HSSS Summer School on MCMC
Rebild, Denmark, August 1998

Tutorial lectures on Markov chain Monte Carlo

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

- more about graphical modelling
- some spatial examples
- variable dimension MCMC
- applications to model determination, change point analysis and mixtures

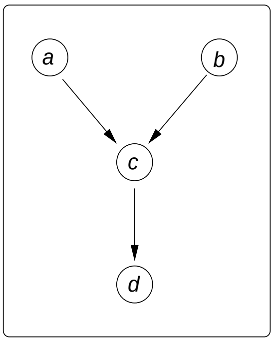
Graphical models and connections to MCMC

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.

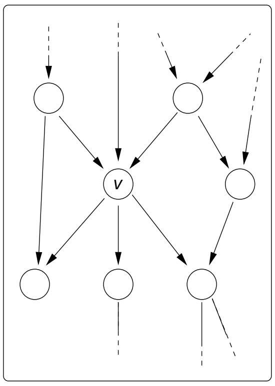
Directed acyclic graphs

For example, the DAG (directed acyclic graph) – a graph in which all edges are directed, and there are no directed loops –



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

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c)$$



In general

$$p(x) = \prod_{v \in V} p(x_v | x_{\mathrm{pa}(v)})$$

which in turn implies a *Markov property*, that variables are conditionally independent of their non-descendants, given their parents.

From the perspective of setting up MCMC methods, the graphical structure assists in identifying which terms need be included in a full conditional.

Since

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

as a function of x_v,

$$p(x) = \prod_{v \in V} p(x_v|x_{\mathrm{pa}(v)})$$

implies

$$p(x_v|x_{-v}) \propto p(x_v|x_{\mathrm{pa}(v)}) \prod_{w: v \in \mathrm{pa}(w)} p(x_w|x_{\mathrm{pa}(w)})$$

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

Graphical modelling, the *construction* of MCMC methods through full conditional distributions, and *good practice* in statistical model building all exploit the same modular structure.

Trivial 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:

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

which is not of standard form.

Normal random sample

$$\pi(\mu|\sigma^2, Y) \propto \pi_0(\mu) f(Y|\mu, \sigma^2)$$

$$= \sqrt{\frac{\kappa}{2\pi}} e^{-\kappa(\mu-\xi)^2/2} \frac{1}{(2\pi)^{n/2}\sigma^n} \prod_{i=1}^{n} e^{-(Y_i-\mu)^2/2\sigma^2}$$

$$\propto \sigma^{-n} \exp\left\{-\frac{\kappa(\mu-\xi)^2}{2} - \frac{\sum(Y_i-\mu)^2}{2\sigma^2}\right\},$$

whence by completing the square:

$$\mu|\sigma, Y \sim N\left(\frac{\sigma^{-2}\sum Y_i + \kappa\xi}{\sigma^{-2}n + \kappa}, \frac{1}{\sigma^{-2}n + \kappa}\right).$$

Similarly,

$$\pi(\sigma^2|\mu, Y) \propto \pi_0(\sigma^2) f(Y|\mu, \sigma^2)$$

$$\propto (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2} \frac{1}{(2\pi)^{n/2}\sigma^n} \prod_{i=1}^{n} e^{-(Y_i-\mu)^2/2\sigma^2}$$

$$\propto (\sigma^2)^{-\alpha-n/2-1} \exp[-\{\beta + \sum(Y_i-\mu)^2/2\}/\sigma^2],$$

that is,

$$\sigma^{-2}|\mu, Y \sim \Gamma(\alpha + n/2, \beta + \sum(Y_i-\mu)^2/2)$$

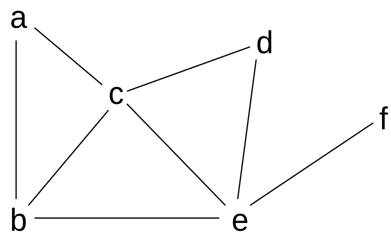
and we can implement a Gibbs sampler by alternately drawing μ and σ^{-2} from these distributions.

More on graphical modelling

Directed acyclic graphs are a natural representation of the way we **usually** *specify* a statistical model (directionally, disease $\rightarrow$ symptom, past $\rightarrow$ future, parameters $\rightarrow$ data), but

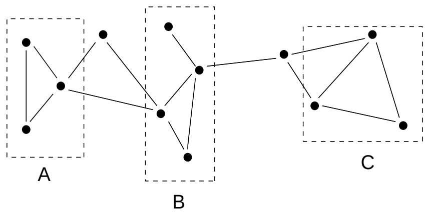
- sometimes (e.g. spatial models) there is *no natural direction*;
- in *understanding associations* between variables implied by a model, however specified, directions confuse;
- and these associations represent the full conditionals needed in *setting up MCMC* methods.

