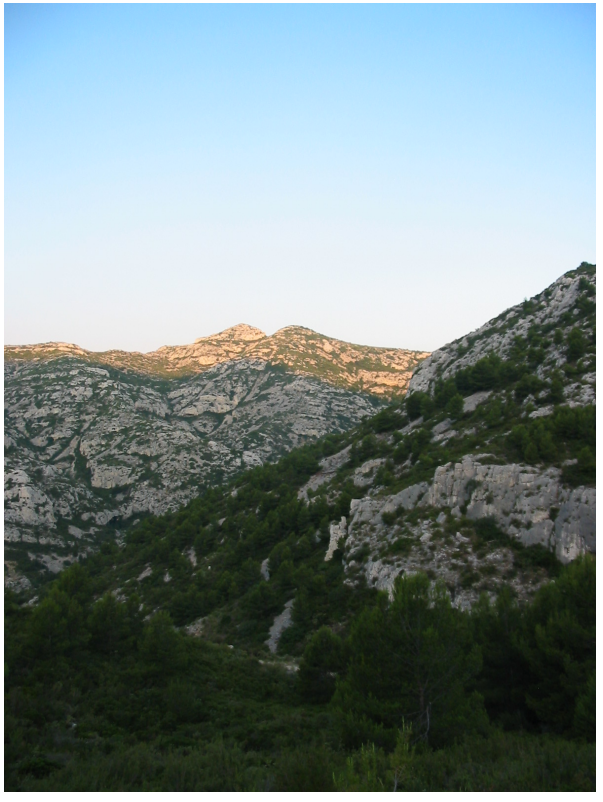


Markov chain based on a random walk with scale $\omega = .5$.

The Gibbs Sampler

MCMC # 2: Gibbs Sampling

- General Principles
- Completion
- Convergence
- The Hammersley-Clifford theorem
- Improper Priors



General Principles

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

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

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

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

General Principles

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

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

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

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

General Principles

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

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

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

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

General Principles

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

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

Properties

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

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

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

Properties

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

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

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

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...

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

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

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)

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...

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

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

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

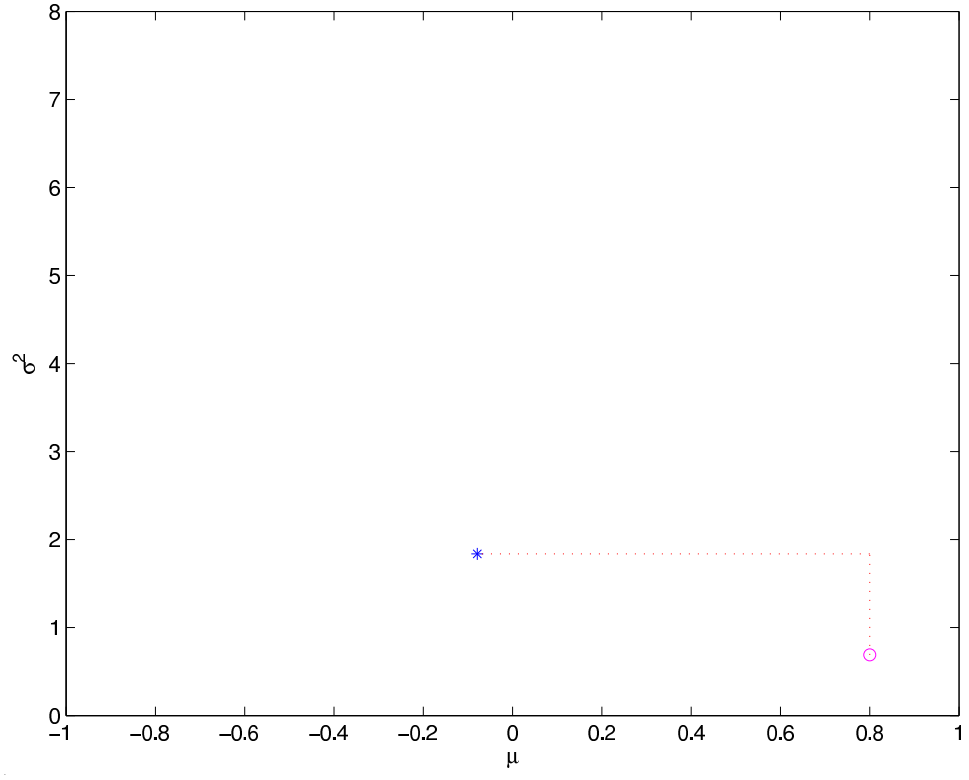
- Hence we may use the Gibbs sampler for simulating from the posterior of (μ, σ^2)

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

R Gibbs Sampler for Gaussian posterior

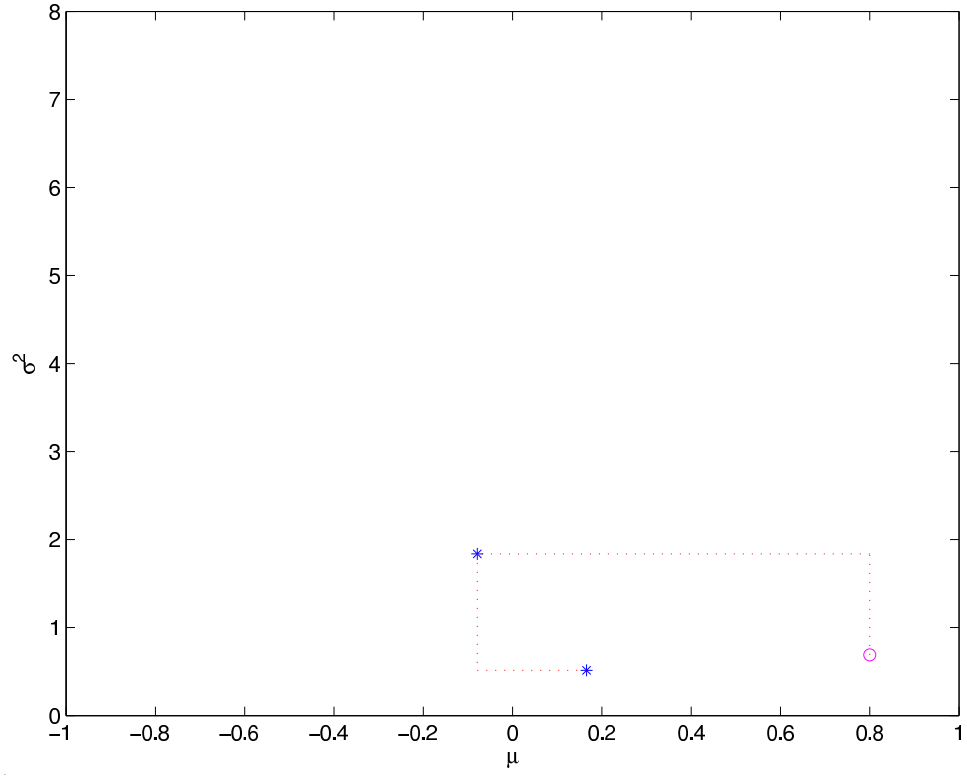
```
n = length(Y);
S = sum(Y);
mu = S/n;
for (i in 1:500)
    S2 = sum((Y-mu)^2);
    sigma2 = 1/rgamma(1,n/2-1,S2/2);
    mu = S/n + sqrt(sigma2/n)*rnorm(1);
```

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



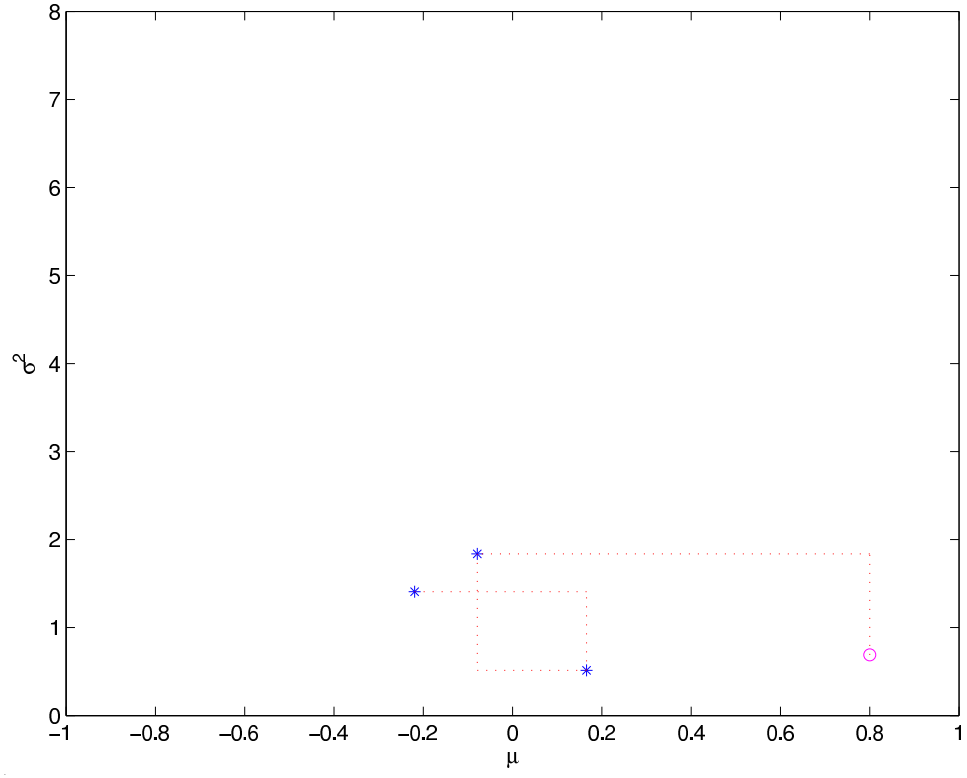
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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



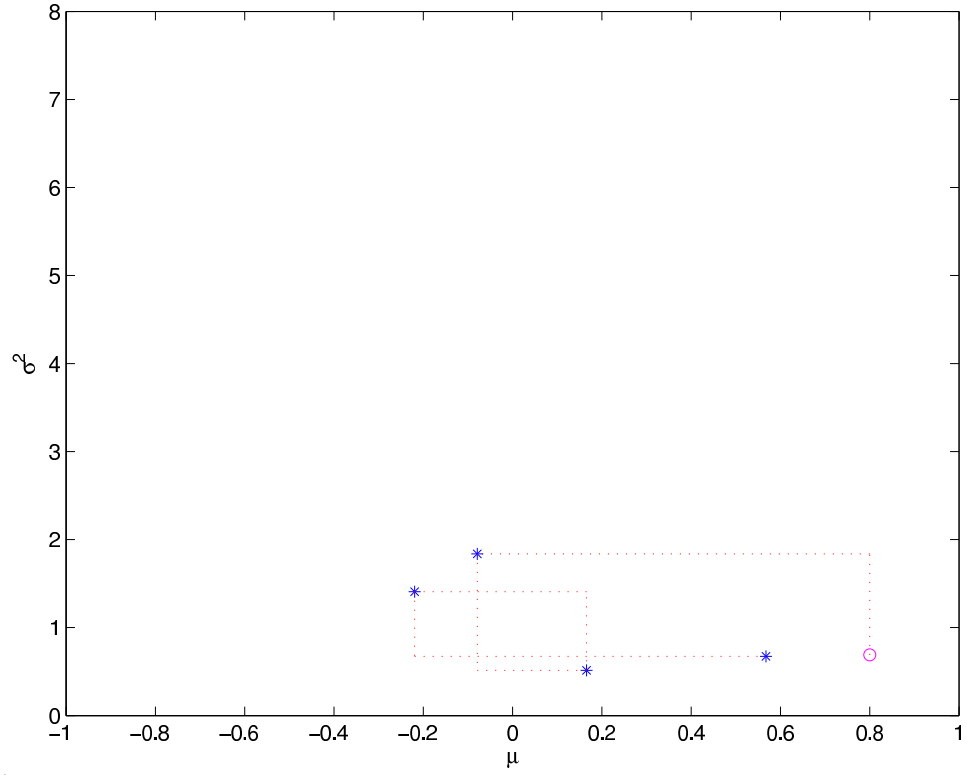
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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



