Stochastic Volatility
An experimental approach

Christian P. Robert

Université Paris Dauphine and CREST-INSEE
http://www.ceremade.dauphine.fr/~xian

22 mars – 19 avril 2007

Outline

Stochastic volatility model

The Metropolis-Hastings Algorithm

The Gibbs Sampler

Monte Carlo Integration

Sequential importance sampling

Stochastic volatility model

Stochastic volatility model

The Metropolis-Hastings Algorithm

The Gibbs Sampler

Monte Carlo Integration

Sequential importance sampling

Latent structures make life harder!

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

$$f(x|\theta) = \int f^{\star}(x, x^{\star}|\theta)\, \mathrm{d}x^{\star}$$

Latent structures make life harder!

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

$$f(x|\theta) = \int f^\star(x, x^\star|\theta)\,\mathrm{d}x^\star$$

If $(x, x^\star)$ observed, fine!

Latent structures make life harder!

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

$$f(x|\theta) = \int f^\star(x, x^\star|\theta)\, \mathrm{d}x^\star$$

If $(x, x^\star)$ observed, fine!
If **only** x observed, trouble!

Stochastic volatility

Observables

$$y_t \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_t^2)$$

Stochastic volatility

Observables

$$y_t \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_t^2)$$

with unobserved variances related by

$$\log \sigma_{t+1}^2 = \mu + \varrho \log \sigma_t^2 + \tau \varepsilon_t \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

Data production

R code

```
# Parameters
N=1000
mu=0
rho=0.1
mu=mu*(1-rho)
sigma=0.1
# Data
y=rnorm(N)
h=rnorm(N)*sigma
for (t in 2:N)
h[t]=mu+h[t-1]*rho+h[t]
y=exp(h/2)*y
```

JPR's 1994 version

Notations

$$
\begin{aligned}
y_t &= \sqrt{h_t} u_t \\
\log h_t &= \alpha + \delta \log h_{t-1} + \sigma_\nu \nu_t \\
(u_t, \nu_t) &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

Note: No stationarity/stability constraint on δ for the AR model to be causal

Stochastic volatility (2)

Likelihood is not available:

$$\mathsf{L}(\mu,\tau,\varrho|y_{1:T}) = \int \mathsf{L}(\mu,\tau,\varrho|y_{1:T},\sigma_{1:T})\,\mathsf{d}\sigma_{1:T}$$

Stochastic volatility (2)

Likelihood is not available:

$$\mathsf{L}(\mu, \tau, \varrho | y_{1:T}) = \int \mathsf{L}(\mu, \tau, \varrho | y_{1:T}, \sigma_{1:T}) \, \mathsf{d}\sigma_{1:T}$$

Impossible to integrate out the σ_t's

Stochastic volatility (3)

Expression of the complete likelihood

$$\mathsf{L}(\mu,\tau,\varrho|y_{1:T},\sigma_{1:T})$$
$$\propto \prod_{t=1}^{T} \exp -\frac{y_t^2}{2\sigma_t^2}\ \exp -\frac{(\log\sigma_{t+1}^2-\mu-\varrho\log\sigma_t^2)^2}{2\tau^2}\,\frac{1}{\tau\sigma_t}$$

JPR's 1994 version

Priors

Standard (non-stationary) conjugate:

$$\alpha, \delta \sim \mathcal{N}(\mu, \sigma^2)$$
$$\sigma_\nu^2 \sim \mathcal{IG}(\nu_0, s_0^2)$$

The Metropolis-Hastings Algorithm

Stochastic volatility model

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

The Gibbs Sampler

Monte Carlo Integration

Sequential importance sampling

Running Monte Carlo via Markov Chains

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

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

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

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.

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**

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

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**

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 (2)

4. If

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

the chain is **irreducible**

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**

Convergence properties (2)

4. If

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

the chain is **irreducible**

5. For M-H, f-irreducibility implies **Harris recurrence**
6. Thus, for M-H satisfying (1) and (2)
   (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.

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

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

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

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,

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

and observables

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

Example (Noisy AR(1))

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

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

and observables

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

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 $\mathscr{N}(\mu_t, \omega_t^2)$ distribution, with

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

Example (Noisy AR(1) too)

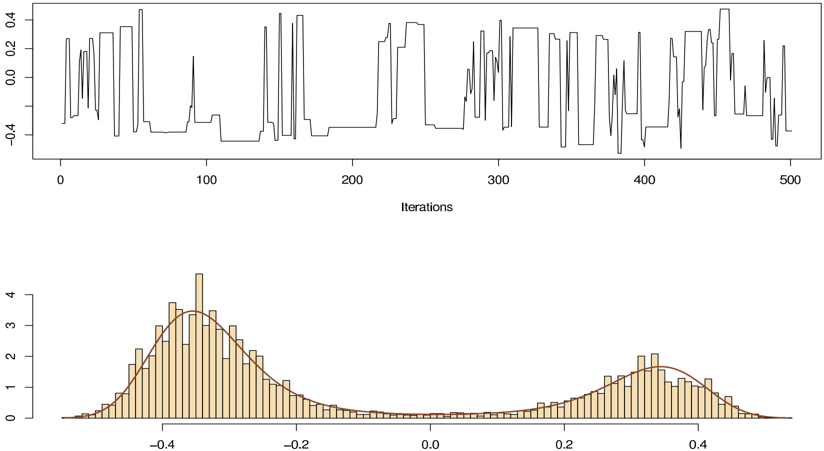
Use for proposal the $\mathscr{N}(\mu_t, \omega_t^2)$ distribution, with

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

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

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

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

Example (Cauchy by normal)

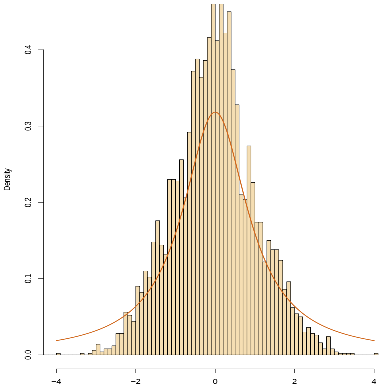
▸ go random W Given a Cauchy $\mathscr{C}(0,1)$ distribution, consider a normal $\mathscr{N}(0,1)$ proposal

The Metropolis–Hastings acceptance ratio is

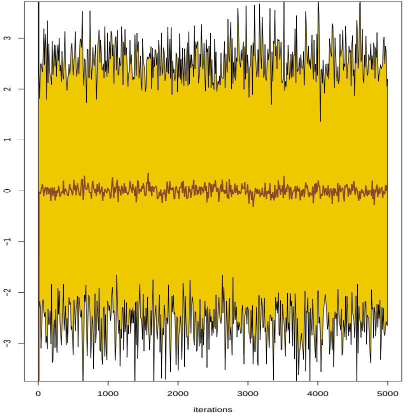
$$\frac{\pi(\xi')/\nu(\xi')}{\pi(\xi)/\nu(\xi))} = \exp\left[\left\{\xi^2 - (\xi')^2\right\}/2\right] \frac{1+(\xi')^2}{(1+\xi^2)}\,.$$

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]



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



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

JPR's 1994 version

Stranger version of independent MH where g is replaced with

$$g_c(x) \propto \min(f(x), cg(x))$$

with c calibrated so that the average acceptance is optimised.

[Tierney, 1994]

JPR's 1994 version

Stranger version of independent MH where g is replaced with

$$g_c(x) \propto \min(f(x), cg(x))$$

with c calibrated so that the average acceptance is optimised.

[Tierney, 1994]

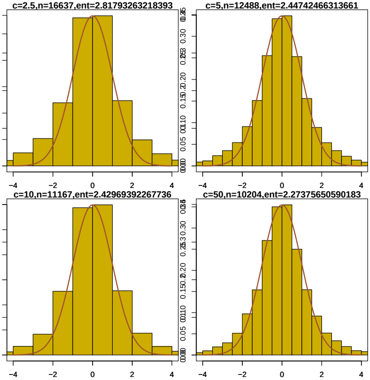
Q.: Is there any advantage in replacing g with g_c?

Maybe!

Larger c lead to better acceptance rates/entropy ratings...

Maybe!

