RSS-98
Strathclyde, September 1998

MCMC in statistics – interpretations and explanations

by Peter Green (University of Bristol,
P.J.Green@bristol.ac.uk).

(Not much on *how* to do it,
but *why* it is important and useful,
and some of *what* to worry about)

with thanks to....

Robert Aykroyd, Julian Besag, Arnoldo Frigessi, Paolo Giudici, Xiao-liang Han, Miles Harkness, David Higdon, Matthew Hodgson, Kerrie Mengersen, Antonietta Mira, Duncan Murdoch, Agostino Nobile, Marco Pievatolo, Sylvia Richardson, Claudia Tarantola, and Iain Weir.

Plan

- basics and a toy example
- variable dimensions: a change-point example
- measurement error in epidemiology
- perfect simulation

1. MCMC: the basic idea

Given a distribution of interest $\pi(\boldsymbol{\theta})$

(*in statistical analysis, more often than not a Bayesian posterior*)

- construct a Markov chain $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, whose limiting distribution is π,
- simulate it, and
- treat the realisation as a sample from π.

For example, estimate

$$\pi(\boldsymbol{\theta} \in A) \quad \text{by} \quad \#\{t \leq N : \boldsymbol{\theta}^{(t)} \in A\}/N$$

or

$$E_\pi(g(\boldsymbol{\theta})) \quad \text{by} \quad \frac{\sum_{t=1}^{N} g(\boldsymbol{\theta}^{(t)})}{N}$$

Toy example: Normal random sample

Data $Y_1, Y_2, \ldots, Y_n$ are a random sample from $N(\mu, \sigma^2)$.

Independent priors on μ and σ:

$$\begin{aligned} \mu &\sim N(\xi, \kappa^{-1}) \\ \sigma^{-2} &\sim \Gamma(\alpha, \beta) \end{aligned}$$

(these are only conditionally conjugate).

Joint posterior:

$$\begin{aligned} \pi(\mu, \sigma^2|Y) &\propto (\sigma^2)^{-\alpha-1-n/2} \\ &\times \exp\left\{-\frac{\beta}{\sigma^2} - \frac{\kappa(\mu-\xi)^2}{2} - \frac{\sum(Y_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

which is not of standard form.

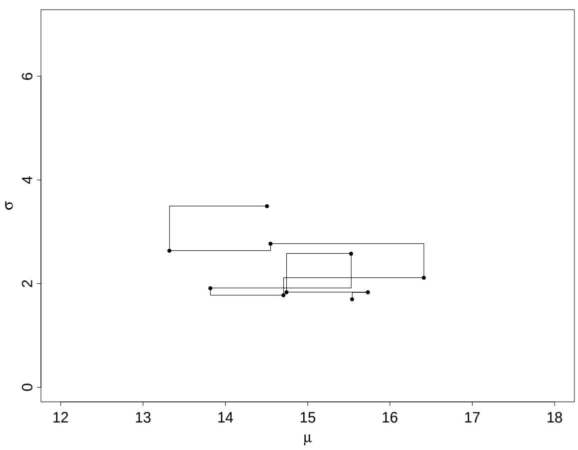
Normal random sample, continued...

Full conditionals are easily found:

$$\begin{aligned} \mu|\sigma, Y &\sim N\left(\frac{\sigma^{-2}\sum Y_i + \kappa\xi}{\sigma^{-2}n + \kappa}, \frac{1}{\sigma^{-2}n + \kappa}\right) \\ \sigma^{-2}|\mu, Y &\sim \Gamma(\alpha + n/2, \beta + \textstyle\sum(Y_i - \mu)^2/2) \end{aligned}$$

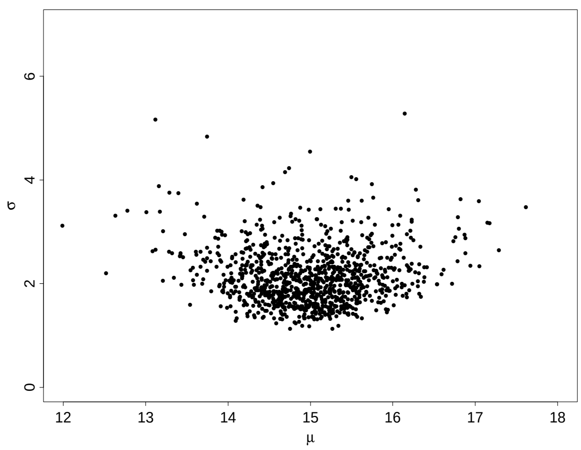
and we can implement the so-called *Gibbs sampler* by alternately drawing μ and σ^{-2} from these distributions.