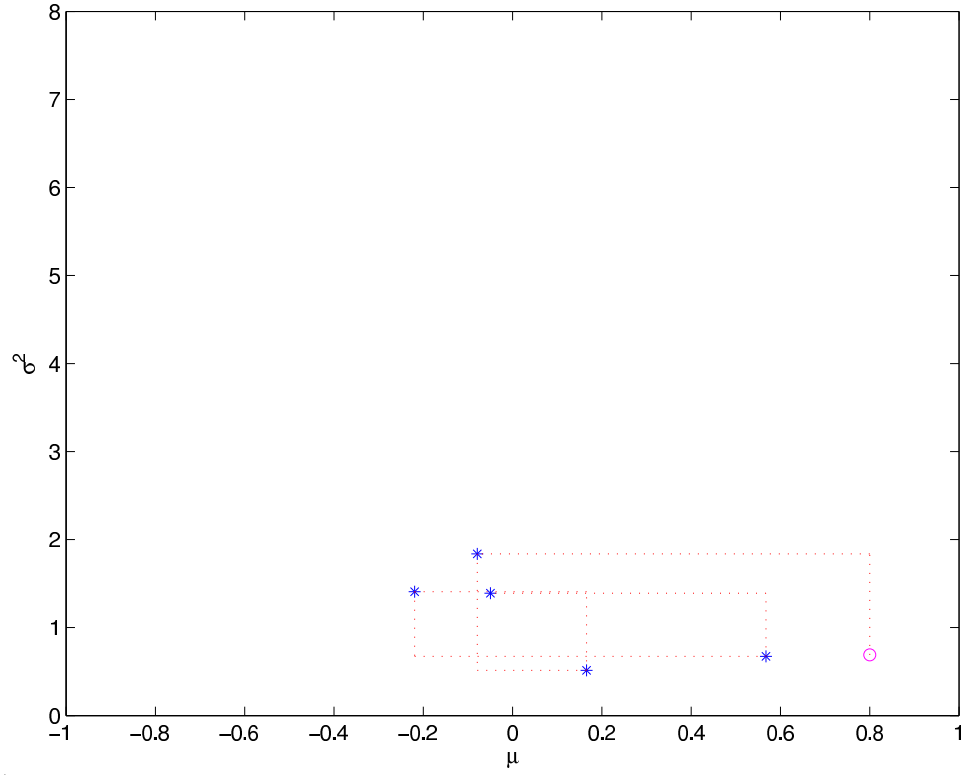
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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



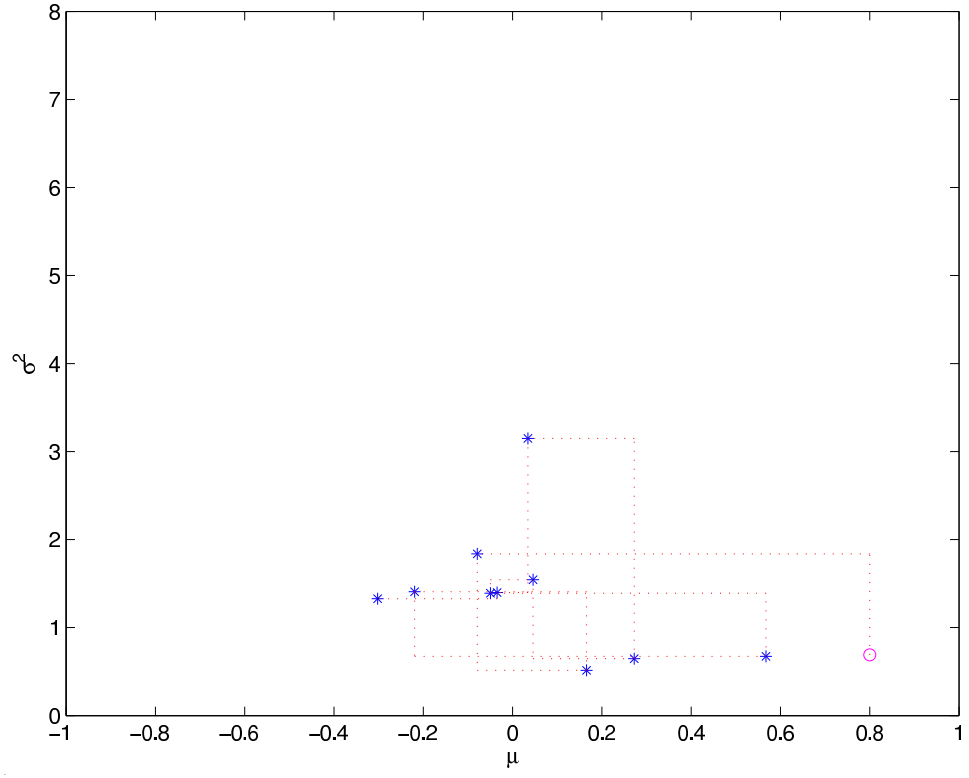
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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



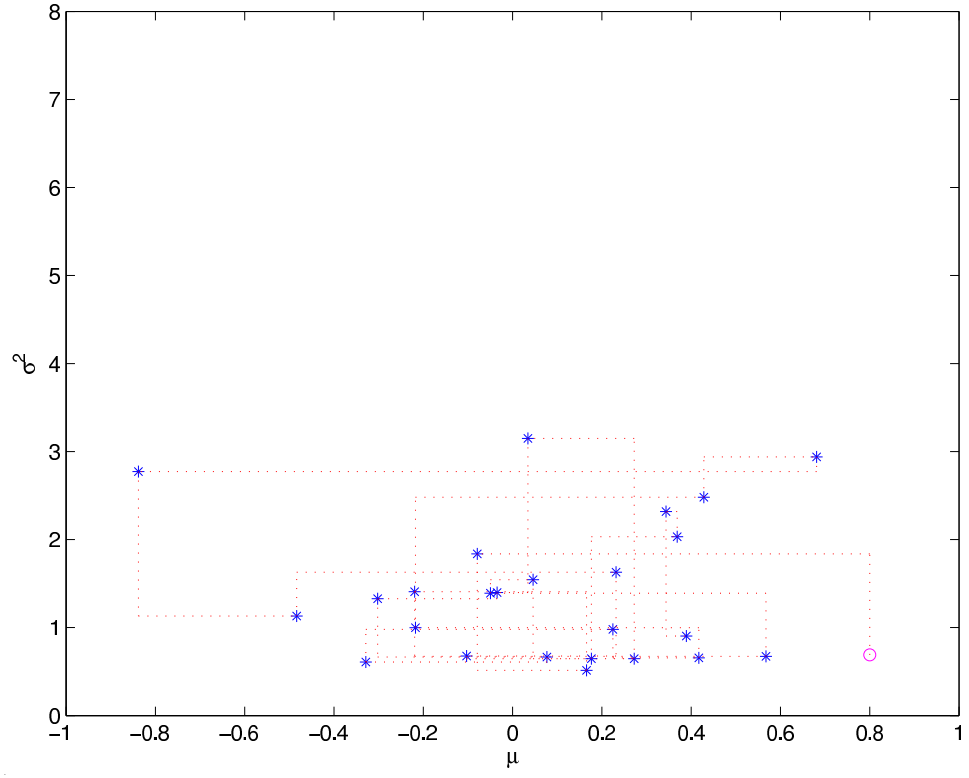
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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



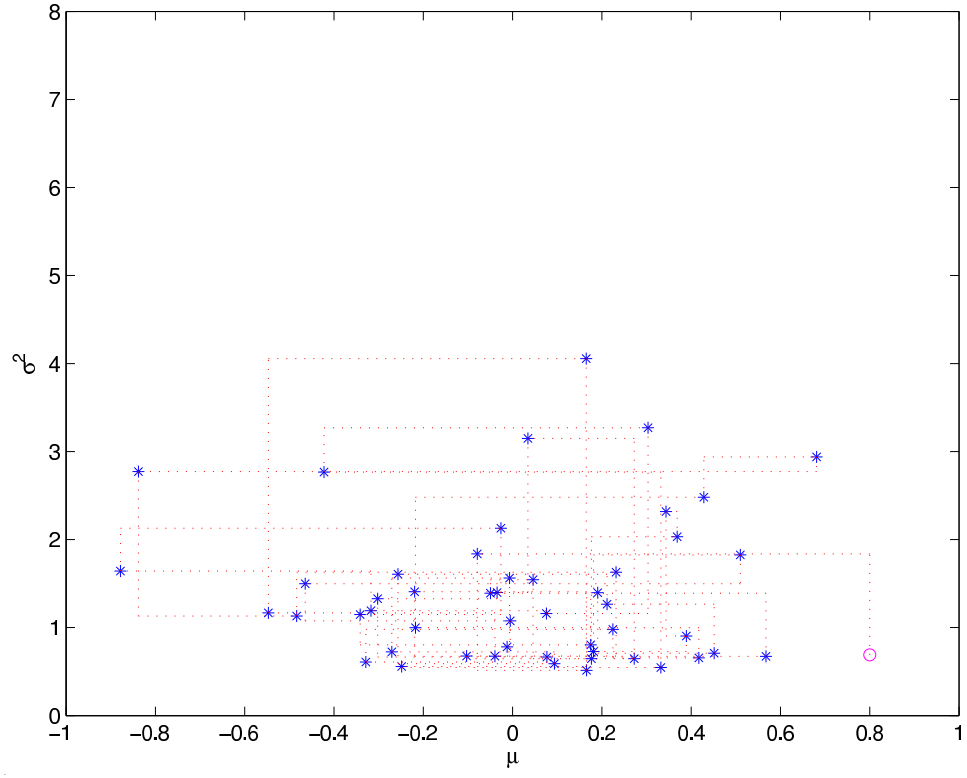
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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



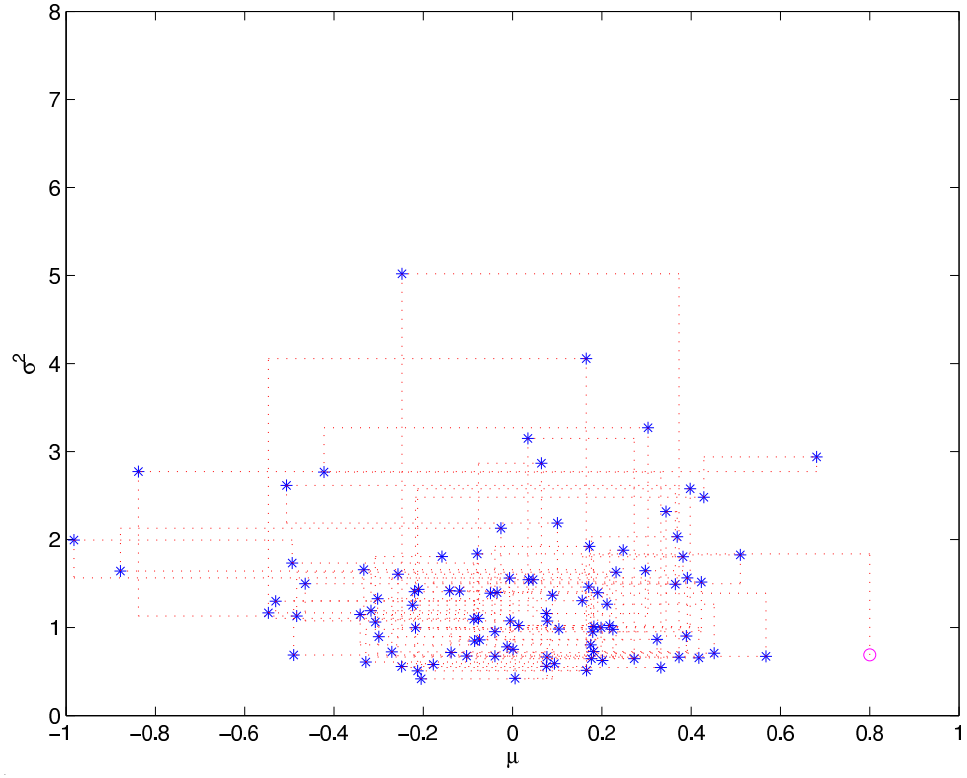
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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



