Markov Chain Monte Carlo Methods

Christian P. Robert

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

3-6 Mayo 2005

Outline

Motivation and leading example

Random variable generation

Monte Carlo Integration

Notions on Markov Chains

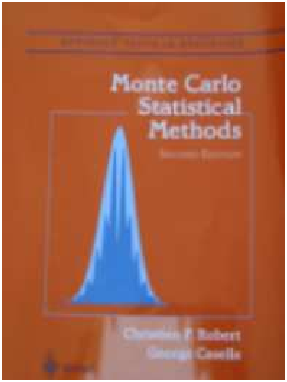
The Metropolis-Hastings Algorithm

The Gibbs Sampler

MCMC tools for variable dimension problems

Sequential importance sampling

New [2004] edition:



Motivation and leading example

Motivation and leading example

- Introduction
- Likelihood methods
- Missing variable models
- Bayesian Methods
- Bayesian troubles

Random variable generation

Monte Carlo Integration

Notions on Markov Chains

The Metropolis-Hastings Algorithm

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!

Example (Mixture models)

Models of *mixtures of distributions*:

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

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

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

Example (Mixture models)

Models of *mixtures of distributions*:

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

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

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

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

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

Example (Mixture models)

Models of *mixtures of distributions*:

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

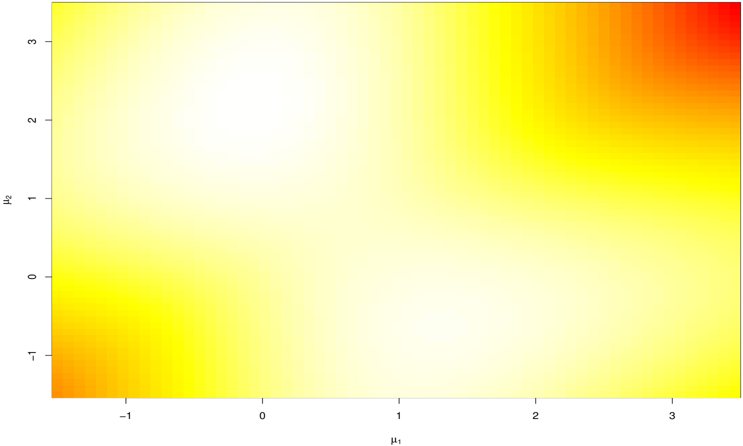
for $j = 1, 2, \ldots, k$, with overall density

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

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

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

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



Case of the $0.3\mathcal{N}(\mu_1, 1) + 0.7\mathcal{N}(\mu_2, 1)$ likelihood

Maximum likelihood methods

▸ Go Bayes!!

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

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

Maximum likelihood methods

▸ Go Bayes!!

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

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

- Global justifications from asymptotics

Maximum likelihood methods

▸ Go Bayes!!

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

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

- Global justifications from asymptotics
- Computational difficulty depends on structure, **eg latent variables**

Example (Mixtures again)

For a mixture of two normal distributions,

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

likelihood proportional to

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

containing 2^n terms.

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

Example

In the special case

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

with $\epsilon > 0$ known,

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

Example

In the special case

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

with $\epsilon > 0$ known, whatever n, the likelihood is unbounded:

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

The special case of missing variable models

Consider again a latent variable representation

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

The special case of missing variable models

Consider again a latent variable representation

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

Define the completed **(but unobserved)** likelihood

$$L^c(\theta|\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\theta)$$

The special case of missing variable models

Consider again a latent variable representation

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

Define the completed **(but unobserved)** likelihood

$$L^c(\theta|\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\theta)$$

Useful for optimisation algorithm

The EM Algorithm

▸ Gibbs connection

▸ Bayes rather than EM

Algorithm (Expectation–Maximisation)

Iterate (in m)

1. (*E step*) Compute

$$Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}) = \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\hat{\theta}_{(m)}, \mathbf{x}]\,,$$

The EM Algorithm

▸ Gibbs connection

▸ Bayes rather than EM

Algorithm (Expectation–Maximisation)

Iterate (in m)

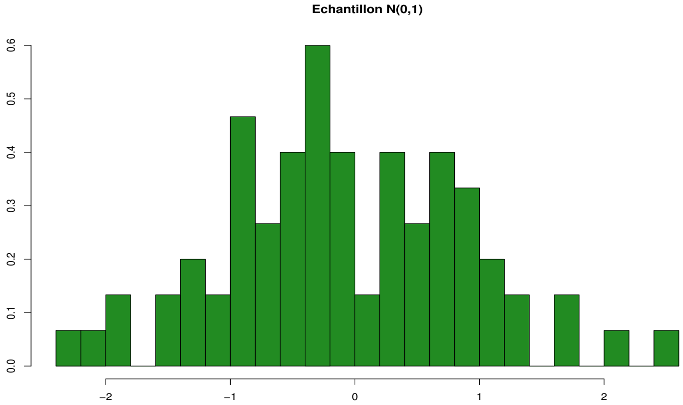
1. (*E step*) Compute

$$Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}) = \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\hat{\theta}_{(m)}, \mathbf{x}] \,,$$

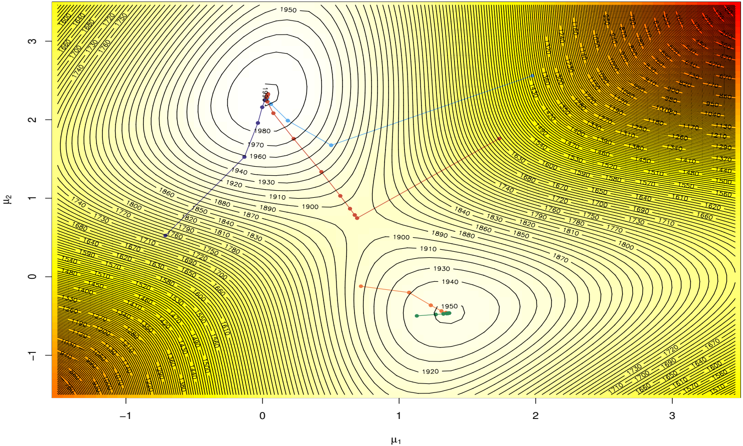
2. (*M step*) Maximise $Q(\theta|\hat{\theta}_{(m)}, \mathbf{x})$ in θ and take

$$\hat{\theta}_{(m+1)} = \arg\max_{\theta} \; Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}).$$

until a fixed point [of Q] is reached



Sample from (1)



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

The Bayesian Perspective

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

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

The Bayesian Perspective

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

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

is combined with **prior information** specified by *prior distribution* with density

$$\pi(\theta)$$

Central tool

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

Central tool

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

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

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

[Bayes Theorem]

Central tool

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

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

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

[Bayes Theorem]

where

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

is the *marginal density* of X

Central tool...central to Bayesian inference

Posterior defined up to a constant as

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations

Central tool...central to Bayesian inference

Posterior defined up to a constant as

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information and information brought by x

Central tool...central to Bayesian inference

Posterior defined up to a constant as

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information and information brought by x
- Avoids averaging over the unobserved values of x

Central tool...central to Bayesian inference

Posterior defined up to a constant as

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information and information brought by x
- Avoids averaging over the unobserved values of x
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected

Central tool...central to Bayesian inference

Posterior defined up to a constant as

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information and information brought by x
- Avoids averaging over the unobserved values of x
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected
- Provides a complete inferential scope and a unique motor of inference

Conjugate bonanza...

Example (Binomial)

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

Conjugate bonanza...

Example (Binomial)

For an observation $X \sim \mathscr{B}(n,p)$ so-called **conjugate prior** is the family of beta $\mathscr{B}e(a,b)$ distributions
The classical Bayes estimator δ^π is the posterior mean

$$\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(n-x+b)} \int_0^1 p\, p^{x+a-1}(1-p)^{n-x+b-1} dp$$
$$= \frac{x+a}{a+b+n}.$$

Example (Normal)

In the normal $\mathcal{N}(\mu, \sigma^2)$ case, with both μ and σ unknown, conjugate prior on $\theta = (\mu, \sigma^2)$ of the form

$$(\sigma^2)^{-\lambda_\sigma} \, \exp - \left\{ \lambda_\mu (\mu - \xi)^2 + \alpha \right\} / \sigma^2$$

Example (Normal)

In the normal $\mathcal{N}(\mu,\sigma^2)$ case, with both μ and σ unknown, conjugate prior on $\theta=(\mu,\sigma^2)$ of the form

$$(\sigma^2)^{-\lambda_\sigma}\ \exp -\left\{\lambda_\mu(\mu-\xi)^2+\alpha\right\}/\sigma^2$$

since

$$\begin{aligned}\pi((\mu,\sigma^2)|x_1,\ldots,x_n) &\propto (\sigma^2)^{-\lambda_\sigma}\ \exp -\left\{\lambda_\mu(\mu-\xi)^2+\alpha\right\}/\sigma^2\\ &\quad\times(\sigma^2)^{-n}\ \exp -\left\{n(\mu-\overline{x})^2+s_x^2\right\}/\sigma^2\\ &\propto (\sigma^2)^{-\lambda_\sigma+n}\ \exp -\Bigg\{(\lambda_\mu+n)(\mu-\xi_x)^2\\ &\quad+\alpha+s_x^2+\frac{n\lambda_\mu}{n+\lambda_\mu}\Bigg\}/\sigma^2\end{aligned}$$

...and conjugate curse

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

- implies a restriction on the modeling of the available prior information

...and conjugate curse

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

- implies a restriction on the modeling of the available prior information
- may be detrimental to the usefulness of the Bayesian approach

...and conjugate curse

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

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

A typology of Bayes computational problems

(i). use of a complex parameter space, as for instance in constrained parameter sets like those resulting from imposing stationarity constraints in dynamic models;

A typology of Bayes computational problems

(i). use of a complex parameter space, as for instance in constrained parameter sets like those resulting from imposing stationarity constraints in dynamic models;

(ii). use of a complex sampling model with an intractable likelihood, as for instance in missing data and graphical models;

A typology of Bayes computational problems

(i). use of a complex parameter space, as for instance in constrained parameter sets like those resulting from imposing stationarity constraints in dynamic models;

(ii). use of a complex sampling model with an intractable likelihood, as for instance in missing data and graphical models;

(iii). use of a huge dataset;

A typology of Bayes computational problems

(i). use of a complex parameter space, as for instance in constrained parameter sets like those resulting from imposing stationarity constraints in dynamic models;

(ii). use of a complex sampling model with an intractable likelihood, as for instance in missing data and graphical models;

(iii). use of a huge dataset;

(iv). use of a complex prior distribution (which may be the posterior distribution associated with an earlier sample);

A typology of Bayes computational problems

(i). use of a complex parameter space, as for instance in constrained parameter sets like those resulting from imposing stationarity constraints in dynamic models;

(ii). use of a complex sampling model with an intractable likelihood, as for instance in missing data and graphical models;

(iii). use of a huge dataset;

(iv). use of a complex prior distribution (which may be the posterior distribution associated with an earlier sample);

(v). use of a complex inferential procedure as for instance, **Bayes factors**

$$B^{\pi}_{01}(x) = \frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)} \bigg/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}.$$

Example (Mixture once again)

Observations from

$$x_1, \ldots, x_n \sim f(x|\theta) = p\varphi(x; \mu_1, \sigma_1) + (1-p)\varphi(x; \mu_2, \sigma_2)$$

Example (Mixture once again)

Observations from

$$x_1, \ldots, x_n \sim f(x|\theta) = p\varphi(x; \mu_1, \sigma_1) + (1-p)\varphi(x; \mu_2, \sigma_2)$$

Prior

$$\mu_i|\sigma_i \sim \mathscr{N}(\xi_i, \sigma_i^2/n_i), \qquad \sigma_i^2 \sim \mathscr{IG}(\nu_i/2, s_i^2/2), \qquad p \sim \mathscr{B}e(\alpha, \beta)$$

Example (Mixture once again)

Observations from

$$x_1, \ldots, x_n \sim f(x|\theta) = p\varphi(x; \mu_1, \sigma_1) + (1-p)\varphi(x; \mu_2, \sigma_2)$$

Prior

$$\mu_i|\sigma_i \sim \mathscr{N}(\xi_i, \sigma_i^2/n_i), \qquad \sigma_i^2 \sim \mathscr{IG}(\nu_i/2, s_i^2/2), \qquad p \sim \mathscr{B}e(\alpha, \beta)$$

Posterior

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

$[O(2^n)]$

Example (Mixture once again (cont'd))

For a given permutation (k_t), conditional posterior distribution

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

Example (Mixture once again (cont'd))

where

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

and

$$\begin{aligned}
\xi_1(k_t) &= \frac{n_1\xi_1 + \ell\bar{x}_1(k_t)}{n_1 + \ell}, \qquad \xi_2(k_t) = \frac{n_2\xi_2 + (n-\ell)\bar{x}_2(k_t)}{n_2 + n - \ell}, \\
s_1(k_t) &= s_1^2 + \hat{s}_1^2(k_t) + \frac{n_1\ell}{n_1 + \ell}(\xi_1 - \bar{x}_1(k_t))^2, \\
s_2(k_t) &= s_2^2 + \hat{s}_2^2(k_t) + \frac{n_2(n-\ell)}{n_2 + n - \ell}(\xi_2 - \bar{x}_2(k_t))^2,
\end{aligned}$$

posterior updates of the hyperparameters

Example (Mixture once again)

Bayes estimator of θ:

$$\delta^\pi(x_1,\ldots,x_n)=\sum_{\ell=0}^{n}\sum_{(k_t)}\omega(k_t)\mathbb{E}^\pi[\theta|\mathbf{x},(k_t)]$$

Too costly: 2^n terms

▸ press for AR

Example (Poly-t priors)

Normal observation $x \sim \mathcal{N}(\theta, 1)$, with conjugate prior

$$\theta \sim \mathcal{N}(\mu, \epsilon)$$

Closed form expression for the posterior mean

$$\int_\Theta \theta\, f(x|\theta)\, \pi(\theta)\, d\theta \quad \Big/ \quad \int_\Theta f(x|\theta)\, \pi(\theta)\, d\theta =$$

$$= \quad \frac{x + \epsilon^{-2}\mu}{1 + \epsilon^{-2}}\,.$$

Example (Poly-t priors (2))

More involved prior distribution:
poly-t distribution

[Bauwens,1985]

$$\pi(\theta) = \prod_{i=1}^{k} \left[\alpha_i + (\theta - \beta_i)^2\right]^{-\nu_i} \qquad \alpha_i, \nu_i > 0$$

Example (Poly-t priors (2))

More involved prior distribution:
poly-t distribution

[Bauwens,1985]

$$\pi(\theta) = \prod_{i=1}^{k} \left[\alpha_i + (\theta - \beta_i)^2\right]^{-\nu_i} \qquad \alpha_i, \nu_i > 0$$

Computation of $\mathbb{E}[\theta|x]$???

Example ($AR(p)$ model)

Auto-regressive representation of a time series,

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + \sigma \varepsilon_t$$

Example ($AR(p)$ model)

Auto-regressive representation of a time series,

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + \sigma \varepsilon_t$$

If order p unknown, predictive distribution of x_{t+1} given by

$$\pi(x_{t+1}|x_t, \ldots, x_1) \propto \int f(x_{t+1}|x_t, \ldots, x_{t-p+1}) \pi(\theta, p|x_t, \ldots, x_1) dp\, d\theta\,,$$

Example ($AR(p)$ model (cont'd))

Integration over the parameters of all models

$$\sum_{p=0}^{\infty} \int f(x_{t+1}|x_t, \ldots, x_{t-p+1}) \pi(\theta|p, x_t, \ldots, x_1) \, d\theta \, \pi(p|x_t, \ldots, x_1) \, .$$

Example ($AR(p)$ model (cont'd))

Multiple layers of complexity

(i). Complex parameter space within each $AR(p)$ model because of stationarity constraint

(ii). if p unbounded, infinity of models

(iii). θ varies between models $AR(p)$ and $AR(p+1)$, with a different stationarity constraint (except for root reparameterisation).

(iv). if prediction used sequentially, every tick/second/hour/day, posterior distribution $\pi(\theta, p|x_t, \ldots, x_1)$ must be re-evaluated

Random variable generation

Motivation and leading example

Random variable generation

- Basic methods
- Uniform pseudo-random generator
- Beyond Uniform distributions
- Transformation methods
- Accept-Reject Methods
- Fundamental theorem of simulation
- Log-concave densities

Monte Carlo Integration

Notions on Markov Chains

Random variable generation

- Rely on the possibility of producing (computer-wise) an endless flow of random variables (usually iid) from well-known distributions

Random variable generation

- Rely on the possibility of producing (computer-wise) an endless flow of random variables (usually iid) from well-known distributions
- Given a uniform random number generator, illustration of methods that produce random variables from both standard and nonstandard distributions

The inverse transform method

For a function F on $\mathbb{R}$, the *generalized inverse* of F, F^{-}, is defined by

$$F^{-}(u) = \inf \left\{x;\ F(x) \geq u\right\} .$$

The inverse transform method

For a function F on $\mathbb{R}$, the *generalized inverse* of F, F^{-}, is defined by

$$F^{-}(u) = \inf \left\{x;\; F(x) \geq u\right\} .$$

Definition (Probability Integral Transform)

If $U \sim \mathcal{U}_{[0,1]}$, then the random variable $F^{-}(U)$ has the distribution F.

The inverse transform method (2)

To generate a random variable $X \sim F$, simply generate

$$U \sim \mathscr{U}_{[0,1]}$$

The inverse transform method (2)

To generate a random variable $X \sim F$, simply generate

$$U \sim \mathscr{U}_{[0,1]}$$

and then make the transform

$$x = F^{-}(u)$$

Desiderata and limitations

▸ skip Uniform

- Production of a *deterministic* sequence of values in $[0, 1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$.

