Tutorial Lectures on MCMC II

Sujit Sahu [a]

Utrecht: August 2000.

You are here because:

- You would like to see more advanced methodology.
- Maybe some computer illustration.

[a] In close association with Gareth Roberts

Outline:

- $\boxed{\text{Convergence Study for Better Design}}$
- Graphical Models
- BUGS illustrations.
- Bayesian Model Choice
- Reversible Jump
- Adaptive Methods

Rate of Convergence

Let $\| \cdot \|$ define a suitable divergence measure. Rate of convergence is given by:

$$\rho = \inf_{\xi \in [0,1]} \xi$$

where

$$\|K^t(x^{(0)}, x) - \pi(x)\| \leq V(x^{(0)})\xi^t$$

where

- K^t is the density of $X^{(t)}$
- $V(\cdot)$ is a non-negative function taking care of the effect of the starting point.

For example, if $\pi(x)$ is the density of bivariate normal with correlation ζ then $\rho = \zeta^2$.

The rate ρ

- is a global measure of convergence
- does not depend on $x^{(0)}$, the starting point
- not only tells how fast it converges
- but also indicates the auto-correlation present in the MCMC sequence.
- the last point is concerned with the nse of ergodic averages.

Use ρ to design MCMC.

Recall: $\rho =$ cor^2 for bivariate normal example.

Can we have faster convergence if correlations are reduced?

Answer: Yes and No !

No, because the correlation structure matters a great deal in high dimensions.

That is why, Roberts and Sahu (1997) investigate to see the effects of

- blocking
- parameterization
- correlation structure
- updating schemes

on the Gibbs sampler for Gaussian cases.

Extrapolation

- Extrapolate results to the non-normal cases with similar correlation structure.
- Some of this should work because of asymptotic normality of the posterior distribution.
- Two further papers Roberts and Sahu (2000) and Sahu and Roberts (1999) investigate the issues in detail.

We show:

Under a Gaussian setup the rate of convergence of the EM/ECM algorithm is the same as the rate of convergence of the Gibbs sampler.

Implications:

- if a given problem is easy (hard) for the EM, then it is easy (hard) for the Gibbs too.
- can use improvement strategy for one algorithm to hasten convergence in the other.

Extra benefits of first running the EM:

- EM convergence is easy to assess.
- Can use the lessons learned to implement the Gibbs.
- Can use the last EM iterate as the starting point in MCMC.

Further, we also prove the intuitive result that *under conditions the Gibbs may converge faster than the EM algorithm which does not use the assumed proper prior distributions.*

♡ Example: Sahu and Roberts (1999)

$$\begin{aligned} y_{ij} &= \mu + b_i + \epsilon_{ij}, \\ b_i &\sim N(0, \sigma_2^2) \\ \epsilon_{ij} &\sim N(0, \sigma_1^2) \end{aligned}$$

This is called NCEN (not centered).

Priors:

- $1/\sigma_2^2 = \tau \sim \Gamma(A, A)$
- $1/\sigma_1^2 = \delta \sim \Gamma(A, A)$
- Flat prior for μ.

HCEN (hierarchically centered)

(Gelfand, Sahu and Carlin, 1995)

$$\begin{aligned} y_{ij} &= \eta_i + \epsilon_{ij}, \\ \eta_i &\sim N(\mu, \sigma_2^2) \\ \epsilon_{ij} &\sim N(0, \sigma_1^2) \end{aligned}$$

AUXA (Auxiliary, Meng and Van Dyk, 1997)

$$\begin{aligned} y_{ij} &= \mu + \tau^{a/2} b_i + \epsilon_{ij} \\ b_i &\sim N(0, \tau^{-(1+a)}) \\ \epsilon_{ij} &\sim N(0, \sigma_1^2) \end{aligned}$$

	NCEN	AUXA	HCEN
	ECM Algorithm		
Rate NIT CPU	0.832 3431 23.32	0.845 150 5.57	0.287 15 0.25
	Gibbs Sampler		
λ	0.786	0.760	0.526

λ= Principal eigenvalue of the lag 1 auto-correlation matrix of $(\mu^{(t)}, \sigma_2^{2(t)}, \sigma_1^{2(t)})$