Conditional independence graph: draw an (undirected) edge between variables a and b if they are **not** conditionally independent given all other variables.



$P : c \perp f|(a,b,d,e) \qquad L : d \perp (a,b,f)|(c,e)$

$$F \quad : \quad \pi(a,b,c,d,e,f) = \psi_1(a,b,c)\psi_2(b,c,e)\psi_3(c,d,e)\psi_4(e,f)$$



$G : A \perp C|B$

Markov properties

The Markov property is familiar from temporal stochastic processes, where we learn that it may be expressed in several equivalent ways. For variables located on an arbitrary graph, the situation is more subtle: can distinguish 4 related properties, each capturing an aspect of Markovness.

Pairwise Non-adjacent pairs of variables are conditionally independent given the rest (see definition of graph).

Local Conditional only on adjacent variables (neighbours), each variable is independent of all others (this simplifies full conditionals).

Global Any two subsets of variables separated by a third are conditionally independent given the values of the third subset.

Factorisation The joint distribution factorises as a product of functions on cliques (=maximal complete subgraphs).

Mathematically, the interesting thing is that these are different, although $F \Rightarrow G \Rightarrow L \Rightarrow P$ always.

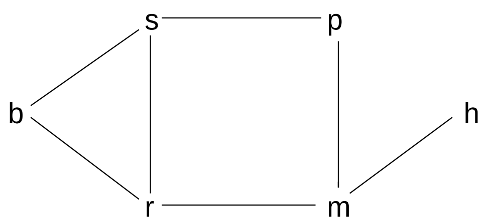
But for most statistical purposes, the important thing is that they are often the same; a sufficient but not necessary condition is that the joint distribution has the positivity property ("any values realisable individually are realisable jointly").

This result includes the Hammersley-Clifford theorem (Markov random field = Gibbs distribution, $L = F$).

For directed acyclic graphs, the situation is simpler: the directed local Markov property is *always* equivalent to the directed graph factorisation criterion: $DL = DF$.

Example: using CI graphs to summarise dependencies

Edwards and Havranek (Biometrics, 1985) summarise their conclusions from a 2^6 contingency table like this:



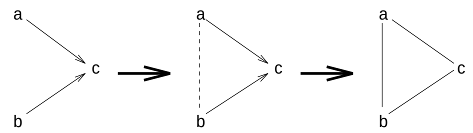
b	blood pressure > 140?
s	smoking?
r	ratio of α and β lipoproteins > 3?
p	strenuous physical work?
m	strenuous mental work?
h	family history of coronary heart disease?

Reading off all conditional dependencies from a model specified as a DAG

(to understand associations, or to help construct a MCMC method)

Need to "moralise" the graph:

1. Marry parents
2. Drop directions



So the neighbours of a variable are its parents, children and spouses.

Caveat: this *may* introduce superfluous dependencies, e.g.

$$c \sim \mathrm{Geom}(a) + \mathrm{Beta}(b, b)$$

Modelling directly with an undirected graph

With a DAG, *any* set of conditional distributions $\pi(v|\mathrm{parents}(v))$ combine to form a consistent joint distribution (because of the acyclicity).

But in an undirected graph, we need consistency conditions on the full conditionals $\pi(v|\mathrm{rest})(= \pi(v|\mathrm{neighbours}(v))$ since $P = L$).

The only safe strategy is to use property F, to model the joint distribution as a product of functions on cliques

$$\pi(V) = \prod_C \psi_C(v_C)$$

which implies

$$\pi(v|\mathrm{rest}) = \prod_{C: v \in C} \psi_C(v_C)$$

Chain graphs

Hybrid modelling: some directed and some undirected edges.

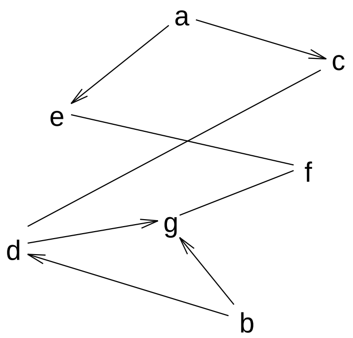
If there are no one-way cycles, graph can be arranged to form a DAG whose nodes are blocks Δ_t of variables connected only by undirected edges: we call this a chain graph.