Gibbs sample of size 10



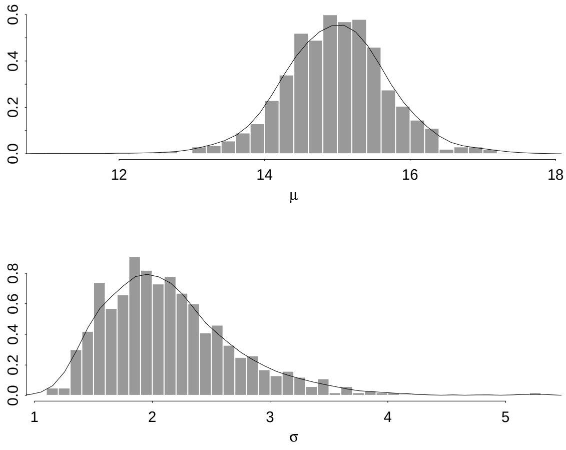
Posterior sample of (μ, σ) from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

Gibbs sample of size 1000



Posterior sample of (μ, σ) from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

Marginal distributions



Posterior distributions of μ and σ from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

Sample-based computation

allows complete freedom of inference.

For example, prediction: we can calculate $P\{Y_{n+1} > 19\}$ by averaging $1 - \Phi(\{19 - \mu\}/\sigma)$:

$$\frac{1}{N}\sum_{t=1}^{N}\left[1 - \Phi(\{19 - \mu^{(t)}\}/\sigma^{(t)})\right] \approx 0.045$$

Why does MCMC work?

MCMC is available even where static simulation isn't, because

- All simulation methods rely on the Law of Large Numbers, and this remains true (the Ergodic theorem) when you have a Markov chain instead of an i.i.d. sequence.
- If you don't mind Markov dependence, then you can update the parameters $\theta_1, \theta_2, \ldots, \theta_p$ one-by-one (or in small groups).

There are many recipes for doing this — the Gibbs sampler is just one of them.

The basic limit theorems

If $\{\boldsymbol{\theta}^{(t)}\}$ is an *irreducible* Markov chain with transition kernel P and *invariant distribution* π, and g is a real valued function with $\int |g(\boldsymbol{\theta})|\pi(d\boldsymbol{\theta}) < \infty$, then

$$\frac{1}{N}\sum_{t=1}^{N} g(\boldsymbol{\theta}^{(t)}) \to \int g(\boldsymbol{\theta})\pi(d\boldsymbol{\theta})$$

for π-almost all $\boldsymbol{\theta}^{(0)}$.

If the chain is also *aperiodic*, then there is convergence in total variation:

$$||P^t(\boldsymbol{\theta}^{(0)}, \cdot) - \pi(\cdot)|| \to 0$$

as $t \to \infty$, for π-almost all $\boldsymbol{\theta}^{(0)}$.

How fast?

Knowing that the chain converges is not the same as knowing that it converges quickly enough to be useful.

Qualitative results: Geometric and uniform ergodicity

The chain is *geometrically ergodic* if

$$||P^t(\boldsymbol{\theta}^{(0)}, \cdot) - \pi(\cdot)|| \leq M(\boldsymbol{\theta})\rho^t$$

for finite $M(\boldsymbol{\theta})$, $\rho < 1$.

It is *uniformly ergodic* if

$$||P^t(\boldsymbol{\theta}^{(0)}, \cdot) - \pi(\cdot)|| \leq M\rho^t.$$

Various conditions are known to imply uniform ergodicity, for example Doeblin's condition: there exists a probability measure ϕ and constants $\epsilon < 1$, $b > 0$, t such that

$$\phi(A) \geq \epsilon \Rightarrow P^t(\boldsymbol{\theta}, A) \geq b \quad \text{for all} \quad \boldsymbol{\theta}$$

Application to MCMC

There are both positive and negative results about uniform or geometric ergodicity of popular MCMC recipes: *e.g.* see papers by Kerrie Mengersen, Gareth Roberts, Jeff Rosenthal and Richard Tweedie.

Quantitative results

Only in very rare cases can numerical bounds be found for rates of convergence, and when they can, they are often very discouraging.