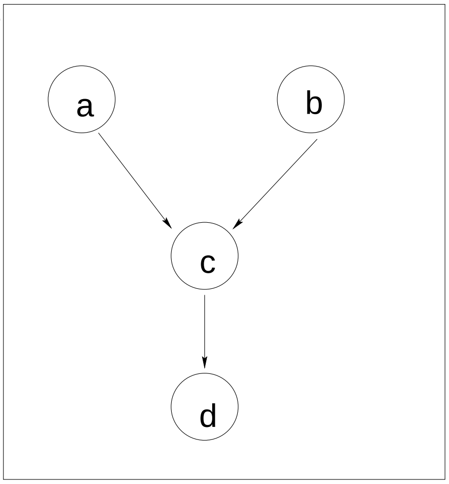
Outline:

- Convergence Study
- Graphical Models
- BUGS illustrations
- Bayesian Model Choice
- Reversible Jump
- Adaptive Methods

Directed acyclic graphs (DAG)

For example, the DAG- a graph in which all edges are directed, and there are no directed loops-

PSfrag replacements



expresses the natural factorisation of a joint distribution into factors each giving the joint distribution of a variable given its *parents*

$$\pi(a,b,c,d) = \pi(a)\pi(b)\pi(c|a,b)\pi(d|c)$$

The role of graphical models

- Graphical modelling provides a powerful language for specifying and understanding statistical models.

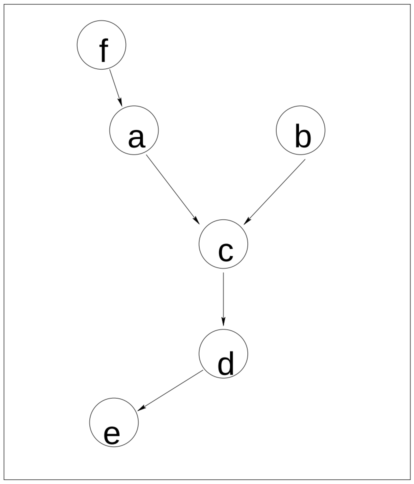
- Graphs consist of vertices representing variables, and edges (directed or otherwise) that express conditional dependence properties.

- For setting up MCMC, the graphical structure assists in identifying which terms need be in a full conditional.

Markov Property

PSfrag replacements

Variables are conditionally independent of their non-descendants, given their parents.



In the graph shown, c is conditionally independent of (f, e) given (a, b, d).

Find other conditional independence relationships yourself.

Since

$$\pi(x_v|x_{-v}) \propto \pi(x)$$

as a function of x_v,

$$\pi(x) = \prod_{v \in V} \pi\left(x_v | x_{\mathsf{pa}(v)}\right),$$

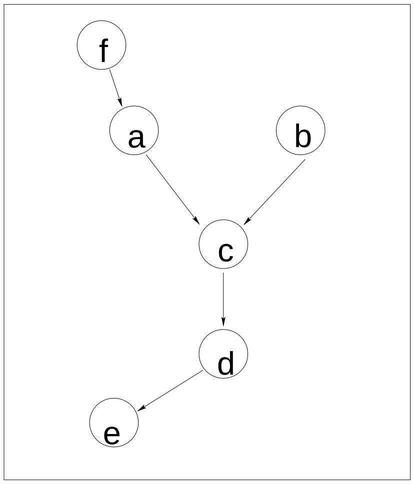
where $V =$ all possible vertices, implies

$$\begin{aligned} \pi(x_v|x_{-v}) &= \pi\left(x_v|x_{\mathsf{pa}(v)}\right) \times \\ & \qquad \textstyle\prod_{w:v\in \mathsf{pa}(w)} \pi\left(x_w|x_{\mathsf{pa}(w)}\right) \end{aligned}$$

pa=parents, $-v =$ everything but v in V.

That is, one term for the variable itself, and one for each of its children.

PSfrag replacements



In the above

$$\pi(c|\text{rest}) \propto \pi(c|a,b)\pi(d|c).$$

Leads us to...

Outline:

- Theoretical Convergence Study
- Graphical Models
- BUGS illustrations
- Bayesian Model Choice
- Reversible Jump
- Adaptive Methods

BUGS

Bayesian inference Using Gibbs Sampling is a general purpose computer software for doing MCMC.