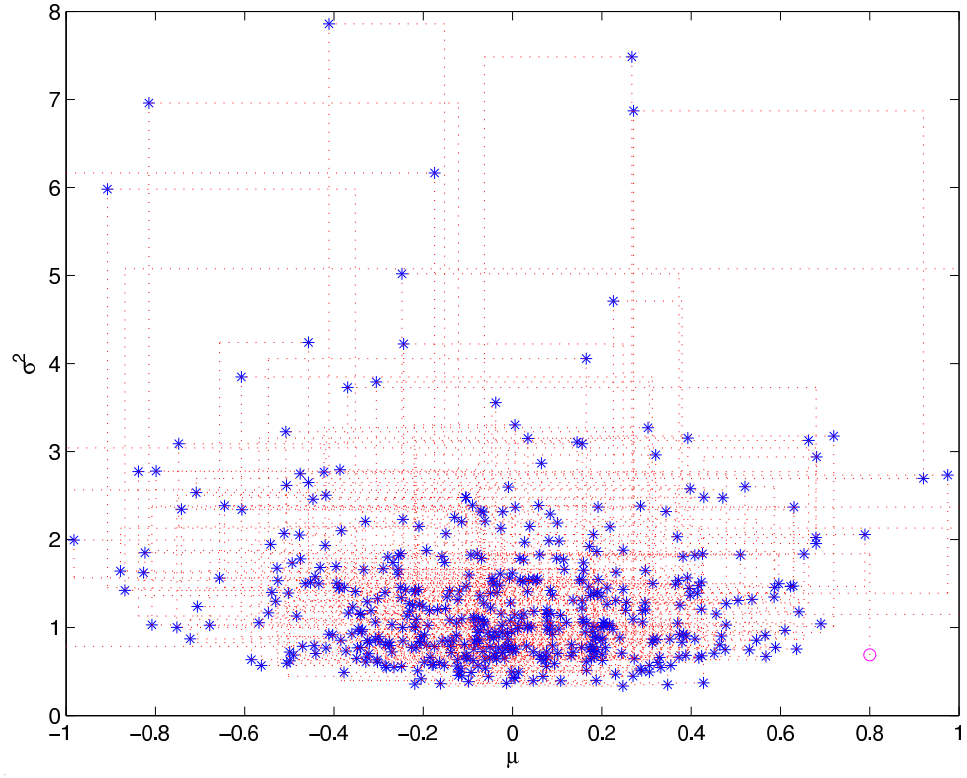
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

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

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.

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.

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.

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 to the rescue

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

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

Note

The variable z may be meaningless for the problem

Latent variables to the rescue

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 be meaningless for the problem

Purpose

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

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

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

Purpose

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

Algorithm (Completion Gibbs sampler)

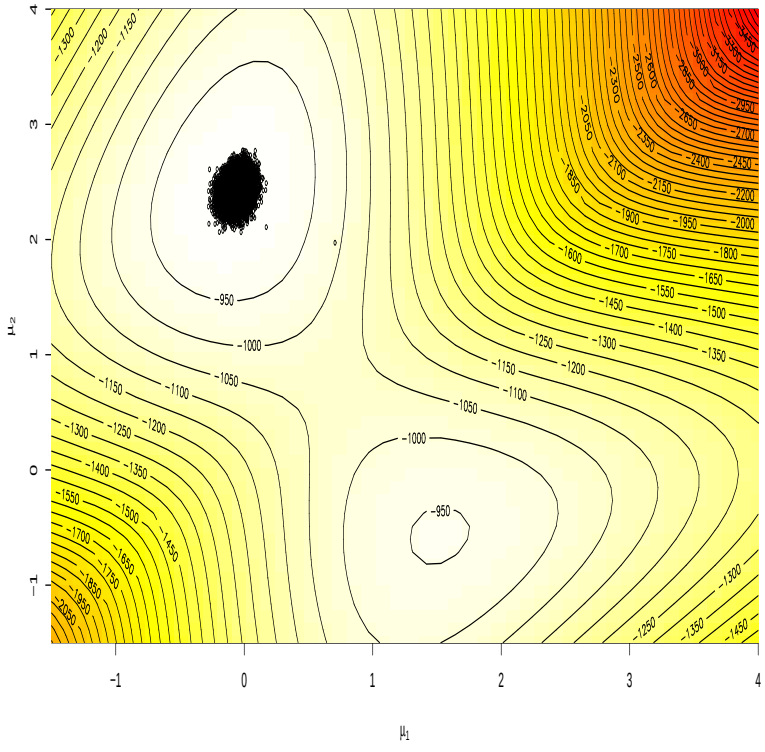
Given $(y_1^{(t)}, \ldots, y_p^{(t)})$, simulate

1. $Y_1^{(t+1)} \sim g_1(y_1 | y_2^{(t)}, \ldots, y_p^{(t)})$,
2. $Y_2^{(t+1)} \sim g_2(y_2 | y_1^{(t+1)}, y_3^{(t)}, \ldots, y_p^{(t)})$,

   $\ldots$

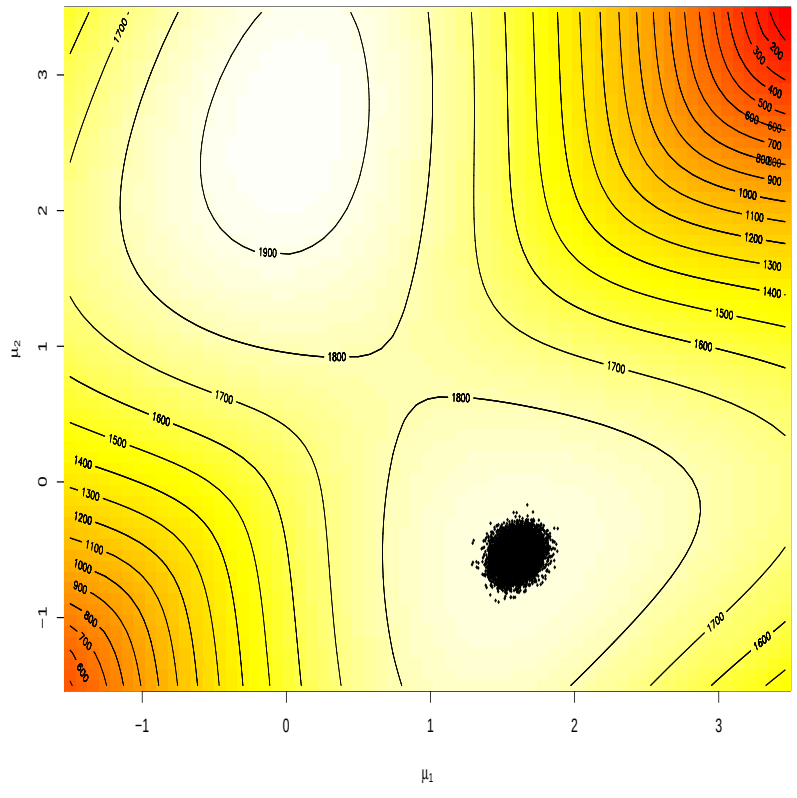
p. $Y_p^{(t+1)} \sim g_p(y_p | y_1^{(t+1)}, \ldots, y_{p-1}^{(t+1)})$.

A wee problem



Gibbs started at random

Gibbs stuck at the wrong model



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:

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:

Slice sampler

Algorithm (Slice sampler)

Simulate

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

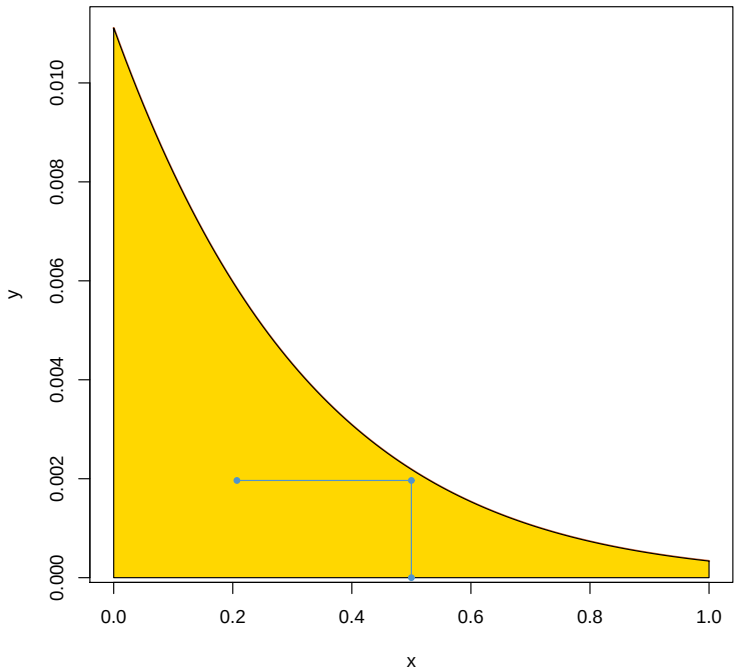
$\ldots$

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

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

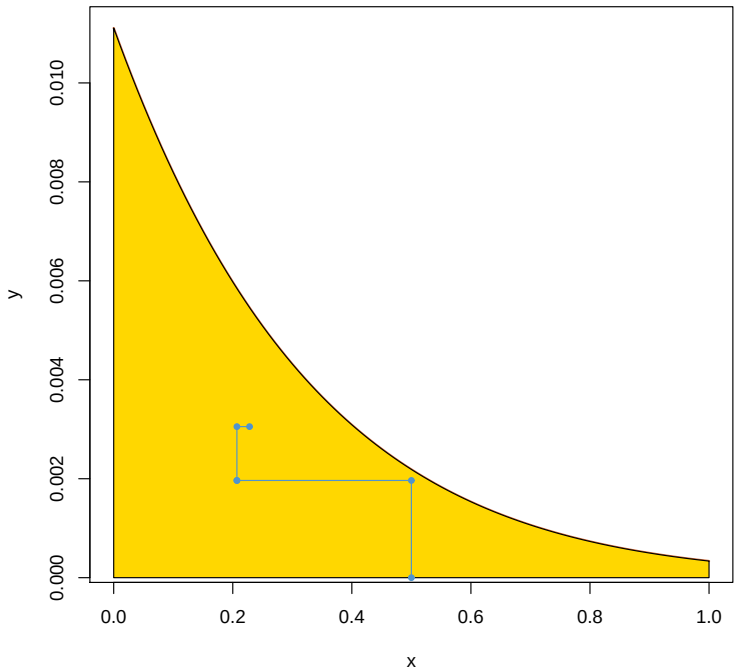
$$A^{(t+1)} = \{y;\ f_i(y) \geq \omega_i^{(t+1)},\ i = 1, \ldots, k\}.$$