Eigenvalues can only rarely be found for interesting problems; there are interesting recent computational complexity results by Frigessi, Martinelli and Stander, but multiplying constants are intractable.

2. Markov chain Monte Carlo for variable dimension problems

What if the number of things you don't know is one of the things you don't know?

Examples:

- variable selection
- finite mixture estimation
- change-point analysis
- nonparametric hazard estimation
- automatic curve fitting (degree, discontinuities)
- ARIMA model fitting
- graphical model determination
- ion channel data
- quantitative trait locus analysis
- model-based CART
- image segmentation
- restoring old movies
- object recognition

Reversible jump MCMC

For complex problems, we need MCMC methods for a general measure space. We can adapt the Metropolis-Hastings recipe:

- *propose* a change from the current state $\boldsymbol{\theta}$ to a new one $\boldsymbol{\theta}'$, using a distribution of convenience
- *accept* or *reject* the proposal with probabilities determined by the two states and the desired target

For this to work, we need the equilibrium joint distribution of current and proposed next states to have a density (Radon-Nikodym derivative) $f(\boldsymbol{\theta}, \boldsymbol{\theta}')$ with respect to *any* symmetric dominating measure – the acceptance probability is then just

$$\min\left\{1, \frac{f(\boldsymbol{\theta}', \boldsymbol{\theta})}{f(\boldsymbol{\theta}, \boldsymbol{\theta}')}\right\}$$

Moves between two simple subspaces

In most cases, we can automate this by *modelling the program* itself.

When in state $\boldsymbol{\theta}$, we generate random numbers u with density $q(u)$, and set the proposed new state $\boldsymbol{\theta}'$ to a deterministic function $\boldsymbol{\theta}'(\boldsymbol{\theta}, u)$. Similarly in reverse: $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\theta}', u')$.

Providing $(\boldsymbol{\theta}, u) \leftrightarrow (\boldsymbol{\theta}', u')$ is a bijection, the acceptance probability is

$$\alpha = \min\left\{1, \frac{p(y|\boldsymbol{\theta}')}{p(y|\boldsymbol{\theta})} \frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta})} \frac{q(u')}{q(u)} \left| \frac{\partial(\boldsymbol{\theta}', u')}{\partial(\boldsymbol{\theta}, u)} \right| \right\},$$

that is, the product of likelihood, prior and proposal ratios and a Jacobian.

Multiple change-point analysis

A Bayesian approach to change-point analysis for point processes $(y_1, y_2, \ldots, y_N)$: combines

- Poisson-process likelihood: $p(y|x) \propto \exp\{\sum_{i=1}^N \log x(y_i) - \int_0^L x(t)dt\}$
- prior model for step function $x(t)$, $0 \le t < L$, representing intensity

Prior model: represent step function by $(k, \{s_j\}_{j=1}^k, \{h_j\}_{j=0}^k)$: $x(t) = \sum_j h_j I_{[s_j, s_{j+1})}(t)$:

- number of steps k: Poisson(λ),
- step heights h_j: Gamma(α, β),
- step positions s_j: $p(s|k) \propto \prod_j (s_{j+1} - s_j)$,

all independent.

MCMC for step functions

I will use four moves:

(a) Metropolis change to a randomly chosen step height h_j.

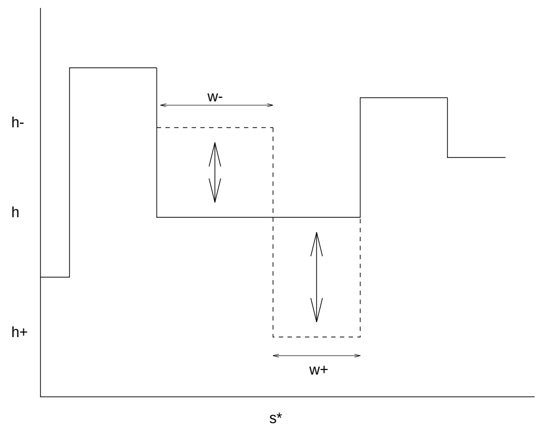
(b) Metropolis change to a randomly chosen step position s_j.