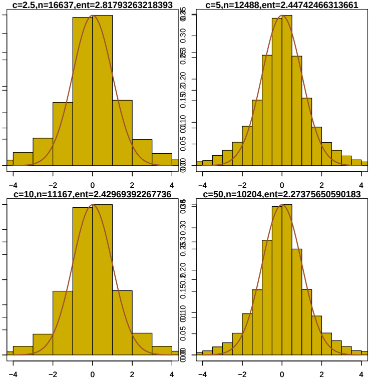
Larger c lead to better acceptance rates/entropy ratings...



Simulation of a $\mathcal{N}(0,1)$ using a Student's $\mathcal{T}(3,0,1)$ proposal and various c's

Maybe!

Larger c lead to better acceptance rates/entropy ratings...



Simulation of a $\mathcal{N}(0,1)$ using a Student's $\mathcal{T}(3,0,1)$ proposal and various c's

Interesting Master project!!!

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*

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

Corresponding pseudo-code

Algorithm (Random walk Metropolis)

Given $x^{(t)}$

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

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

Convergence properties

Uniform ergodicity prohibited by random walk structure

Convergence properties

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

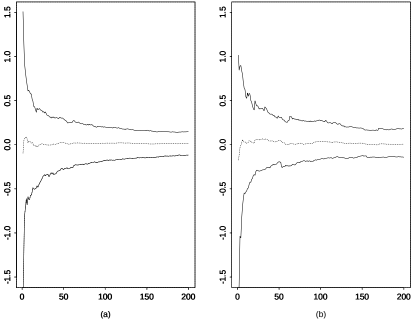
Theorem (Sufficient ergodicity)

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 $\mathscr{N}(0,1)$ 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

JPR's Metropolis-Hastings scheme

Use of a taylored proposal for the target

$$\begin{aligned} p(h_t|h_{t-1},h_{t+1},y_t) \propto \\ & h_t^{-.5} \exp\left\{-.5y_t^2/h_t\right\} 1/h_t \\ & \exp\left\{-(\log h_t - \mu_t)^2/(2\sigma^2)\right\} \end{aligned}$$

where

$$\mu_t = [\alpha(1-\delta) + \delta(\log h_{t+1} + \log h_{t-1})]/(1+\delta^2)$$

and

$$\sigma^2 = \sigma_\nu^2/(1+\delta^2)$$

JPR's Metropolis-Hastings proposal

Choice of an inverse Gamma density

$$h_t \sim \lambda^{\varphi} h^{-\varphi-1} \exp -\lambda/h$$

with

$$\varphi = (1 - 2\exp\sigma^2))/(1 - \exp(\sigma^2)) + .5$$

and

$$\lambda = (\varphi - 1)\exp(\mu_t + .5\sigma^2) + .5y_t^2$$

obtained by moment matching

JPR's Metropolis-Hastings proposal

Choice of an inverse Gamma density

$$h_t \sim \lambda^{\varphi} h^{-\varphi-1} \exp -\lambda/h$$

with

$$\varphi = (1 - 2\exp\sigma^2))/(1 - \exp(\sigma^2)) + .5$$

and

$$\lambda = (\varphi - 1)\exp(\mu_t + .5\sigma^2) + .5y_t^2$$

obtained by moment matching

JPR: distinguished from the inverted gamma density?!

Direct implementation

Not convinced that JPR's use of pseudo accept reject is relevant

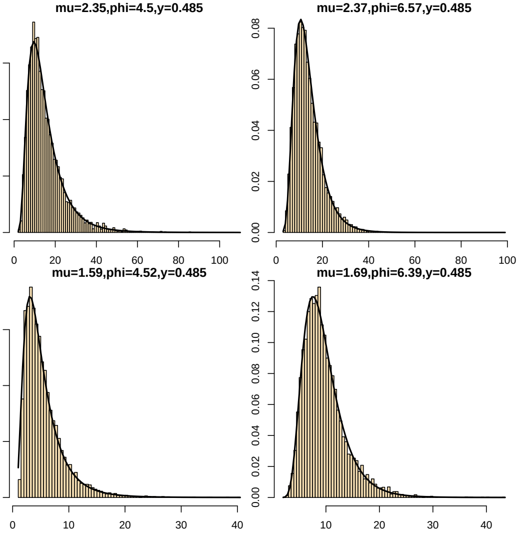
Direct implementation

Not convinced that JPR's use of pseudo accept reject is relevant

R code : iteration i

```
# Previous value
y=sample[i-1]
# Proposal
z=rgamma(1,phi)
if (runif(1) < f(lambda/z)*dgamma(1/lambda*y,phi)/
   (f(y)*dgamma(z,phi)))
sample[i]=z
else
sample[i]=y
```

Direct result



The Gibbs Sampler

The Gibbs Sampler

- General Principles
- Completion
- Convergence
- Data Augmentation
- 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

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 $\mathbf{X} = (X_1, \ldots, X_p)$

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

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

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

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

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

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

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

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

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

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

Latent variables are back

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

Purpose

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

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

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

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

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

JPR's Gibbs sampler

JPR's approach is a simple completion Gibbs sampler

JPR's Gibbs sampler

JPR's approach is a simple completion Gibbs sampler

Conditionals

$$\begin{aligned} \alpha, \delta, \sigma &\sim \pi(\alpha, \delta, \sigma | y_{1:T}, h_{1:T} \\ h_{1:T} &\sim \pi(h_{1:T} | y_{1:T}, \alpha, \delta, \sigma) \end{aligned}$$

JPR's Gibbs sampler

JPR's approach is a simple completion Gibbs sampler

Conditionals

$$\begin{aligned} \alpha, \delta, \sigma &\sim \pi(\alpha, \delta, \sigma | y_{1:T}, h_{1:T} \\ h_{1:T} &\sim \pi(h_{1:T} | y_{1:T}, \alpha, \delta, \sigma) \end{aligned}$$

except that $\pi(h_{1:T} | y_{1:T}, \alpha, \delta, \sigma)$ is not available!

More on JPR's Gibbs sampler

Instead, use of the full conditionals

$$\pi(h_t|h_{-t}, y_{1:T}, \alpha, \delta, \sigma) = \pi(h_t|h_{t-1}, h_{t+1}, y_t, \alpha, \delta, \sigma)$$

[thanks to the Markov property]

More on JPR's Gibbs sampler

Instead, use of the full conditionals

$$\pi(h_t|h_{-t}, y_{1:T}, \alpha, \delta, \sigma) = \pi(h_t|h_{t-1}, h_{t+1}, y_t, \alpha, \delta, \sigma)$$

[thanks to the Markov property]
and replacement of an exact simulation from $\pi(h_t|h_{t-1}, h_{t+1}, y_t, \alpha, \delta, \sigma)$ with *one single* hybrid Metropolis-Hastings step based on the Inverse Gamma approximation

◂ Use in the full Gibbs

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

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

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

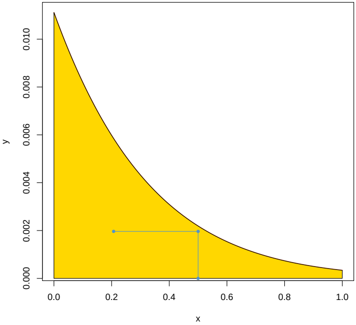
   $\ldots$

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

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

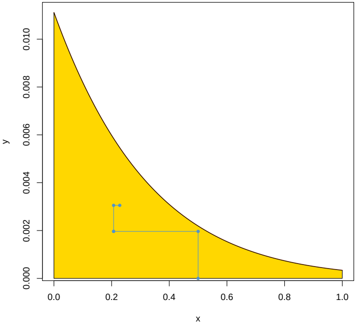
$$A^{(t+1)} = \{y;\ f_i(y) \geq \omega_i^{(t+1)},\ i = 1, \ldots, k\}.$$