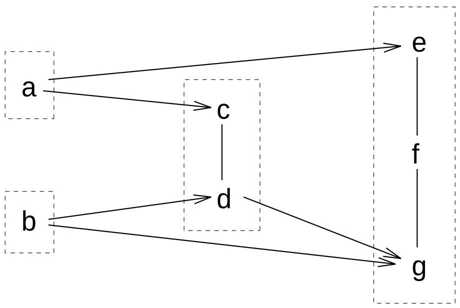
Model specification is by the appropriate combination of the two approaches; this builds a joint distribution

$$\begin{aligned} \pi(V) &= \prod_t \pi(\Delta_t|\mathrm{parents}(\Delta_t)) \\ &= \prod_t \prod_{C \in \mathcal{C}_t} \psi_C(v_C) \end{aligned}$$

where $\mathcal{C}_t$ are the cliques in an undirected graph with nodes $(\Delta_t, \mathrm{parents}(\Delta_t))$ and undirected edges consisting of (a) those already in Δ_t, (b) the links between Δ_t and parents, with directions dropped, and (c) links between all members of $\mathrm{parents}(\Delta_t)$.

Manipulating the graph into block-DAG form:





Full conditionals are constructed as usual, by reading off those factors in the product defining the joint distribution that refer to the variable in question:

$$\begin{aligned}\pi(v|\text{rest}) &\propto \pi(v|\text{neighbours}(v), \text{parents}(v)) \\ &\times \prod_t \pi(\Delta_{t:\text{child}(v)}|\text{parents}(\Delta_t))\end{aligned}$$

where $\Delta_{t:\text{child}(v)}$ are the children of v in block Δ_t.

Prostate cancer, and logistic regression with correlated random effects

Holford (*Biometrics*, 1983) data on mortality from prostate cancer among non-white US males, classified by

- 7 age groups: 50–54, …, 80–84
- 7 time periods: 1935–39, …, 1965–69

Let $i = 1, 2, \ldots, I$ denote age group; $j = 1, 2, \ldots, J$ time period; and $k = [ij] = I - i + j = 1, 2, \ldots, K$ the corresponding *cohort*.
Number at risk: n_{ij}
Deaths from prostate cancer: $y_{ij} \sim \text{Bin}(n_{ij}, p_{ij})$.

Logistic regression with age, period and cohort effects is inadequate (see also Breslow, *Appl. Stat.*, 1984), so allow for additional unobserved covariates:

$$\xi_{ij} = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mu + \theta_i + \phi_j + \psi_{[ij]} + z_{ij}$$

Priors

Use prior for θ reflecting anticipated smoothness with time:

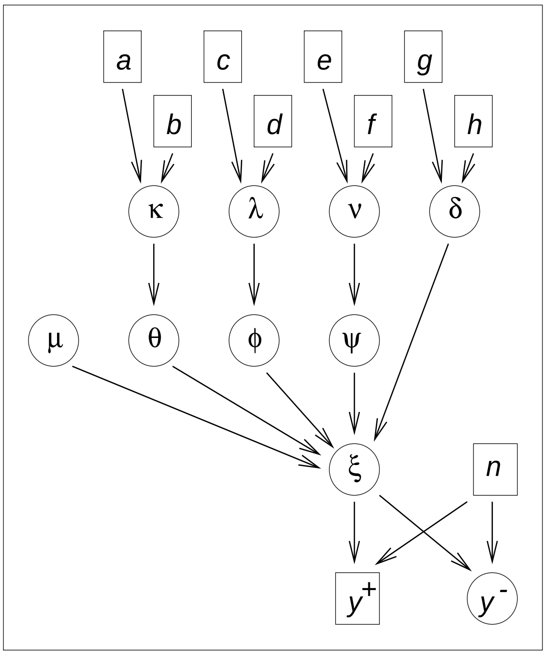
$$\pi(\theta|\kappa) \propto \kappa^{\frac{1}{2}I} \exp\left(-\frac{\kappa}{2}\sum_{i=2}^{I-1}(\theta_{i-1} - 2\theta_i + \theta_{i+1})^2\right)$$

Similarly for ϕ and ψ, with different dispersion parameters λ and ν. These are particular **2nd order** autoregressive processes.