Desiderata and limitations

▸ skip Uniform

- Production of a *deterministic* sequence of values in $[0, 1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$.
- Can't use the physical imitation of a "random draw" `[no guarantee of uniformity, no reproducibility]`

Desiderata and limitations

▸ skip Uniform

- Production of a *deterministic* sequence of values in $[0,1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$.
- Can't use the physical imitation of a "random draw" `[no guarantee of uniformity, no reproducibility]`
- *Random* sequence in the sense: Having generated $(X_1, \cdots, X_n)$, knowledge of X_n [or of $(X_1, \cdots, X_n)$] imparts no discernible knowledge of the value of X_{n+1}.

Desiderata and limitations

▸ skip Uniform

- Production of a *deterministic* sequence of values in $[0,1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$.
- Can't use the physical imitation of a "random draw" `[no guarantee of uniformity, no reproducibility]`
- *Random* sequence in the sense: Having generated $(X_1, \cdots, X_n)$, knowledge of X_n [or of $(X_1, \cdots, X_n)$] imparts no discernible knowledge of the value of X_{n+1}.
- Deterministic: Given the initial value X_0, sample $(X_1, \cdots, X_n)$ always the same

Desiderata and limitations

▸ skip Uniform

- Production of a *deterministic* sequence of values in $[0, 1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$.
- Can't use the physical imitation of a "random draw" `[no guarantee of uniformity, no reproducibility]`
- *Random* sequence in the sense: Having generated $(X_1, \cdots, X_n)$, knowledge of X_n [or of $(X_1, \cdots, X_n)$] imparts no discernible knowledge of the value of X_{n+1}.
- Deterministic: Given the initial value X_0, sample $(X_1, \cdots, X_n)$ always the same
- Validity of a random number generator based on a single sample $X_1, \cdots, X_n$ when n tends to $+\infty$, **not** on replications

$$(X_{11}, \cdots, X_{1n}), (X_{21}, \cdots, X_{2n}), \ldots (X_{k1}, \cdots, X_{kn})$$

where n fixed and k tends to infinity.

Uniform pseudo-random generator

Algorithm starting from an initial value $0 \leq u_0 \leq 1$ and a transformation D, which produces a sequence

$$(u_i) = (D^i(u_0))$$

in $[0, 1]$.

Uniform pseudo-random generator

Algorithm starting from an initial value $0 \leq u_0 \leq 1$ and a transformation D, which produces a sequence

$$(u_i) = (D^i(u_0))$$

in $[0, 1]$.
For all n,

$$(u_1, \cdots, u_n)$$

reproduces the behavior of an **iid $\mathscr{U}_{[0,1]}$ sample** $(V_1, \cdots, V_n)$ when compared through usual tests

Uniform pseudo-random generator (2)

- Validity means the sequence $U_1, \cdots, U_n$ leads to accept the hypothesis

$$\text{H} : U_1, \cdots, U_n \quad \text{are iid} \quad \mathscr{U}_{[0,1]}.$$

Uniform pseudo-random generator (2)

- Validity means the sequence $U_1, \cdots, U_n$ leads to accept the hypothesis

$$\text{H} : U_1, \cdots, U_n \quad \text{are iid} \quad \mathscr{U}_{[0,1]}.$$

- The set of tests used is generally of some consequence
    - Kolmogorov–Smirnov and other nonparametric tests
    - Time series methods, for correlation between U_i and $(U_{i-1}, \cdots, U_{i-k})$
    - Marsaglia's battery of tests called *Die Hard* (!)

Usual generators

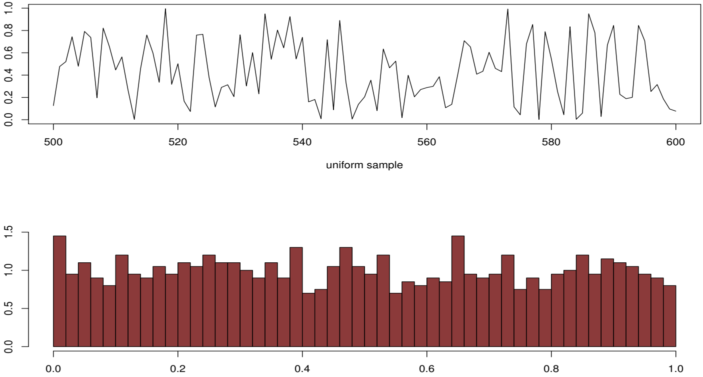
In R and S-plus, procedure `runif()`

```
The Uniform Distribution

Description:
‘runif’ generates random deviates.

Example:
     u <- runif(20)

‘.Random.seed’ is an integer vector, containing
the random  number  generator  state for random
number  generation  in  R.  It can be saved and
restored, but should not be altered by users.
```


1.0
0.8
0.6
0.4
0.2
0.0
500
520
540
560
580
600
uniform sample
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Usual generators (2)

In C, procedure `rand()` or `random()`

```
SYNOPSIS
       #include <stdlib.h>
       long int random(void);
DESCRIPTION
The random() function uses a non-linear additive
feedback  random  number  generator  employing a
default table of size 31 long integers to return
successive  pseudo-random  numbers  in the range
from  0  to  RAND_MAX. The period of this random
generator is very large, approximately
16*((2**31)-1).
RETURN VALUE
random() returns a value between 0 and RAND_MAX.
```

Usual generators(3)

In Scilab, procedure `rand()`

```
rand()  :  with no arguments gives a scalar whose
value  changes  each  time  it  is referenced. By
default, random numbers are uniformly distributed
in the interval (0,1). rand('normal') switches to
a normal distribution with mean 0 and variance 1.

EXAMPLE
  x=rand(10,10,'uniform')
```

Beyond Uniform generators

- Generation of any sequence of random variables can be formally implemented through a uniform generator
  - Distributions with explicit F^- (for instance, exponential, and Weibull distributions), use the probability integral transform ◂ here

Beyond Uniform generators

- Generation of any sequence of random variables can be formally implemented through a uniform generator
  - Distributions with explicit F^{-} (for instance, exponential, and Weibull distributions), use the probability integral transform ◂ here
  - Case specific methods rely on properties of the distribution (for instance, normal distribution, Poisson distribution)

Beyond Uniform generators

- Generation of any sequence of random variables can be formally implemented through a uniform generator
  - Distributions with explicit F^{-} (for instance, exponential, and Weibull distributions), use the probability integral transform ◂ here
  - Case specific methods rely on properties of the distribution (for instance, normal distribution, Poisson distribution)
  - More generic methods (for instance, accept-reject and ratio-of-uniform)

Beyond Uniform generators

- Generation of any sequence of random variables can be formally implemented through a uniform generator
    - Distributions with explicit F^- (for instance, exponential, and Weibull distributions), use the probability integral transform ◂ here
    - Case specific methods rely on properties of the distribution (for instance, normal distribution, Poisson distribution)
    - More generic methods (for instance, accept-reject and ratio-of-uniform)
- Simulation of the standard distributions is accomplished quite efficiently by many numerical and statistical programming packages.

Transformation methods

Case where a distribution F is linked in a simple way to another distribution easy to simulate.

Example (Exponential variables)

If $U \sim \mathcal{U}_{[0,1]}$, the random variable

$$X = -\log U/\lambda$$

has distribution

$$\begin{aligned} P(X \leq x) &= P(-\log U \leq \lambda x) \\ &= P(U \geq e^{-\lambda x}) = 1 - e^{-\lambda x}, \end{aligned}$$

the exponential distribution $\mathscr{E}xp(\lambda)$.

Other random variables that can be generated starting from an exponential include

$$Y = -2 \sum_{j=1}^{\nu} \log(U_j) \sim \chi^2_{2\nu}$$

$$Y = -\frac{1}{\beta} \sum_{j=1}^{a} \log(U_j) \sim \mathscr{G}a(a, \beta)$$

$$Y = \frac{\sum_{j=1}^{a} \log(U_j)}{\sum_{j=1}^{a+b} \log(U_j)} \sim \mathscr{B}e(a, b)$$

Points to note

- Transformation quite simple to use
- There are more efficient algorithms for gamma and beta random variables
- Cannot generate gamma random variables with a non-integer shape parameter
- For instance, cannot get a χ^2_1 variable, which would get us a $\mathcal{N}(0,1)$ variable.

Box-Muller Algorithm

Example (Normal variables)

If r, θ polar coordinates of (X_1, X_2), then,

$$r^2 = X_1^2 + X_2^2 \sim \chi_2^2 = \mathscr{E}(1/2) \quad \text{and} \quad \theta \sim \mathscr{U}[0, 2\pi]$$

Box-Muller Algorithm

Example (Normal variables)

If r, θ polar coordinates of (X_1, X_2), then,

$$r^2 = X_1^2 + X_2^2 \sim \chi_2^2 = \mathscr{E}(1/2) \quad \text{and} \quad \theta \sim \mathscr{U}[0, 2\pi]$$

Consequence: If U_1, U_2 iid $\mathcal{U}_{[0,1]}$,

$$\begin{aligned} X_1 &= \sqrt{-2\log(U_1)}\ \cos(2\pi U_2) \\ X_2 &= \sqrt{-2\log(U_1)}\ \sin(2\pi U_2) \end{aligned}$$

iid $\mathcal{N}(0, 1)$.

Box-Muller Algorithm (2)

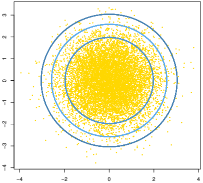
1. Generate U_1, U_2 iid $\mathcal{U}_{[0,1]}$;
2. Define

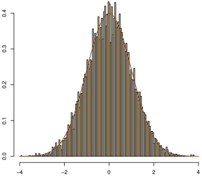
$$\begin{aligned} x_1 &= \sqrt{-2\log(u_1)}\cos(2\pi u_2)\ , \\ x_2 &= \sqrt{-2\log(u_1)}\sin(2\pi u_2)\ ; \end{aligned}$$

3. Take x_1 and x_2 as two independent draws from $\mathcal{N}(0,1)$.

Box-Muller Algorithm (3)

- Unlike algorithms based on the CLT, this algorithm is exact
- Get two normals for the price of two uniforms





- Drawback (in speed) in calculating log, cos and sin.

More transforms

▸ Reject

Example (Poisson generation)

Poisson–exponential connection:
If $N \sim \mathcal{P}(\lambda)$ and $X_i \sim \mathscr{E}xp(\lambda)$, $i \in \mathbb{N}^*$,

$$P_\lambda(N = k) = \\ P_\lambda(X_1 + \cdots + X_k \leq 1 < X_1 + \cdots + X_{k+1}) \, .$$

More Poisson

▸ Skip Poisson

- A Poisson can be simulated by generating $\mathrm{E}xp(1)$ till their sum exceeds 1.
- This method is simple, but is really practical only for smaller values of λ.
- On average, the number of exponential variables required is λ.
- Other approaches are more suitable for large λ's.

Atkinson's Poisson

To generate $N \sim \mathcal{P}(\lambda)$:

1. Define

   $$\beta = \pi/\sqrt{3\lambda}, \quad \alpha = \lambda\beta \quad \text{and} \quad k = \log c - \lambda - \log\beta;$$

2. Generate $U_1 \sim \mathscr{U}_{[0,1]}$ and calculate

   $$x = \{\alpha - \log\{(1 - u_1)/u_1\}\}/\beta$$

   until $x > -0.5$;
3. Define $N = \lfloor x + 0.5 \rfloor$ and generate $U_2 \sim \mathscr{U}_{[0,1]}$;
4. Accept N if

   $$\alpha - \beta x + \log\left(u_2/\{1 + \exp(\alpha - \beta x)\}^2\right) \leq k + N\log\lambda - \log N! \,.$$

Negative extension

- A generator of Poisson random variables can produce negative binomial random variables since,

$$Y \sim \mathcal{G}a(n,(1-p)/p) \quad X|y \sim \mathcal{P}(y)$$

implies

$$X \sim \mathcal{N}eg(n,p)$$

Mixture representation

- The representation of the negative binomial is a particular case of a *mixture distribution*
- The principle of a mixture representation is to represent a density f as the marginal of another distribution, for example

$$f(x) = \sum_{i \in \mathscr{Y}} p_i \; f_i(x) \; ,$$

- If the component distributions $f_i(x)$ can be easily generated, X can be obtained by first choosing f_i with probability p_i and then generating an observation from f_i.

Partitioned sampling

Special case of mixture sampling when

$$f_i(x) = f(x)\,\mathbb{I}_{A_i}(x) \bigg/ \int_{A_i} f(x)\,dx$$

and

$$p_i = \Pr(X \in A_i)$$

for a partition $(A_i)_i$

Accept-Reject algorithm

- Many distributions from which it is difficult, or even impossible, to **directly** simulate.
- Another class of methods that only require us to know the functional form of the density f of interest **only** up to a multiplicative constant.
- The key to this method is to use a simpler (simulation-wise) density g, the *instrumental density*, from which the simulation from the *target density* f is actually done.

Fundamental theorem of simulation

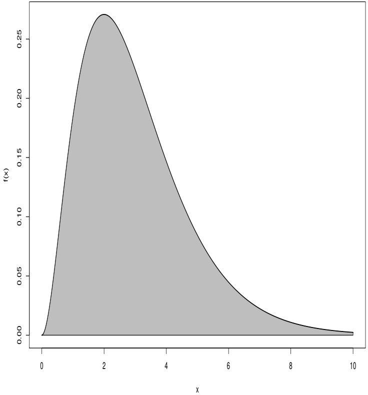
Lemma

Simulating

$$X \sim f(x)$$

equivalent to simulating

$$(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f(x)\}$$



The Accept-Reject algorithm

Given a density of interest f, find a density g and a constant M such that

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

on the support of f.

The Accept-Reject algorithm

Given a density of interest f, find a density g and a constant M such that

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

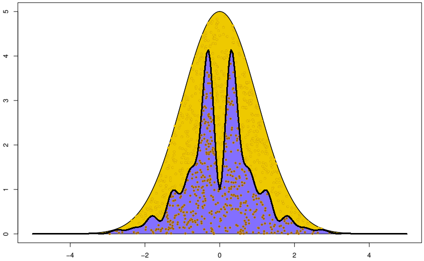
on the support of f.

```
1. Generate $X \sim g$, $U \sim \mathcal{U}_{[0,1]}$ ;
2. Accept $Y = X$ if $U \leq f(X)/Mg(X)$ ;
3. Return to 1. otherwise.
```

Validation of the Accept-Reject method

Warranty:

This algorithm produces a variable Y distributed according to f



Two interesting properties

- First, it provides a generic method to simulate from any density f that is known *up to a multiplicative factor*
Property particularly important in Bayesian calculations where the posterior distribution

$$\pi(\theta|x) \propto \pi(\theta)\, f(x|\theta)\,.$$

is specified up to a normalizing constant

Two interesting properties

- First, it provides a generic method to simulate from any density f that is known *up to a multiplicative factor* Property particularly important in Bayesian calculations where the posterior distribution

$$\pi(\theta|x) \propto \pi(\theta)\, f(x|\theta)\,.$$

is specified up to a normalizing constant

- Second, the probability of acceptance in the algorithm is $1/M$, e.g., expected number of trials until a variable is accepted is M

More interesting properties

- In cases f and g both probability densities, the constant M is necessarily larger that 1.

More interesting properties

- In cases f and g both probability densities, the constant M is necessarily larger that 1.
- The size of M, and thus the efficiency of the algorithm, are functions of how closely g can imitate f, especially in the tails

More interesting properties

- In cases f and g both probability densities, the constant M is necessarily larger that 1.
- The size of M, and thus the efficiency of the algorithm, are functions of how closely g can imitate f, especially in the tails
- For f/g to remain bounded, necessary for g to have tails thicker than those of f.
  It is therefore impossible to use the A-R algorithm to simulate a Cauchy distribution f using a normal distribution g, however the reverse works quite well.

▸ No Cauchy!

Example (Normal from a Cauchy)

Take

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

and

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

densities of the normal and Cauchy distributions.

▸ No Cauchy!

Example (Normal from a Cauchy)

Take

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

and

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

densities of the normal and Cauchy distributions.
Then

$$\frac{f(x)}{g(x)} = \sqrt{\frac{\pi}{2}}(1+x^2)\, e^{-x^2/2} \leq \sqrt{\frac{2\pi}{e}} = 1.52$$

attained at $x = \pm 1$.

Example (Normal from a Cauchy (2))

So probability of acceptance

$$1/1.52 = 0.66,$$

and, on the average, one out of every three simulated Cauchy variables is rejected.

▸ No Double!

Example (Normal/Double Exponential)

Generate a $\mathscr{N}(0,1)$ by using a double-exponential distribution with density

$$g(x|\alpha) = (\alpha/2)\,\text{exp}(-\alpha|x|)$$

Then

$$\frac{f(x)}{g(x|\alpha)} \leq \sqrt{\frac{2}{\pi}}\alpha^{-1}e^{-\alpha^2/2}$$

and minimum of this bound (in α) attained for

$$\alpha^\star = 1$$

Example (Normal/Double Exponential (2))

Probability of acceptance

$$\sqrt{\pi/2e} = .76$$

To produce one normal random variable requires on the average $1/.76 \approx 1.3$ uniform variables.

▸ truncate

Example (Gamma generation)

Illustrates a real advantage of the Accept-Reject algorithm
The gamma distribution $\mathcal{G}a(\alpha, \beta)$ represented as the sum of α exponential random variables, only if α is an integer

Example (Gamma generation (2))

Can use the Accept-Reject algorithm with instrumental distribution

$$\mathcal{G}a(a, b), \text{ with } a = [\alpha], \quad \alpha \geq 0.$$

(Without loss of generality, $\beta = 1$.)

Example (Gamma generation (2))

Can use the Accept-Reject algorithm with instrumental distribution

$$\mathcal{G}a(a,b),\ \text{with } a=[\alpha], \quad \alpha \geq 0.$$

(Without loss of generality, $\beta = 1$.)
Up to a normalizing constant,

$$f/g_b = b^{-a} x^{\alpha-a}\ \exp\{-(1-b)x\} \leq b^{-a}\left(\frac{\alpha-a}{(1-b)e}\right)^{\alpha-a}$$

for $b \leq 1$.
The maximum is attained at $b = a/\alpha$.