- It exploits the structure of the graphical models to setup the Gibbs sampler.
- It is (still!) freely available from
- www.mrc-bsu.cam.ac.uk

We will go through two non-trivial examples.

(Sorry, no introduction to BUGS).

♡ Example: Regression models with fat tails and skewness.

Assume

$$y_i = \alpha + \beta(x_i - \bar{x}) + \delta z_i + \epsilon_i.$$

and

- Specification for ϵ.

  1. $\epsilon_i \sim N(0, \sigma^2)$
  2. $\epsilon_i \sim t_{\nu+1}(0, \sigma^2)$

- Specification for z.

  1. $z_i \sim N(0, 1) I(z_i > 0)$
  2. $z_i \sim t_{\nu}(0, 1) I(z_i > 0)$

- δ, α and β are unrestricted.
- σ^2 and ν are given appropriate prior distributions.

All details including a recent technical report, computer programs are available from my home-page.

www.maths.soton.ac.uk/staff/Sahu/utrecht

Please make your own notes for BUGS in here.

Basic 5 steps:

- check model
- load data
- compile
- load inits
- gen inits

Then update (from the Model menu) and Samples (from the Inference menu).

$\heartsuit$ Example: A Bayesian test for determining the number of components in mixtures.

Develop further the ideas of Mengersen and Robert (1996).

Let

$$d(f,g) = \int \log \frac{f(x)}{g(x)} f(x) dx$$

be the Kullback-Leibler distance between f and g.

Let

$$f^{(k)}(x) = \sum_{j=1}^{k} w_j f_j(x|\mu_j).$$

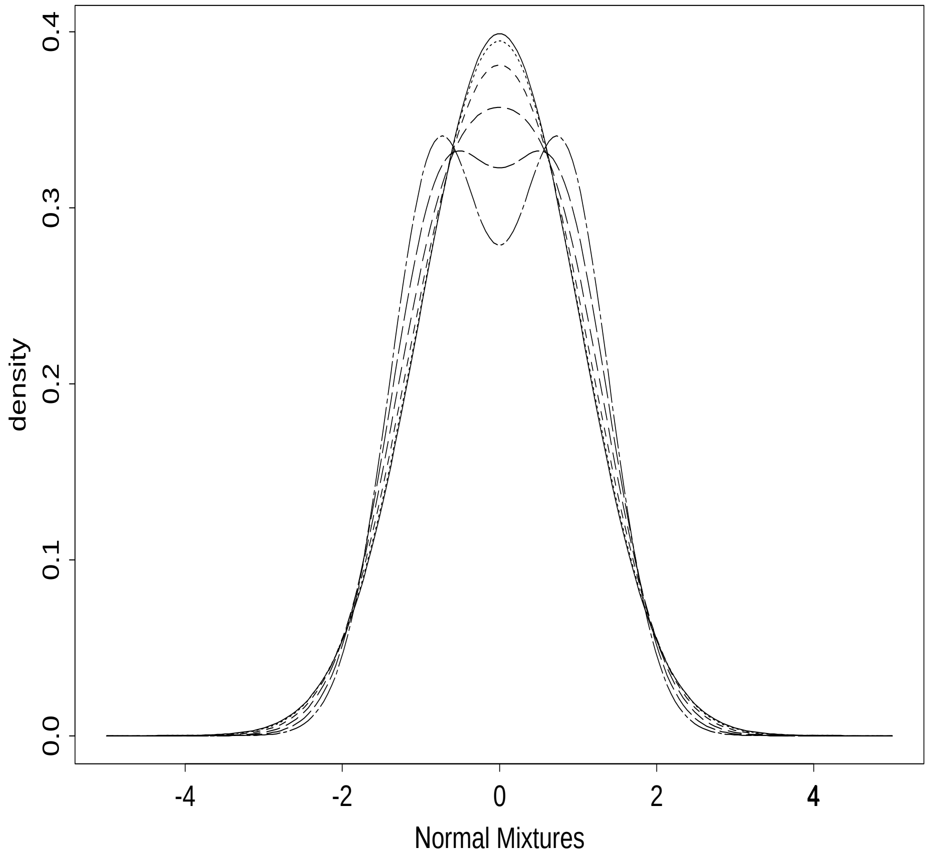
We develop an easy to use approximation for $d(f^{(k)}, f^{(k-1)})$.