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



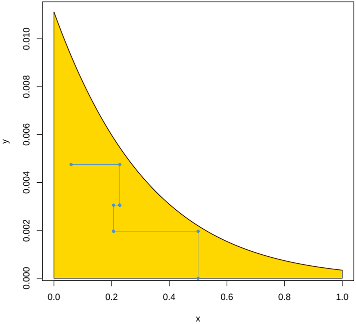
Number of Iterations 2

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



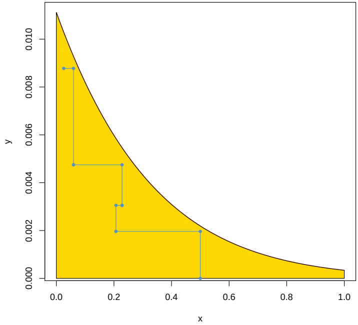
Number of Iterations 2, 3

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



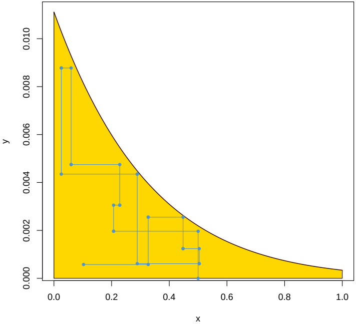
Number of Iterations 2, 3, 4

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



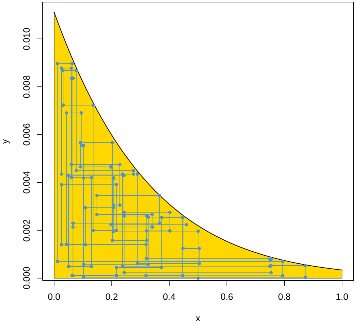
Number of Iterations 2, 3, 4, 5

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



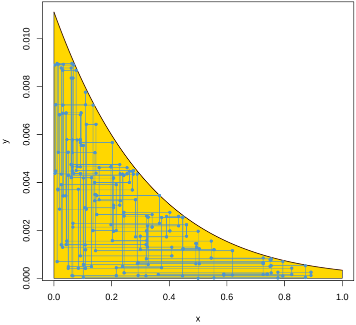
Number of Iterations 2, 3, 4, 5, 10

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



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

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.

Example (Stochastic volatility core distribution)

Difficult part of the stochastic volatility model is the distribution of the volatility vector $\sigma_{1:T}$
If we consider only σ_t conditional on σ_{t-1} and on y_t, we obtain a density of the kind (in $\log \sigma_t$)

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

simplified in

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

by a change of variable

Example (Stochastic volatility core distribution (2))

Slice sampling $\exp - \left\{x^2 + \alpha \exp(-x)\right\}$ means simulating 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 $\omega = -2 \log u$.

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

Example (Stochastic volatility core distribution (3))

Alternative with two uniforms

$$\exp - \left\{x^2 + \alpha \exp(-x)\right\} = \int \mathbb{I}_{u_1 \leq \exp_x 2} \mathbb{I}_{u_1 \leq \exp - \alpha \exp(-x)} \mathsf{d} u_1 \mathsf{d} u_2$$

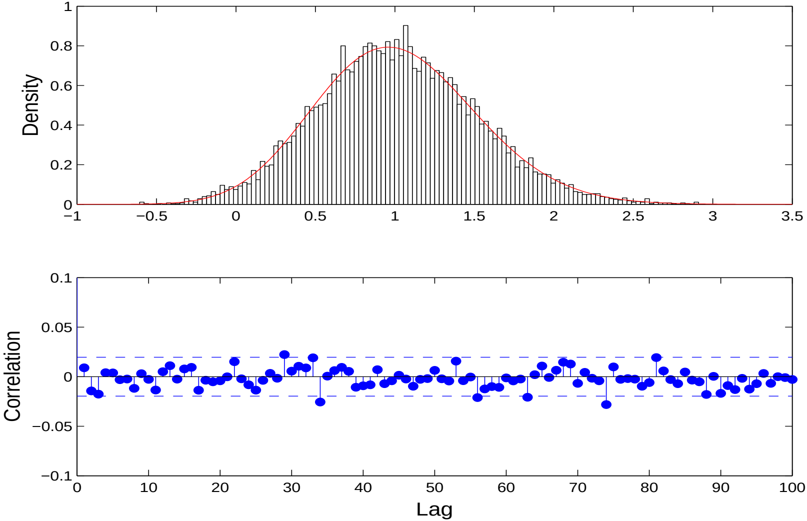
Example (Stochastic volatility core distribution (3))

Alternative with two uniforms

$$\exp - \left\{x^2 + \alpha \exp(-x)\right\} = \int \mathbb{I}_{u_1 \le \exp_x 2} \mathbb{I}_{u_1 \le \exp -\alpha \exp(-x)} \mathsf{d}u_1 \mathsf{d}u_2$$

R code

```
alpha=3
u=log(runif(2)*c(exp(-x*x),exp(-alpha*exp(-x))))
upa=sqrt(-u[1])
low=max(-upa,-log(-u[2]/alpha)))
x=runif(1,low,upa)
```



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.

A poor slice sampler

Example

Consider

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

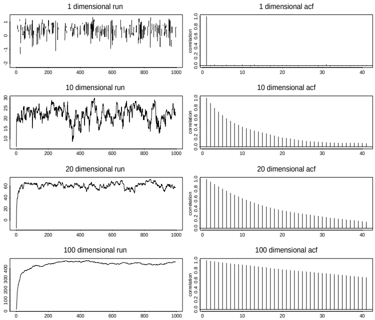
Slice sampler equivalent to one-dimensional slice sampler on

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

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)

Data Augmentation

The Gibbs sampler with only two steps is particularly useful

Algorithm (Data Augmentation)

Given $y^{(t)}$,

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

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

Data Augmentation

The Gibbs sampler with only two steps is particularly useful

Algorithm (Data Augmentation)

Given $y^{(t)}$,

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

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

Theorem (Markov property)

Both $(Y_1^{(t)})$ and $(Y_2^{(t)})$ are Markov chains, with transitions

$$\mathfrak{K}_i(x, x^*) = \int g_i(y|x) g_{3-i}(x^*|y)\, dy,$$

Example (Grouped counting data)

360 consecutive records of the number of passages per unit time

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

Example (Grouped counting data (2))

Feature Observations with 4 passages and more are grouped
If observations are Poisson $\mathscr{P}(\lambda)$, the likelihood is

$$\ell(\lambda|x_1,\ldots,x_5) \\ \propto e^{-347\lambda}\lambda^{128+55\times2+25\times3}\left(1-e^{-\lambda}\sum_{i=0}^{3}\frac{\lambda^i}{i!}\right)^{13},$$

which can be difficult to work with.

Example (Grouped counting data (2))

Feature Observations with 4 passages and more are grouped

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

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

which can be difficult to work with.

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

Algorithm (Poisson-Gamma Gibbs)

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

b Simulate

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

Algorithm (Poisson-Gamma Gibbs)

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

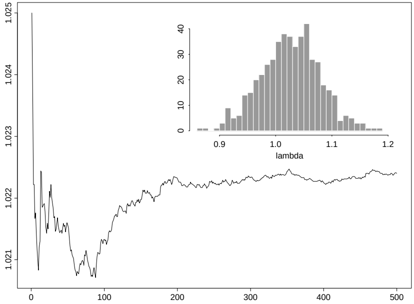
b Simulate

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

The Bayes estimator

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

converges quite rapidly ▸ to R& B



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.

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 (2)

Then

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

Rao-Blackwellization (2)

Then

- Both estimators converge to $\mathbb{E}[h(Y_1)]$
- Both are unbiased,
- and
$$\text{var}\left(\mathbb{E}\left[h(Y_1)|Y_2^{(t)},\ldots,Y_p^{(t)}\right]\right)\le\text{var}(h(Y_1)),$$
so δ_{rb} is uniformly better (for Data Augmentation)

Improper Priors

↯ Unsuspected danger resulting from careless use of MCMC algorithms:

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...

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

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

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

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**

Example (Improper random effects)

Consider

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

where

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

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

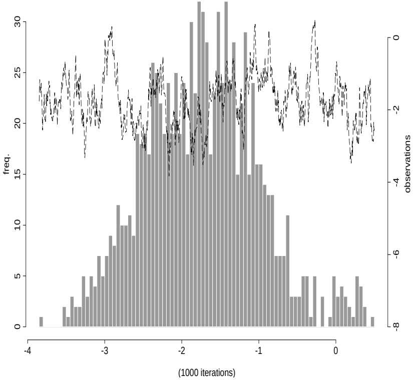
$$\pi(\mu, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2}\ .$$

Example (Improper random effects 2)

The conditional distributions

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

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

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.