Cheng and Feast's Gamma generator

Gamma $\mathscr{G}a(\alpha, 1)$, $\alpha > 1$ distribution

```
1. Define $c_1 = \alpha - 1$, $c_2 = (\alpha - (1/6\alpha))/c_1$,
   $c_3 = 2/c_1$, $c_4 = 1 + c_3$, and $c_5 = 1/\sqrt{\alpha}$.
2. Repeat
          generate $U_1, U_2$
   take $U_1 = U_2 + c_5(1 - 1.86U_1)$ if $\alpha > 2.5$
   until $0 < U_1 < 1$.
3. Set $W = c_2 U_2 / U_1$.
4. If $c_3 U_1 + W + W^{-1} \leq c_4$ or
   $c_3 \log U_1 - \log W + W \leq 1$,
   take $c_1 W$;
   otherwise, repeat.
```

Truncated Normal simulation

Example (Truncated Normal distributions)

Constraint $x \geq \underline{\mu}$ produces density proportional to

$$e^{-(x-\mu)^2/2\sigma^2}\ \mathbb{I}_{x \geq \underline{\mu}}$$

for a bound $\underline{\mu}$ large compared with μ

Truncated Normal simulation

Example (Truncated Normal distributions)

Constraint $x \geq \underline{\mu}$ produces density proportional to

$$e^{-(x-\mu)^2/2\sigma^2} \, \mathbb{I}_{x \geq \underline{\mu}}$$

for a bound $\underline{\mu}$ large compared with μ
There exists alternatives far superior to the naïve method of generating a $\mathcal{N}(\mu, \sigma^2)$ until exceeding $\underline{\mu}$, which requires an average number of

$$1/\Phi((\mu - \underline{\mu})/\sigma)$$

simulations from $\mathcal{N}(\mu, \sigma^2)$ for a single acceptance.

Example (Truncated Normal distributions (2))

Instrumental distribution: translated exponential distribution, $\mathscr{E}(\alpha, \underline{\mu})$, with density

$$g_\alpha(z) = \alpha e^{-\alpha(z-\underline{\mu})} \, \mathbb{I}_{z \geq \underline{\mu}} \ .$$

Example (Truncated Normal distributions (2))

Instrumental distribution: translated exponential distribution, $\mathscr{E}(\alpha, \underline{\mu})$, with density

$$g_\alpha(z) = \alpha e^{-\alpha(z-\underline{\mu})} \, \mathbb{I}_{z \geq \underline{\mu}} \, .$$

The ratio f/g_α is bounded by

$$f/g_\alpha \leq \begin{cases} 1/\alpha \ \exp(\alpha^2/2 - \alpha\underline{\mu}) & \text{if } \alpha > \underline{\mu}\,, \\ 1/\alpha \ \exp(-\underline{\mu}^2/2) & \text{otherwise.} \end{cases}$$

Log-concave densities (1)

▸ move to next chapter Densities f whose logarithm is concave, for instance Bayesian posterior distributions such that

$$\log\ \pi(\theta|x) = \log\ \pi(\theta) + \log\ f(x|\theta) + c$$

concave

Log-concave densities (2)

Take

$$\mathfrak{S}_n = \{x_i, i = 0, 1, \ldots, n+1\} \subset \mathrm{supp}(f)$$

such that $h(x_i) = \log f(x_i)$ known up to the same constant.
By concavity of h, line $L_{i,i+1}$ through $(x_i, h(x_i))$ and $(x_{i+1}, h(x_{i+1}))$

Log-concave densities (2)

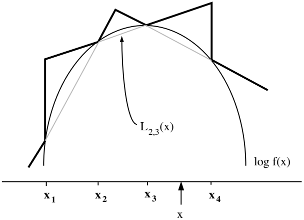
Take

$$\mathfrak{S}_n = \{x_i, i = 0, 1, \ldots, n+1\} \subset \mathsf{supp}(f)$$

such that $h(x_i) = \log f(x_i)$ known up to the same constant.

By concavity of h, line $L_{i,i+1}$ through $(x_i, h(x_i))$ and $(x_{i+1}, h(x_{i+1}))$

- below h in $[x_i, x_{i+1}]$ and
- above this graph outside this interval



Log-concave densities (3)

For $x \in [x_i, x_{i+1}]$, if

$$\overline{h}_n(x) = \min\{L_{i-1,i}(x), L_{i+1,i+2}(x)\} \quad \text{and} \quad \underline{h}_n(x) = L_{i,i+1}(x)\,,$$

the envelopes are

$$\underline{h}_n(x) \le h(x) \le \overline{h}_n(x)$$

uniformly on the support of f, with

$$\underline{h}_n(x) = -\infty \quad \text{and} \quad \overline{h}_n(x) = \min(L_{0,1}(x), L_{n,n+1}(x))$$

on $[x_0, x_{n+1}]^c$.

Log-concave densities (4)

Therefore, if

$$\underline{f}_n(x) = \exp \underline{h}_n(x) \text{ and } \overline{f}_n(x) = \exp \overline{h}_n(x)$$

then

$$\underline{f}_n(x) \leq f(x) \leq \overline{f}_n(x) = \varpi_n \, g_n(x) \; ,$$

where ϖ_n normalizing constant of f_n

ARS Algorithm

1. Initialize n and $\mathfrak{S}_n$.
2. Generate $X \sim g_n(x)$, $U \sim \mathcal{U}_{[0,1]}$.
3. If $U \leq \underline{f}_n(X)/\varpi_n \, g_n(X)$, accept X;
   otherwise, if $U \leq f(X)/\varpi_n \, g_n(X)$, accept X

▸ kill ducks

Example (Northern Pintail ducks)

Ducks captured at time i with both probability p_i and size N of the population unknown.

Dataset

$$(n_1, \ldots, n_{11}) = (32, 20, 8, 5, 1, 2, 0, 2, 1, 1, 0)$$

Number of recoveries over the years 1957–1968 of $N = 1612$ Northern Pintail ducks banded in 1956

Example (Northern Pintail ducks (2))

Corresponding conditional likelihood

$$L(p_1, \dots, p_I | N, n_1, \dots, n_I) = \frac{N!}{(N-r)!} \prod_{i=1}^{I} p_i^{n_i}(1-p_i)^{N-n_i},$$

where I number of captures, n_i number of captured animals during the ith capture, and r is the total number of different captured animals.

Example (Northern Pintail ducks (3))

Prior selection

If

$$N \sim \mathscr{P}(\lambda)$$

and

$$\alpha_i = \log\left(\frac{p_i}{1-p_i}\right) \sim \mathcal{N}(\mu_i, \sigma^2),$$

[Normal logistic]

Example (Northern Pintail ducks (4))

Posterior distribution

$$\begin{aligned}\pi(\alpha, N|, n_1, \ldots, n_I) \quad &\propto \quad \frac{N!}{(N-r)!}\,\frac{\lambda^N}{N!}\prod_{i=1}^{I}(1+e^{\alpha_i})^{-N} \\ &\prod_{i=1}^{I}\exp\left\{\alpha_i n_i - \frac{1}{2\sigma^2}(\alpha_i-\mu_i)^2\right\}\end{aligned}$$

Example (Northern Pintail ducks (5))

For the conditional posterior distribution

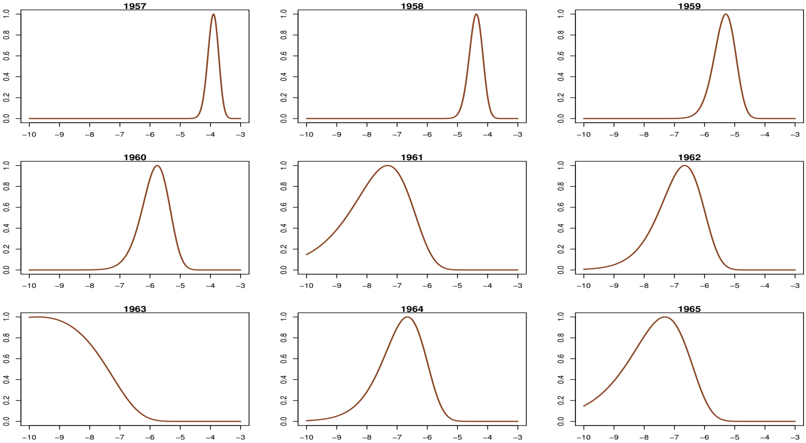
$$\pi(\alpha_i|N,n_1,\ldots,n_I)\;\propto\;\exp\left\{\alpha_i n_i-\frac{1}{2\sigma^2}(\alpha_i-\mu_i)^2\right\}\Big/(1{+}e^{\alpha_i})^N\;,$$

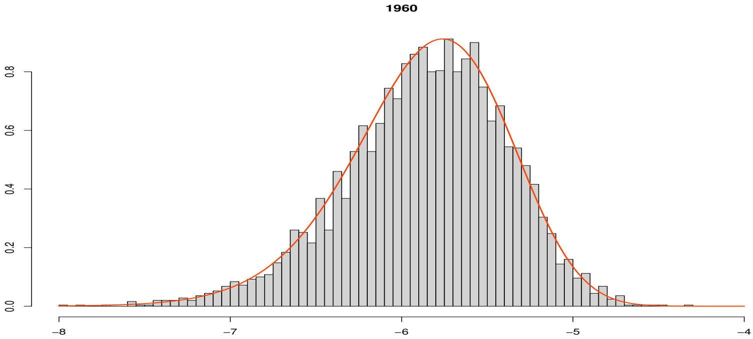
the ARS algorithm can be implemented since

$$\alpha_i n_i-\frac{1}{2\sigma^2}\,(\alpha_i-\mu_i)^2-N\;\log(1+e^{\alpha_i})$$

is concave in α_i.

Posterior distributions of capture log-odds ratios for the years 1957–1965.





True distribution versus histogram of simulated sample

Monte Carlo integration

Motivation and leading example

Random variable generation

Monte Carlo Integration

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

Notions on Markov Chains

The Metropolis-Hastings Algorithm

Quick reminder

Two major classes of numerical problems that arise in statistical inference

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

Quick reminder

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

▸ skip Example!

Example (Bayesian decision theory)

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

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

Proper loss:

For $\mathsf{L}(\theta, \delta) = (\theta - \delta)^2$, the Bayes estimator is the **posterior mean**

▸ skip Example!

Example (Bayesian decision theory)

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

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

Proper loss:
For $\mathsf{L}(\theta,\delta) = (\theta-\delta)^2$, the Bayes estimator is the **posterior mean**
Absolute error loss:
For $\mathsf{L}(\theta,\delta) = |\theta-\delta|$, the Bayes estimator is the **posterior median**

▸ skip Example!

Example (Bayesian decision theory)

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

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

Proper loss:

For $\mathrm{L}(\theta,\delta) = (\theta-\delta)^2$, the Bayes estimator is the **posterior mean**

Absolute error loss:

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

With no loss function

use the maximum a posteriori (MAP) estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

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,

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

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

Example (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

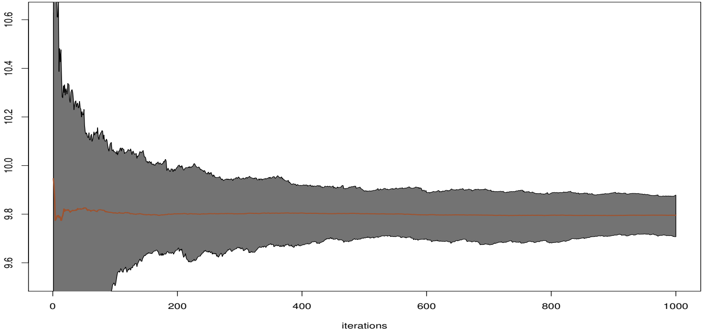
$$\theta_1, \cdots, \theta_m \sim \mathscr{N}(x, 1)$$

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

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

- Finite variance only when

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

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.

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

- Finite variance only when

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

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

Example (Cauchy target)

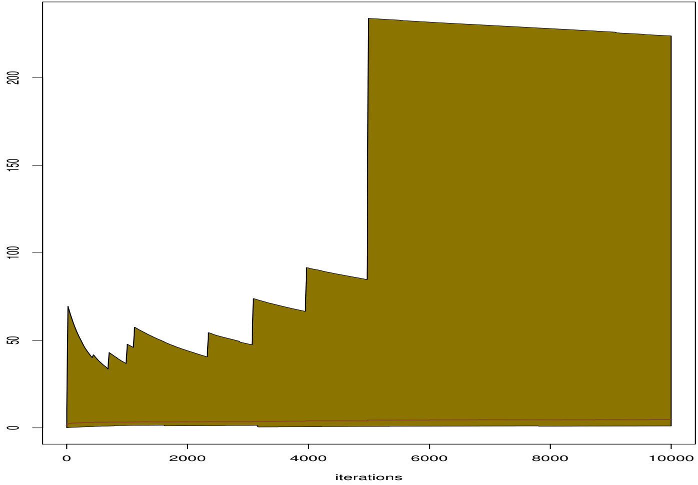
Case of Cauchy distribution $\mathrm{C}(0,1)$ when importance function is Gaussian $\mathscr{N}(0,1)$.
Ratio of the densities

$$\varrho(x) = \frac{p^\star(x)}{p_0(x)} = \sqrt{2\pi}\,\frac{\exp x^2/2}{\pi\,(1+x^2)}$$

very badly behaved: e.g.,

$$\int_{-\infty}^{\infty} \varrho(x)^2 p_0(x)dx = \infty\,.$$

Poor performances of the associated importance sampling estimator



Range and average of 500 replications of IS estimate of $\mathbb{E}[\exp -X]$ over $10,000$ iterations.

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

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

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

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

Example (Student's t distribution)

$X \sim \mathcal{T}(\nu, \theta, \sigma^2)$, with density

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

Without loss of generality, take $\theta = 0$, $\sigma = 1$.
Problem: Calculate the integral

$$\int_{2.1}^{\infty} \left(\frac{\sin(x)}{x}\right)^n f_\nu(x) dx.$$

Example (Student's t distribution (2))

- Simulation possibilities
  - Directly from f_ν, since $f_\nu = \frac{\mathscr{N}(0,1)}{\sqrt{\chi^2_\nu}}$

Example (Student's t distribution (2))

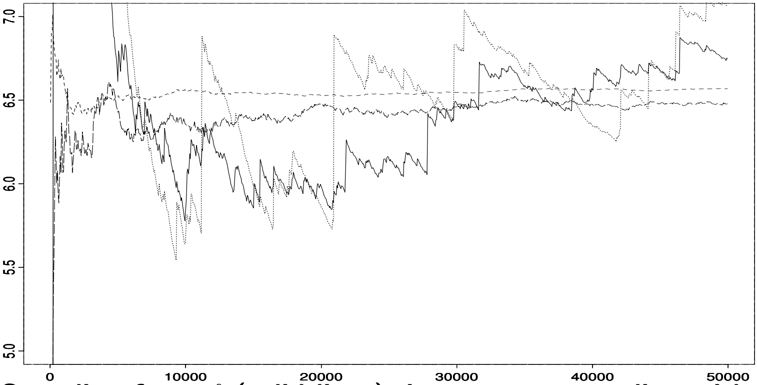
- Simulation possibilities
  - Directly from f_ν, since $f_\nu = \frac{\mathscr{N}(0{,}1)}{\sqrt{\chi^2_\nu}}$
  - Importance sampling using Cauchy $\mathscr{C}(0,1)$

Example (Student's t distribution (2))

- Simulation possibilities
  - Directly from f_ν, since $f_\nu = \frac{\mathscr{N}(0,1)}{\sqrt{\chi^2_\nu}}$
  - Importance sampling using Cauchy $\mathscr{C}(0,1)$
  - Importance sampling using a normal $\mathscr{N}(0,1)$ (expected to be nonoptimal)

Example (Student's t distribution (2))

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



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

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

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

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

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

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

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

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

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

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!

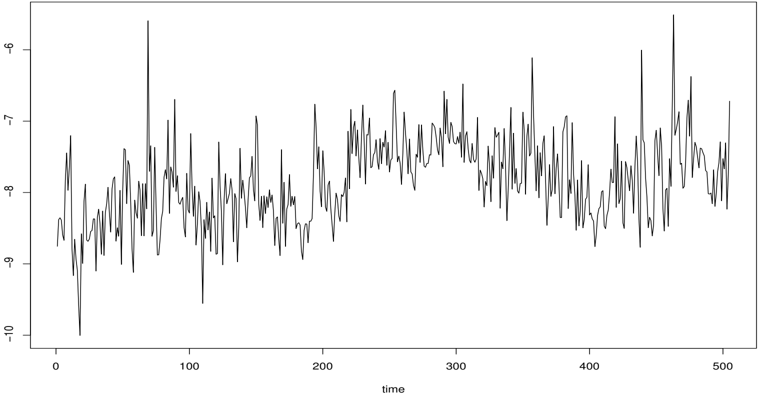
Example (Stochastic volatility model)

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

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

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

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 (2).

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

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

and variance

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

Prior proposal on X_1,

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

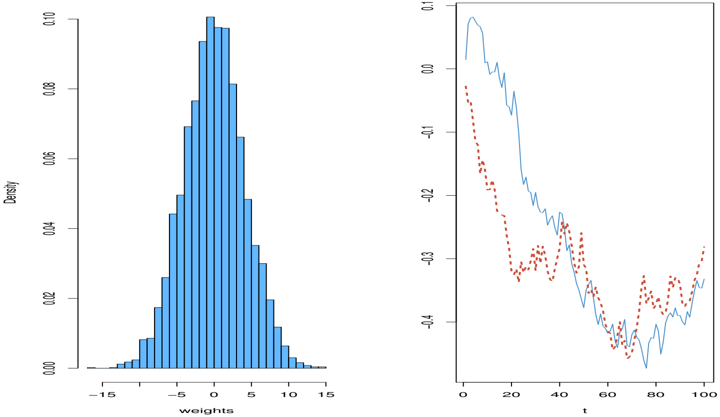
Example (Stochastic volatility model (6))

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}.