The Bayesian test evaluates the posterior probability

$$P_c(k) = Pr\{d(f^{(k)}, f^{(k-1)}) \leq c| \text{ data } \}.$$

A simple decision rule is to select the $k-1$ component model if the above posterior probability is high, (greater than 0.5, for example).

Again the paper and computer programs are available from my home-page.



How to choose c?

Height at 0 decreases as distance increases. Five normal mixtures corresponding to distance = 0.1, 0.2, 0.3, 0.4 and 0.5.

Outline:

- Convergence Study
- Graphical Models
- BUGS illustrations.
- Bayesian Model Choice
- Reversible Jump
- Adaptive Methods

Will discuss

1. A very simple method: Cross-validation
2. A very complex method: Reversible Jump

Assume: $Y_1, \ldots, Y_n \sim i.i.d. f(y|\theta)$. Parameter θ.

Examine the influence of y_j to the ‘fit’.

Done by using the cross-validatory predictive distribution of y_j.

$$f(y_j|y_{-j}) = \int f(y_j|\theta, y_{-j})\pi(\theta|y_{-j})d\theta$$

This is called the conditional predictive ordinate or CPO. Can calculate residuals

$$y_j - \mathbb{E}(y_j|y_{-j}).$$

How to estimate the CPO? Gelfand and Dey (1994).

Can be shown:

$$f(y_j|y_{-j}) = \left\{ \mathbb{E}_{\theta|y} \left[\frac{1}{f(y_j|y_{-j}, \theta)} \right] \right\}^{-1}.$$

Let $\left\{ \theta^{(t)} \right\}$ be MCMC samples from the posterior. Then:

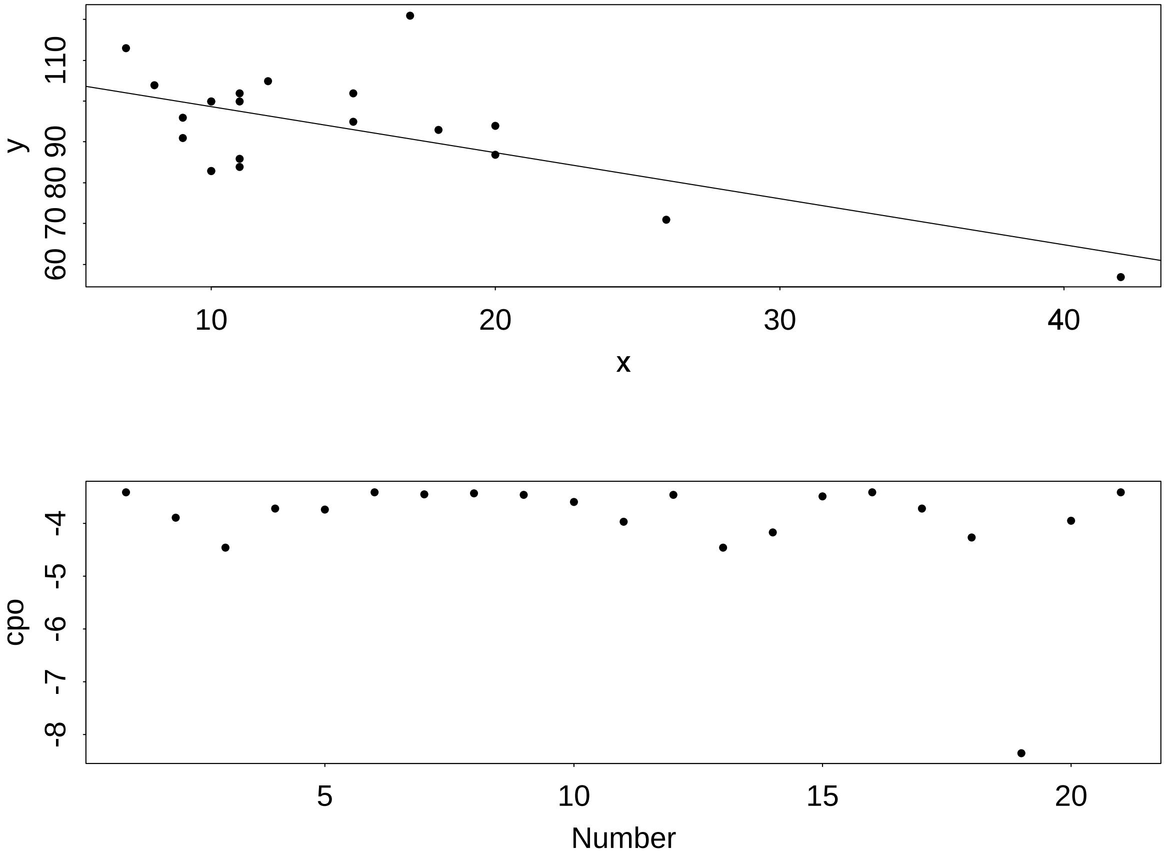
$$\hat{f}(y_j|y_{-j}) = \left\{ \frac{1}{N-M} \sum_{t=M+1}^{N} \frac{1}{f(y_j|\theta^{(t)})} \right\}^{-1}$$