The z_{ij} are i.i.d. $N(0, \delta^{-1})$, or equivalently $\xi_{ij}|\theta, \phi, \psi, \delta \sim N(\mu + \theta_i + \phi_j + \psi_{[ij]}, \delta^{-1})$.

We take a diffuse prior for μ ($U(\Re)$), and proper Gamma priors: $\kappa \sim \Gamma(a, b)$, $\lambda \sim \Gamma(c, d)$, $\nu \sim \Gamma(e, f)$ and $\delta \sim \Gamma(g, h)$, with $a = c = e = g = 1$, $b = d = f = h = 0.005$.

DAG for the prostate cancer model:



If we also include any missing values y^- as unknowns in the model, the **posterior/predictive density**, given the observations y^+, is

$$\pi(\mu, \theta, \phi, \psi, z, \kappa, \lambda, \nu, \delta, y^-|y^+)$$

$$\propto \prod_{i,j} \frac{\exp(\xi_{ij} y_{ij})}{\{1+\exp(\xi_{ij})\}^{n_{ij}}}$$

$$\times \kappa^{\frac{1}{2}I} \exp(-\frac{\kappa}{2} \sum_{i=2}^{I-1} (\theta_{i-1} - 2\theta_i + \theta_{i+1})^2)$$

$$\times \lambda^{\frac{1}{2}J} \exp(-\frac{\lambda}{2} \sum_{j=2}^{J-1} (\phi_{j-1} - 2\phi_j + \phi_{j+1})^2)$$

$$\times \nu^{\frac{1}{2}K} \exp(-\frac{\nu}{2} \sum_{k=2}^{K-1} (\psi_{k-1} - 2\psi_k + \psi_{k+1})^2)$$

$$\times \delta^{\frac{1}{2}IJ} \exp(-\frac{\delta}{2} \sum_{i,j} z_{ij}^2)$$

$$\times \kappa^{a-1} e^{-b\kappa} \times \lambda^{c-1} e^{-d\lambda} \times \nu^{e-1} e^{-f\nu} \times \delta^{g-1} e^{-h\delta}.$$

The **MCMC calculations** are arranged blockwise, using 7 blocks: $[\mu]$, $[\theta]$, $[\phi]$, $[\psi]$, $[\xi]$, $[\kappa, \lambda, \nu, \delta]$ and $[y^-]$, which are updated in random order each sweep. Blocks 2, 3, and 4 use multivariate normal Gibbs sampler via Cholesky decomposition, and in 5, 6 and 7, note the components are conditionally independent.

The **full conditionals** we need are:

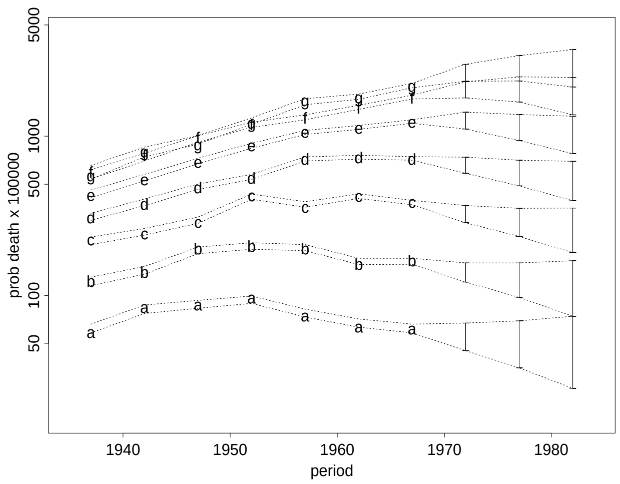
$$\begin{aligned}
\pi(\mu|\ldots) &\propto e^{-\frac{1}{2}\delta \sum z_{ij}^2} \\
\pi(\theta|\ldots) &\propto e^{-\frac{1}{2}\kappa \sum (\theta_{i-1} - 2\theta_i + \theta_{i+1})^2 - \frac{1}{2}\delta \sum z_{ij}^2} \\
\pi(\phi|\ldots) &\propto e^{-\frac{1}{2}\lambda \sum (\phi_{j-1} - 2\phi_j + \phi_{j+1})^2 - \frac{1}{2}\delta \sum z_{ij}^2} \\
\pi(\psi|\ldots) &\propto e^{-\frac{1}{2}\nu \sum (\psi_{k-1} - 2\psi_k + \psi_{k+1})^2 - \frac{1}{2}\delta \sum z_{ij}^2} \\
\pi(\xi_{ij}|\ldots) &\propto \frac{e^{\xi_{ij} y_{ij} - \frac{1}{2}\delta z_{ij}^2}}{(1+e^{\xi_{ij}})^{n_{ij}}}
\end{aligned}$$