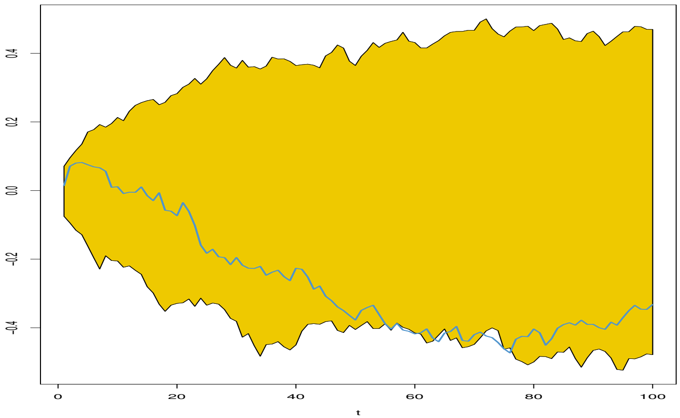
Importance weight computed sequentially as the product of

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

$(1 \leq k \leq K)$



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 $\{X_k\}_{1\leq k\leq 100}$, compared with the true value.

Correlated simulations

Negative correlation reduces variance

Special technique — but efficient when it applies

Two samples $(X_1, \dots, X_m)$ and $(Y_1, \dots, Y_m)$ 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

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

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

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

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

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

Optimal control

Optimal choice of β

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

with

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

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

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

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

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

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

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

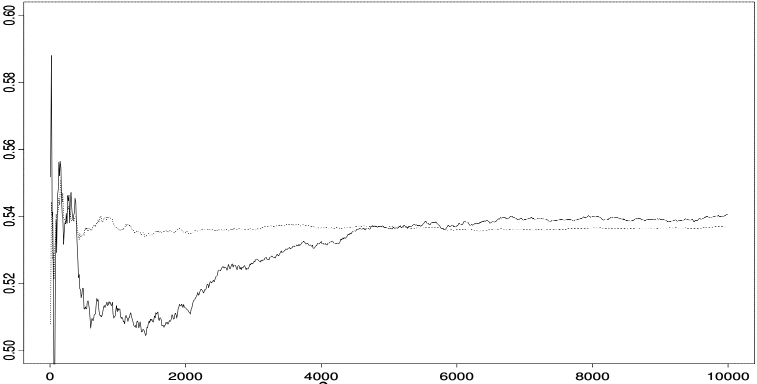
can be improved from the joint sample

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

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

▸ directly Markovian

Recall algorithm:

1. `Generate` $\theta_1^{(1)}, \cdots, \theta_1^{(T)}$ `from` $cg(\theta)$ `with`

$$c^{-1} = \int g(\theta) d\theta$$

2. `Take`

$$\begin{aligned} \int f(x|\theta)\pi(\theta)d\theta &\approx \frac{1}{T}\sum_{t=1}^{T} f(x|\theta^{(t)}) \frac{\pi(\theta^{(t)})}{cg(\theta^{(t)})} \\ &\approx \frac{\sum_{t=1}^{T} f(x|\theta^{(t)}) \frac{\pi(\theta^{(t)})}{g(\theta^{(t)})}}{\sum_{t=1}^{T} \frac{\pi(\theta^{(t)})}{g(\theta^{(t)})}} = m^{IS}(x) \end{aligned}$$

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:

$$\begin{array}{rcl}\pi_1(\theta_1|x) & \propto & \tilde\pi_1(\theta_1)\\ \pi_2(\theta_2|x) & \propto & \tilde\pi_2(\theta_2)\end{array}$$

$$B_{12} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde\pi_1(\theta_i)}{\tilde\pi_2(\theta_i)} \qquad \theta_i \sim \pi_2$$

Bridge sampling (3)

(ii) Still missing normalising constants:

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

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

Bridge sampling (4)

Optimal choice

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

[Chen, Meng & Wong, 2000]

Notions on Markov Chains

Notions on Markov Chains

- Basics
- Irreducibility
- Transience and Recurrence
- Invariant measures
- Ergodicity and convergence
- Limit theorems
- Quantitative convergence rates
- Coupling
- Renewal and CLT

Basics

Definition (Markov chain)

A sequence of random variables whose distribution evolves over **time** as a function of past realizations

Basics

Definition (Markov chain)

A sequence of random variables whose distribution evolves over **time** as a function of past realizations

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

- $\forall x \in \mathscr{X}$, $K(x, \cdot)$ is a probability measure;
- $\forall A \in \mathscr{B}(\mathscr{X})$, $K(\cdot, A)$ is measurable.

▸ no discrete

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

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

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

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

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

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

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

Then, for any bounded ϕ, we may define

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

Note that

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

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

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

Markov chains

skip definition

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

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

Note that the entire structure of the chain only depends on

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

Example (Random walk)

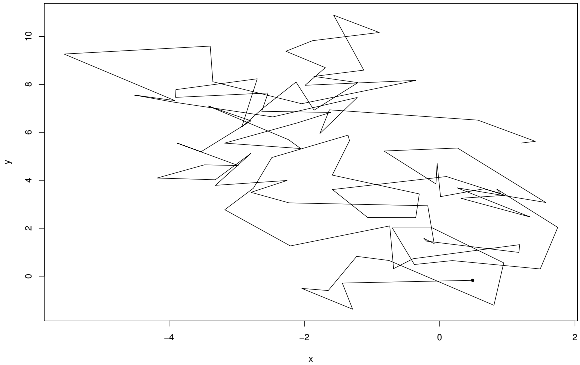
The normal random walk is the kernel $K(x,\cdot)$ associated with the distribution

$$\mathcal{N}_p(x,\tau^2 I_p)$$

which means

$$X_{t+1} = X_t + \tau\epsilon_t$$

ϵ_t being an iid additional noise



100 consecutive realisations of the random walk in $\mathbb{R}^2$ with $\tau = 1$

▸ bypass remarks

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

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

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

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

▸ bypass remarks

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

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

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

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

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

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

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

Composition of kernels

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

$$Q_1 Q_2(x, A) = \int_{\mathscr{X}} \mathfrak{Q}_1(x, dy) \mathfrak{Q}_2(y, A)$$

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

Irreducibility

Irreducibility is one measure of the sensitivity of the Markov chain to initial conditions
It leads to a guarantee of convergence for MCMC algorithms

Irreducibility

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

It leads to a guarantee of convergence for MCMC algorithms

Definition (Irreducibility)

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

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

τ_y being the first (positive) time y is visited

Irreducibility for a continuous chain

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

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

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

Minoration condition

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

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

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

Small sets

Definition (Small set)

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

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

C is called a **small set**

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

Towards further stability

- Irreducibility: every set A has a chance to be visited by the Markov chain (X_n)
- This property is too weak to ensure that the trajectory of (X_n) will enter A often enough.

Towards further stability

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

Transience and Recurrence

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

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

Transience and Recurrence

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

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

If $\mathbb{E}_\omega[\eta_\omega] = \infty$, the state is *recurrent*
If $\mathbb{E}_\omega[\eta_\omega] < \infty$, the state is *transient*
For irreducible chains, recurrence/transience is **property of the chain**, not of a particular state

Transience and Recurrence

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

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

If $\mathbb{E}_\omega[\eta_\omega] = \infty$, the state is *recurrent*
If $\mathbb{E}_\omega[\eta_\omega] < \infty$, the state is *transient*
For irreducible chains, recurrence/transience is **property of the chain**, not of a particular state
Similar definitions for the continuous case.

Harris recurrence

Stronger form of recurrence:

Definition (Harris recurrence)

A set A is *Harris recurrent* if

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

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

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

Harris recurrence

Stronger form of recurrence:

Definition (Harris recurrence)

A set A is *Harris recurrent* if

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

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

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

Note that

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

Invariant measures

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

Requires the existence of a probability distribution π such that

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

Invariant measures

Stability increases for the chain (X_n) if marginal distribution of X_n independent of n
Requires the existence of a probability distribution π such that

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

Definition (Invariant measure)

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

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

Stability properties and invariance

- The chain is **positive recurrent** if π is a probability measure.
- Otherwise it is **null recurrent** or **transient**

Stability properties and invariance

- The chain is **positive recurrent** if π is a probability measure.
- Otherwise it is **null recurrent** or **transient**

- If π probability measure, π also called *stationary distribution* since

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

- The stationary distribution is unique

Insights

▸ no time for that!

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

To understand why, consider $P_\mu(X_n \in \cdot)$ for a starting distribution μ. If a limiting measure γ_μ exists such as

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

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

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

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

Ergodicity and convergence

We finally consider: **to what is the chain converging?**
The invariant distribution π is a natural candidate for the *limiting distribution*

Ergodicity and convergence

We finally consider: **to what is the chain converging?**
The invariant distribution π is a natural candidate for the *limiting distribution*
A fundamental property is **ergodicity**, or independence of initial conditions. In the discrete case, a state ω is *ergodic* if

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

Norm and convergence

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

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

and we want

$$\left\| \int K^n(x, \cdot)\mu(dx) - \pi \right\|_{\mathsf{TV}}$$

$$= \quad \sup_A \left| \int K^n(x, A)\mu(dx) - \pi(A) \right|$$

to be small.

▸ skip minoration TV

Total variation distance and minoration

Lemma

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

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

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

Total variation distance and minoration (2)

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

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

Then, for all I and all partitions A_1, A_2, …, A_I,

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

and the previous result thus implies that

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

Harris recurrence and ergodicity

Theorem

If (X_n) Harris positive recurrent and aperiodic, then

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

for every initial distribution μ.

Harris recurrence and ergodicity

Theorem

If (X_n) Harris positive recurrent and aperiodic, then

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

for every initial distribution μ.

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

[Meyn & Tweedie, 1993]

Harris recurrence and ergodicity

Theorem

If (X_n) Harris positive recurrent and aperiodic, then

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

for every initial distribution μ.

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

[Meyn & Tweedie, 1993]

Convergence in total variation implies

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

for every *bounded function* h.

▸ no detail of convergence

Convergences

There are difference speeds of convergence

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

Geometric ergodicity

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

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

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

Geometric ergodicity implies a lot of important results

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

Geometric ergodicity implies a lot of important results

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

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

Geometric ergodicity implies a lot of important results

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

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

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

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

Minoration condition and uniform ergodicity

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

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

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

Uniform ergodicity

Theorem (S&n ergodicity)

The following conditions are equivalent:

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

$$\|P^n(x,\cdot) - \pi\|_{\mathsf{TV}} \leq R\rho^n\,,$$

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

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

[Meyn and Tweedie, 1993]

Limit theorems

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

Limit theorems

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

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

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

$$X_n \sim P_x^n$$

© **We need LLN's and CLT's!!!**

Limit theorems

Ergodicity determines the probabilistic properties of **average** behavior of the chain.
But also need of *statistical inference*, made by induction from the observed sample.
If $\|P_x^n - \pi\|$ close to 0, no direct information about

$$X_n \sim P_x^n$$

© **We need LLN's and CLT's!!!**
Classical LLN's and CLT's not directly applicable due to:

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

The Theorem

Theorem (Ergodic Theorem)

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

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

Central Limit Theorem

To get a CLT, we need more assumptions.

▸ skip conditions and results

Central Limit Theorem

To get a CLT, we need more assumptions.

▸ skip conditions and results

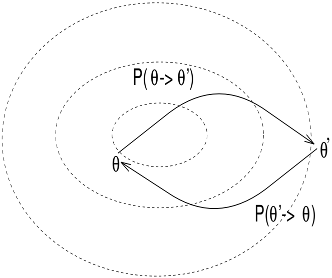
For MCMC, the easiest is

Definition (reversibility)

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

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

The direction of time does not matter



[Green, 1995]

The CLT

Theorem

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

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

where

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

[Kipnis & Varadhan, 1986]

Quantitative convergence rates

skip detailed results

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

Quantitative convergence rates

▸ skip detailed results

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

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

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

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

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

Practical purposes?

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

[Douc, Moulines and Rosenthal, 2001]

[Jones and Hobert, 2001]

Tools at hand

For MCMC algorithms, kernels are “explicitly” known.
Type of quantities (more or less directly) available:

- Minoration constants

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

- Foster-Lyapunov Drift conditions,

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

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

Coupling

▸ skip coupling

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

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

Coupling

skip coupling

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

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

The **basic coupling construction**

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

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

  respectively.

[Thorisson, 2000]

Coupling inequality

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

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

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

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

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

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

Small set and coupling sets

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

$$K(x, A) \geq \epsilon\nu(A), \qquad \forall x \in C. \tag{3}$$

Small sets always exist when the MC is φ-irreducible

[Jain and Jamieson, 1967]

Small set and coupling sets

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

$$K(x, A) \geq \epsilon \nu(A), \qquad \forall x \in C. \tag{3}$$

Small sets always exist when the MC is φ-irreducible [Jain and Jamieson, 1967]

For MCMC kernels, small sets in general easy to find.
If C is a small set, then $\bar{C} = C \times C$ is a coupling set:

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

Coupling for Markov chains

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

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

Coupling for Markov chains

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

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

For example,

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

$$\begin{aligned}\bar{R}(x, x'; A \times \mathscr{X}) &= (1-\epsilon)^{-1}(K(x, A) - \epsilon\nu_{x,x'}(A)) \\ \bar{R}(x, x'; \mathscr{X} \times A') &= (1-\epsilon)^{-1}(K(x', A) - \epsilon\nu_{x,x'}(A')).\end{aligned}$$

Coupling algorithm

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

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

Coupling inequality (again!)

Define the coupling time T as

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

Coupling inequality

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

[Pitman, 1976; Lindvall, 1992]

Drift conditions

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

Drift conditions

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

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

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

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

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

Drift conditions

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

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

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

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

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

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

[Meyn and Tweedie, 1993]

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

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

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

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

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

Explicit bound

Theorem

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

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

where

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

[DMR,2001]

Renewal and CLT

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

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

is

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

Renewal and CLT

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

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

is

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

Standard MC **if** CLT

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

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

Minoration

▸ skip construction

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

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

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

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

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

Split chain

Let $\delta_0, \delta_1, \delta_2, \ldots$ be iid $\mathscr{B}(\varepsilon)$. Then the *split chain*

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

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

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

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

Renewals

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

Then

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

where N_t length of the tth tour, and S_t sum of the $g(X_j)$'s over the tth tour.

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

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

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

Moment conditions

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

If

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

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

[Hobert & al., 2002]

Note that drift + minoration ensures geometric ergodicity

[Rosenthal, 1995; Roberts & Tweedie, 1999]

The Metropolis-Hastings Algorithm

Motivation and leading example

Random variable generation

Monte Carlo Integration

Notions on Markov Chains

The Metropolis-Hastings Algorithm

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

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

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

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

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.

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

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

In this case,

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

[Mengersen & Tweedie, 1996]

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)

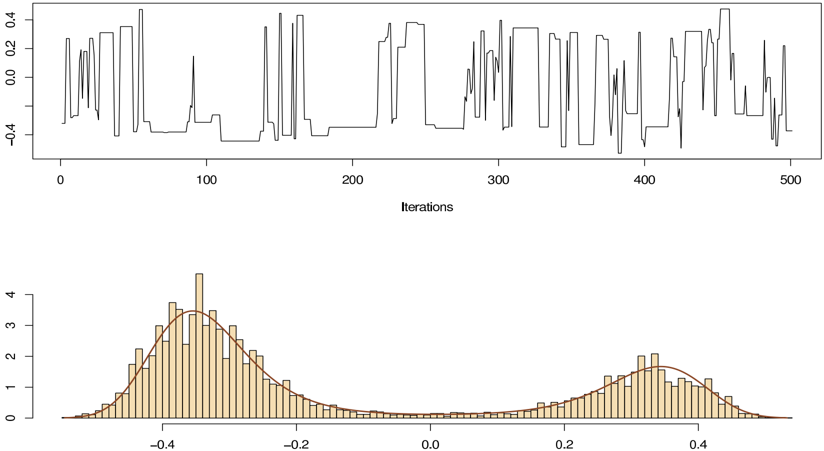
Use for proposal the $\mathscr{N}(\mu_t, \omega_t^2)$ distribution, with

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

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)

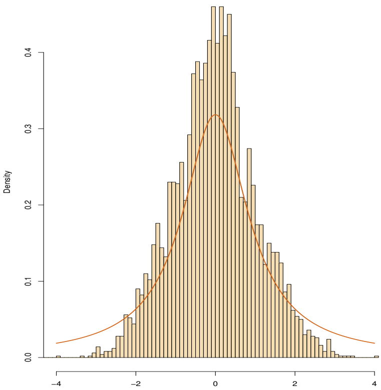
▸ go random W Given a Cauchy $\mathscr{C}(0,1)$ distribution, consider a normal $\mathscr{N}(0,1)$ proposal

The Metropolis–Hastings acceptance ratio is

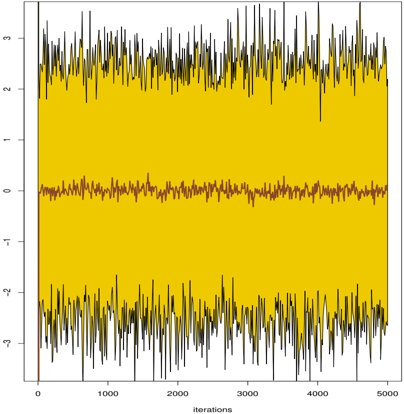
$$\frac{\pi(\xi')/\nu(\xi')}{\pi(\xi)/\nu(\xi))} = \exp\left[\left\{\xi^2 - (\xi')^2\right\}/2\right] \frac{1+(\xi')^2}{(1+\xi^2)} \,.$$

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

[Mengersen & Tweedie, 1996]



Histogram of Markov chain $(\xi_t)_{1\le t\le 5000}$ 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* $g(x) = g(-x)$

Algorithm (Random walk Metropolis)

Given $x^{(t)}$

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

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

Example (Random walk and normal target)

▸ forget History! Generate $\mathcal{N}(0,1)$ based on the uniform proposal $[-\delta,\delta]$
[Hastings (1970)]

The probability of acceptance is then

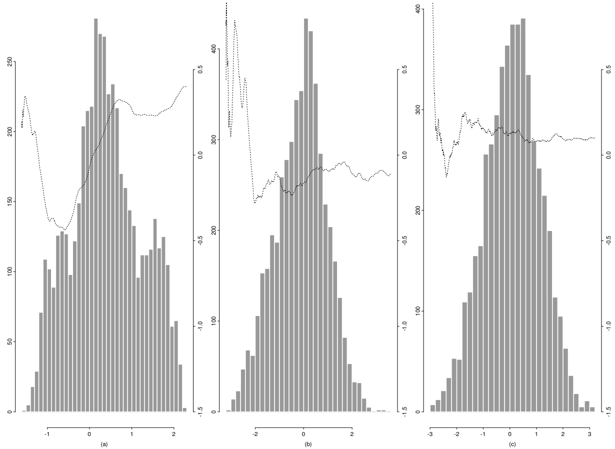
$$\rho(x^{(t)}, y_t) = \exp\{(x^{(t)^2} - y_t^2)/2\} \wedge 1.$$

Example (Random walk & normal (2))

Sample statistics

δ	0.1	0.5	1.0
mean	0.399	-0.111	0.10
variance	0.698	1.11	1.06

© As $\delta \uparrow$, we get better histograms and a faster exploration of the support of f.