(c) Jump move: birth/death of steps

- birth: choose new step position s^* at random, split current step height h into two: (h_-, h_+)
- death: choose step at random to kill, combine current step heights (h_-, h_+) into one: h

(d) Update hyperparameters α, β

Birth and death of steps

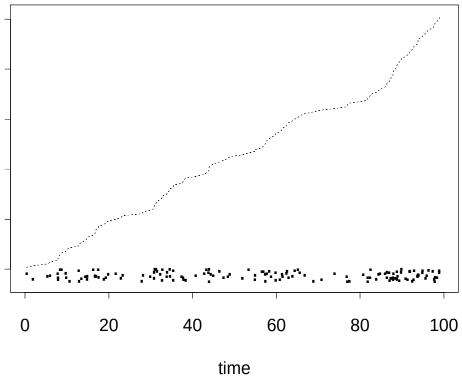


$$h_-^{w_-} h_+^{w_+} = h^{w_- + w_+}$$

$$(h, w, s^\star, u) \leftrightarrow (h_-, h_+, w_-, w_+)$$

Example:
cyclones hitting the Bay of Bengal

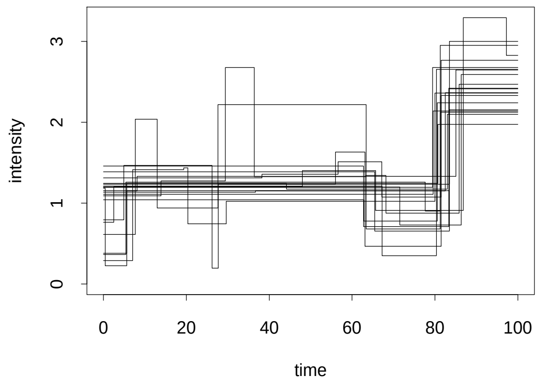
141 cyclones over a period of 100 years
(a cyclone is a storm with winds > 88 km h^{-1}).



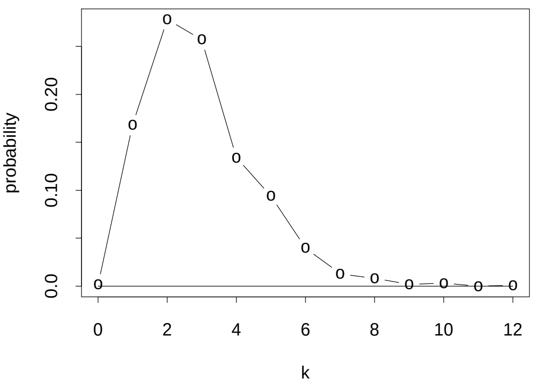
Choices of hyperparameters:

- Prior on k: Poisson(λ), with $\lambda = 3$.
- Prior on h_j: Gamma(α,β), with
  – $\alpha \sim \Gamma(2, 2)$
  – $\beta \sim \Gamma(1, N/L)$

Sample of step functions from the posterior:

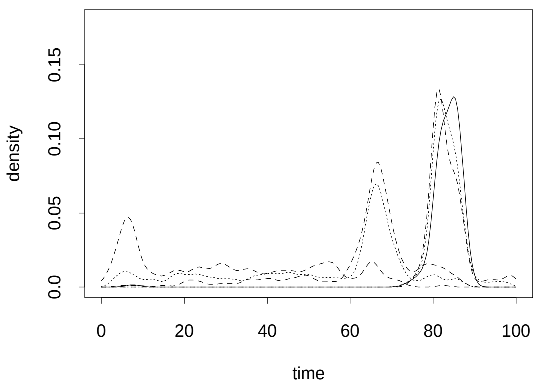


Posterior for the number of change points k

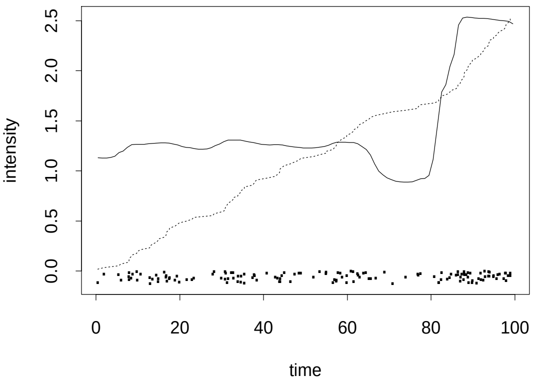


Zero change points is ruled out; $k = 1$ or 2 more probable than under the prior.