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



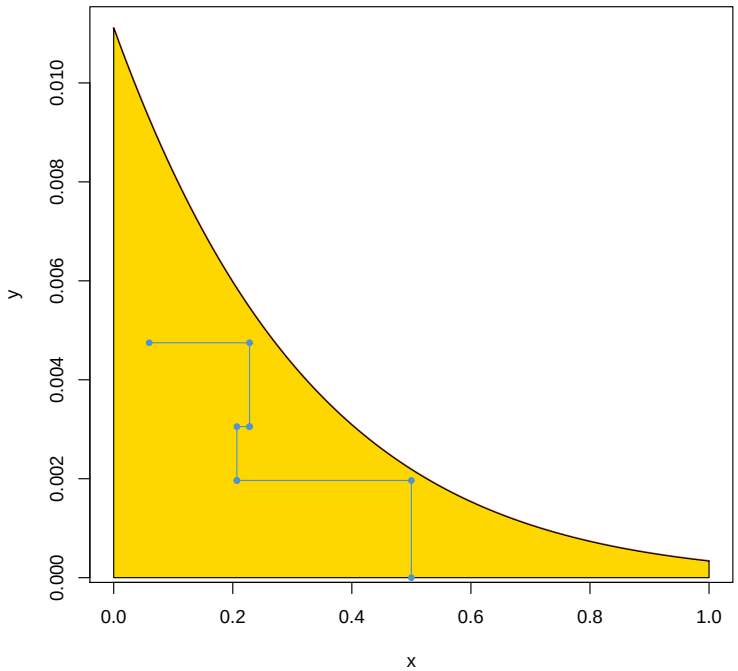
Number of Iterations 2, 3, 4, 5, 10, 50, 100

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



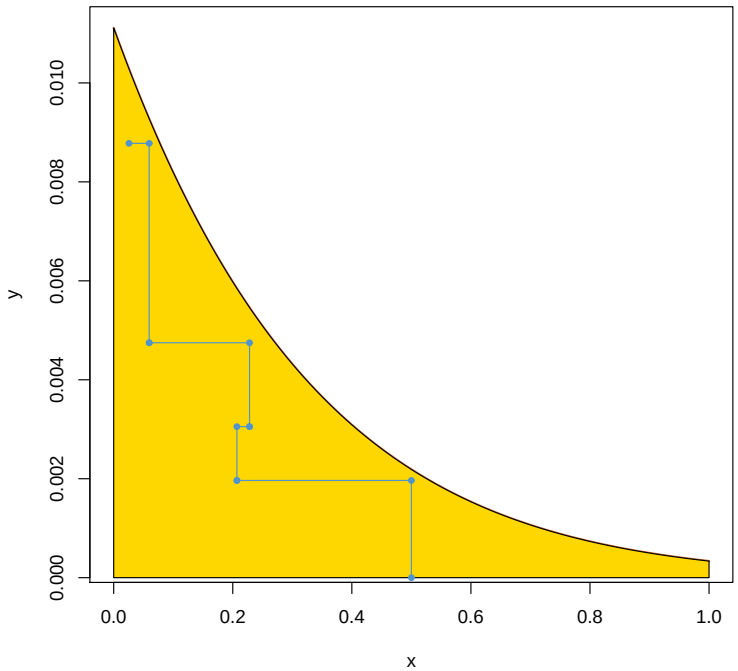
Number of Iterations 2, 3, 4, 5, 10, 50, 100

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



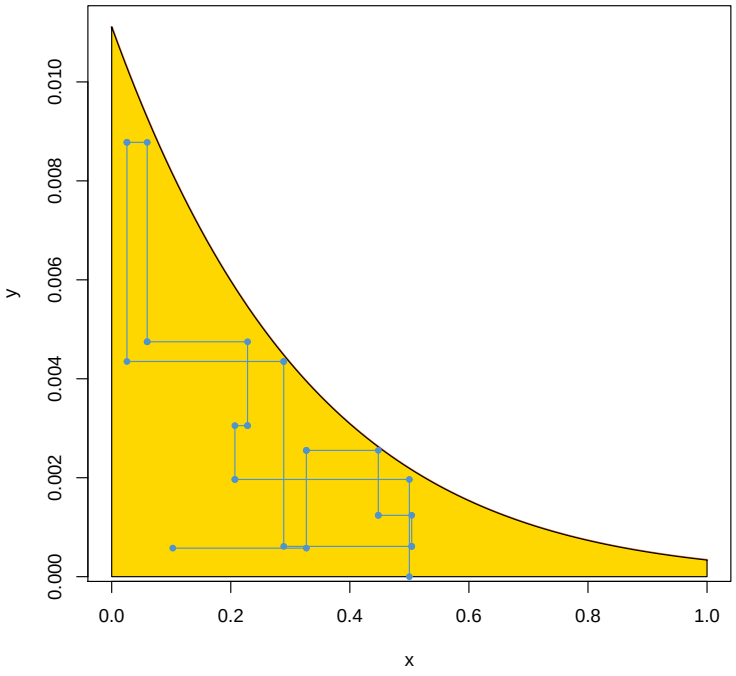
Number of Iterations 2, 3, 4, 5, 10, 50, 100

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



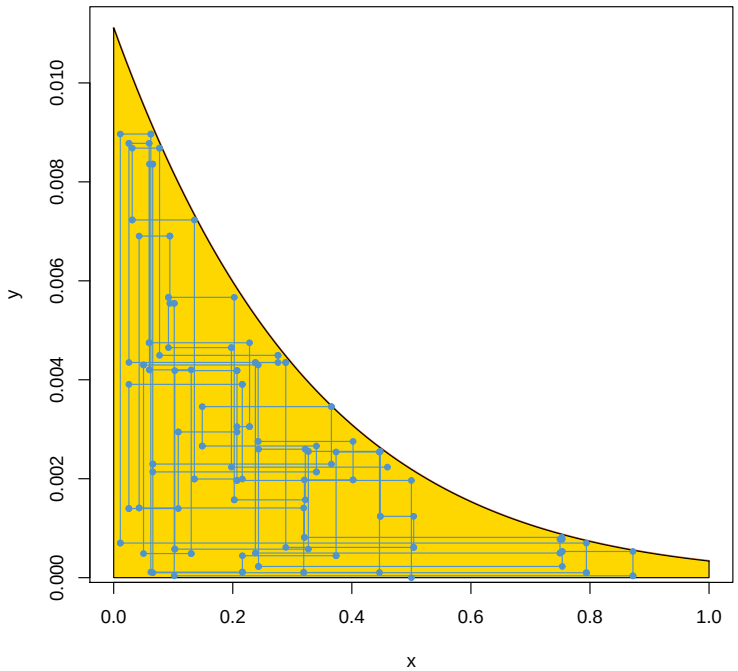
Number of Iterations 2, 3, 4, 5, 10, 50, 100

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



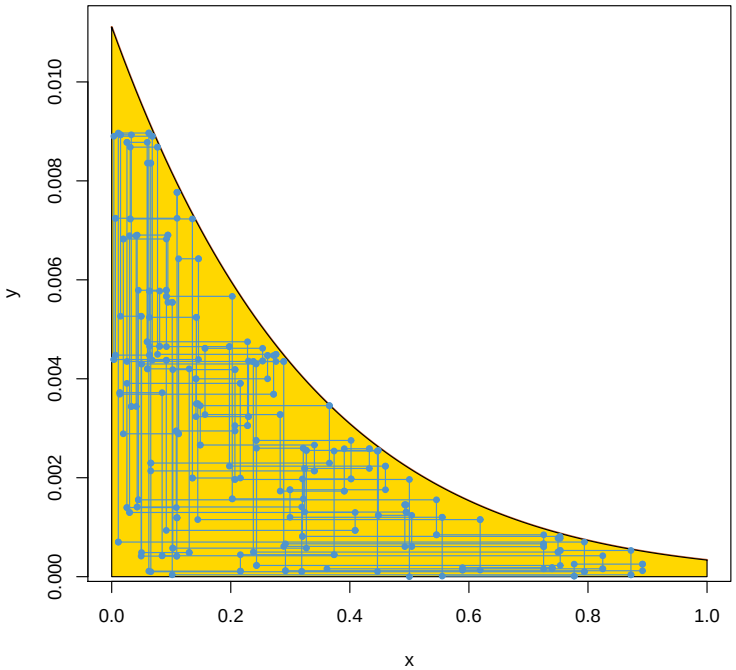
Number of Iterations 2, 3, 4, 5, 10, 50, 100

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



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

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.

Slice sampler: illustration

Example (Stochastic volatility core distribution)

Difficult part of the stochastic volatility model

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

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

Slice sampling means simulation from a uniform distribution on

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

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

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

Slice sampler: illustration

Example (Stochastic volatility core distribution)

Difficult part of the stochastic volatility model

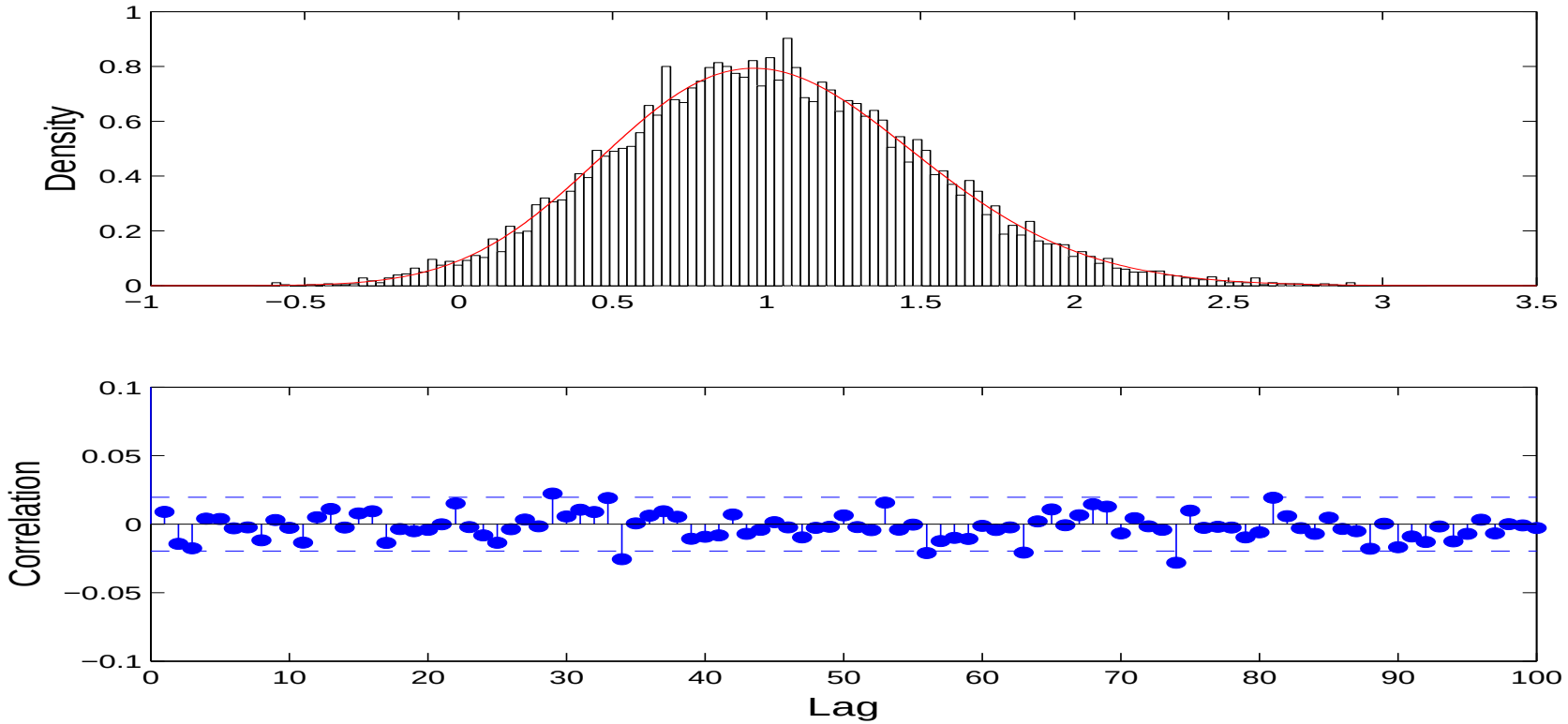
$$\pi(x) \propto \exp - \left\{\sigma^2(x-\mu)^2 + \beta^2 \exp(-x) y^2 + x\right\}/2\,,$$

simplified in $\exp - \left\{x^2 + \alpha \exp(-x)\right\}$
Slice sampling means simulation from a uniform distribution on