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

Example (Mixture models (again!))

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

Example (Mixture models (again!))

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

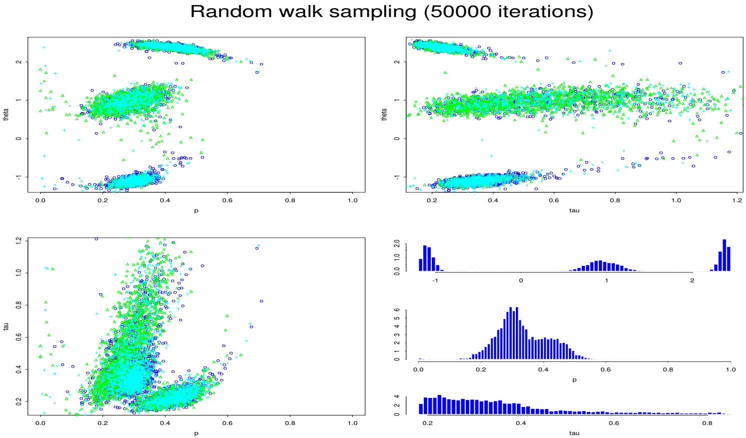
Metropolis-Hastings proposal:

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

where

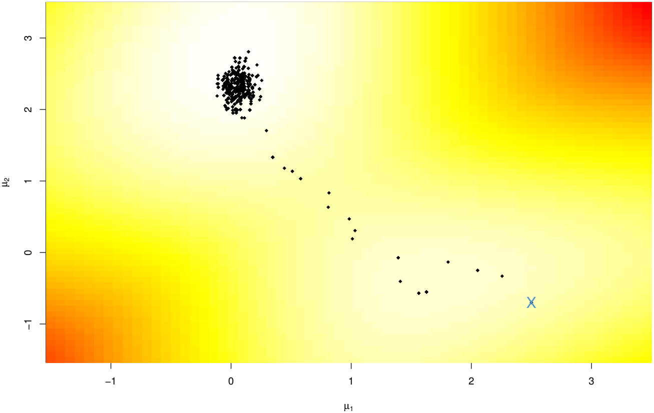
$$\rho^{(t)} = \frac{\pi(\theta^{(t)} + \omega\varepsilon^{(t)}|x)}{\pi(\theta^{(t)}|x)} \wedge 1$$

and ω scaled for good acceptance rate



General case of a 3 component normal mixture

[Celeux & al., 2000]



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

Example (probit model)

▸ skip probit

Likelihood of the **probit model**

$$\prod_{i=1}^{n} \Phi(y_i^{\mathsf{T}}\beta)^{x_i}\,\Phi(-y_i^{\mathsf{T}}\beta)^{1-x_i}$$

Example (probit model)

▸ skip probit

Likelihood of the **probit model**

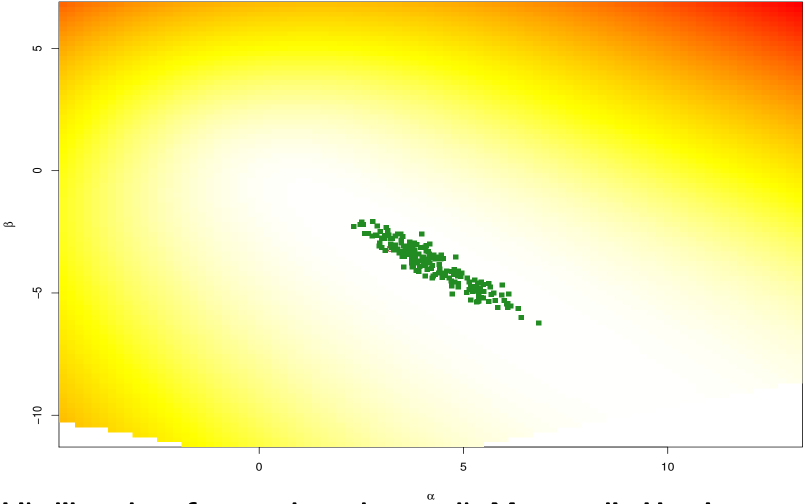
$$\prod_{i=1}^{n} \Phi(y_i^{\mathsf{T}}\beta)^{x_i}\,\Phi(-y_i^{\mathsf{T}}\beta)^{1-x_i}$$

Random walk proposal

$$\beta^{(t+1)} = \beta^{(t)} + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}_p(0, \Sigma)$$

where, for instance,

$$\Sigma = \alpha(YY^{\mathsf{T}})^{-1}$$



Likelihood surface and random walk Metropolis-Hastings steps

Convergence properties

Uniform ergodicity prohibited by random walk structure

Convergence properties

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

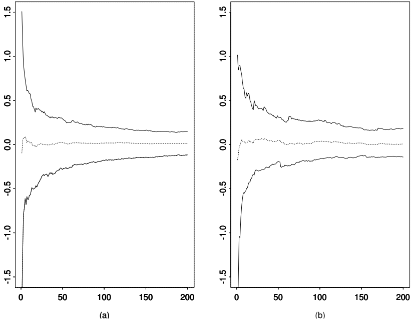
Theorem (Sufficient ergodicity)

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**

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

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

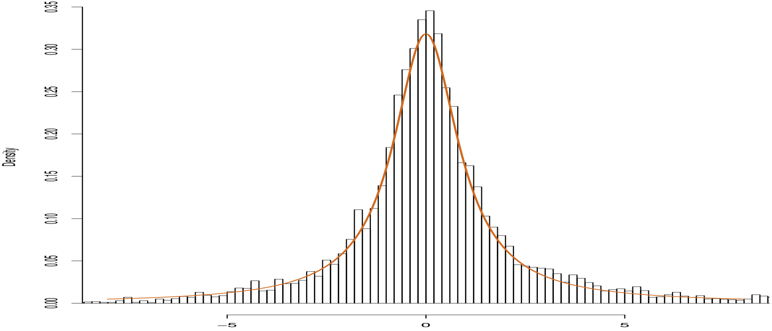
Overall fit of the Cauchy density by the histogram satisfactory, but poor exploration of the tails: 99% quantile of $\mathscr{C}(0,1)$ equal to 3, but no simulation exceeds 14 out of $10,000$!

[Roberts & Tweedie, 2004]

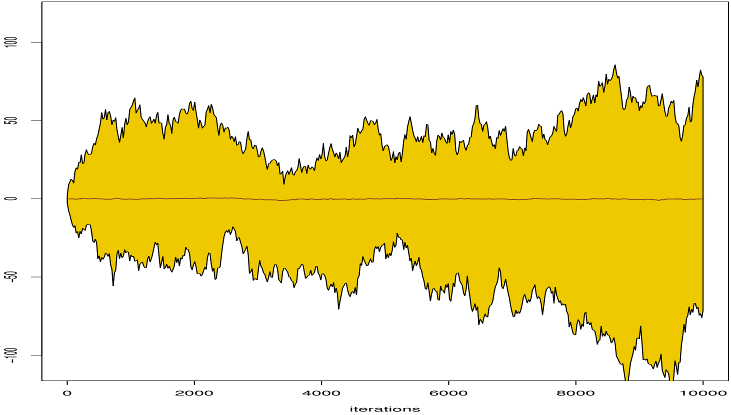
Again, lack of geometric ergodicity!

[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 $\mathscr{N}(\xi, 1)$ proposal



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

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

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) $PV_{n+1} \leq V_n - r(n)h + br(n)\mathbb{I}_C$.

(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

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

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

Extensions

There are many other families of HM algorithms

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

to name just a few...

Langevin Algorithms

Proposal based on the *Langevin diffusion* L_t is defined by the stochastic differential equation

$$dL_t = dB_t + \frac{1}{2}\nabla \log f(L_t)dt,$$

where B_t is the standard *Brownian motion*

Theorem

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

Discretization

Instead, consider the sequence

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

where σ^2 corresponds to the discretization step

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 be **transient**, for instance when

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

MH correction

Accept the new value Y_t with probability

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

Choice of the scaling factor σ
Should lead to an acceptance rate of 0.574 to achieve optimal convergence rates (when the components of x are uncorrelated)
[Roberts & Rosenthal, 1998]

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 (2)

Practical implementation

Choose a parameterized instrumental distribution $g(\cdot|\theta)$ and adjusting the corresponding parameters θ based on the evaluated acceptance rate

$$\hat{\rho}(\theta) = \frac{2}{m} \sum_{i=1}^{m} \mathbb{I}_{\{f(y_i)g(x_i) > f(x_i)g(y_i)\}} ,$$

where $x_1, \ldots, x_m$ sample from f and $y_1, \ldots, y_m$ iid sample from g.

Example (Inverse Gaussian distribution)

▸ no inverse

Simulation from

$$f(z|\theta_1,\theta_2) \propto z^{-3/2}\ \exp\left\{-\theta_1 z - \frac{\theta_2}{z} + 2\sqrt{\theta_1\theta_2} + \log\ \sqrt{2\theta_2}\right\}\ \mathbb{I}_{\mathbb{R}_+}(z)$$

based on the Gamma distribution $\mathcal{G}a(\alpha,\beta)$ with $\alpha = \beta\sqrt{\theta_2/\theta_1}$

Example (Inverse Gaussian distribution)

▸ no inverse

Simulation from

$$f(z|\theta_1,\theta_2) \propto z^{-3/2}\ \exp\left\{-\theta_1 z - \frac{\theta_2}{z} + 2\sqrt{\theta_1\theta_2} + \log\ \sqrt{2\theta_2}\right\} \mathbb{I}_{\mathbb{R}_+}(z)$$

based on the Gamma distribution $\mathcal{G}a(\alpha,\beta)$ with $\alpha = \beta\sqrt{\theta_2/\theta_1}$
Since

$$\frac{f(x)}{g(x)} \propto x^{-\alpha-1/2}\ \exp\left\{(\beta-\theta_1)x - \frac{\theta_2}{x}\right\}\ ,$$

the maximum is attained at

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

Example (Inverse Gaussian distribution (2))

The analytical optimization (in β) of

$$M(\beta) = (x_\beta^*)^{-\alpha-1/2} \ \exp\left\{(\beta-\theta_1)x_\beta^* - \frac{\theta_2}{x_\beta^*}\right\}$$

is impossible

β	0.2	0.5	0.8	0.9	1	1.1	1.2	1.5
$\hat\rho(\beta)$	0.22	0.41	0.54	0.56	0.60	0.63	0.64	0.71
$\mathbb{E}[Z]$	1.137	1.158	1.164	1.154	1.133	1.148	1.181	1.148
$\mathbb{E}[1/Z]$	1.116	1.108	1.116	1.115	1.120	1.126	1.095	1.115

$(\theta_1 = 1.5, \theta_2 = 2$, and $m = 5000)$.

Case of the random walk

Different approach to acceptance rates
A high acceptance rate does not indicate that the algorithm is moving correctly since it indicates that the random walk is moving too slowly on the surface of f.

Case of the random walk

Different approach to acceptance rates
A high acceptance rate does not indicate that the algorithm is moving correctly since it indicates that the random walk is moving too slowly on the surface of f.
If $x^{(t)}$ and y_t are close, i.e. $f(x^{(t)}) \simeq f(y_t)$ y is accepted with probability

$$\min\left(\frac{f(y_t)}{f(x^{(t)})}, 1\right) \simeq 1 \, .$$

For multimodal densities with well separated modes, the negative effect of limited moves on the surface of f clearly shows.

Case of the random walk (2)

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

Rule of thumb

In small dimensions, aim at an average acceptance rate of 50%. In large dimensions, at an average acceptance rate of 25%.

[Gelman,Gilks and Roberts, 1995]

Rule of thumb

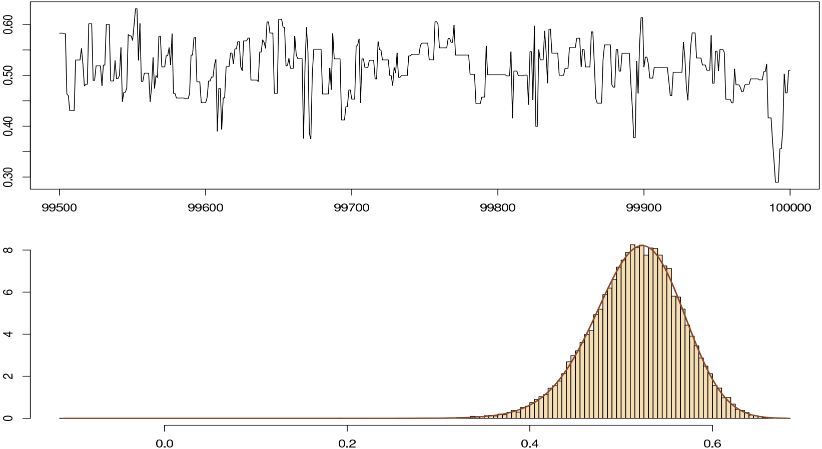
In small dimensions, aim at an average acceptance rate of 50%. In large dimensions, at an average acceptance rate of 25%.

[Gelman,Gilks and Roberts, 1995]

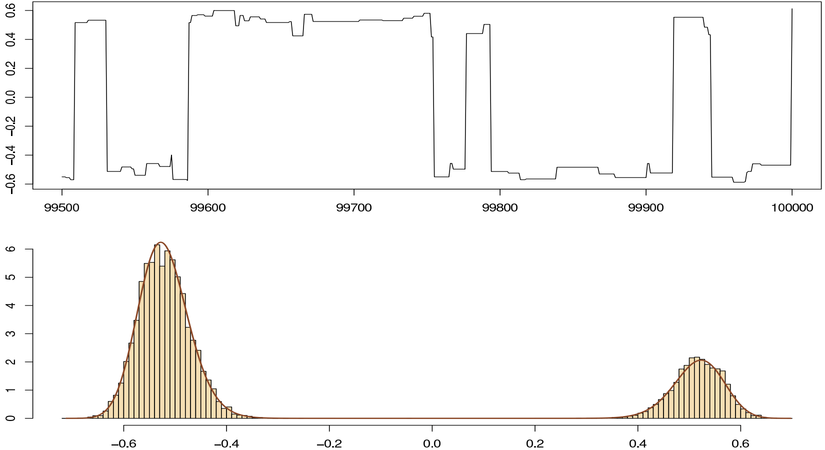
This rule is to be taken with a pinch of salt!

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

The Gibbs Sampler

- General Principles
- Completion
- Convergence
- The Hammersley-Clifford theorem
- Hierarchical models
- 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

A Very Simple Example: Independent N(μ, σ^2) Observations

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

A Very Simple Example: Independent N(μ, σ^2) Observations

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

But...

$$\mu|Y_{0:n}, \sigma^2 \sim \text{N}\left(\mu \,\middle|\, \frac{1}{n}\sum_{i=1}^n Y_i, \frac{\sigma^2}{n}\ \right)$$

$$\sigma^2|Y_{1:n}, \mu \sim \text{IG}\left(\sigma^2 \,\middle|\, \frac{n}{2} - 1, \frac{1}{2}\sum_{i=1}^n (Y_i - \mu)^2\right)$$

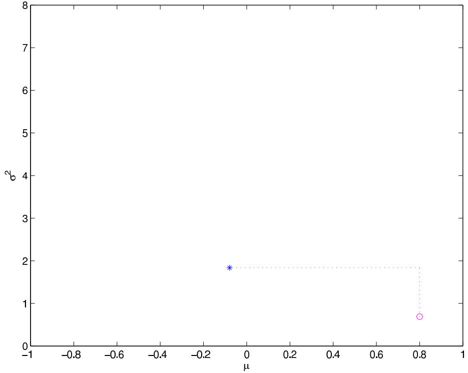
assuming constant (improper) priors on both μ and σ^2

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

R Gibbs Sampler for Gaussian posterior

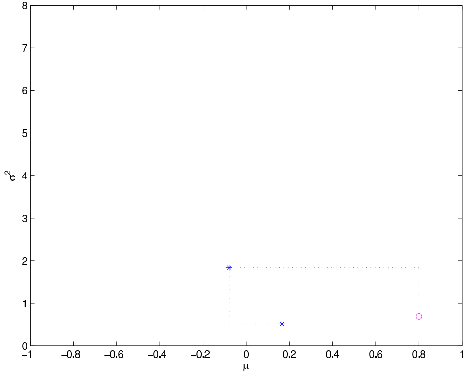
```
n = length(Y);
S = sum(Y);
mu = S/n;
for (i in 1:500)
   S2 = sum((Y-mu)^2);
   sigma2 = 1/rgamma(1,n/2-1,S2/2);
   mu = S/n + sqrt(sigma2/n)*rnorm(1);
```