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

Monte Carlo integration

Stochastic volatility model

The Metropolis-Hastings Algorithm

The Gibbs Sampler

Monte Carlo Integration

- Introduction
- Monte Carlo integration
- Importance Sampling
- Acceleration methods
- Bayesian importance sampling

Sequential importance sampling

Problems with numerical solutions

Two major classes of numerical problems that arise in statistical inference

- **Optimization** - generally associated with the likelihood approach

Problems with numerical solutions

Two major classes of numerical problems that arise in statistical inference

- **Optimization** - generally associated with the likelihood approach
- **Integration**- generally associated with the Bayesian approach

Monte Carlo integration

Theme:

Generic problem of evaluating the integral

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

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

Monte Carlo integration (2)

Monte Carlo solution

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

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

Monte Carlo integration (2)

Monte Carlo solution

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

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

which converges

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

by the **Strong Law of Large Numbers**

Monte Carlo precision

Estimate the variance with

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

and for m large,

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

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

Example (Cauchy prior/normal sample)

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

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

Under squared error loss, posterior mean

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

Example (Cauchy prior/normal sample (2))

Form of δ^π suggests simulating iid variables

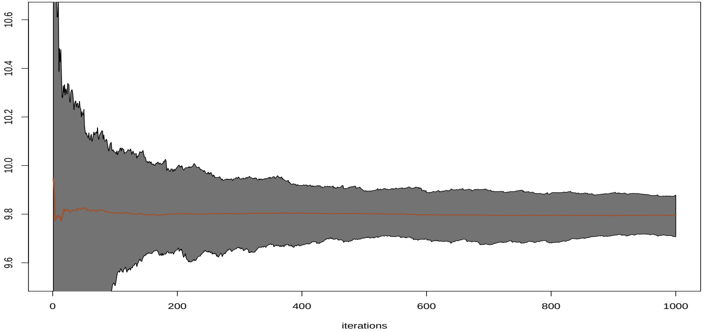
$$\theta_1, \cdots, \theta_m \sim \mathscr{N}(x, 1)$$

and calculating

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

The Law of Large Numbers implies

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



Range of estimators δ_m^π for 100 runs and $x = 10$

Importance sampling

Paradox

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

Importance sampling

Paradox

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

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

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

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

Importance sampling algorithm

Evaluation of

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

by

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

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

Same thing as before!!!

Convergence of the estimator

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

Same thing as before!!!

Convergence of the estimator

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

converges for any choice of the distribution g
[as long as $supp(g) \supset supp(f)$**]**

Important details

- Instrumental distribution g chosen from distributions easy to simulate
- The same sample (generated from g) can be used repeatedly, not only for different functions h, but also for different densities f
- Even dependent proposals can be used, as seen later

▸ PMC chapter

Important choice

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

- Finite variance only when

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

Important choice

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

- Finite variance only when

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

- Instrumental distributions with tails lighter than those of f (that is, with $\sup f/g = \infty$) not appropriate.
- If $\sup f/g = \infty$, the weights $f(x_j)/g(x_j)$ vary widely, giving too much importance to a few values x_j.

Important choice

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

- Finite variance only when

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

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

Optimal importance function

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

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

Optimal importance function

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

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

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

Practical impact

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

where f and g are known up to constants.

- ○ Also converges to $\mathfrak{I}$ by the Strong Law of Large Numbers.
- ○ Biased, but the bias is quite small

Practical impact

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

where f and g are known up to constants.

- Also converges to $\mathfrak{I}$ by the Strong Law of Large Numbers.
- Biased, but the bias is quite small
- In some settings beats the unbiased estimator in squared error loss.
- Using the ‘optimal’ solution does not always work:

$$\frac{\sum_{j=1}^{m} h(x_j)\; f(x_j)/|h(x_j)|\; f(x_j)}{\sum_{j=1}^{m} f(x_j)/|h(x_j)|\; f(x_j)} = \frac{\#\text{positive } h - \#\text{negative } h}{\sum_{j=1}^{m} 1/|h(x_j)|}$$

Selfnormalised importance sampling

For ratio estimator

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

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

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

Selfnormalised variance

then

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

for

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

Rough approximation

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

IS suffers from curse of dimensionality

As dimension increases, discrepancy between importance and target worsens

▸ skip explanation

IS suffers from curse of dimensionality

As dimension increases, discrepancy between importance and target worsens

▸ skip explanation

Explanation:

Take target distribution μ and instrumental distribution ν
Simulation of a sample of iid samples of size n $x_{1:n}$ from $\mu_n = \mu^{\bigotimes n}$
Importance sampling estimator for $\mu_n(f_n) = \int f_n(x_{1:n})\mu_n(dx_{1:n})$

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

where $W_k^i = \frac{d\mu}{d\nu}(\xi_k^i)$, and ξ_j^i are iid with distribution ν.
For $\{V_k\}_{k\geq 0}$, sequence of iid nonnegative random variables and for $n \geq 1$, $\mathcal{F}_n = \sigma(V_k; k \leq n)$, set

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

IS suffers (2)

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

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

and thus $\{U_n\}_{n\geq 0}$ **martingale**
Since $x \mapsto \sqrt{x}$ concave, by Jensen's inequality,

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

and thus $\{\sqrt{U_n}\}_{n\geq 0}$ **supermartingale**
Assume $\mathbb{E}(\sqrt{V_{n+1}}) < 1$. Then

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

IS suffers (3)

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

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

Hence, $Z = 0$ and $U_n \to 0$ a.s., which implies that the martingale $\{U_n\}_{n\geq 0}$ is not regular.
Apply these results to $V_k = \frac{d\mu}{d\nu}(\xi_k^i)$, $i \in \{1, \ldots, N\}$:

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

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

Thus all importance weights converge to 0

▸ too volatile!

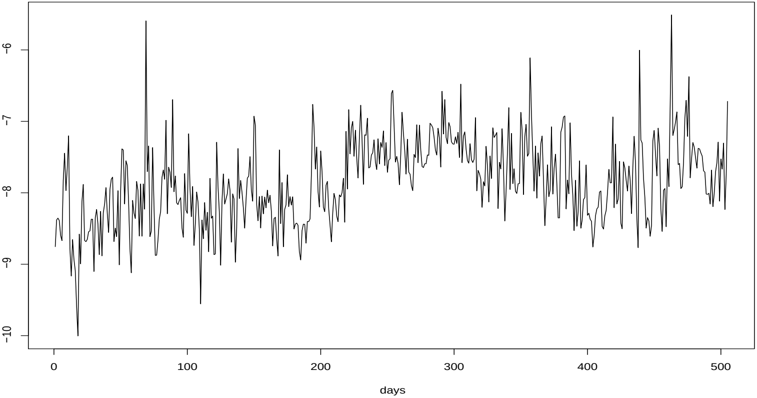
Example (Stochastic volatility model)

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

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

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

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



Example (Stochastic volatility model (2))

Observed likelihood unavailable in closed from.
Joint posterior (or conditional) distribution of the hidden state sequence $\{X_k\}_{1\leq k\leq K}$ can be evaluated explicitly

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

up to a normalizing constant.

Computational problems

Example (Stochastic volatility model (3))

Direct simulation from this distribution impossible because of

(a) dependence among the X_k's,

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

(c) exponential term $\exp(-x_k)y_k^2$ within (1).