Posterior density estimates for change-point positions

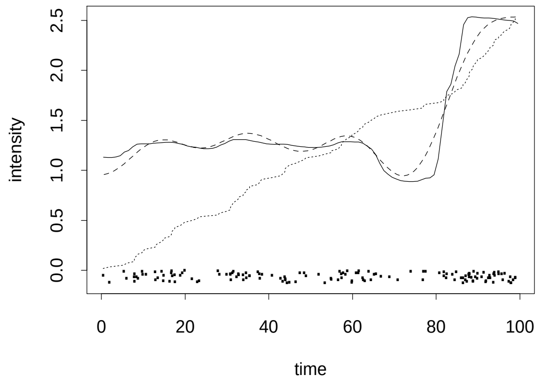


Model-averaged estimate: $E(x(\cdot)|y)$



(the expectation of a random step function is not a step function).

Ordinary smoothing methods (in this case a kernel smoother) can't match that mean curve



– fixed-bandwidth smoothers either over-smooth the steps, or under-smooth the plateaux.

Introducing a cyclic component

(As further illustration of the flexibility in modelling allowed by MCMC)

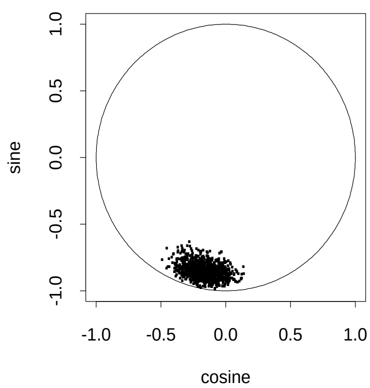
The simplest way to model periodicity is to take a simple sinusoid: assume data generated from a Poisson process with instantaneous rate

$$x(t)\left\{1+\gamma\cos(2\pi f t)+\delta\sin(2\pi f t)\right\},$$

where $x(t)$ is the step function defined above, and f denotes the assumed (known) frequency.

If *a priori* (γ,δ) are taken uniform on the unit disc, then a simultaneous Metropolis update is easily implemented.

Sample from posterior of (γ,δ)



3. Bayesian measurement error modelling

with Sylvia Richardson (INSERM, Paris)

Aim: to quantify the association between an outcome Y and a set of covariates X
where covariates are imperfectly observed and only measured through "surrogates".

Ignoring measurement error and treating the surrogate as the true covariate may produce biased results.

Why be Bayesian here?

- latent covariates with imprecisely specified prior distributions
- combining information on measurement process from several sources
- propagating uncertainty

Model building – structural specifications

- Y known outcome
- X true (latent) covariate
- U observed surrogate for X
- C known covariates

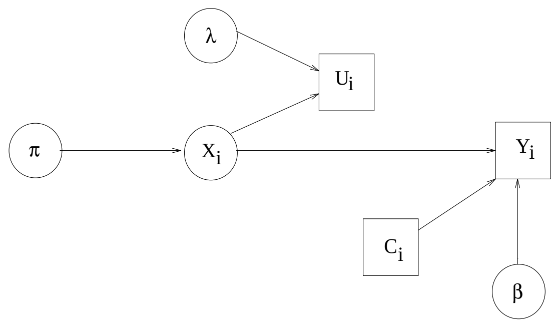
Formulation of local submodels between components using
– conditional independence assumptions
– prior information on the structure of the measurement process

Submodels:

- $p(Y|X, C, \beta)$ regression model
- $p(U|X, \lambda)$ measurement model
- $p(X|\pi)$ prior model

Bayesian analysis using graphical models

Non differential measurement error assumption:
$Y \perp U|X$



Joint distribution:

$$p(\beta)p(\lambda)p(\pi)\prod_i p(X_i|\pi)$$

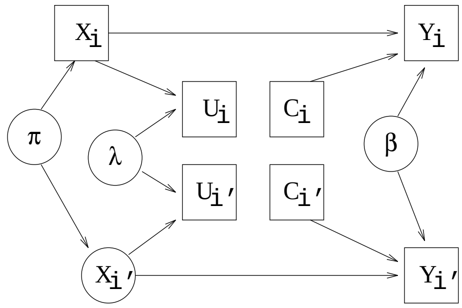
$$\times \prod_i p(U_i|X_i, \lambda) \prod_i p(Y_i|X_i, C_i, \beta)$$

Where does quantitative information on measurement model come from ?

One possibility: design with a validation group:

reference method which can be used to get information on X from a subgroup where both X and U are recorded.