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

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

Slice sampler: illustration



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

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

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

(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: 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

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

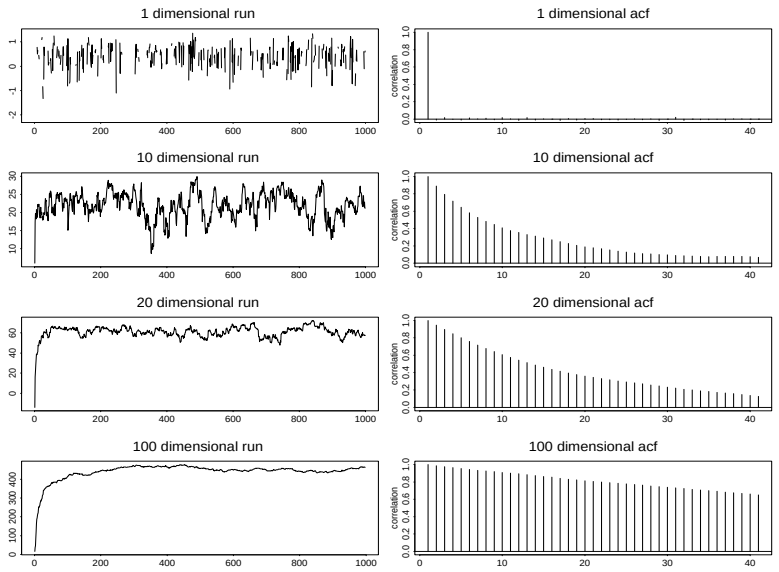
Slice sampler equivalent to one-dimensional slice sampler on

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

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

Rao-Blackwellization

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

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

and is unbiased.

The Rao-Blackwellization replaces δ_0 with its conditional expectation

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

Rao-Blackwellization

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

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

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

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

Rao-Blackwellization (2)

Then

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

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

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

Rao-Blackwellization (2)

Then

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

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

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

Examples of Rao-Blackwellization

Example

Bivariate normal Gibbs sampler

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

Then

$$\delta_0 = \frac{1}{T}\sum_{i=1}^{T} X^{(i)} \quad \text{and} \quad \delta_1 = \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}[X^{(i)}|Y^{(i)}] = \frac{1}{T}\sum_{i=1}^{T} \varrho Y^{(i)},$$

estimate $\mathbb{E}[X]$ and $\sigma^2_{\delta_0}/\sigma^2_{\delta_1} = \frac{1}{\rho^2} > 1$.

Examples of Rao-Blackwellization (2)

Example (Poisson-Gamma Gibbs cont'd)

Naïve estimate

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

and Rao-Blackwellized version

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

back to graph

NP Rao-Blackwellization & Rao-Blackwellized NP

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

Lemma

The estimator

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

is unbiased.

NP Rao-Blackwellization & Rao-Blackwellized NP

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

Lemma

The estimator

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

is unbiased.

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

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

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

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

Improper posteriors

Example (Conditional exponential distributions)

For the model

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

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**

Improper posteriors

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

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

Improper posteriors

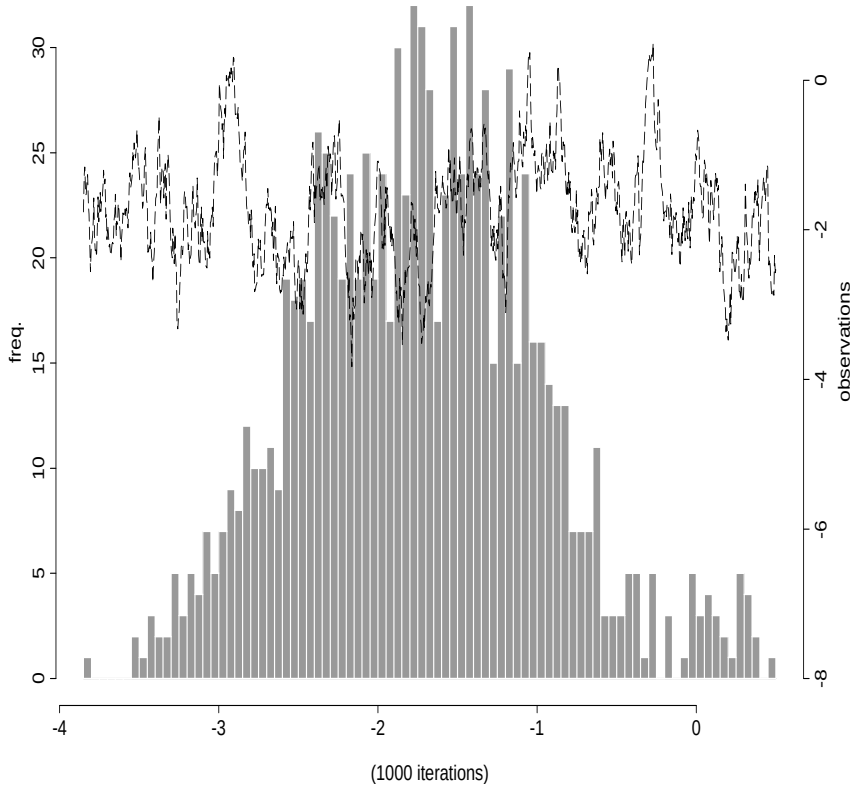
Example (Improper random effects 2)

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.

Improper posteriors



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 & al (1990) as a legitimate model

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 & al (1990) as a legitimate model

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.

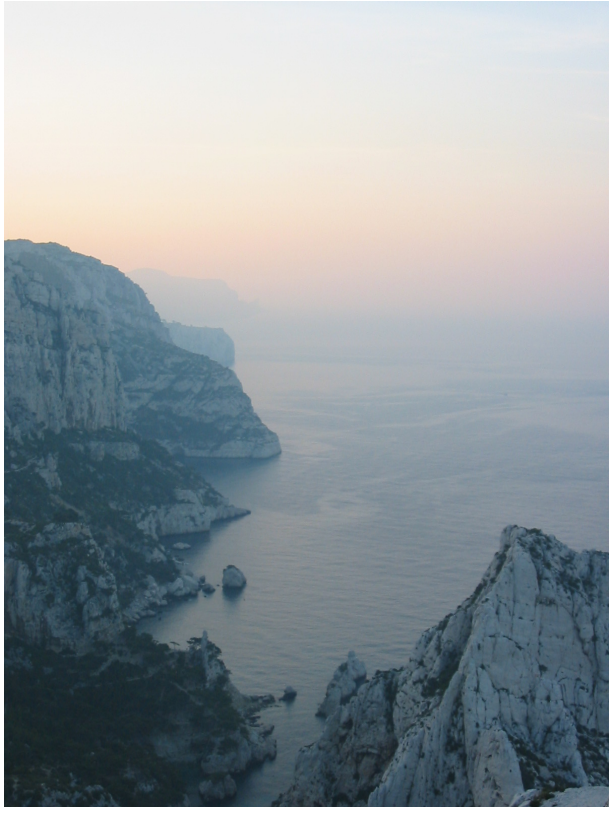
Example

The random effects model was initially treated in Gelfand & al (1990) as a legitimate model

Sequential importance sampling

MCMC # 3: Sequential Monte Carlo

- Adaptive MCMC
- Importance sampling revisited
- Population Monte Carlo
- pseudo-marginal extension



Adaptive MCMC may be hazardous to your ergodicity!

↯ **Algorithms trained on-line usually invalid:**

using the whole past of the "chain" implies that this is no longer a Markov chain! !

Adaptive MCMC may be hazardous to your ergodicity!

↯ **Algorithms trained on-line usually invalid:**
using the whole past of the "chain" implies that this is no longer a Markov chain! !

Illustration

Example (Poly t distribution)

Consider a t-distribution $\mathcal{T}(3, \theta, 1)$ sample $(x_1, \ldots, x_n)$ with a flat prior $\pi(\theta) = 1$
If we try fit a normal proposal from empirical mean and variance of the chain so far,

$$\mu_t = \frac{1}{t}\sum_{i=1}^{t}\theta^{(i)} \quad \text{and} \quad \sigma_t^2 = \frac{1}{t}\sum_{i=1}^{t}(\theta^{(i)} - \mu_t)^2\,,$$

Metropolis–Hastings algorithm with acceptance probability

$$\prod_{j=2}^{n}\left[\frac{\nu + (x_j - \theta^{(t)})^2}{\nu + (x_j - \xi)^2}\right]^{-(\nu+1)/2} \frac{\exp -(\mu_t - \theta^{(t)})^2/2\sigma_t^2}{\exp -(\mu_t - \xi)^2/2\sigma_t^2}\,,$$

where $\xi \sim \mathcal{N}(\mu_t, \sigma_t^2)$.

Illustration

Example (Poly t distribution)

Consider a t-distribution $\mathcal{T}(3, \theta, 1)$ sample $(x_1, \ldots, x_n)$ with a flat prior $\pi(\theta) = 1$
If we try fit a normal proposal from empirical mean and variance of the chain so far,

$$\mu_t = \frac{1}{t}\sum_{i=1}^{t}\theta^{(i)} \quad \text{and} \quad \sigma_t^2 = \frac{1}{t}\sum_{i=1}^{t}(\theta^{(i)} - \mu_t)^2\,,$$

Metropolis–Hastings algorithm with acceptance probability

$$\prod_{j=2}^{n}\left[\frac{\nu + (x_j - \theta^{(t)})^2}{\nu + (x_j - \xi)^2}\right]^{-(\nu+1)/2} \frac{\exp -(\mu_t - \theta^{(t)})^2/2\sigma_t^2}{\exp -(\mu_t - \xi)^2/2\sigma_t^2}\,,$$

where $\xi \sim \mathcal{N}(\mu_t, \sigma_t^2)$.