$$\begin{aligned}
\kappa|\ldots &\sim \Gamma(a + \frac{I}{2}, b + \frac{1}{2} \sum (\theta_{i-1} - 2\theta_i + \theta_{i+1})^2) \\
\lambda|\ldots &\sim \Gamma(c + \frac{J}{2}, d + \frac{1}{2} \sum (\phi_{j-1} - 2\phi_j + \phi_{j+1})^2) \\
\nu|\ldots &\sim \Gamma(e + \frac{K}{2}, f + \frac{1}{2} \sum (\psi_{k-1} - 2\psi_k + \psi_{k+1})^2) \\
\delta|\ldots &\sim \Gamma(g + \frac{IJ}{2}, h + \frac{1}{2} \sum z_{ij}^2) \\
y_{ij}^-|\ldots &\sim \mathrm{bin}(n_{ij}^-, p_{ij})
\end{aligned}$$

Note that individual location parameters such as θ_i and ϕ_j are unidentifiable in both likelihood and prior: we must confine inference to quantities such as ξ_{ij} and p_{ij} that do have proper posteriors.

In MCMC, you don't need to think about this marginalisation, although to avoid accumulating numerical errors, we re-centre θ, etc., each sweep.

Observations, and 80% credibility/prediction intervals for next 3 time periods:



Pixel-based modelling in SPECT

An interesting example of pixel-based image analysis: analysing emission tomography data (PET or SPECT), based on the Poisson regression model introduced by Shepp and Vardi (1982).

In SPECT (single photon emission computed tomography), the pixel values in the recorded image $y = \{y_t : t \in T\}$, representing counts of detected gamma ray photons, satisfy

$$y_t \sim \text{Poisson}(\sum_{s \in S} a_{ts} x_s), \text{ independently}$$

where $x = \{x_s : s \in S\}$ denotes the true concentration of a radioisotope as a function of location in the patient, the object of inference, and $\{a_{ts}\}$ are constant coefficients.

Modelling the coefficients $\{a_{ts}\}$ accurately is important and intricate; determined by:

- geometrical and physical properties of the instrumentation
- attenuation within the body of the subject
- decay in the strength of the radioisotope,

they account for the geometrical transformation between the 'body-space' in which the true scene x is located and the 'projection-space' where the recorded data y are obtained.

Here we briefly describe how to set up a Metropolis-Hastings sampler for the problem: more detail in my chapter in 'Markov chain Monte Carlo in practice'.

Prior model: a pairwise-difference MRF on the log scale:

$$p(x) \propto \exp(-\sum_{s \sim r} \phi(\gamma(\log x_s - \log x_r)))/\prod x_s$$

where $\phi(u) = \delta(1+\delta) \log\cosh(u/\delta)$.

This expression gives absolute value and gaussian priors as $\delta \to 0$ and ∞ respectively (resulting in $\phi(u) = |u|$ or $\frac{1}{2}u^2$).

This leads to full conditionals:

$$
\begin{aligned}
p(x_s|x_{-s},y) &\propto \exp\left(-\sum_{r\in\partial s}\phi(\gamma\log(x_s/x_r))\right.\\
&+ \sum_{t:a_{ts}>0}(y_t\log\sum_r a_{tr}x_r-\sum_r a_{tr}x_r)\\
&- \left.\log x_s\right).
\end{aligned}
$$

Even for a small-scale 2-D example, there are around 300 terms in each of the sums over t and r above: so a heavy burden in computing even one value of the probability density above – partially offset by storing also $\{\mu_t=\sum_s a_{ts}x_s\}$ and keeping these up-to-date.

The density is also of awkward, non standard form (and is neither convex in x_s or $\log x_s$) so the Gibbs sampler is not at all practical here.