for conditionally independent data.

Plot the CPO's against observation number. Then detect outliers etc for model checking.

♡ Example: Adaptive score data, Cook and Weisberg (1994).

Simple Linear regression.



The log(cpo) plot indicates one outliers.

The psedo-Bayes factor

$$\text{PsBF} = \frac{\prod_j f(y_j|y_{-j}, M_1)}{\prod_j f(y_j|y_{-j}, M_2)}$$

is a variant of the

$$\text{Bayes factor} = \frac{\text{Marginal Likelihood under } M_1}{\text{Marginal Likelihood under } M_2}.$$

There are many other excellent methods available to calculate the Bayes factor, see DiCicio *et al.* (1997) for a review.

We next turn to the second, more complex method.

Outline:

- Convergence Study
- Graphical Models
- BUGS illustrations.
- Bayesian Model Choice
- Reversible Jump
- Adaptive Methods

Reversibility

A Markov chain with transition kernel $P(y|x)$ and invariant distribution $\pi(x)$ is reversible if:

$$\pi(x)P(y|x) = \pi(y)P(x|y).$$

- This is a sufficient condition for $\pi(x)$ to be the invariant distribution of $P(\cdot|\cdot)$.
- Reversible chains are easier to analyse.
- All eigenvalues of P are real.
- Reversibility is not a necessary condition for MCMC to work.
- The Gibbs sampler when updated in a deterministic order is not reversible.

Jump

Green (1995) extended the Metropolis-Hastings algorithm for varying dimensional state space.

Let $k \in \mathbb{Z}$. Conditional on k, the state space is assumed to be n_k dimensional.

Recall that the acceptance ratio is:

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

Now require that the numerator and denominator have densities with respect to a common dominating measure (“dimension-balancing”). That is, a dominating measure for the joint distribution of the current state (in equilibrium) and next one.

We need to find a bijection

$$(x, u) \leftrightarrow (y, v)$$

where u and v are random numbers of appropriate dimensions so that

dim$(x, u) =$ dim(y, v) and $(y, v) = T(x, u)$

where the transformation is one-to-one. Now the acceptance probability becomes:

$$\alpha \left\{(x, u), (y, v)\right\} =$$

$$\min\left\{1, \frac{\pi(y)g_2(v)}{\pi(x)g_1(u)}\left|\frac{\partial T(x, u)}{\partial(x, u)}\right|\right\}.$$

where $g_1(u)$ and $g_2(v)$ are the densities of u and v.

How can we understand this?

The usual M-H ratio:

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)} \frac{q(x|y)}{q(y|x)}\right\}. \tag{1}$$

Now reversible jump ratio:

$$\alpha\{(x, u), (y, v)\} =$$

$$\min\left\{1, \frac{\pi(y)g_2(v)}{\pi(x)g_1(u)} \left|\frac{\partial T(x, u)}{\partial(x, u)}\right|\right\}. \tag{2}$$

The ratios in the last expression matches with those of the first.

- The first ratio in (2) exactly comes from the first ratio in (1). Look at the arguments of α in both cases.