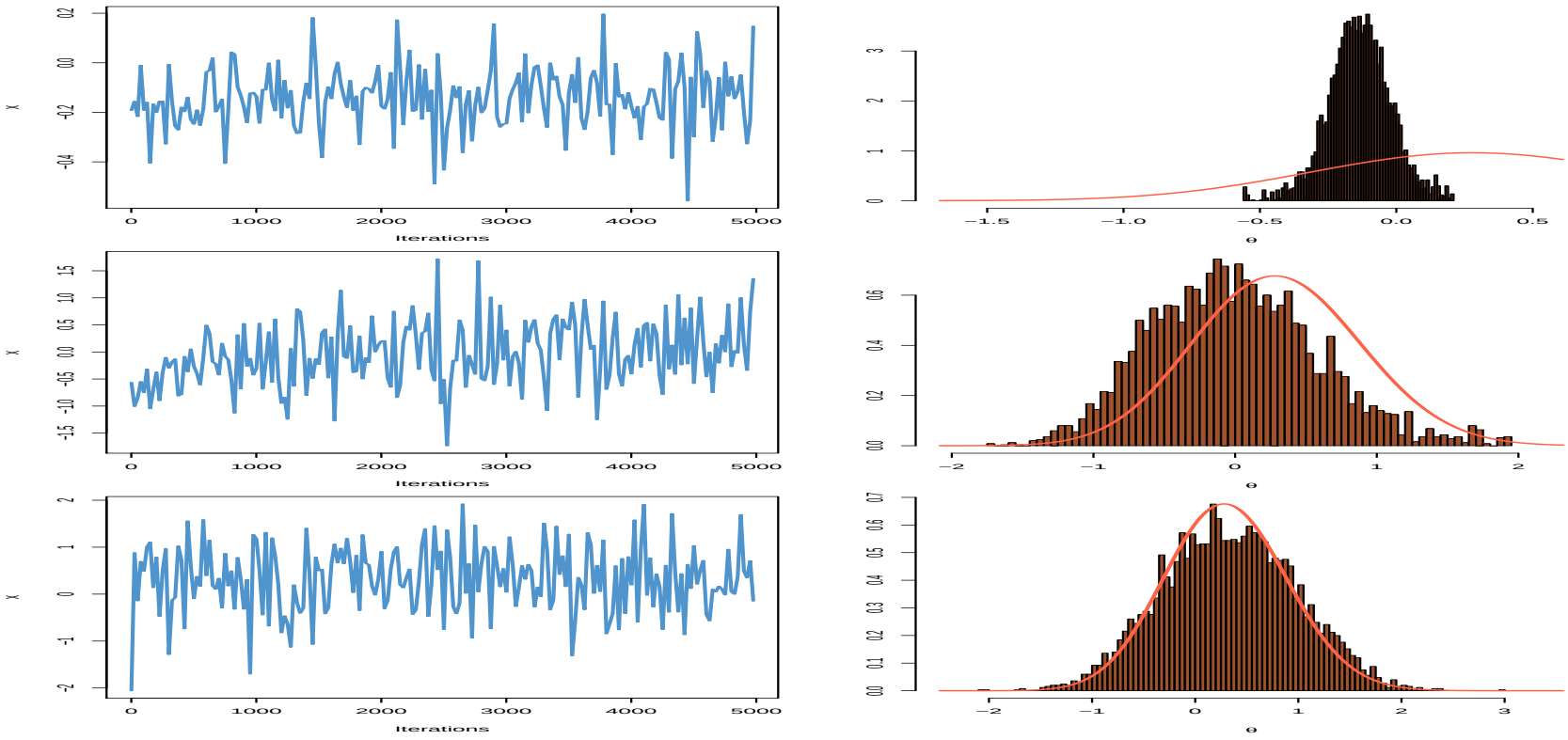
Illustration

Example (Poly t distribution (2))

Invalid scheme:

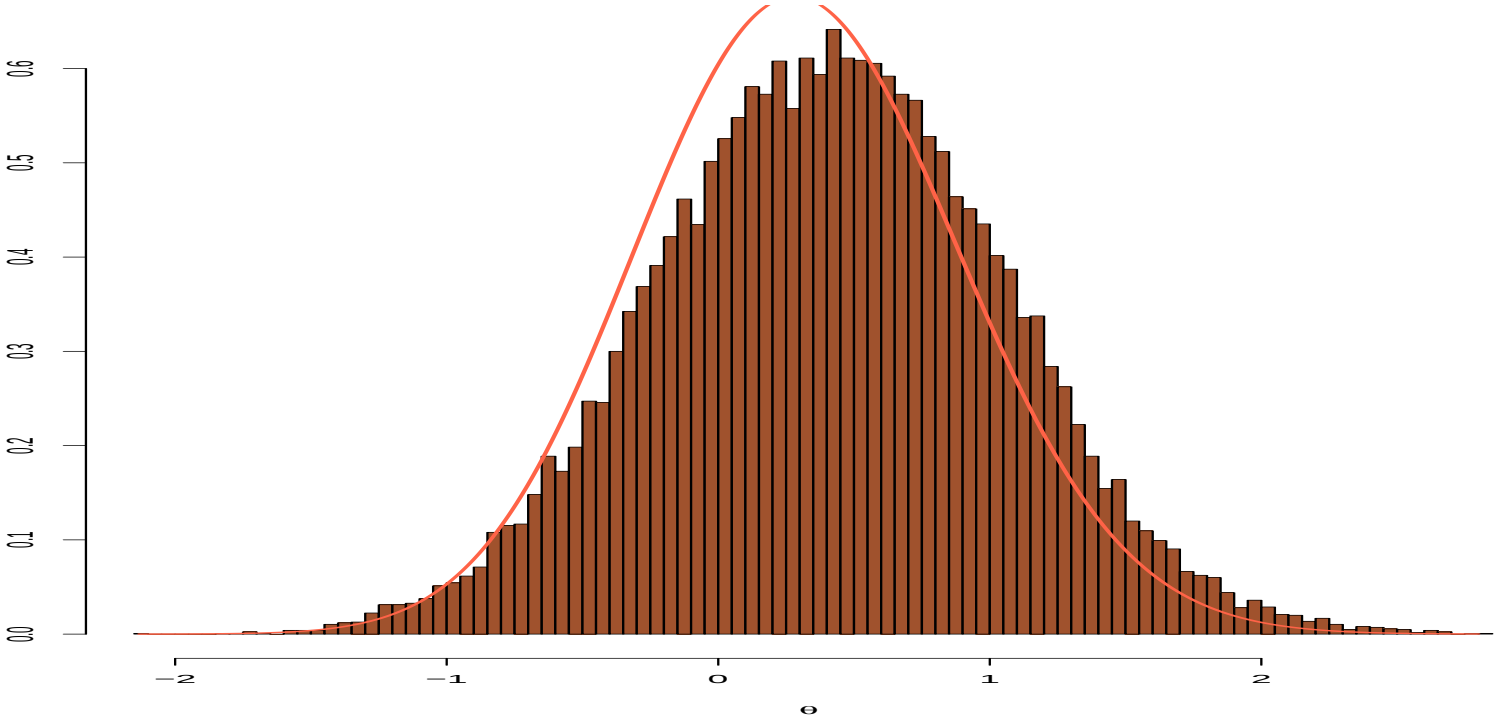
- when range of initial values too small, the $\theta^{(i)}$'s cannot converge to the target distribution and concentrates on too small a support.
- long-range dependence on past values modifies the distribution of the sequence.
- using past simulations to create a non-parametric approximation to the target distribution does not work either

Illustration



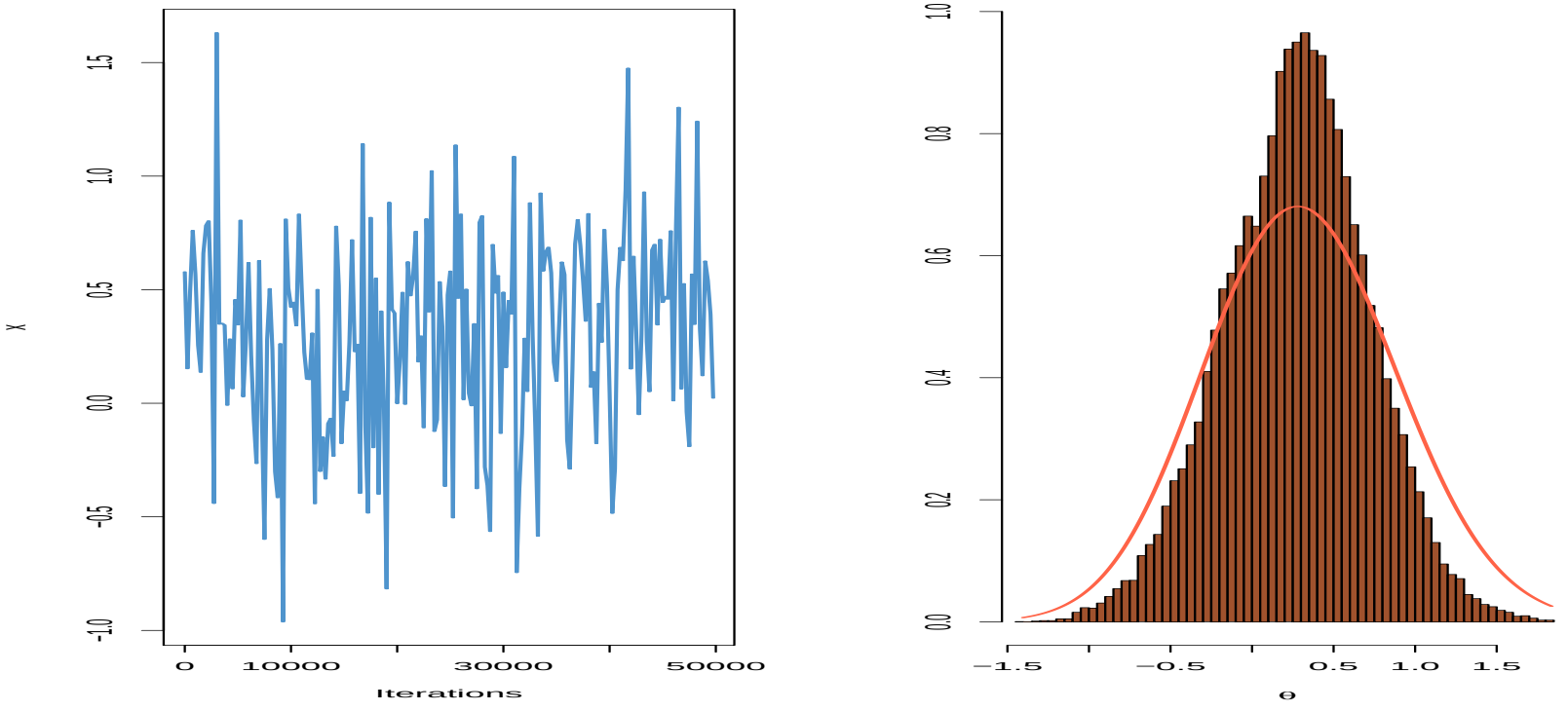
Adaptive scheme for a sample of 10 $x_j \sim \mathcal{T}_\ni$ and initial variances of (top) 0.1, (middle) 0.5, and (bottom) 2.5.

Illustration



Comparison of the distribution of an adaptive scheme sample of $25,000$ points with initial variance of 2.5 and of the target distribution.

Illustration



Sample produced by $50,000$ iterations of a nonparametric adaptive MCMC scheme and comparison of its distribution with the target distribution.

Simply forget about it!

Warning:

One should not constantly adapt the proposal on past performances

Either adaptation ceases after a period of *burnin*
or the adaptive scheme must be theoretically assessed on its own right.