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



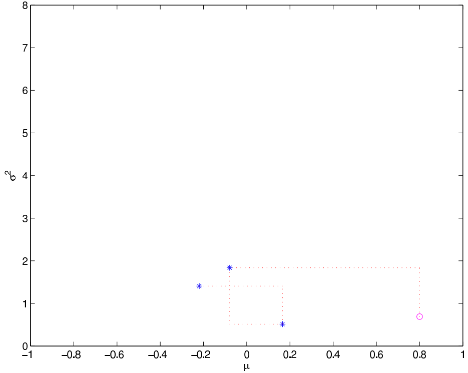
Number of Iterations 1

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



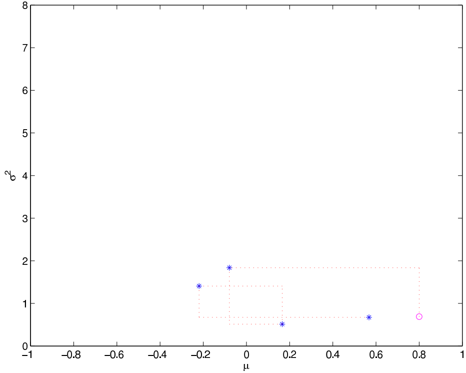
Number of Iterations 1, 2

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



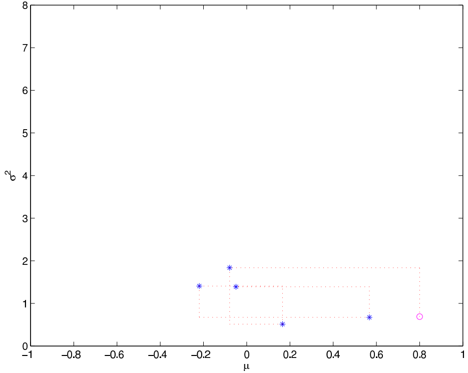
Number of Iterations 1, 2, 3

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



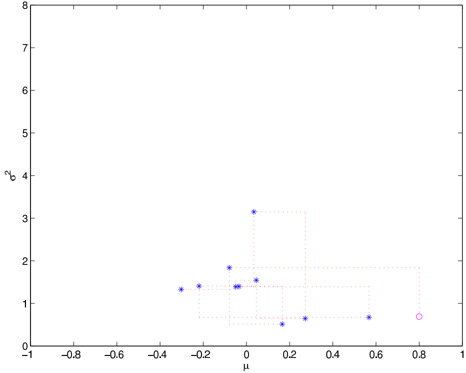
Number of Iterations 1, 2, 3, 4

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



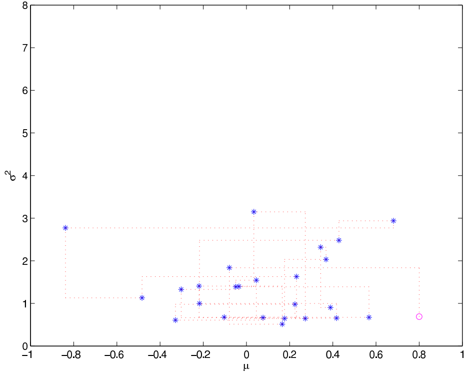
Number of Iterations 1, 2, 3, 4, 5

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



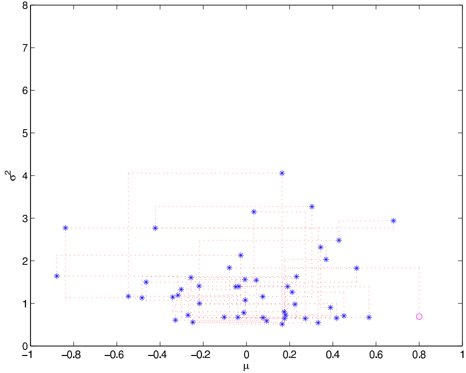
Number of Iterations 1, 2, 3, 4, 5, 10

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



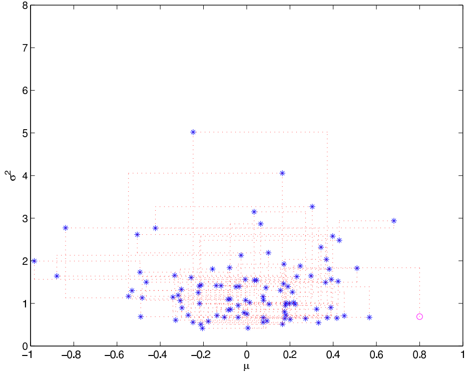
Number of Iterations 1, 2, 3, 4, 5, 10, 25

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



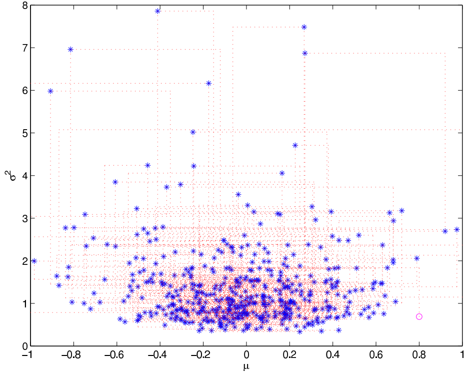
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50

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



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

Example of results with $n = 10$ observations from the 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

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

Example (Mixtures all over again)

Hierarchical missing data structure:

If

$$X_1, \ldots, X_n \sim \sum_{i=1}^{k} p_i f(x|\theta_i),$$

then

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

Z is the component indicator associated with observation x

Example (Mixtures (2))

Conditionally on $(Z_1, \ldots, Z_n) = (z_1, \ldots, z_n)$:

$$\begin{aligned}&\pi(p_1, \ldots, p_k, \theta_1, \ldots, \theta_k | x_1, \ldots, x_n, z_1, \ldots, z_n)\\ &\propto p_1^{\alpha_1+n_1-1} \ldots p_k^{\alpha_k+n_k-1}\\ &\quad \times \pi(\theta_1 | y_1 + n_1\bar{x}_1, \lambda_1 + n_1) \ldots \pi(\theta_k | y_k + n_k\bar{x}_k, \lambda_k + n_k),\end{aligned}$$

with

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

Algorithm (Mixture Gibbs sampler)

1. Simulate

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

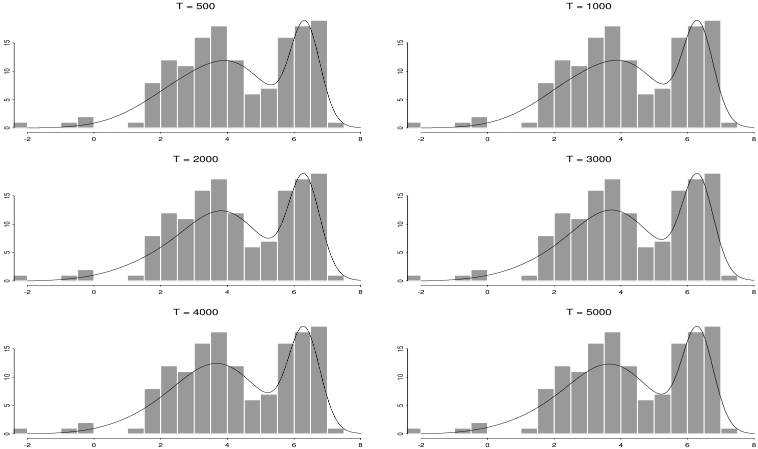
2. Simulate $(j = 1, \ldots, n)$

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

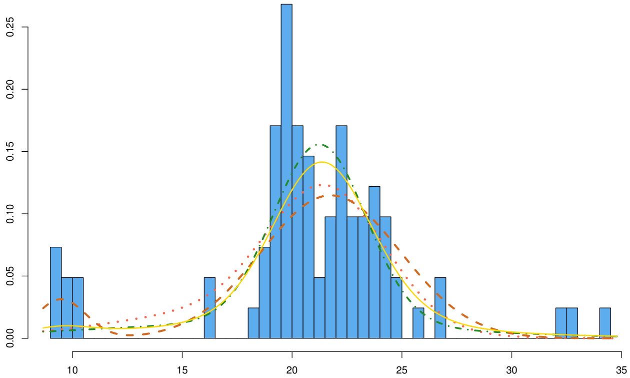
   with $(i = 1, \ldots, k)$

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

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

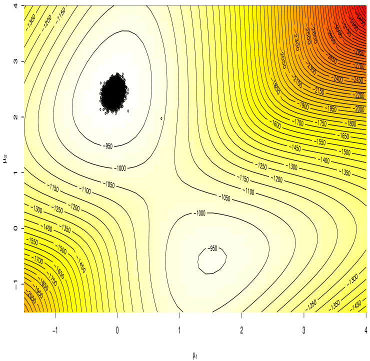


Estimation of the pluggin density for 3 components and T iterations for 149 observations of acidity levels in US lakes



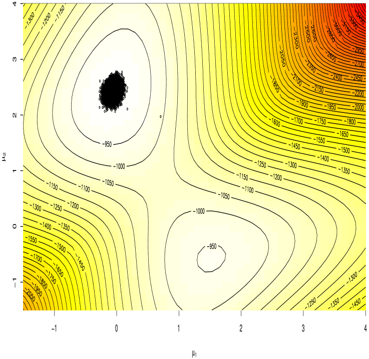
Galaxy dataset (82 **observations) with** $k = 2$ **components**
average density (yellow), and pluggins:
average (tomato), marginal MAP (green), MAP (marroon)

A wee problem



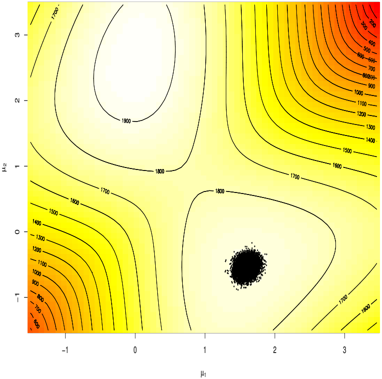
Gibbs started at random

A wee problem



Gibbs started at random

Gibbs stuck at the wrong mode



Random Scan Gibbs sampler

◂ back to basics

▸ don't do random

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

Motivation

The Random Scan Gibbs sampler is **reversible**.

Slice sampler as generic Gibbs

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

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

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

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

leading to the following Gibbs algorithm:

Algorithm (Slice sampler)

Simulate

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

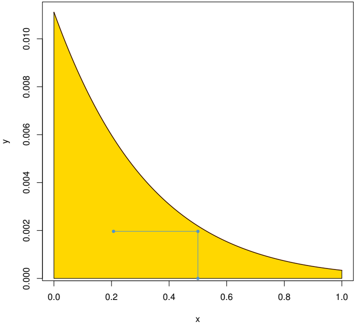
   $\dots$

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

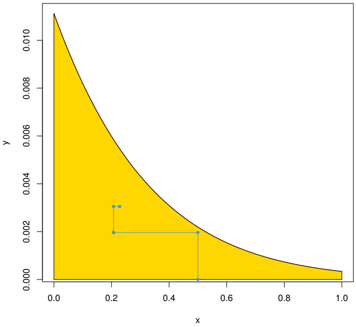
$$A^{(t+1)} = \{y;\ f_i(y) \geq \omega_i^{(t+1)},\ i = 1, \dots, k\}.$$

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



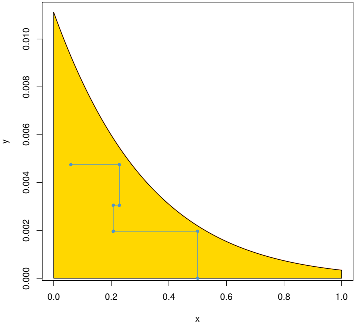
Number of Iterations 2

Example of results with a truncated $N(-3,1)$ distribution



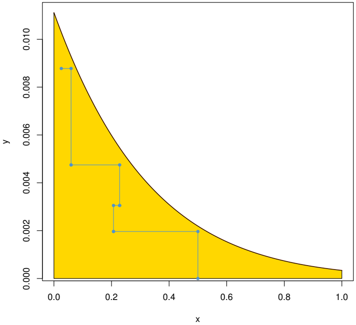
Number of Iterations 2, 3

Example of results with a truncated $N(-3,1)$ distribution



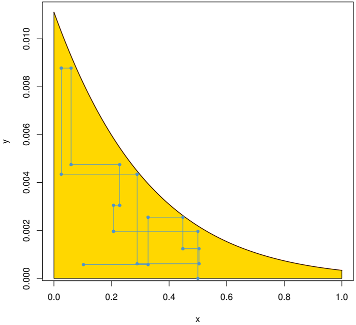
Number of Iterations 2, 3, 4

Example of results with a truncated $N(-3, 1)$ distribution



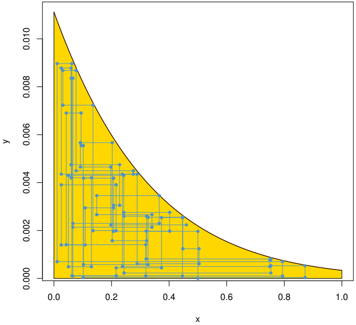
Number of Iterations 2, 3, 4, 5

Example of results with a truncated $N(-3, 1)$ distribution



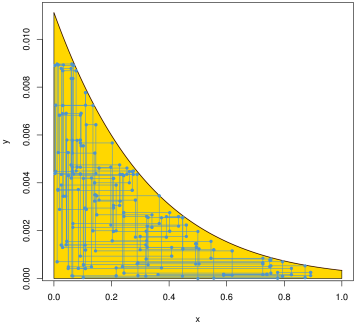
Number of Iterations 2, 3, 4, 5, 10

Example of results with a truncated $N(-3,1)$ distribution



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

Example of results with a truncated $N(-3,1)$ 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

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

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

Example (Stochastic volatility core distribution)

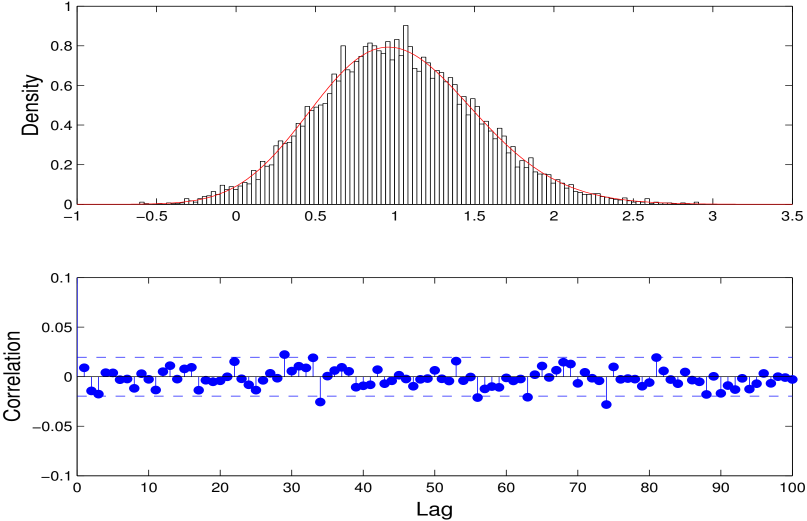
Difficult part of the stochastic volatility model

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

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.



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 $\text{supp} f$ 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]

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

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

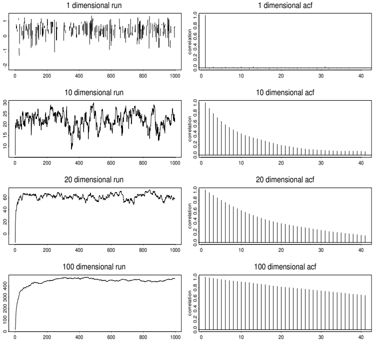
Slice sampler equivalent to one-dimensional slice sampler on

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

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

Hierarchical models

▸ no hierarchy

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

Example (Animal epidemiology)

Counts of the number of cases of clinical mastitis in 127 dairy cattle herds over a one year period

Number of cases in herd i

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

where λ_i is the underlying rate of infection in herd i

Lack of independence might manifest itself as overdispersion.

Example (Animal epidemiology (2))

Modified model

$$\begin{aligned} X_i &\sim \mathscr{P}(\lambda_i) \\ \lambda_i &\sim \mathscr{G}a(\alpha, \beta_i) \\ \beta_i &\sim \mathscr{I}\mathscr{G}(a, b), \end{aligned}$$

Example (Animal epidemiology (2))

Modified model

$$\begin{aligned} X_i &\sim \mathscr{P}(\lambda_i) \\ \lambda_i &\sim \mathscr{G}a(\alpha, \beta_i) \\ \beta_i &\sim \mathscr{I}\mathscr{G}(a, b), \end{aligned}$$

The Gibbs sampler corresponds to conditionals

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

▸ if you hate rats

Example (Rats)

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

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

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

Example (Rats (2))

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

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

and

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

if ith rat treated with a placebo (P) or a drug (D)

Example (Rats (2))

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

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

and

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

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

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

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

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

Conditional decompositions

Easy decomposition of the posterior distribution
For instance, if

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

then

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

Conditional decompositions (2)

where

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

Conditional decompositions (3)

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

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

where

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

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

Posterior complete distribution given by

$$\pi((\theta_i,\delta_i,\xi_i)_i,\mu_\theta,\ldots,\sigma_c,\ldots|\mathscr{D}) \propto$$

$$\prod_{i=1}^{I}\Big\{\exp-\{(\theta_i-\mu_\theta)^2/2\sigma_\theta^2+(\delta_i-\mu_\delta)^2/2\sigma_\delta^2\}$$

$$\prod_{j=1}^{J_i^c}\exp-\{(x_{ij}-\theta_i)^2/2\sigma_c^2\}\prod_{j=1}^{J_i^a}\exp-\{(y_{ij}-\theta_i-\delta_i)^2/2\sigma_a^2\}$$

$$\prod_{j=1}^{J_i^t}\exp-\{(z_{ij}-\theta_i-\delta_i-\xi_i)^2/2\sigma_t^2\}\Big\}$$

$$\prod_{\ell_i=0}\exp-\{(\xi_i-\mu_P)^2/2\sigma_P^2\}\prod_{\ell_i=1}\exp-\{(\xi_i-\mu_D)^2/2\sigma_D^2\}$$

$$\sigma_c^{-\sum_i J_i^c-1}\sigma_a^{-\sum_i J_i^a-1}\sigma_t^{-\sum_i J_i^t-1}(\sigma_\theta\sigma_\delta)^{-I-1}\sigma_D^{-I_D-1}\sigma_P^{-I_P-1}\,,$$

Local conditioning property

For the hierarchical model

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

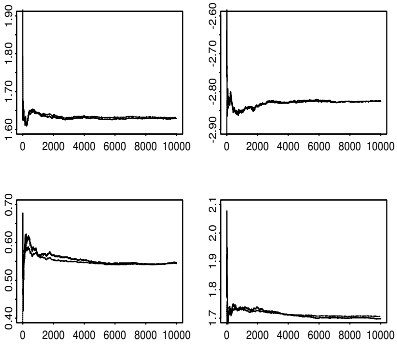
we have

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

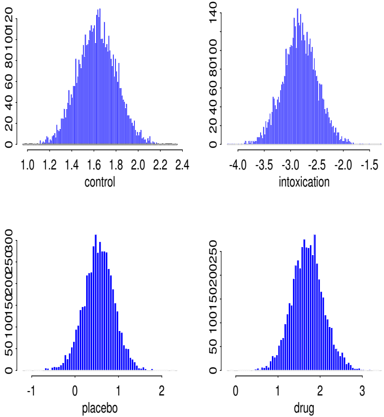
with the convention $\theta_0 = \theta$ and $\theta_{n+1} = 0$.