So we use a Metropolis-Hastings sampler, using a proposal in which the new value of $\log x_s$ is uniformly distributed over an interval centred at the old value. Thus the proposal density is

$$q_s(x_s\to x'_s)\propto(x'_s)^{-1}$$

on an interval $|\log(x'_s/x_s)|<c$. The corresponding acceptance probability is $\alpha=\min\{1,R\}$ where

$$
\begin{aligned}
R &= \exp\left[-\sum_{r\in\partial s}(\phi(\gamma\log x_r/x'_s)-\phi(\gamma\log x_r/x_s))\right.\\
&+ \sum_{t:a_{ts}>0}\left(y_t\log\left\{1+a_{ts}(x'_s-x_s)/\sum_r a_{tr}x_r\right\}\right.\\
&\quad \left.\left.-a_{ts}(x'_s-x_s)\right)\right].
\end{aligned}
$$

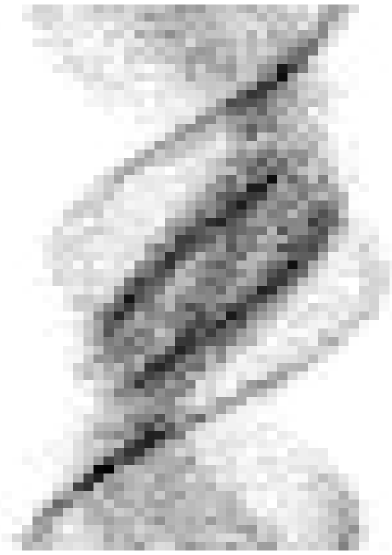
A multi-pixel update can be treated similarly: if we update all x_s, $s\in A$, using proposals symmetric about the current values, on the log scale, the corresponding acceptance probability is $\alpha=\min\{1,R\}$ where

$$
\begin{aligned}
R &= \exp\left[-\sum_{(r,s)\in\partial A}(\phi(\gamma\log x_r/x'_s)-\phi(\gamma\log x_r/x_s))\right.\\
&+ \sum_{t\in\nu A}\left(y_t\log\left\{1+\sum_{s\in A}a_{ts}(x'_s-x_s)/\sum_r a_{tr}x_r\right\}\right.\\
&\quad \left.\left.-\sum_{s\in A}a_{ts}(x'_s-x_s)\right)\right],
\end{aligned}
$$

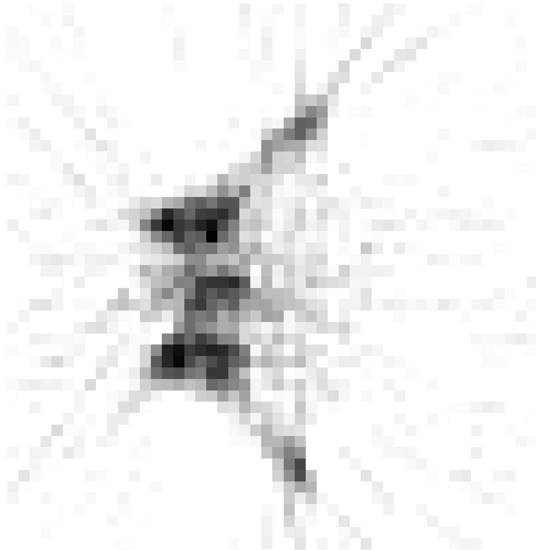
where $\partial A=\{(r,s):r\notin A,s\in A\}$ and $\nu A=\{t:a_{ts}>0$ for some $s\in A\}$.

In a fully Bayesian treatment, the hyperparameters γ and δ in the prior should themselves be considered drawn from appropriate hyperpriors, but here, for simplicity, we take γ and δ as fixed, with values chosen by trial and error.

Application to pelvis scan data. Here are data relevant to a particular horizontal slice:

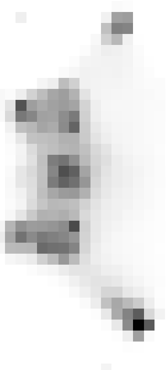


and this is a filtered back-projection reconstruction from this slice:

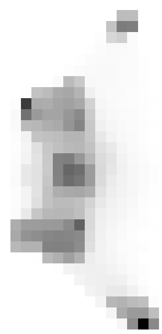


Note the noisy character of this reconstruction, and the presence of radial artifacts.