Importance sampling revisited

Approximation of integrals

back to basic importance

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

by *unbiased estimators*

$$\hat{\mathfrak{I}} = \frac{1}{n}\sum_{i=1}^{n} \varrho_i h(x_i)$$

when

$$x_1, \ldots, x_n \overset{iid}{\sim} q(x) \qquad \text{and} \qquad \varrho_i \overset{\text{def}}{=} \frac{\pi(x_i)}{q(x_i)}$$

Pros and cons of importance sampling vs. MCMC

- Production of a weighted sample (IS) vs. of a Markov chain (MCMC)
- Dependence on importance function (IS) vs. on previous value (MCMC)
- Unbiasedness (IS) vs. convergence to the true distribution (MCMC)
- Variance control (IS) vs. learning costs (MCMC)
- Recycling of past simulations (IS) vs. progressive adaptability (MCMC)
- Processing of moving targets (IS) vs. handling large dimensional problems (MCMC)
- curse of dimensionality (IS) vs. tall and big data (MCMC)

Population Monte Carlo

Idea

Simulate from the product distribution

$$\pi^{\otimes n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \pi(x_i)$$

and apply dynamic importance sampling to the sample (*a.k.a.* population)

$$\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)})$$

Iterated importance sampling

As in Markov Chain Monte Carlo (MCMC) algorithms, introduction of a *temporal dimension* :

$$x_i^{(t)} \sim q_t(x|x_i^{(t-1)}) \qquad i = 1, \ldots, n, \quad t = 1, \ldots$$

and

$$\hat{\mathfrak{I}}_t = \frac{1}{n} \sum_{i=1}^{n} \varrho_i^{(t)} h(x_i^{(t)})$$

is still unbiased for

$$\varrho_i^{(t)} = \frac{\pi_t(x_i^{(t)})}{q_t(x_i^{(t)}|x_i^{(t-1)})}\,, \qquad i = 1, \ldots, n$$

Fundamental importance equality

Preservation of unbiasedness

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

for **any distribution** g on $X^{(t-1)}$

Sequential variance decomposition

Furthermore,

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

if $\mathrm{var}\left(\varrho_i^{(t)}\right)$ exists, because the $x_i^{(t)}$'s are conditionally uncorrelated

Note

This decomposition is still valid for correlated [in i] $x_i^{(t)}$'s when incorporating weights $\varrho_i^{(t)}$

Simulation of a population

The importance distribution of the sample (*a.k.a.* particles) $\mathbf{x}^{(t)}$

$$q_t(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$$

can depend on the previous sample $\mathbf{x}^{(t-1)}$ in any possible way as long as marginal distributions

$$q_{it}(x) = \int q_t(\mathbf{x}^{(t)})\, d\mathbf{x}_{-i}^{(t)}$$

can be expressed to build importance weights

$$\varrho_{it} = \frac{\pi(x_i^{(t)})}{q_{it}(x_i^{(t)})}$$

Special case of the product proposal

If

$$q_t(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \prod_{i=1}^{n} q_{it}(x_i^{(t)}|\mathbf{x}^{(t-1)})$$

[Independent proposals]

then

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

Validation

▸ skip validation

$$\mathbb{E}\left[\varrho_i^{(t)} h(X_i^{(t)})\ \varrho_j^{(t)} h(X_j^{(t)})\right]$$

$$= \int h(x_i) \frac{\pi(x_i)}{q_{it}(x_i|\mathbf{x}^{(t-1)})} \frac{\pi(x_j)}{q_{jt}(x_j|\mathbf{x}^{(t-1)})} h(x_j)$$

$$q_{it}(x_i|\mathbf{x}^{(t-1)})\, q_{jt}(x_j|\mathbf{x}^{(t-1)})\, dx_i\, dx_j\, g(\mathbf{x}^{(t-1)}) d\mathbf{x}^{(t-1)}$$

$$= \mathbb{E}_\pi\left[h(X)\right]^2$$

whatever the distribution g on $\mathbf{x}^{(t-1)}$

Self-normalised version

In general, π is unscaled and the weight

$$\varrho_i^{(t)} \propto \frac{\pi(x_i^{(t)})}{q_{it}(x_i^{(t)})}\,, \qquad i = 1, \ldots, n\,,$$

is scaled so that

$$\sum_i \varrho_i^{(t)} = 1$$

Self-normalised version properties

- Loss of the unbiasedness property and the variance decomposition
- Normalising constant can be estimated by

$$\varpi_t = \frac{1}{tn} \sum_{\tau=1}^{t} \sum_{i=1}^{n} \frac{\pi(x_i^{(\tau)})}{q_{i\tau}(x_i^{(\tau)})}$$

- Variance decomposition (approximately) recovered if ϖ_{t-1} is used instead

Sampling importance resampling

Importance sampling from g can **also** produce samples from the target π

[Rubin, 1987]

Theorem (Bootstraped importance sampling)

If a sample $(x_i^\star)_{1\le i\le m}$ is derived from the weighted sample $(x_i, \varrho_i)_{1\le i\le n}$ by multinomial sampling with weights ϱ_i, then

$$x_i^\star \sim \pi(x)$$

Note

Obviously, the $x_i^\star$'s are **not iid**

Sampling importance resampling

Importance sampling from g can **also** produce samples from the target π

[Rubin, 1987]

Theorem (Bootstraped importance sampling)

If a sample $(x_i^\star)_{1\leq i\leq m}$ is derived from the weighted sample $(x_i, \varrho_i)_{1\leq i\leq n}$ by multinomial sampling with weights ϱ_i, then

$$x_i^\star \sim \pi(x)$$

Note

Obviously, the $x_i^\star$'s are **not iid**

Sampling importance resampling

Importance sampling from g can **also** produce samples from the target π

[Rubin, 1987]

Theorem (Bootstraped importance sampling)

If a sample $(x_i^\star)_{1\leq i\leq m}$ is derived from the weighted sample $(x_i, \varrho_i)_{1\leq i\leq n}$ by multinomial sampling with weights ϱ_i, then

$$x_i^\star \sim \pi(x)$$

Note

Obviously, the $x_i^\star$'s are **not iid**

Iterated sampling importance resampling

This principle can be extended to iterated importance sampling:
After each iteration, resampling produces a sample from π

[Again, not iid!]

Incentive

Use previous sample(s) to learn about π and q

Iterated sampling importance resampling

This principle can be extended to iterated importance sampling:
After each iteration, resampling produces a sample from π

[Again, not iid!]

Incentive

Use previous sample(s) to learn about π and q

Generic Population Monte Carlo

Algorithm (Population Monte Carlo Algorithm)

For $t = 1, \ldots, T$

For $i = 1, \ldots, n$,

1. Select the generating distribution $q_{it}(\cdot)$
2. Generate $\tilde{x}_i^{(t)} \sim q_{it}(x)$
3. Compute $\varrho_i^{(t)} = \pi(\tilde{x}_i^{(t)})/q_{it}(\tilde{x}_i^{(t)})$

Normalise the $\varrho_i^{(t)}$'s into $\bar{\varrho}_i^{(t)}$'s

Generate $J_{i,t} \sim \mathcal{M}((\bar{\varrho}_i^{(t)})_{1 \le i \le N})$ and set $x_{i,t} = \tilde{x}_{J_{i,t}}^{(t)}$

D-kernels in competition

A general adaptive construction:

Construct $q_{i,t}$ as a mixture of D different transition kernels depending on $x_i^{(t-1)}$

$$q_{i,t} = \sum_{\ell=1}^{D} p_{t,\ell} \mathfrak{K}_\ell(x_i^{(t-1)}, x), \qquad \sum_{\ell=1}^{D} p_{t,\ell} = 1\,,$$

and adapt the weights $p_{t,\ell}$.

Example

Take $p_{t,\ell}$ proportional to the survival rate of the points (*a.k.a.* particles) $x_i^{(t)}$ generated from $\mathfrak{K}_\ell$

D-kernels in competition

A general adaptive construction:

Construct $q_{i,t}$ as a mixture of D different transition kernels depending on $x_i^{(t-1)}$

$$q_{i,t} = \sum_{\ell=1}^{D} p_{t,\ell} \mathfrak{K}_\ell(x_i^{(t-1)}, x), \qquad \sum_{\ell=1}^{D} p_{t,\ell} = 1\,,$$

and adapt the weights $p_{t,\ell}$.

Example

Take $p_{t,\ell}$ proportional to the survival rate of the points (*a.k.a.* particles) $x_i^{(t)}$ generated from $\mathfrak{K}_\ell$

Implementation

Algorithm (D-kernel PMC)

For $t = 1, \ldots, T$

generate $(K_{i,t})_{1 \leq i \leq N} \sim \mathscr{M}((p_{t,k})_{1 \leq k \leq D})$

for $1 \leq i \leq N$, generate

$$\tilde{x}_{i,t} \sim \mathfrak{K}_{K_{i,t}}(x)$$

compute and renormalize the importance weights $\omega_{i,t}$

generate $(J_{i,t})_{1 \leq i \leq N} \sim \mathscr{M}((\overline{\omega}_{i,t})_{1 \leq i \leq N})$

take $x_{i,t} = \tilde{x}_{J_{i,t},t}$ and $p_{t+1,d} = \sum_{i=1}^{N} \bar{\omega}_{i,t} \mathbb{I}_d(K_{i,t})$

Links with particle filters

- Usually setting where $\pi = \pi_t$ changes with t: Population Monte Carlo also adapts to this case
- Can be traced back all the way to Hammersley and Morton (1954) and the self-avoiding random walk problem
- Gilks and Berzuini (2001) produce iterated samples with (SIR) resampling steps, and add an MCMC step: this step must use a π_t invariant kernel
- Chopin (2001) uses iterated importance sampling to handle large datasets: this is a special case of PMC where the q_{it}'s are the posterior distributions associated with a portion k_t of the observed dataset

Links with particle filters (2)

- Rubinstein and Kroese's (2004) *cross-entropy* method is parameterised importance sampling targeted at rare events
- Stavropoulos and Titterington's (1999) *smooth bootstrap* and Warnes' (2001) *kernel coupler* use nonparametric kernels on the previous importance sample to build an improved proposal: this is a special case of PMC
- West (1992) mixture approximation is a precursor of smooth bootstrap
- Mengersen and Robert (2002) "pinball sampler" is an MCMC attempt at population sampling
- Del Moral and Doucet (2003) sequential Monte Carlo samplers also relates to PMC, with a Markovian dependence on the past sample $\mathbf{x}^{(t)}$ but (limited) stationarity constraints

Things can go wrong

Unexpected behaviour of the mixture weights when the number of particles increases

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} \mathbb{I}_{K_{i,t}=d} \longrightarrow P \frac{1}{D}$$

Conclusion

At *each* iteration, every weight converges to $1/D$:
the algorithm fails to learn from experience!!

Things can go wrong

Unexpected behaviour of the mixture weights when the number of particles increases

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} \mathbb{I}_{K_{i,t}=d} \longrightarrow P \frac{1}{D}$$

Conclusion

At *each* iteration, every weight converges to $1/D$:
the algorithm fails to learn from experience!!

Saved by Rao-Blackwell!!

Modification: Rao-Blackwellisation (=conditioning)

Use the whole mixture in the importance weight:

$$\omega_{i,t} = \pi(\tilde{x}_{i,t}) \Big/ \sum_{d=1}^{D} p_{t,d} \mathfrak{K}_d(x_{i,t-1}, \tilde{x}_{i,t})$$

instead of

$$\omega_{i,t} = \frac{\pi(\tilde{x}_{i,t})}{\mathfrak{K}_{K_{i,t}}(x_{i,t-1}, \tilde{x}_{i,t})}$$

Saved by Rao-Blackwell!!