Importance sampling

Example (Stochastic volatility model (4))

Natural candidate: replace the exponential term with a quadratic approximation to preserve Gaussianity.
E.g., expand $\exp(-x_k)$ around its conditional expectation ϕx_{k-1} as

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

Example (Stochastic volatility model (5))

Corresponding Gaussian importance distribution with mean

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

and variance

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

Prior proposal on X_1,

$$X_1 \sim \mathcal{N}(0, \sigma^2)$$

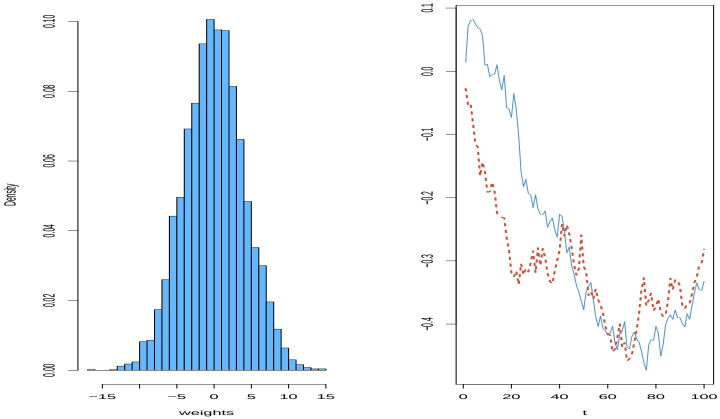
Example (Stochastic volatility model (6))

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

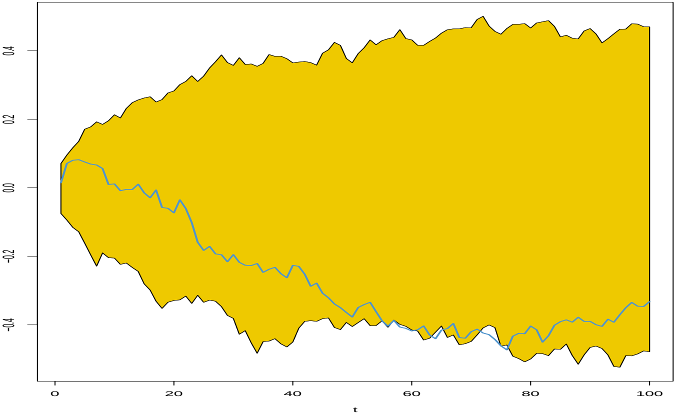
Importance weight computed sequentially as the product of

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

$(1 \leq k \leq K)$



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



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

Correlated simulations

Negative correlation reduces variance

Special technique — but efficient when it applies

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

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

by

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

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

Variance reduction

Variance of the average

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

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

$$\mathrm{cov}(\widehat{\mathfrak{I}}_1, \widehat{\mathfrak{I}}_2) \le 0\,,$$

they improve on two independent samples of same size

Antithetic variables

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

Control variates

▸ out of control!

For

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

unknown and

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

known,

$\mathfrak{I}_0$ estimated by $\widehat{\mathfrak{I}}_0$ and

$\mathfrak{I}$ estimated by $\widehat{\mathfrak{I}}$

Control variates (2)

Combined estimator

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

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

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

Optimal control

Optimal choice of β

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

with

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

where ρ correlation between $\widehat{\mathfrak{I}}$ and $\widehat{\mathfrak{I}}_0$
Usual solution: **regression coefficient of** $h(x_i)$ **over** $h_0(x_i)$

Example (Quantile Approximation)

Evaluate

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

by

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

with X_i iid f.
If $\mathsf{Pr}(X > \mu) = \frac{1}{2}$ known

Example (Quantile Approximation (2))

Control variate

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

improves upon $\widehat{\varrho}$ if

Example (Quantile Approximation (2))

Control variate

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

improves upon $\widehat{\varrho}$ if

$$\beta < 0 \quad \text{and} \quad |\beta| < 2\frac{\mathrm{cov}(\widehat{\varrho}, \widehat{\varrho}_0)}{\mathrm{var}(\widehat{\varrho}_0)}2\frac{\mathsf{Pr}(X > a)}{\mathsf{Pr}(X > \mu)}.$$

Integration by conditioning

Use **Rao-Blackwell Theorem**

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

Consequence

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

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

the estimator

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

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

▸ skip expectation

Example (Student's t expectation)

For

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

a Student's t distribution can be simulated as

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

Example (Student's t expectation (2))

Empirical distribution

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

can be improved from the joint sample

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

Example (Student's t expectation (2))

Empirical distribution

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

can be improved from the joint sample

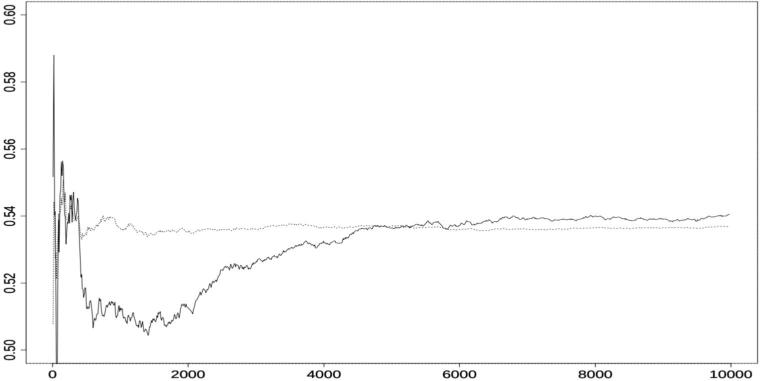
$$((X_1, Y_1), \ldots, (X_m, Y_m))$$

since

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

is the conditional expectation.

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



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

Bayesian importance functions

Recall algorithm:

1. Generate $\theta_1^{(1)}, \cdots, \theta_1^{(T)}$ from $cg(\theta)$ with $c^{-1} = \int g(\theta) d\theta$
2. Take

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

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

[Marginal approximation]

Choice of g

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

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

- often inefficient if data informative
- impossible if π is improper

Choice of g

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

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

- often inefficient if data informative
- impossible if π is improper

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

- c unknown
- $m^{IS}(x) = 1 \Bigg/ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{f(x|\theta^{(t)})}$
- improper priors allowed

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

- defensive mixture
- $\rho \ll 1$ Ok

[Hestenberg, 1998]

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

- defensive mixture
- $\rho \ll 1$ Ok

[Hestenberg, 1998]

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

- $m^h(x) = \dfrac{1}{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\dfrac{h(\theta)}{f(x|\theta)\pi(\theta)}}$
- works for any h
- finite variance if

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

Bridge sampling

[Chen & Shao, 1997]

Given two models $f_1(x|\theta_1)$ and $f_2(x|\theta_2)$,

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

Bayes factor:

$$B_{12}(x) = \frac{m_1(x)}{m_2(x)}$$

ratio of normalising constants

Bridge sampling (2)

(i) Missing normalising constants:

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

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

Bridge sampling (3)

(ii) Still missing normalising constants:

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

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

Bridge sampling (4)

Optimal choice

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

[Chen, Meng & Wong, 2000]

Sequential importance sampling

◂ basic importance

Sequential importance sampling

- Adaptive MCMC
- Importance sampling revisited
- Dynamic extensions
- Population Monte Carlo

Adaptive MCMC is not possible

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

Adaptive MCMC is not possible

↯ **Algorithms trained on-line usually invalid:**
using the whole past of the “chain” implies that this is not a Markov chain any longer!

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,

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

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

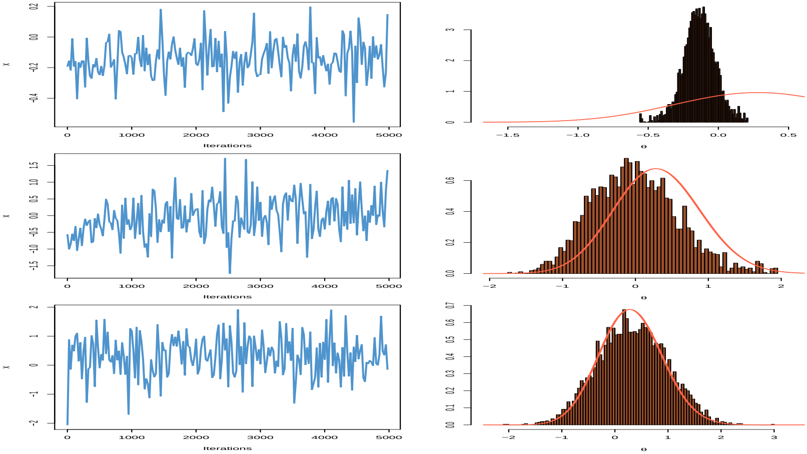
$$\prod_{j=2}^{n} \left[\frac{\nu + (x_j - \theta^{(t)})^2}{\nu + (x_j - \xi)^2} \right]^{-(\nu+1)/2} \frac{\exp -(\mu_t - \theta^{(t)})^2 / 2\sigma_t^2}{\exp -(\mu_t - \xi)^2 / 2\sigma_t^2} \,,$$