- According to (1) the second ratio in (2) should be

  $$\frac{q(x,u|y,v)}{q(y,v|x,u)} \equiv \frac{f(x,u)}{f(y,v)}.$$

  where $f(\cdot,\cdot)$ is the joint density.

- Recall that $(y,v) = T(x,u)$ and T is one-to-one.

- Hence the density ratio will be just the appropriate Jacobian.

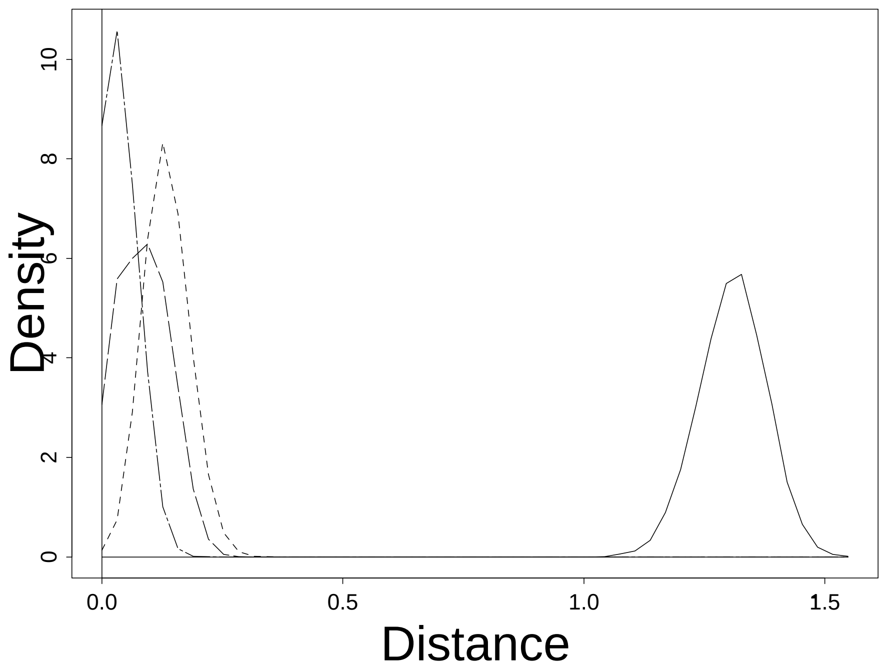
- To see this consider your favourite example of transformation and obtain the ratio of the original and transformed densities.

Peter Green's explanation is much more rigorous.

♡ Example: Enzyme data, Richardson and Green (1997) Probabilities (after the correction in 1998)

k	2	3	4	5	≥ 6
$\pi(k\|y)$	0.047	0.343	0.307	0.200	0.103

Our distance based approach:



Outline:

- Convergence Study
- Graphical Models
- BUGS illustrations.
- Bayesian Model Choice
- Reversible Jump
- Adaptive Methods

Problems with MCMC

- Slow convergence...
- Problem in estimating nse due to dependence.

A way forward is to consider Regeneration.

Regeneration points divide the Markov sample path in i.i.d. tours.

So the tour means can be used as independent batch means.

Using renewal theory and ratio estimation can approximate expectations and nse that are valid without quantifying Markov dependence.

Mykland, Tierney and Yu (1995), Robert (1995).

Regeneration also provides a framework for adaptation.

Background: Infinite adaptation leads to bias in the estimates of expectations. Gelfand and Sahu (1994).

But the bias disappears if adaptation is performed at regeneration points.

Gilks, Roberts and Sahu (1998) prove a theorem to this effect and illustrates.

How do you see regeneration points?

Nummelin's Splitting.

Suppose that the transition kernel $P(x, A)$ satisfies

$$P(x, A) \geq s(x)\nu(A)$$

where ν is a probability measure and s is a non-negative function such that $\int s(x)\pi(dx) > 0$.

Let

$$r(x, x') = \frac{s(x)\nu(dx')}{P(x, dx')} \leq 1,$$

Now, given a realisation $x^{(0)}, x^{(1)}, \ldots$ from P, construct conditionally independent 0/1 random variables $S^{(0)}, S^{(1)}, \ldots$ with

$$Pr(S^{(t)} = 1|\dots) = r\left(x^{(t)}, x^{(t+1)}\right)$$

From our experience, this is little hard to do in multi-dimension.