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

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



Convergence of the posterior means



Posteriors of the effects

Posterior Gibbs inference

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

Posterior probabilities of significant effects

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

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

which can be difficult to work with.

Example (Grouped counting data (2))

Feature Observations with 4 passages and more are grouped

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

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

which can be difficult to work with.

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

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

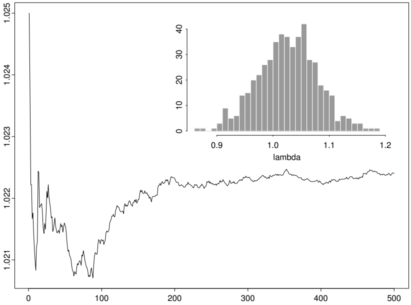
b Simulate

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

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

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

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

Examples of Rao-Blackwellization

Example

Bivariate normal Gibbs sampler

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

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.

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.

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.

The Duality Principle

▸ skip dual part

Ties together the properties of the two Markov chains in Data Augmentation

Consider a Markov chain $(X^{(t)})$ and a sequence $(Y^{(t)})$ of random variables generated from the conditional distributions

$$\begin{aligned} X^{(t)}|y^{(t)} &\sim \pi(x|y^{(t)}) \\ Y^{(t+1)}|x^{(t)}, y^{(t)} &\sim f(y|x^{(t)}, y^{(t)})\ . \end{aligned}$$

The Duality Principle

▸ skip dual part

Ties together the properties of the two Markov chains in Data Augmentation

Consider a Markov chain $(X^{(t)})$ and a sequence $(Y^{(t)})$ of random variables generated from the conditional distributions

$$\begin{aligned} X^{(t)}|y^{(t)} &\sim \pi(x|y^{(t)}) \\ Y^{(t+1)}|x^{(t)}, y^{(t)} &\sim f(y|x^{(t)}, y^{(t)}) \ . \end{aligned}$$

Theorem (Duality properties)

If the chain $(Y^{(t)})$ is ergodic then so is $(X^{(t)})$ and the duality also holds for geometric or uniform ergodicity.

Note

The chain $(Y^{(t)})$ can be discrete, and the chain $(X^{(t)})$ continuous.

Improper Priors

$\lightning$ 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

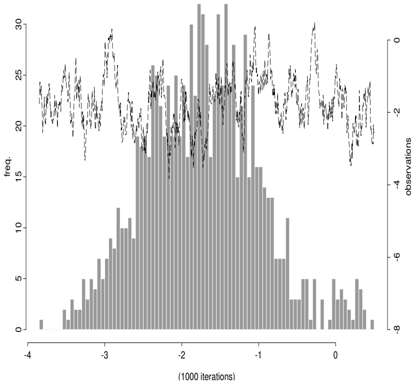
$$\pi(\mu, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2} \; .$$

Example (Improper random effects 2)

The conditional distributions

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

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

MCMC tools for variable dimension problems

MCMC tools for variable dimension problems

- Introduction
- Green's method
- Birth and Death processes

A new brand of problems

There exist setups where

> **One of the things we do not know is the number of things we do not know**
>
> [Peter Green]

Bayesian Model Choice

Typical in model choice settings

- **model construction (nonparametrics)**
- **model checking (goodness of fit)**
- **model improvement (expansion)**
- **model prunning (contraction)**
- **model comparison**
- *hypothesis* **testing (Science)**
- **prediction (finance)**

Bayesian Model Choice II

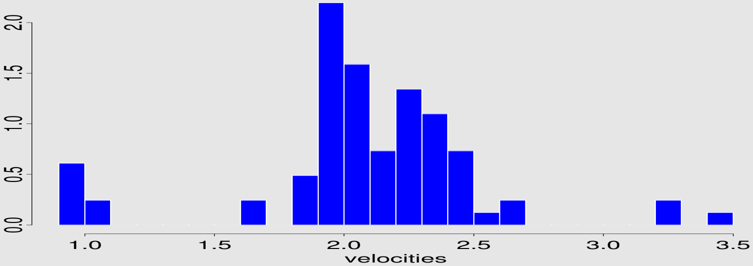
Many areas of application

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

Example (Mixture again, yes!)

Benchmark dataset: Speed of galaxies

[Roeder, 1990; Richardson & Green, 1997]



Example (Mixture again (2))

Modelling by a mixture model

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

i?

Bayesian variable dimension model

Definition

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

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

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

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

and a prior distribution on the indices of these models,

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

Alternative notation:

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

Bayesian solution

Formally over:

1. Compute

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

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

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

as predictive

[Different decision theoretic perspectives]

Difficulties

Not at

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

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

[even if there are parameters common to several models]

Difficulties

Not at

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

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

[even if there are parameters common to several models]

Rather at

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

Green's resolution

Setting up a proper measure–theoretic framework for designing moves *between* models $\mathfrak{M}_k$

[Green, 1995]

Green's resolution

Setting up a proper measure–theoretic framework for designing moves *between* models $\mathfrak{M}_k$

[Green, 1995]

Create a **reversible kernel** $\mathfrak{K}$ on $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ such that

$$\int_A \int_B \mathfrak{K}(x, dy)\pi(x)dx = \int_B \int_A \mathfrak{K}(y, dx)\pi(y)dy$$

for the invariant density π [x is of the form $(k, \theta^{(k)})$]

Green's resolution (2)

Write $\mathfrak{K}$ as

$$\mathfrak{K}(x,B) = \sum_{m=1}^{\infty} \int \rho_m(x,y)\mathfrak{q}_m(x,dy) + \omega(x)\mathbb{I}_B(x)$$

where $\mathfrak{q}_m(x,dy)$ is a transition measure to model $\mathfrak{M}_m$ and $\rho_m(x,y)$ the corresponding acceptance probability.

Green's resolution (2)

Write $\mathfrak{K}$ as

$$\mathfrak{K}(x,B) = \sum_{m=1}^{\infty} \int \rho_m(x,y)\mathfrak{q}_m(x,dy) + \omega(x)\mathbb{I}_B(x)$$

where $\mathfrak{q}_m(x,dy)$ is a transition measure to model $\mathfrak{M}_m$ and $\rho_m(x,y)$ the corresponding acceptance probability.

Introduce a **symmetric** measure $\xi_m(dx,dy)$ on $\mathfrak{H}^2$ and impose on $\pi(dx)\mathfrak{q}_m(x,dy)$ to be absolutely continuous wrt ξ_m,

$$\frac{\pi(dx)\mathfrak{q}_m(x,dy)}{\xi_m(dx,dy)} = g_m(x,y)$$

Green's resolution (2)

Write $\mathfrak{K}$ as

$$\mathfrak{K}(x, B) = \sum_{m=1}^{\infty} \int \rho_m(x, y)\mathfrak{q}_m(x, dy) + \omega(x)\mathbb{I}_B(x)$$

where $\mathfrak{q}_m(x, dy)$ is a transition measure to model $\mathfrak{M}_m$ and $\rho_m(x, y)$ the corresponding acceptance probability.

Introduce a **symmetric** measure $\xi_m(dx, dy)$ on $\mathfrak{H}^2$ and impose on $\pi(dx)\mathfrak{q}_m(x, dy)$ to be absolutely continuous wrt ξ_m,

$$\frac{\pi(dx)\mathfrak{q}_m(x, dy)}{\xi_m(dx, dy)} = g_m(x, y)$$

Then

$$\rho_m(x, y) = \min\left\{1, \frac{g_m(y, x)}{g_m(x, y)}\right\}$$

ensures reversibility

Special case

When contemplating a move between two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the Markov chain being in state $\theta_1 \in \mathfrak{M}_1$, denote by $\mathfrak{K}_{1\to 2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to 1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

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

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

Special case

When contemplating a move between two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the Markov chain being in state $\theta_1 \in \mathfrak{M}_1$, denote by $\mathfrak{K}_{1\rightarrow 2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\rightarrow 1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

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

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

Proposal expressed as

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

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

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

Special case (2)

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

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

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

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

Interpretation (1)

The representation puts us back in a fixed dimension setting:

- $\mathfrak{M}_1 \times \mathfrak{V}_{1\to 2}$ and $\mathfrak{M}_2$ are in one-to-one relation

Interpretation (1)

The representation puts us back in a fixed dimension setting:

- $\mathfrak{M}_1 \times \mathfrak{V}_{1\to 2}$ and $\mathfrak{M}_2$ are in one-to-one relation
- *regular* Metropolis–Hastings move from the couple $(\theta_1, v_{1\to 2})$ to θ_2 when stationary distributions are

$$\pi(\mathfrak{M}_1, \theta_1) \times \varphi_{1\to 2}(v_{1\to 2})$$

  and $\pi(\mathfrak{M}_2, \theta_2)$, and when proposal distribution is *deterministic* (??)

Interpretation (2)

Consider, instead, the proposals

$$\theta_2 \sim \mathcal{N}(\Psi_{1\to2}(\theta_1, v_{1\to2}), \varepsilon) \qquad \text{and} \qquad \Psi_{1\to2}(\theta_1, v_{1\to2}) \sim \mathcal{N}(\theta_2, \varepsilon)$$

Interpretation (2)

Consider, instead, the proposals

$$\theta_2 \sim \mathcal{N}(\Psi_{1\to2}(\theta_1, v_{1\to2}), \varepsilon) \qquad \text{and} \qquad \Psi_{1\to2}(\theta_1, v_{1\to2}) \sim \mathcal{N}(\theta_2, \varepsilon)$$

Reciprocal proposal has density

$$\frac{\exp\left\{-(\theta_2 - \Psi_{1\to2}(\theta_1, v_{1\to2}))^2/2\varepsilon\right\}}{\sqrt{2\pi\varepsilon}} \times \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|$$

by the Jacobian rule.

Interpretation (2)

Consider, instead, the proposals

$$\theta_2 \sim \mathcal{N}(\Psi_{1\to2}(\theta_1, v_{1\to2}), \varepsilon) \qquad \text{and} \qquad \Psi_{1\to2}(\theta_1, v_{1\to2}) \sim \mathcal{N}(\theta_2, \varepsilon)$$

Reciprocal proposal has density

$$\frac{\exp\left\{-(\theta_2 - \Psi_{1\to2}(\theta_1, v_{1\to2}))^2/2\varepsilon\right\}}{\sqrt{2\pi\varepsilon}} \times \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|$$

by the Jacobian rule.
Thus Metropolis–Hastings acceptance probability is

$$1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)}{\pi(\mathfrak{M}_1, \theta_1)\,\varphi_{1\to2}(v_{1\to2})}\left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|$$

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

Saturation

[Brooks, Giudici, Roberts, 2003]

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

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

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

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

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

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

Back to fixed dimension

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

Back to fixed dimension

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

Algorithm (Three stage MCMC update)

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

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

Example (Mixture of normal distributions)

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

[Richardson & Green, 1997]

Moves:

Example (Mixture of normal distributions)

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

[Richardson & Green, 1997]

Moves:

(i) Split

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

(ii) Merge (*reverse*)

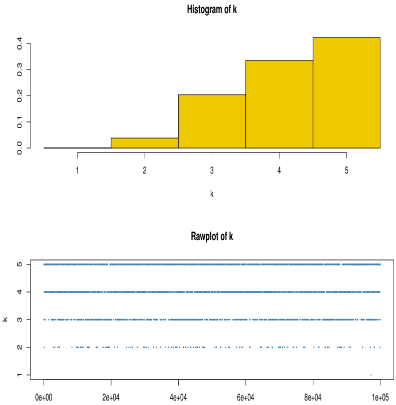
Example (Mixture (2))

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

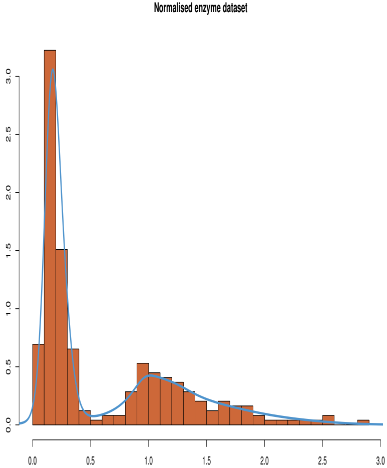
Equivalent

(i). Split

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



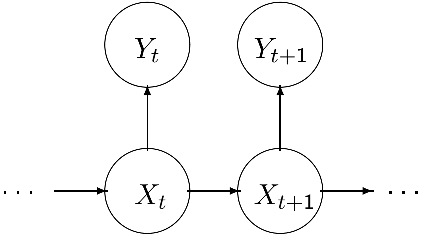
Histogram and rawplot of $100,000$ k's under the constraint $k \leq 5$.



Example (Hidden Markov model)

▸ move to birth Extension of the mixture model

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


Y_t
Y_{t+1}
$\cdots$
X_t
X_{t+1}
$\cdots$

Example (Hidden Markov model (2))

Move to split component $j_\star$ into j_1 and j_2:

$$\omega_{ij_1} = \omega_{ij_\star}\varepsilon_i, \quad \omega_{ij_2} = \omega_{ij_\star}(1-\varepsilon_i), \quad \varepsilon_i \sim \mathcal{U}(0,1);$$

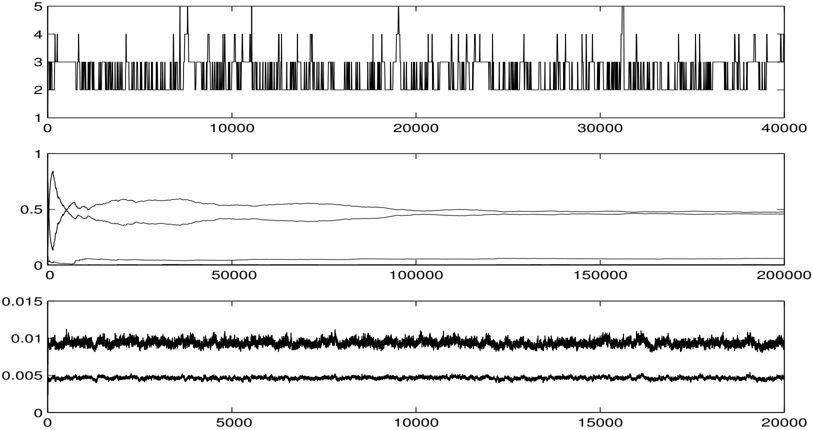
$$\omega_{j_1 j} = \omega_{j_\star j}\xi_j, \quad \omega_{j_2 j} = \omega_{j_\star j}/\xi_j, \quad \xi_j \sim \log\mathcal{N}(0,1);$$

similar ideas give $\omega_{j_1 j_2}$ etc.;

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

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

[Robert & al., 2000]



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

Example (Autoregressive model)

▸ move to birth

Typical setting for model choice: determine order p of $AR(p)$ model

Example (Autoregressive model)

▸ move to birth

Typical setting for model choice: determine order p of $AR(p)$ model
Consider the (less standard) representation

$$\prod_{i=1}^{p} (1 - \lambda_i B) \; X_t = \epsilon_t \,, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

where the λ_i's are within the unit circle if complex and within $[-1, 1]$ if real.

[Huerta and West, 1998]

$AR(p)$ reversible jump algorithm

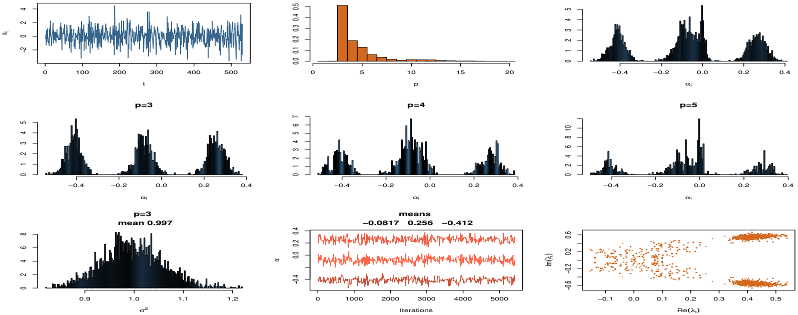
Example (Autoregressive (2))

Uniform priors for the real and complex roots λ_j,

$$\frac{1}{\lfloor k/2 \rfloor + 1} \prod_{\lambda_i \in \mathbb{R}} \frac{1}{2} \mathbb{I}_{|\lambda_i|<1} \prod_{\lambda_i \notin \mathbb{R}} \frac{1}{\pi} \mathbb{I}_{|\lambda_i|<1}$$

and (purely birth-and-death) proposals based on these priors

- k → k+1 [Creation of real root]
- k → k+2 [Creation of complex root]
- k → k-1 [Deletion of real root]
- k → k-2 [Deletion of complex root]



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

Birth and Death processes

▸ instant death!

Use of an alternative methodology based on a Birth–&-Death (point) process
[Preston, 1976; Ripley, 1977; Geyer & Møller, 1994; Stevens, 1999]

Birth and Death processes

▸ instant death!

Use of an alternative methodology based on a Birth–&-Death (point) process
[Preston, 1976; Ripley, 1977; Geyer & Møller, 1994; Stevens, 1999]

Idea: Create a Markov chain in *continuous time*, i.e. a *Markov jump process*, moving between models $\mathfrak{M}_k$, by births (to increase the dimension), deaths (to decrease the dimension), and other moves.