Modification: Rao-Blackwellisation (=conditioning)

Use the whole mixture in the importance weight:

$$\omega_{i,t} = \pi(\tilde{x}_{i,t}) \Big/ \sum_{d=1}^{D} p_{t,d} \mathfrak{K}_d(x_{i,t-1}, \tilde{x}_{i,t})$$

instead of

$$\omega_{i,t} = \frac{\pi(\tilde{x}_{i,t})}{\mathfrak{K}_{K_{i,t}}(x_{i,t-1}, \tilde{x}_{i,t})}$$

Adapted algorithm

Algorithm (Rao-Blackwellised D-kernel PMC)

At time t ($t = 1, \ldots, T$),

Generate
$$(K_{i,t})_{1 \leq i \leq N} \overset{iid}{\sim} \mathcal{M}((p_{t,d})_{1 \leq d \leq D});$$

Generate
$$(\tilde{x}_{i,t})_{1 \leq i \leq N} \overset{\text{ind}}{\sim} \mathfrak{K}_{K_{i,t}}(x_{i,t-1}, x)$$

and set $\omega_{i,t} = \pi(\tilde{x}_{i,t}) \Big/ \sum_{d=1}^{D} p_{t,d} \mathfrak{K}_d(x_{i,t-1}, \tilde{x}_{i,t})$;

Generate
$$(J_{i,t})_{1 \leq i \leq N} \overset{iid}{\sim} \mathcal{M}((\bar{\omega}_{i,t})_{1 \leq i \leq N})$$

and set $x_{i,t} = \tilde{x}_{J_{i,t},t}$ and $p_{t+1,d} = \sum_{i=1}^{N} \bar{\omega}_{i,t} p_{t,d}$.

Convergence properties

Theorem (LLN)

Under regularity assumptions, for $h \in L^1_\Pi$ *and for every* $t \geq 1$,

$$\frac{1}{N}\sum_{k=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}) \overset{N\to\infty}{\longrightarrow}_P \Pi(h)$$

and

$$p_{t,d} \overset{N\to\infty}{\longrightarrow}_P \alpha_d^t$$

The limiting coefficients $(\alpha_d^t)_{1\leq d\leq D}$ *are defined recursively as*

$$\alpha_d^t = \alpha_d^{t-1} \int \left(\frac{\mathfrak{K}_d(x,x')}{\sum_{j=1}^{D} \alpha_j^{t-1} \mathfrak{K}_j(x,x')} \right) \Pi \otimes \Pi(dx, dx').$$

Recursion on the weights

Set F as

$$F(\alpha) = \left(\alpha_d \int \left[\frac{\mathfrak{K}_d(x, x')}{\sum_{j=1}^{D} \alpha_j \mathfrak{K}_j(x, x')} \right] \Pi \otimes \Pi(dx, dx') \right)_{1 \leq d \leq D}$$

on the simplex

$$S = \left\{ \alpha = (\alpha_1, \ldots, \alpha_D);\ \forall d \in \{1, \ldots, D\},\ \alpha_d \geq 0 \quad \text{and} \sum_{d=1}^{D} \alpha_d = 1 \right\}.$$

and define the sequence

$$\boldsymbol{\alpha}^{t+1} = F(\boldsymbol{\alpha}^t)$$

Kullback divergence

Definition (Kullback divergence)

For $\alpha \in S$,

$$\mathrm{KL}(\boldsymbol{\alpha}) = \int \left[\log \left(\frac{\pi(x)\pi(x')}{\pi(x) \sum_{d=1}^{D} \alpha_d \mathfrak{K}_d(x, x')} \right) \right] \Pi \otimes \Pi(dx, dx').$$

Kullback divergence between Π and the mixture.

Goal: Obtain the mixture closest to Π, i.e., that minimises $\mathrm{KL}(\boldsymbol{\alpha})$

Connection with RBDPMCA ??

Theorem

Under the assumption

$$\forall d \in \{1, \ldots, D\}, -\infty < \int \log(\mathfrak{K}_d(x, x'))\Pi \otimes \Pi(dx, dx') < \infty$$

for every $\boldsymbol{\alpha} \in \mathfrak{S}_D$,

$$KL(F(\boldsymbol{\alpha})) \leq KL(\boldsymbol{\alpha}).$$

Conclusion

The Kullback divergence decreases at every iteration of RBDPMCA

Connection with RBDPMCA ??

Theorem

Under the assumption

$$\forall d \in \{1, \ldots, D\}, -\infty < \int \qquad \log(\mathfrak{K}_d(x, x'))\Pi \otimes \Pi(dx, dx') < \infty$$

for every $\boldsymbol{\alpha} \in \mathfrak{S}_D$,

$$KL(F(\boldsymbol{\alpha})) \leq KL(\boldsymbol{\alpha}).$$

Conclusion

The Kullback divergence decreases at every iteration of RBDPMCA

Illustration

Example (A toy example)

Take the target

$$1/4\mathcal{N}(-1, 0.3)(x) + 1/4\mathcal{N}(0, 1)(x) + 1/2\mathcal{N}(3, 2)(x)$$

and use 3 proposals: $\mathcal{N}(-1, 0.3)$, $\mathcal{N}(0, 1)$ and $\mathcal{N}(3, 2)$

[Surprise!!!]

Then

1	0.0500000	0.05000000	0.9000000
2	0.2605712	0.09970292	0.6397259
6	0.2740816	0.19160178	0.5343166
10	0.2989651	0.19200904	0.5090259
16	0.2651511	0.24129039	0.4935585

Weight evolution

Illustration

Example (A toy example)

Take the target

$$1/4\mathcal{N}(-1, 0.3)(x) + 1/4\mathcal{N}(0, 1)(x) + 1/2\mathcal{N}(3, 2)(x)$$

and use 3 proposals: $\mathcal{N}(-1, 0.3)$, $\mathcal{N}(0, 1)$ and $\mathcal{N}(3, 2)$

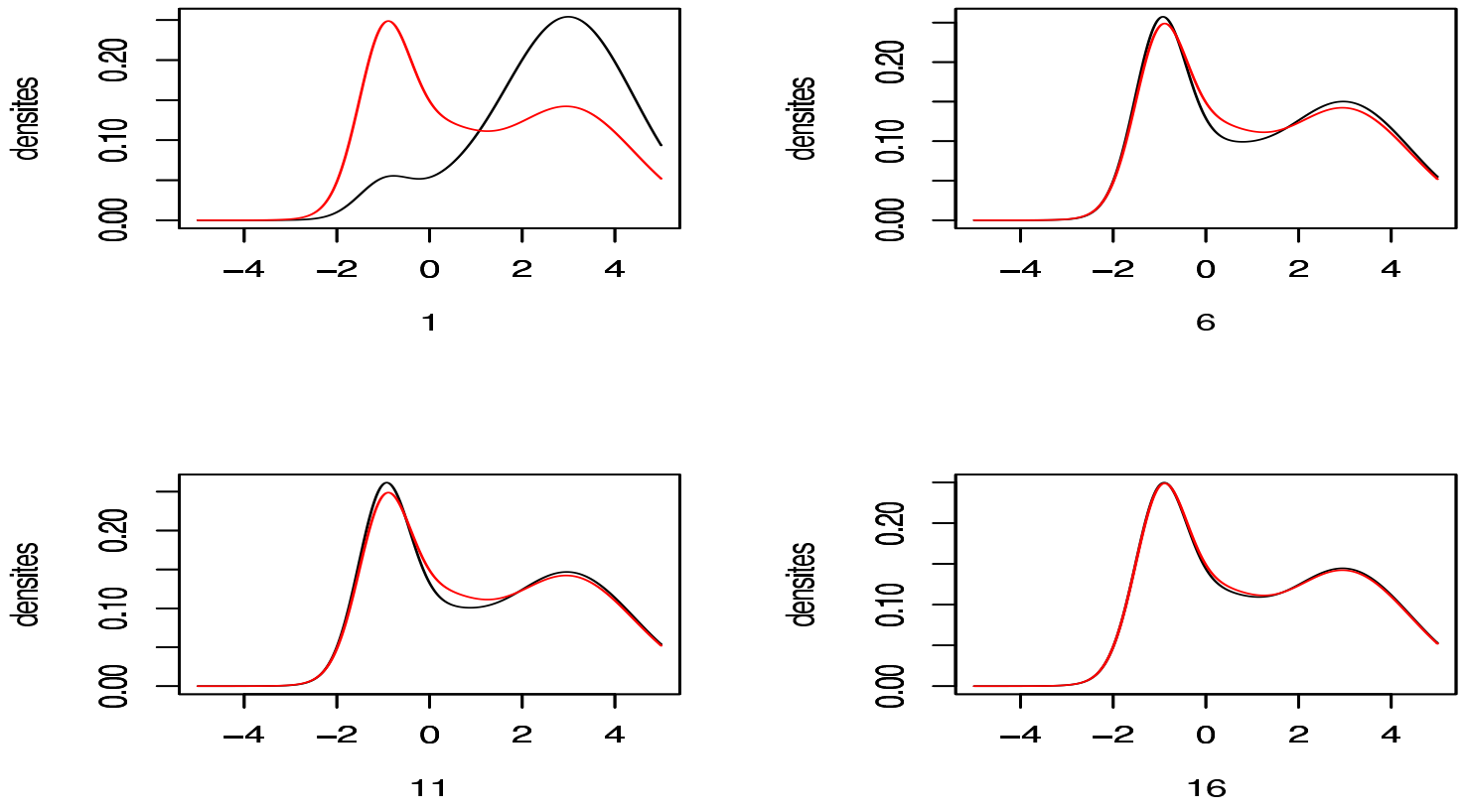
[Surprise!!!]

Then

1	0.0500000	0.05000000	0.9000000
2	0.2605712	0.09970292	0.6397259
6	0.2740816	0.19160178	0.5343166
10	0.2989651	0.19200904	0.5090259
16	0.2651511	0.24129039	0.4935585

Weight evolution

Illustration



Target and mixture evolution

intractable and doubly-intractable likelihoods

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

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

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 $z \sim q(\cdot|\theta)$

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

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

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 $z \sim q(\cdot|\theta)$

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

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

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 $z \sim q(\cdot|\theta)$

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

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]

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 $z \sim q(\cdot|\theta)$

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

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
Performances depend on quality of estimators $\hat{\pi}$ but always poorer than when using the exact target π

[Andrieu & Vihola, 2012]

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 $\theta' \sim \mathfrak{h}(\theta|\theta^{(t)})$

[Andrieu, Doucet & Holenstein, 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 of latent series $x'_{0:T}$ via particle filter approximation of $p(x_{0:T}|\theta', y_{1:T})$

[Andrieu, Doucet & Holenstein, 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

- derive unbiased estimator of marginal posterior of $y_{1:T}$, $\hat{q}(y_{1:T}|\theta')$

[Andrieu, Doucet & Holenstein, 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

- use estimator in Metropolis–Hastings ratio

$$\frac{\hat{q}(y_{1:T}|\theta')\pi(\theta')\mathfrak{h}(\theta^{(t)}|\theta')}{\hat{q}(y_{1:T}|\theta)\pi(\theta^{(t)})\mathfrak{h}(\theta'|\theta^{(t)})} \wedge 1\,.$$

[Andrieu, Doucet & Holenstein, 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

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

[Andrieu, Doucet & Holenstein, 2010]

Extension of pMCMC called SMC^2 that approximates sequential filtering distribution proposed in Chopin et al (2013)

Next iterations for MCMC?

- intractability or double intractability of the target: is ABC the only solution?
- non reversible versions of MCMC like NUTS and Hamiltonian Monte Carlo (STAN)
- scalable MCMC (divide-and-conquer)
- approximate and noisy MCMC
- asynchronous Gibbs samplers
- zero variance MCMC