Designs with a validation group



- transfer of information on λ from the validation group to the main study
- strengthens inference about regression parameters β

Problems in specifying prior for $p(X|\pi)$

Some approaches

- pseudo-likelihood *(Carroll, 1993)* based on plugging an empirical estimate of $p(X|\pi)$ based on the validation subgroup
- non parametric modelling of $p(X|\pi)$ via NPML *(Roeder, Carroll, Lindsay, JASA 1996)*
- joint modelling of $p(X, U|\lambda)$ as a Multivariate normal where λ specified in terms of a Dirichlet Process *(Müller and Roeder, Biometrika, 1997)*
- semi-parametric model for $p(X|\pi)$ via a mixture of gaussian distributions with an unknown number of components

Mixture model for $p(X|\pi)$

$$X_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \quad \text{independently for } i = 1, 2, \ldots, n$$

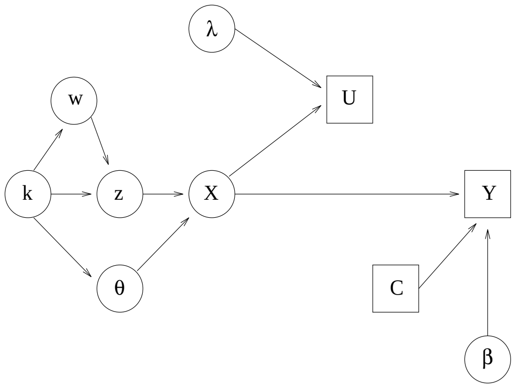
$f(\cdot|\theta)$ is a given parametric family
$\{\theta_j\}, \{w_j\}, k$ **unknown**

The model can be formulated using latent allocation variables:

$$p(z_i = j) = w_j \quad \text{independently for } i = 1, 2, \ldots, n$$

$$X_i|z \sim f(\cdot|\theta_{z_i}) \quad \text{independently for } i = 1, 2, \ldots, n$$

Measurement error model with mixture prior



Of course, computing in such models would be quite impossible by conventional methods.

With MCMC, most of the variables can be updated singly or in small groups, by Gibbs or Metropolis moves.

We update k (with consequent changes to w, z and θ) by reversible jump *split/merge* moves.

Implementation in the case of a logistic regression with validation group design

Prior for X: normal mixture model

$$X \sim \sum_{j=1}^{k} w_j \phi(\cdot|\mu_j, \sigma_j^2), \quad k \text{ unknown}$$

Measurement error: e.g., lognormal

$$\log U_i \sim N(\alpha_0 + \alpha_1 \log X_i, \lambda^{-1})$$

Regression model for disease status: logistic model

$$Y \sim \text{Bernoulli}(\{1 + \exp[-\beta^T(X, C)]\}^{-1})$$

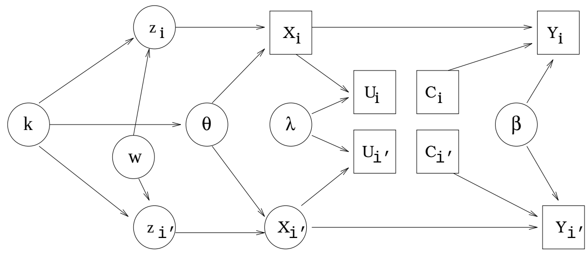


Illustration on a study of the risk of coronary heart disease (CHD) as a function of blood cholesterol

Total cholesterol (TC) and Low density cholesterol (LDL) on 256 subjects: 113 cases, 143 controls.

$\rightarrow$ can we use TC = U as a surrogate for LDL = X?
$\rightarrow$ a validation subgroup with 32 cases and 40 controls is chosen at random

Logistic regressions of CHD on cholesterol level:

- regression on X, complete data set ($n = 256$)

$$\beta_1 = 0.66 \quad (0.34)$$

- regression on U, complete data set ($n = 256$)

$$\beta_1 = 0.54 \quad (0.31)$$

- regression on X, validation group ($n = 72$)

$$\beta_1 = 0.93 \quad (0.68)$$

- Bayesian analysis (validation and main study)

$$\beta_1 = 0.62 \quad (0.44)$$

Performance of mixture priors in measurement error models

What is influence of misspecifying a gaussian prior for X?