Birth and Death processes

Time till next modification **(jump)** is exponentially distributed with rate depending on current state

Remember: if $\xi_1, \ldots, \xi_v$ are exponentially distributed, $\xi_i \sim \mathcal{E}(\lambda_i)$,

$$\min \xi_i \sim \mathcal{E}\left(\sum_i \lambda_i\right)$$

Birth and Death processes

Time till next modification **(jump)** is exponentially distributed with rate depending on current state

Remember: if $\xi_1, \ldots, \xi_v$ are exponentially distributed, $\xi_i \sim \mathcal{E}(\lambda_i)$,

$$\min \xi_i \sim \mathcal{E}\left(\sum_i \lambda_i\right)$$

Difference with MH-MCMC: Whenever a jump occurs, the corresponding move *is always accepted*. Acceptance probabilities replaced with holding times.

Implausible configurations

$$L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ll 1$$

die quickly.

Balance condition

Sufficient to have **detailed balance**

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

for $\tilde{\pi}(\boldsymbol{\theta}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ to be stationary.
Here $q(\boldsymbol{\theta},\boldsymbol{\theta}')$ rate of moving from state $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$.
Possibility to add split/merge and fixed-k processes if balance condition satisfied.

Example (Mixture cont'd)

Stephen's original modelling:

- Representation as a (marked) point process

$$\Phi = \left\{ \{p_j, (\mu_j, \sigma_j)\} \right\}_j$$

- Birth rate λ_0 (constant)
- Birth proposal from the prior
- Death rate $\delta_j(\Phi)$ for removal of point j
- Death proposal removes component and modifies weights

Example (Mixture cont'd (2))

- Overall death rate

$$\sum_{j=1}^{k} \delta_j(\Phi) = \delta(\Phi)$$

- Balance condition

$$(k{+}1)\; d(\Phi \cup \{p,(\mu,\sigma)\})\; L(\Phi \cup \{p,(\mu,\sigma)\}) = \lambda_0 L(\Phi) \frac{\pi(k)}{\pi(k+1)}$$

with

$$d(\Phi \setminus \{p_j,(\mu_j,\sigma_j)\}) = \delta_j(\Phi)$$

- Case of Poisson prior $k \sim \mathcal{P}oi(\lambda_1)$

$$\delta_j(\Phi) = \frac{\lambda_0}{\lambda_1} \frac{L(\Phi \setminus \{p_j,(\mu_j,\sigma_j)\})}{L(\Phi)}$$

Stephen's original algorithm

Algorithm (Mixture Birth& Death)

```
For $v = 0, 1, \cdots, V$
    $t \leftarrow v$
    Run till $t > v + 1$
        1. Compute $\delta_j(\Phi) = \frac{L(\Phi|\Phi_j)}{L(\Phi)} \frac{\lambda_0}{\lambda_1}$
        2. $\delta(\Phi) \leftarrow \sum_{j=1}^{k} \delta_j(\Phi_j)$, $\xi \leftarrow \lambda_0 + \delta(\Phi)$, $u \sim \mathcal{U}([0,1])$
        3. $t \leftarrow t - u \log(u)$
```

Algorithm (Mixture Birth& Death (cont'd))

```
4. With probability $\delta(\Phi)/\xi$
       Remove component $j$ with probability $\delta_j(\Phi)/\delta(\Phi)$
       $k \leftarrow k-1$
       $p_\ell \leftarrow p_\ell/(1-p_j)\ \ (\ell \neq j)$
   Otherwise,
       Add component $j$ from the prior $\pi(\mu_j, \sigma_j)\ p_j \sim \mathcal{B}e(\gamma, k\gamma)$
       $p_\ell \leftarrow p_\ell(1-p_j)\ \ (\ell \neq j)$
       $k \leftarrow k+1$
5. Run $I$ MCMC$(k, \beta, p)$
```

Rescaling time

▸ move to HMM In discrete-time RJMCMC, let the time unit be $1/N$, put

$$\beta_k = \lambda_k/N \quad \text{and} \quad \delta_k = 1 - \lambda_k/N$$

As $N \to \infty$, each birth proposal will be accepted, and having k components births occur according to a Poisson process with rate λ_k while component (w, ϕ) dies with rate

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

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

Example (HMM models (cont'd))

Implementation of the split-and-combine rule of Richardson and Green (1997) in continuous time

Example (HMM models (cont'd))

Implementation of the split-and-combine rule of Richardson and Green (1997) in continuous time

Move to split component j_* into j_1 and j_2:

$$\omega_{ij_1} = \omega_{ij_*}\epsilon_i, \quad \omega_{ij_2} = \omega_{ij_*}(1-\epsilon_i), \quad \epsilon_i \sim \mathcal{U}(0,1);$$

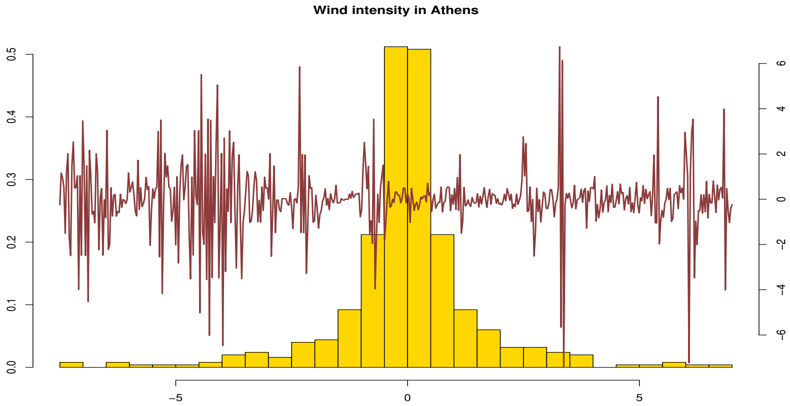
$$\omega_{j_1 j} = \omega_{j_* j}\xi_j, \quad \omega_{j_2 j} = \omega_{j_* j}/\xi_j, \quad \xi_j \sim \log\mathcal{N}(0,1);$$

similar ideas give $\omega_{j_1 j_2}$ etc.;

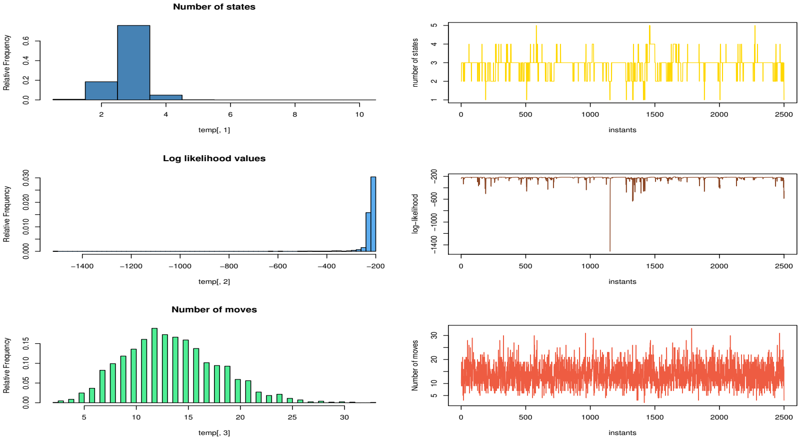
$$\mu_{j_1} = \mu_{j_*} - 3\sigma_{j_*}\epsilon_\mu, \quad \mu_{j_2} = \mu_{j_*} + 3\sigma_{j_*}\epsilon_\mu, \quad \epsilon_\mu \sim \mathcal{N}(0,1);$$

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

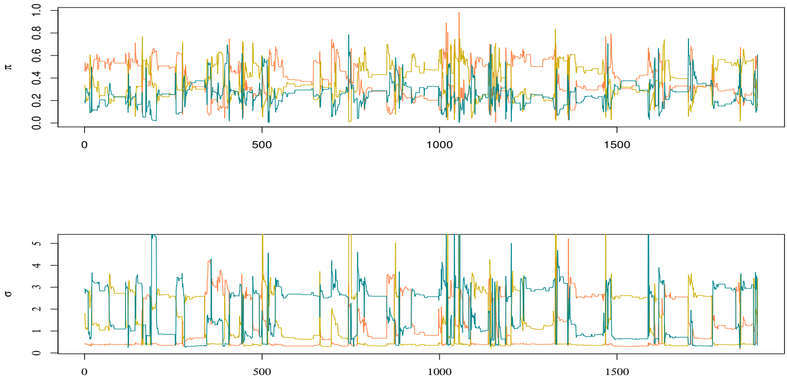
[Cappé & al, 2001]



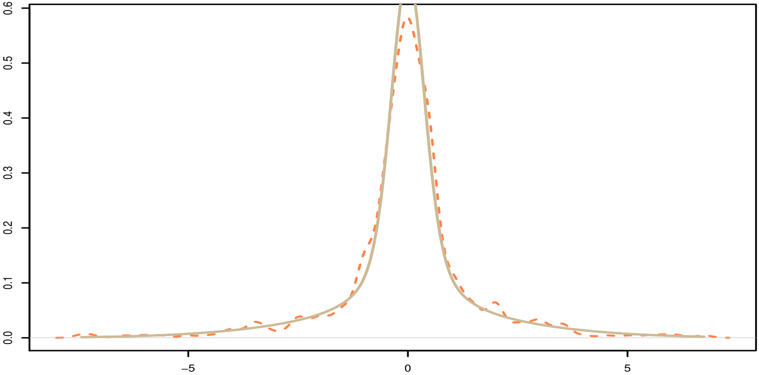
Histogram and rawplot of 500 wind intensities in Athens



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



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



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

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

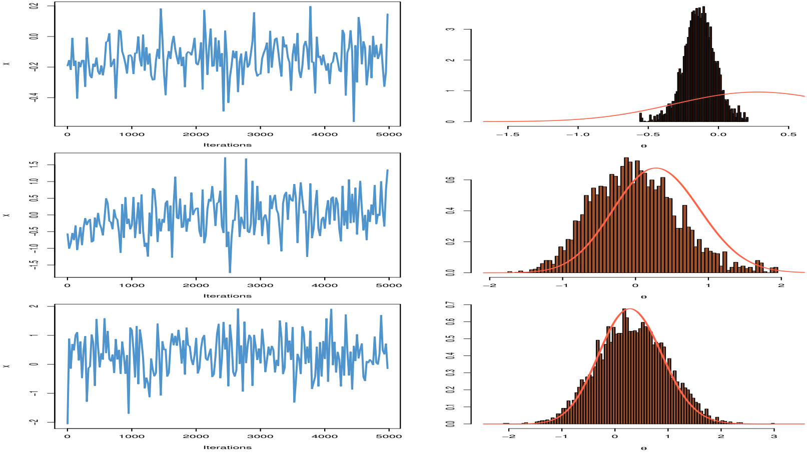
$$\prod_{j=2}^{n} \left[\frac{\nu + (x_j - \theta^{(t)})^2}{\nu + (x_j - \xi)^2} \right]^{-(\nu+1)/2} \frac{\exp -(\mu_t - \theta^{(t)})^2 / 2\sigma_t^2}{\exp -(\mu_t - \xi)^2 / 2\sigma_t^2} \,,$$

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

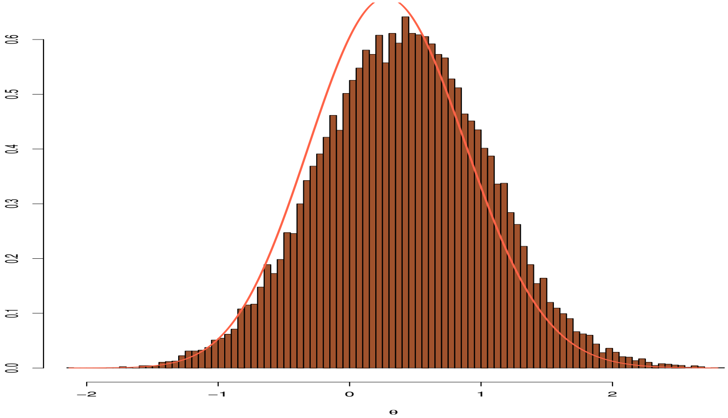
Example (Poly t distribution (2))

Invalid scheme:

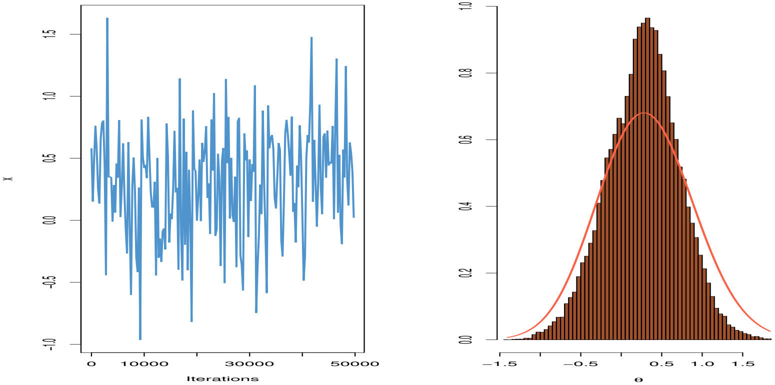
- when range of initial values too small, the $\theta^{(i)}$'s cannot converge to the target distribution and concentrates on too small a support.
- long-range dependence on past values modifies the distribution of the sequence.
- using past simulations to create a non-parametric approximation to the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

$$\sum_n W_n h(Y_n) / \sum_n W_n \tag{4}$$

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

$$\mathbb{E}[\omega_t|\theta_t] \propto \pi(\theta_t) \tag{5}$$

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

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

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

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

and

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

for the R scheme

Notes (2)

- Geometric structure of the weights

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

and

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

for the R scheme

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

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

Alternative scheme

Preservation of weight expectation:

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

Alternative scheme (2)

Then

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

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

$$\min(1, \omega_0 r(x_0, y_1)) \triangleq \alpha(x_0, y_1)$$

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,

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

if 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

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

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

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

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 $t \geq 1$,

$$\frac{1}{N}\sum_{k=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}) \xrightarrow[P]{N\to\infty} \Pi(h)$$

and

$$p_{t,d} \xrightarrow[P]{N\to\infty} \alpha_d^t$$

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

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

and define the sequence

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

Kullback divergence

Definition (Kullback divergence)

For $\alpha \in S$,

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

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

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

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

$$\begin{aligned}\boldsymbol{\alpha}^{\text{min}} = \arg\min_{\boldsymbol{\alpha}\in S} KL(\boldsymbol{\alpha}) &= \arg\max_{\boldsymbol{\alpha}\in S} \int \log p_{\boldsymbol{\alpha}}(\bar{x}) \Pi \otimes \Pi(d\bar{x}) \\ &= \arg\max_{\boldsymbol{\alpha}\in S} \int \log \int p_{\boldsymbol{\alpha}}(\bar{x}, K) dK \, \Pi \otimes \Pi(d\bar{x})\end{aligned}$$

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

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

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

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

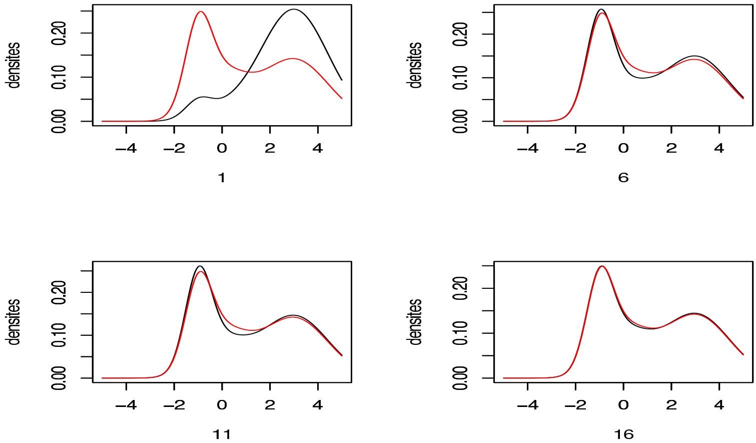
and use 3 proposals: $\mathscr{N}(-1,0.3)$, $\mathscr{N}(0,1)$ and $\mathscr{N}(3,2)$

[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

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

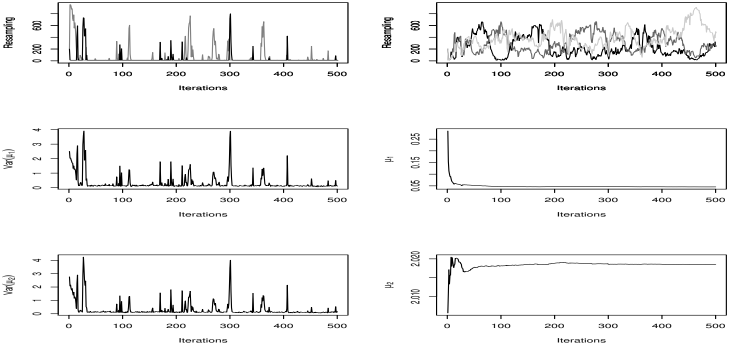
Resample the $\left((\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right)_j$ using the weights ϱ_j,

Details

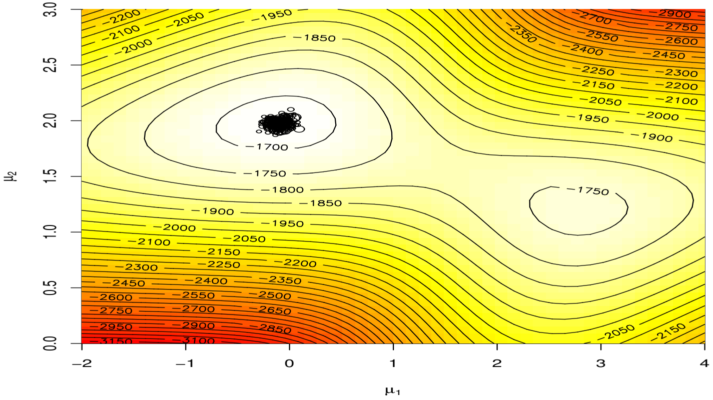
After an arbitrary initialisation, use of the previous (importance) sample (after resampling) to build random walk proposals,

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

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