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

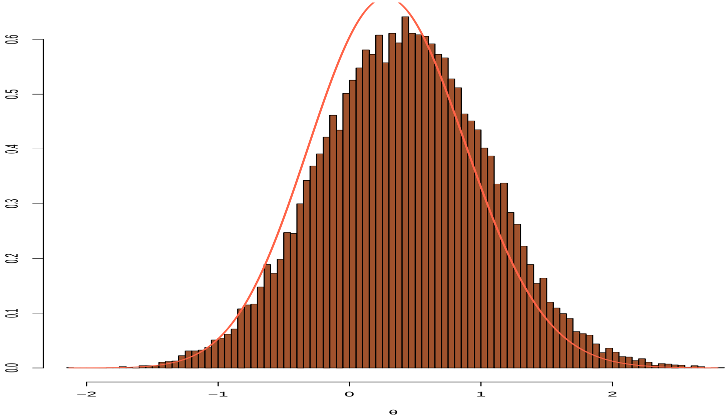
Example (Poly t distribution (2))

Invalid scheme:

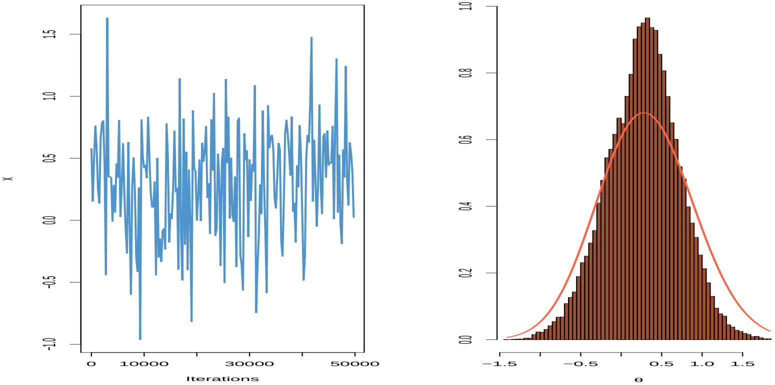
- when range of initial values too small, the $\theta^{(i)}$'s cannot converge to the target distribution and concentrates on too small a support.
- long-range dependence on past values modifies the distribution of the sequence.
- using past simulations to create a non-parametric approximation to the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



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.



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



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

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

Markov extension

For densities f and g, and importance weight

$$\omega(x) = f(x)/g(x)\,,$$

for any kernel $K(x, x')$ with stationary distribution f,

$$\int \omega(x)\,K(x, x')\,g(x)dx = f(x')\,.$$

[McEachern, Clyde, and Liu, 1999]

Markov extension

For densities f and g, and importance weight

$$\omega(x) = f(x)/g(x)\,,$$

for any kernel $K(x,x')$ with stationary distribution f,

$$\int \omega(x)\, K(x,x')\, g(x) dx = f(x')\,.$$

[McEachern, Clyde, and Liu, 1999]

Consequence: An importance sample transformed by MCMC transitions keeps its weights

Unbiasedness preservation:

$$\begin{aligned}\mathbb{E}\left[\omega(X)h(X')\right] &= \int \omega(x)\, h(x')\, K(x,x')\, g(x)\, dx\, dx' \\ &= \mathbb{E}_f\left[h(X)\right]\end{aligned}$$

Not so exciting!

The weights do not change!

Not so exciting!

The weights do not change!

If x has small weight

$$\omega(x) = f(x)/g(x)\,,$$

then

$$x' \sim K(x, x')$$

keeps this small weight.

Pros and cons of importance sampling vs. MCMC

- Production of a sample (IS) vs. of a Markov chain (MCMC)
- Dependence on importance function (IS) vs. on previous value (MCMC)
- Unbiasedness (IS) vs. convergence to the true distribution (MCMC)
- Variance control (IS) vs. learning costs (MCMC)
- Recycling of past simulations (IS) vs. progressive adaptability (MCMC)
- Processing of moving targets (IS) vs. handling large dimensional problems (MCMC)
- **Non-asymptotic validity (IS) vs. difficult asymptotia for adaptive algorithms (MCMC)**

Dynamic importance sampling

Idea

It is possible to generalise importance sampling using random weights ω_t

Dynamic importance sampling

Idea

It is possible to generalise importance sampling using random weights ω_t such that

$$\mathbb{E}[\omega_t|x_t] = \pi(x_t)/g(x_t)$$

(a) Self-regenerative chains

[Sahu & Zhigljavsky, 1998; Gasemyr, 2002]

Proposal

$$Y \sim p(y) \propto \tilde{p}(y)$$

and target distribution $\pi(y) \propto \tilde{\pi}(y)$

Ratios

$$\underbrace{\omega(x) = \pi(x)/p(x)}_{\textbf{Unknown}} \qquad \text{and} \qquad \underbrace{\tilde{\omega}(x) = \tilde{\pi}(x)/\tilde{p}(x)}_{\textbf{Known}}$$

Acceptance function

$$\alpha(x) = \frac{1}{1 + \kappa\tilde{\omega}(x)} \qquad \kappa > 0$$

Geometric jumps

Theorem

If

$$Y \sim p(y)$$

and

$$W|Y = y \sim \mathscr{G}(\alpha(y))\,,$$

then

$$X_t = \cdots = X_{t+W-1} = Y \neq X_{t+W}$$

defines a Markov chain with stationary distribution π

Plusses

- Valid for any choice of κ [κ small $=$ large variance and κ large $=$ slow convergence]
- Only depends on current value [Difference with Metropolis]
- Random integer weight W [Similarity with Metropolis]
- Saves on the rejections: always accept [Difference with Metropolis]
- Introduces geometric noise compared with importance sampling

  $$\sigma^2_{SZ} = 2\,\sigma^2_{IS} + (1/\kappa)\sigma^2_\pi$$

- Can be used with a sequence of proposals p_k and constants κ_k [Adaptativity]

A generalisation

[Gåsemyr, 2002]

Proposal density $p(y)$ and probability $q(y)$ of accepting a jump.

A generalisation

[Gåsemyr, 2002]

Proposal density $p(y)$ and probability $q(y)$ of accepting a jump.

Algorithm (Gåsemyr's dynamic weights)

Generate a sequence of **random weights** W_n by

1. Generate $Y_n \sim p(y)$
2. Generate $V_n \sim \mathcal{B}(q(y_n))$
3. Generate $S_n \sim \mathcal{G}eo(\alpha(y_n))$
4. Take $W_n = V_n S_n$

Validation

▸ direct to PMC

$$\phi(y) = \frac{p(y)q(y)}{\int p(y)q(y)dy},$$

the chain (X_t) associated with the sequence (Y_n, W_n) by

$$Y_1 = X_1 = \cdots = X_{1+W_1-1}, Y_2 = X_{1+W_1} = \cdots$$

is a Markov chain with transition

$$K(x, y) = \alpha(x)\phi(y)$$

which has a point mass at $y = x$ with weight $1 - \alpha(x)$.

Ergodicity for Gåsemyr's scheme

Necessary and sufficient condition

π is stationary for (X_t) iff

$$\alpha(y) = q(y)/(\kappa\pi(y)/p(y)) = q(y)/(\kappa w(y))$$

for some constant κ.

Ergodicity for Gåsemyr's scheme

Necessary and sufficient condition

π is stationary for (X_t) iff

$$\alpha(y) = q(y)/(\kappa\pi(y)/p(y)) = q(y)/(\kappa w(y))$$

for some constant κ.