Sahu and Zhigljavsky (1998) provide an alternative MCMC sampler where regeneration points are automatically identified.

Let $\alpha(x) = \frac{1}{1+\kappa\, w(x)}$ where $w(x) = \frac{\pi(x)}{\psi(x)}$ and $\kappa > 0$ is a tuning constant.

Self Regenerative algorithm:

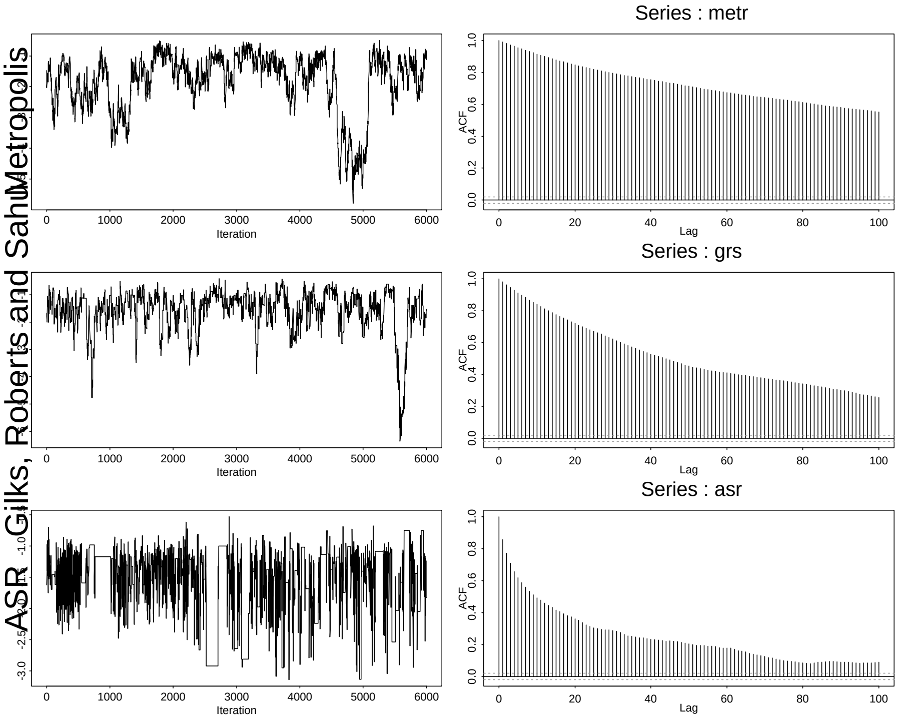
Let $n = 1$ and $m = 0$.

- Generate $Z^{(n)} \sim \psi$.
- Generate $\xi_n \sim$ Geometric$(\alpha(Z^{(n)}))$. If $\xi_n > 0$ set $X^{(m+j)} = Z^{(n)}$, for $j = 1, \ldots, \xi_n$ and $m = m + \xi_n$.
- If $n = N$ then stop, else set $n = n + 1$ and return to Step I.

Every time $\xi_n > 0$ is a regeneration point.

♡ Example: Sahu and Zhigljavsky (1998).

$$y_i = \beta_1 + \frac{\beta_2}{1 + \exp\{-\beta_4(x_i - \beta_3)\}} + \epsilon_i.$$



Time series and acf plots of β_4.

Strengths of MCMC

- Freedom in modelling
- Freedom in inference
- Opportunities for simultaneous inference
- Allows/encourages sensitivity analysis
- Model comparison/criticism/choice

Weaknesses of MCMC

- Order $N^{-1/2}$ precision
- Possibility of slow convergence

Ideas not talked about

Plenty!

- perfect simulation
- particle filtering
- Langevin type methodology
- Hybrid Monte Carlo
- MCMC methods for dynamically evolving data sets,

Definitely need more fool-proof automatic methods for large data analysis!

Online resources

Some links are available from:

www.maths.soton.ac.uk/staff/Sahu/utrecht

- MCMC preprint service (Cambridge)
- Debug: BUGS user group in Germany
- MRC Bio-statistics Unit Cambridge

The Debug site

http://userpage.ukbf.fu-berlin.de/ debug/

maintains a data base of references.