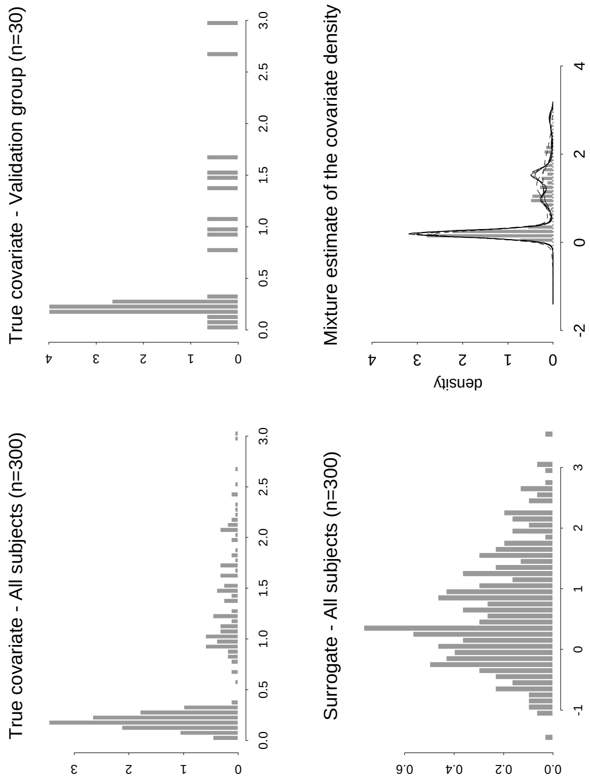
Simulation set up: 50 replications
270 subjects in main study, 30 in validation group.
X drawn from an asymmetric normal mixture :

$$0.6N(0.19, (0.08)^2) + 0.2N(1.05, (0.2)^2) + 0.2N(1.63, (0.48)^2)$$

Measurement model : $U \sim N(X, \lambda^{-1})$

Logistic disease model :
logit $P(Y = 1|X) = \beta_0 + \beta_1 X$

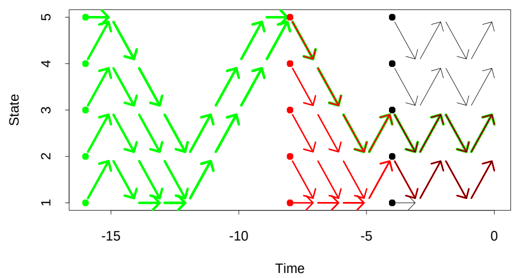
		analysis	
	true	mixture prior	gaussian prior
$\overline{\lambda}$	3	2.82 (0.41)	2.37 (0.51)
$\overline{\beta_0}$	-0.8	-0.86 (0.19)	-1.05 (0.27)
$\overline{\beta_1}$	0.4	0.52 (0.25)	0.76 (0.32)
mse(β_1)		0.053	0.092



4. Coupling from the Past (CFTP)

Here is a beautiful idea due to Propp and Wilson (1995): a way of organising a Markov chain simulation so that after a finite but random amount of work, it *exactly* delivers a sample from the target distribution!

For example, look at this partial simulation of a symmetric random walk with reflecting barriers:



You can see that *for these random numbers*, **all** paths of the chain starting at time $-\infty$ have the same state (3) at time 0. $\boldsymbol{\theta}^{(0)}$ must be drawn from π!

Generally, imagine the process run from all initial states in the indefinite past, and look at the state at time 0, $\boldsymbol{\theta}^{(0)}$. If this is unique, then $\boldsymbol{\theta}^{(0)} \sim \pi$.

So, **if there exists a (random) initial time $-T$ such that for all initial states X_{-T}, $\boldsymbol{\theta}^{(0)}$ is the same then $\boldsymbol{\theta}^{(0)} \sim \pi$.**

We don't need to find T exactly, since coalescence occurs from all initial times $< -T$. So we just try $T = 1, 2, 4, 8, \ldots$

Is CFTP any use?

There have been some spectacular successes in finding CFTP implementations for certain models in statistical physics and spatial processes possessing a lot of symmetry, even with huge numbers of variables (4 million in one case).

But it seems much harder to make it work for even quite low-dimensional continuous distributions without symmetry.

The Rejection Coupler

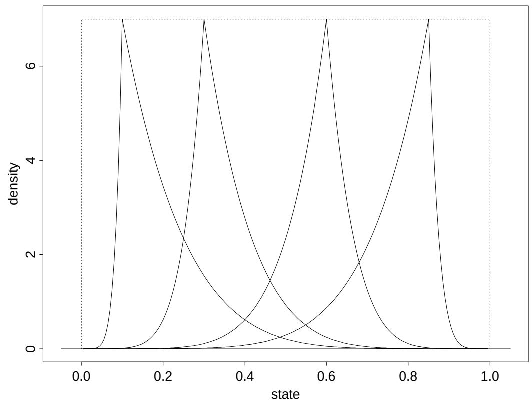
(work with Duncan Murdoch (Western Ontario))

Here is a simple approach to CFTP for a continuous state space.