Implies that

$$\mathbb{E}[W^n|Y^n = y] = \kappa w(y)\,.$$

[Average importance sampling]

Special case: $\alpha(y) = 1/(1 + \kappa w(y))$ of Sahu and Zhigljavski (2001)

Properties

Constraint on κ: for $\alpha(y) \leq 1$, κ must be such that

$$\frac{p(y)q(y)}{\pi(y)} \leq \kappa$$

Reverse of accept-reject conditions (!)
Variance of

$$\sum_n W_n h(Y_n) / \sum_n W_n \tag{2}$$

is

$$2 \int \frac{(h(y)-\mu)^2}{q(y)} w(y)\pi(y) dy - (1/\kappa)\sigma_\pi^2 \,,$$

by Cramer-Wold/Slutsky
Still worse than importance sampling.

(b) Dynamic weighting

[Wong & Liang, 1997; Liu, Liang & Wong, 2001; Liang, 2002]

direct to PMC

Generalisation of the above: simultaneous generation of points and weights, (θ_t, ω_t), under the constraint

$$\mathbb{E}[\omega_t|\theta_t] \propto \pi(\theta_t) \tag{3}$$

Same use as importance sampling weights

Algorithm (Liang's dynamic importance sampling)

1. Generate $y \sim K(x, y)$ and compute

$$\varrho = \omega \, \frac{\pi(y) K(y, x)}{\pi(x) K(x, y)}$$

2. Generate $u \sim \mathcal{U}(0, 1)$ and take

$$(x', \omega') = \begin{cases} (y, (1+\delta)\varrho/a) & \text{if } u < a \\ (x, (1+\delta)\omega/(1-a) & \text{otherwise} \end{cases}$$

where $a = \varrho/(\varrho + \theta)$, $\theta = \theta(x, \omega)$, and $\delta > 0$ constant or independent rv

Preservation of the equilibrium equation

If g_- and g_+ denote the distributions of the augmented variable (X, W) before the step and after the step, respectively, then

$$
\begin{aligned}
&\int_0^\infty \omega' \, g_+(x', \omega') \, d\omega' = \\
&\int (1+\delta) \left[\varrho(\omega, x, x') + \theta\right] \, g_-(x, \omega) \, K(x, x') \frac{\varrho(\omega, x, x')}{\varrho(\omega, x, x') + \theta} \, dx \, d\omega \\
&+ \int (1+\delta) \frac{\omega(\varrho(\omega, x', z) + \theta)}{\theta} \, g_-(x', \omega) \, K(x, z) \frac{\theta}{\varrho(\omega, x', z) + \theta} \, dz \, d\omega \\
&= (1+\delta) \left\{ \int \omega \, g_-(x, \omega) \, \frac{\pi(x') K(x', x)}{\pi(x)} \, dx \, d\omega \right. \\
&+ \left. \int \omega \, g_-(x', \omega) \, K(x', z) \, dz \, d\omega \right\} \\
&= (1+\delta) \left\{ \pi(x') \int c_0 \, K(x', x) \, dx + c_0 \pi(x') \right\} \\
&= 2(1+\delta) c_0 \pi(x') \, ,
\end{aligned}
$$

where c_0 proportionality constant

Special case: R-move

[Liang, 2002]

$\delta = 0$ and $\theta \equiv 1$, and thus

$$(x', \omega') = \begin{cases} (y, \varrho + 1) & \text{if } u < \varrho/(\varrho + 1) \\ (x, \omega(\varrho + 1)) & \text{otherwise,} \end{cases}$$

[Importance sampling]

Special case: W-move

$\theta \equiv 0$, thus $a = 1$ and

$$(x', \omega') = (y, \varrho)\,.$$

Q-move

[Liu & al, 2001]

$$(x', \omega') = \begin{cases} (y, \theta \vee \varrho) & \text{if } u < 1 \wedge \varrho/\theta\,, \\ (x, a\omega) & \text{otherwise,} \end{cases}$$

with $a \geq 1$ either a constant or an independent random variable.

Notes

- Updating step in Q and R schemes written as

$$(x_{t+1}, \omega_{t+1}) = \{x_t, \omega_t/\mathsf{Pr}(R_t = 0)\}$$

with probability $\mathsf{Pr}(R_t = 0)$ and

$$(x_{t+1}, \omega_{t+1}) = \{y_{t+1}, \omega_t r(x_t, y_{t+1})/\mathsf{Pr}(R_t = 1)\}$$

with probability $\mathsf{Pr}(R_t = 1)$, where R_t is the move indicator and

$$y_{t+1} \sim K(x_t, y)$$

Notes (2)

- Geometric structure of the weights

$$\mathrm{Pr}(R_t = 0) = \frac{\omega_t}{\omega_{t+1}}\,.$$

and

$$\mathrm{Pr}(R_t = 0) = \frac{\omega_t\, r(x_t, y_t)}{\omega_t\, r(x_t, y_t) + \theta}, \quad \theta > 0\,,$$

for the R scheme

Notes (2)

- Geometric structure of the weights

$$\mathsf{Pr}(R_t = 0) = \frac{\omega_t}{\omega_{t+1}}\,.$$

  and

$$\mathsf{Pr}(R_t = 0) = \frac{\omega_t\, r(x_t, y_t)}{\omega_t\, r(x_t, y_t) + \theta}, \quad \theta > 0\,,$$

  for the R scheme

- Number of steps T before an acceptance (a jump) such that

$$\begin{aligned} \mathsf{Pr}\,(T \geq t) &= P(R_1 = 0, \ldots, R_{t-1} = 0) \\ &= \mathbb{E}\left[\prod_{j=0}^{t-1} \frac{\omega_j}{\omega_{j+1}}\right] \propto \mathbb{E}[1/\omega_t]\,. \end{aligned}$$

Alternative scheme

Preservation of weight expectation:

$$(x_{t+1}, \omega_{t+1}) = \begin{cases} (x_t, \alpha_t \omega_t / \mathsf{Pr}(R_t = 0)) \\ \quad \text{with probability } \mathsf{Pr}(R_t = 0) \text{ and} \\ (y_{t+1}, (1-\alpha_t)\omega_t r(x_t, y_{t+1}) / \mathsf{Pr}(R_t = 1)) \\ \quad \text{with probability } \mathsf{Pr}(R_t = 1). \end{cases}$$

Alternative scheme (2)

Then

$$\begin{aligned}\Pr(T=t) &= P(R_1=0,\dots,R_{t-1}=0,R_t=1)\\ &= \mathbb{E}\left[\prod_{j=0}^{t-1}\alpha_j\frac{\omega_j}{\omega_{j+1}}(1-\alpha_t)\frac{\omega_{t-1}r(x_0,Y_t)}{\omega_t}\right]\end{aligned}$$

which is equal to

$$\alpha^{t-1}(1-\alpha)\mathbb{E}[\omega_o\,r(x,Y_t)/\omega_t]$$

when α_j constant and deterministic.

Example

Choose a function $0 < \beta(\cdot,\cdot) < 1$ and to take, while in (x_0, ω_0),

$$(x_1, \omega_1) = \left(y_1, \frac{\omega_0 r(x_0, y_1)}{\alpha(x_0, y_1)} (1 - \beta(x_0, y_1) \right)$$

with probability

$$\min(1, \omega_0 r(x_0, y_1)) \overset{\Delta}{=} \alpha(x_0, y_1)$$

and

$$(x_1, \omega_1) = \left(x_0, \frac{\omega_0}{1 - \alpha(x_0, y_1)} \times \beta(x_0, y_1) \right)$$

with probability $1 - \alpha(x_0, y_1)$.

Population Monte Carlo

Idea

Simulate from the product distribution

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

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

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

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

Sequential variance decomposition

Furthermore,

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

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

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

Validation

▸ skip validation

$$\begin{aligned}
&\mathbb{E}\left[\varrho_i^{(t)} h(X_i^{(t)})\ \varrho_j^{(t)} h(X_j^{(t)})\right] \\
&\quad = \int h(x_i) \frac{\pi(x_i)}{q_{it}(x_i|\mathbf{x}^{(t-1)})}\, \frac{\pi(x_j)}{q_{jt}(x_j|\mathbf{x}^{(t-1)})}\, h(x_j) \\
&\qquad\qquad q_{it}(x_i|\mathbf{x}^{(t-1)})\, q_{jt}(x_j|\mathbf{x}^{(t-1)})\, dx_i\, dx_j\, g(\mathbf{x}^{(t-1)}) d\mathbf{x}^{(t-1)} \\
&\quad = \mathbb{E}_\pi\left[h(X)\right]^2
\end{aligned}$$

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

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]

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

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

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**

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!]