For alternative Bayesian reconstructions, the hyperparameters were set at $\gamma = 1.1$ and $\delta = 0.18$ and the proposal spread parameter $c = 0.35$ (from pilot runs). Here is the reconstruction resulting from averaging over 2000 sweeps of the Hastings sampler:



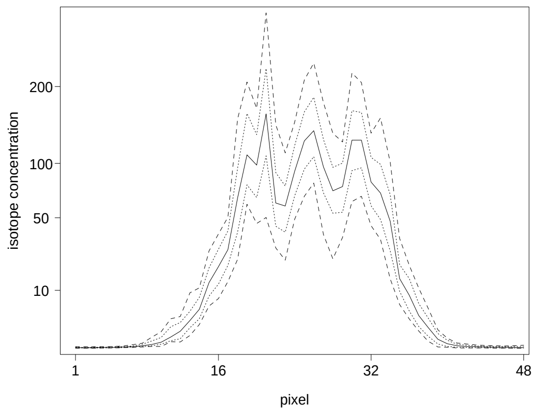
while here is an approximate MAP reconstruction, under the same prior, obtained using the one-step-late algorithm:



The two Bayesian methods have provided reconstructions of comparable quality.

The reconstruction obtained using MCMC was some 20 times more expensive to compute than the MAP; however, the advantage of the MCMC approach is seen when we wish to assess variability.

Here is a one-dimensional transect through the body-space that corresponds to a horizontal left-right line in the patient and to a vertical line in the reconstructions shown before, just to the left of the centre of the diagram:



Solid curve: pixelwise posterior median isotope concentration.

Dotted curves: 10% and 90% quantiles.

Broken lines display order-statistics-based *simultaneous* symmetric credible bounds, covering 80% of the posterior for the whole transect.

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

Bayesian model determination

It is wrong to behave as if the statistical model for our data was not subject to question.

Suppose we have a (countable) collection of models that we wish to entertain: $M_1, M_2, \ldots, M_k, \ldots$.

A priori, we assign probabilities to these: $p(k)$.

For each model, there is a parameter vector $\theta = \theta^{(k)} \in \mathcal{R}^{n_k}$ say, with a prior: $p(\theta^{(k)}|k)$, and a likelihood for the observed data Y: $P(Y|k, \theta^{(k)})$.

The joint distribution of all variables is

$$p(k, \theta^{(k)}, Y) = p(k)p(\theta^{(k)})p(Y|k, \theta^{(k)})$$

Observing Y tells us something about both the model indicator k and the corresponding parameter vector $\theta^{(k)}$, through their posterior distributions:

$$p(k|Y) = \frac{\int p(k, \theta^{(k)}, Y) d\theta^{(k)}}{\sum_k \int p(k, \theta^{(k)}, Y) d\theta^{(k)}}$$

and

$$p(\theta^{(k)}|Y, k) = \frac{p(k, \theta^{(k)}, Y)}{\int p(k, \theta^{(k)}, Y) d\theta^{(k)}}$$

involving integrals that as usual seem to need MCMC!

There are two main approaches: within-model and across-model simulation.

Within-model simulation

Here we treat each model M_k separately.

The posterior for the parameters $\theta^{(k)}$ is in any case a within-model notion:

$$p(\theta^{(k)}|Y, k) = \frac{p(\theta^{(k)}|k) p(Y|k, \theta^{(k)})}{\int p(\theta^{(k)}|k) p(Y|k, \theta^{(k)}) d\theta^{(k)}}$$

As for the posterior model probabilities, since

$$\frac{p(k_1|Y)}{p(k_0|Y)} = \frac{p(k_1)}{p(k_0)} \frac{p(Y|k_1)}{p(Y|k_0)}$$

(the *Bayes factor* for model M_{k_1} vs. M_{k_0}), it is sufficient to estimate the *marginal likelihoods*

$$p(Y|k) = \int p(\theta^{(k)}, Y|k) d\theta^{(k)}$$

separately for each k, using individual MCMC runs.

Estimating the marginal likelihood

There are many possible estimates based on importance sampling, some of which are well-studied, for example

$$\widehat{p}_1(Y|k) = N \Bigg/ \sum_{t=1}^{N} p(Y|k, \theta_t^{(k)})$$

based on a MCMC sample $\theta_1^{(k)}, \theta_2^{(k)}, \ldots$ from the posterior $p(\theta^{(k)}|Y, k)$, or

$$\widehat{p}_2(Y|k) = N^{-1} \sum_{t=1}^{N} p(Y|k, \theta_t^{(k)})$$

based on a sample from the *prior* $p(\theta^{(k)}|k)$.

Both of these has its faults, and composite estimates can perform better.

Across-model simulation

Here we conduct a *single* simulation that traverses the entire $(k, \theta^{(k)})$ space. Since the dimension n_k of $\theta^{(k)}$ typically varies with k, this requires a MCMC sampler that works in more general spaces than $\mathcal{R}^d$.