Suppose we know $f(y|x)$, where

$$P\{X_{t+1} \leq y | X_t = x\} \propto \int_{-\infty}^{y} f(u|x)du,$$

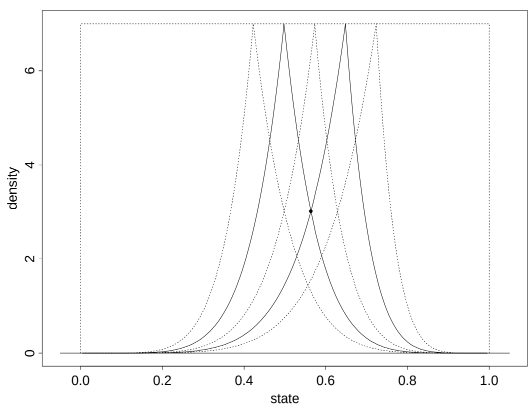
and that the (not necessarily normalised) densities $f(y|x)$ are bounded above by an integrable $h(y)$.



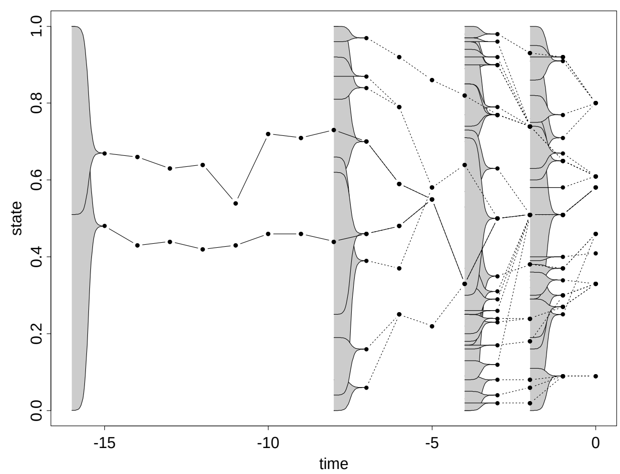
Toy example:
$f(y|x) = (\delta + 1)\min(\{y/x\}^{\delta}, \{(1-y)/(1-x)\}^{\delta})$
with $\delta = 6$

Rejection sampler: To sample from the (not necessarily normalised) $f(\cdot|x)$, draw (Y, Z) uniformly under the graph of h until $Z < f(Y|x)$.

Rejection coupler: To sample from $f(\cdot|x)$ for all x, draw (Y_i, Z_i) uniformly under the graph of h. Let $A_i = \{x : Z_i < f(Y_i|x)\}$; then Y_i is a valid update for all $x \in A_i$. Continue until $\cup_i A_i = \chi$.



Example realisation of rejection coupler for
$f(y|x) = (\delta + 1)\min(\{y/x\}^{\delta}, \{(1-y)/(1-x)\}^{\delta})$
with $\delta = 6$.



Towards methods that might be useful in Bayesian statistics

Our current most promising approach is more generic: we can couple the random walk Metropolis sampler.

It is promising for adaptation to CFTP

- we can separately couple the proposals and the acceptance decisions
- the difficult continuous part can be done once and for all, in a model-independent way, for a proposal distribution chosen at our convenience.

Some strengths of MCMC

- Freedom in modelling
  - in principle, no limits
  - well-adapted for models defined on sparse graphs
- Freedom in inference
  - in principle, no limits
  - can address questions only posed after simulation completed (e.g. ranking and selection)
  - opportunities for simultaneous inference
- Allows/encourages sensitivity analysis
- Model comparison/criticism/choice
- Coherently integrates uncertainty
- Only available method for complex problems

Some weaknesses and dangers

- Order $\sqrt{N}$ precision
- Possibility of slow convergence, especially when not diagnosable (meta-stability)
- Risk that fitting technology runs ahead of statistical science
- Risk of undisciplined, selective presentation
- Difficulty of validating code

Some important lines of continuing research

- Adaptive methods, and other possibilities for automation
- Perfect simulation: useful in practice?
- Getting quantitative results from theoretical analysis
- Learn (even) more from physics

To follow up

My web page:

`http://www.stats.bris.ac.uk/~peter`

My email address:

`P.J.Green@bristol.ac.uk`