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 $J_{i,t} \sim \mathcal{M}((\bar{\varrho}_i^{(t)})_{1 \leq i \leq N})$ 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 $p_{t,\ell}$.

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 $(K_{i,t})_{1 \le i \le N} \sim \mathscr{M}((p_{t,k})_{1 \le k \le D})$
    for $1 \le i \le N$, generate
        $$\tilde{x}_{i,t} \sim \mathfrak{K}_{K_{i,t}}(x)$$
    compute and renormalize the importance weights $\omega_{i,t}$
    generate $(J_{i,t})_{1 \le i \le N} \sim \mathscr{M}((\overline{\omega}_{i,t})_{1 \le i \le N})$
    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

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

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)

Saved by Rao-Blackwell!!

Modification: Rao-Blackwellisation (=conditioning)

Use the whole mixture in the importance weight:

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

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 $(t = 1, \dots, T)$,

Generate

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

Generate

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

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

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

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

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

on the simplex

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

and define the sequence

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

Kullback divergence

Definition (Kullback divergence)

For $\alpha \in S$,

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

Kullback divergence between Π and the mixture.

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

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 $\boldsymbol{\alpha} \in \mathfrak{S}_D$*,*

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

Connection with RBDPMCA ??

Theorem

Under the assumption

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

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

An integrated EM interpretation

▸ skip interpretation

We have

$$\boldsymbol{\alpha}^{\min} = \arg\min_{\boldsymbol{\alpha}\in S} KL(\boldsymbol{\alpha}) = \arg\max_{\boldsymbol{\alpha}\in S} \int \log p_{\boldsymbol{\alpha}}(\bar{x}) \Pi \otimes \Pi(d\bar{x})$$

$$= \arg\max_{\boldsymbol{\alpha}\in S} \int \log \int p_{\boldsymbol{\alpha}}(\bar{x}, K) dK \, \Pi \otimes \Pi(d\bar{x})$$

for $\bar{x} = (x, x')$ and $K \sim \mathcal{M}((\alpha_d)_{1\leq d\leq D})$. Then $\boldsymbol{\alpha}^{t+1} = F(\boldsymbol{\alpha}^t)$ means

$$\boldsymbol{\alpha}^{t+1} = \arg\max_{\boldsymbol{\alpha}} \iint \mathbb{E}_{\boldsymbol{\alpha}^t}(\log p_{\boldsymbol{\alpha}}(\bar{X}, K) | \bar{X} = \bar{x}) \Pi \otimes \Pi(d\bar{x})$$

and

$$\lim_{t\to\infty} \boldsymbol{\alpha}^t = \boldsymbol{\alpha}^{\min}$$

Illustration

Example (A toy example)

Take the target

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

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

[Surprise!!!]

Illustration

Example (A toy example)

Take the target

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

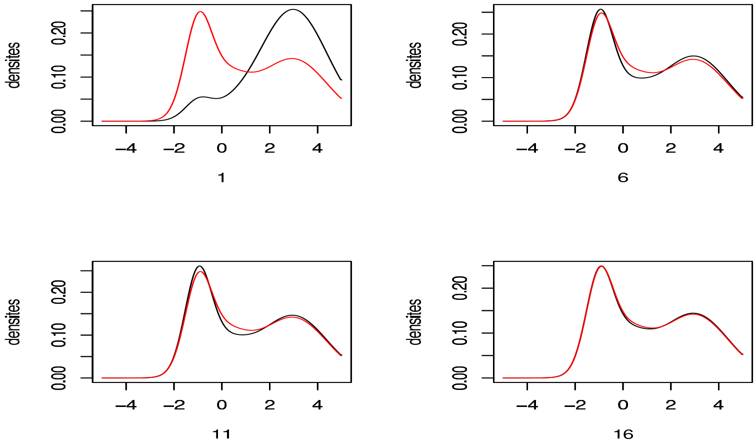
and use 3 proposals: $\mathscr{N}(-1,0.3)$, $\mathscr{N}(0,1)$ and $\mathscr{N}(3,2)$

[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



Target and mixture evolution

Example : PMC for mixtures

Observation of an iid sample $\mathbf{x} = (x_1, \ldots, x_n)$ from

$$p\mathcal{N}(\mu_1, \sigma^2) + (1-p)\mathcal{N}(\mu_2, \sigma^2),$$

with $p \neq 1/2$ and $\sigma > 0$ known.
Usual $\mathcal{N}(\theta, \sigma^2/\lambda)$ prior on μ_1 and μ_2:

$$\pi(\mu_1, \mu_2 | \mathbf{x}) \propto f(\mathbf{x} | \mu_1, \mu_2)\, \pi(\mu_1, \mu_2)$$

Algorithm (Mixture PMC)

Step 0: Initialisation

For $j = 1, \ldots, n = pm$, choose $(\mu_1)_j^{(0)}, (\mu_2)_j^{(0)}$

For $k = 1, \ldots, p$, set $r_k = m$

Step i: Update $(i = 1, \ldots, I)$

For $k = 1, \ldots, p$,

1. generate a sample of size r_k as

$$(\mu_1)_j^{(i)} \sim \mathcal{N}\left((\mu_1)_j^{(i-1)}, v_k\right) \quad \text{and} \quad (\mu_2)_j^{(i)} \sim \mathcal{N}\left((\mu_2)_j^{(i-1)}, v_k\right)$$

2. compute the weights

$$\varrho_j \propto \frac{f\left(\mathbf{x}\middle|(\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right)\pi\left((\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right)}{\varphi\left((\mu_1)_j^{(i)}\middle|(\mu_1)_j^{(i-1)}, v_k\right)\varphi\left((\mu_2)_j^{(i)}\middle|(\mu_2)_j^{(i-1)}, v_k\right)}$$

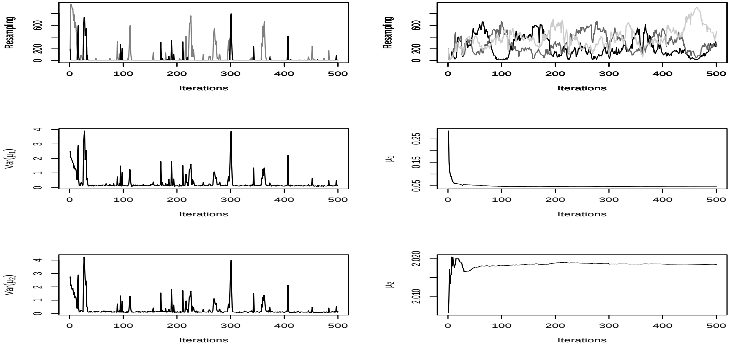
Resample the $\left((\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right)_j$ using the weights ϱ_j,

Details

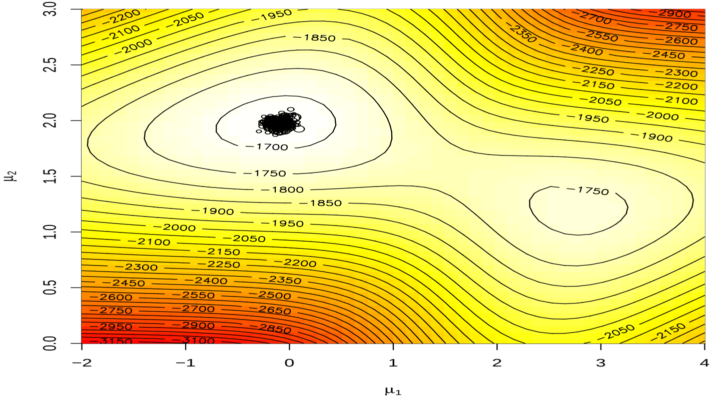
After an arbitrary initialisation, use of the previous (importance) sample (after resampling) to build random walk proposals,

$$\mathcal{N}((\mu)_j^{(i-1)}, v_j)$$

with a multiscale variance v_j within a predetermined set of p scales ranging from 10^3 down to 10^{-3}, whose importance is proportional to its survival rate in the resampling step.



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



Log-posterior distribution and sample of means