MCMC for variable dimension problems

For a fully Bayesian analysis based on a single simulation run, we need a MCMC sampler that jumps between parameter subspaces of differing dimensions. We will use a range of *move types* m, each providing a transition kernel P_m, and insist on detailed balance for each.

Ordinary Metropolis-Hastings MCMC

Unknowns x, e.g. $x = (\rho, \kappa)$, data y
Write $\pi(x)$ for $p(x|y)$
Construct MC kernel P with limiting distribution π
Detailed balance:

$$\pi(x)P(x, x') = \pi(x')P(x', x) \qquad \forall x, x'$$

When at x, propose move to x' with density $q(x, x')$
Accept with probability

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

Reversible jump MCMC

Unknowns $x \in \mathcal{C}$, e.g. $x = (k, \theta^{(k)})$
Detailed balance:

$$\int_A \int_B \pi(dx)P_m(x, dx') = \int_B \int_A \pi(dx')P_m(x', dx)$$

for all $A, B \subset \mathcal{C}$, for each move type m.
When at x, propose move of type m to dx' with probability measure $q_m(x, dx')$
Accept with probability

$$\alpha = \min\left\{1, \frac{\pi(dx')q_m(x', dx)}{\pi(dx)q_m(x, dx')}\right\}$$

Providing proposal degrees of freedom are matched, can make this ratio make sense *automatically*.

Moves between two simple subspaces

In most cases encountered, the matching of degrees of freedom is attained by *modelling the program* itself.

When in state x, we generate random numbers u, and set the proposed new state x' to a deterministic function $x'(x, u)$. Similarly in reverse: $x = x(x', u')$.

For matching, $(x, u) \leftrightarrow (x', u')$ is a bijection, and the acceptance probability is

$$\alpha = \min\left\{1, \frac{p(y|x')}{p(y|x)} \frac{p(x')}{p(x)} \frac{g(m, u'|x')}{g(m, u|x)} \left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\},$$

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

Here $g(m, u|x)$ is the joint distribution of m and u when at x.

Toy example

..... of no statistical use at all!

Suppose x lies in $\mathcal{R} \cup \mathcal{R}^2$: $\pi(x)$ is a mixture:
with probability p_1, x is $U(0, 1)$,
with probability p_2, it is Uniform on the triangle $0 < x_2 < x_1 < 1$.

I will use three moves:

(a) within $\mathcal{R}$: $x \to U(x - \epsilon, x + \epsilon)$, suppressing moves outside $(0, 1)$.

(b) within $\mathcal{R}^2$: $(x_1, x_2) \to (1 - x_2, 1 - x_1)$.

(c) between $\mathcal{R}$ and $\mathcal{R}^2$

In $\mathcal{R}$, choose (a) or (c) with probabilities $1 - r_1, r_1$.
In $\mathcal{R}^2$, choose (b) or (c) with probabilities $1 - r_2, r_2$.

Dimension-changing with move (c)

Proposal:
To go from $x \in \mathcal{R}$ to $(x_1, x_2) \in \mathcal{R}^2$, draw u from $U(0,1)$, and propose $(x_1, x_2) = (x, u)$. For reverse move, no u' required, set $x = x_1$. This certainly gives a bijection: $(x, u) \leftrightarrow (x_1, x_2)$, with Jacobian $= 1$.

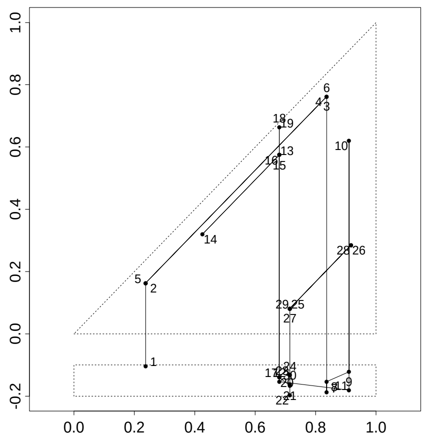
Acceptance decision:

$$\begin{aligned} \alpha &= \min\left\{1, \frac{p(y|x')}{p(y|x)} \frac{p(x')}{p(x)} \frac{g(m, u'|x')}{g(m, u|x)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right| \right\} \\ &= \min\left\{1, \frac{p_2 2I[x_2 < x_1]}{p_1} \frac{r_2}{r_1 \times 1} |1| \right\} \\ &= \min\left\{1, \frac{2p_2 r_2}{p_1 r_1}\right\} I[u < x] \end{aligned}$$

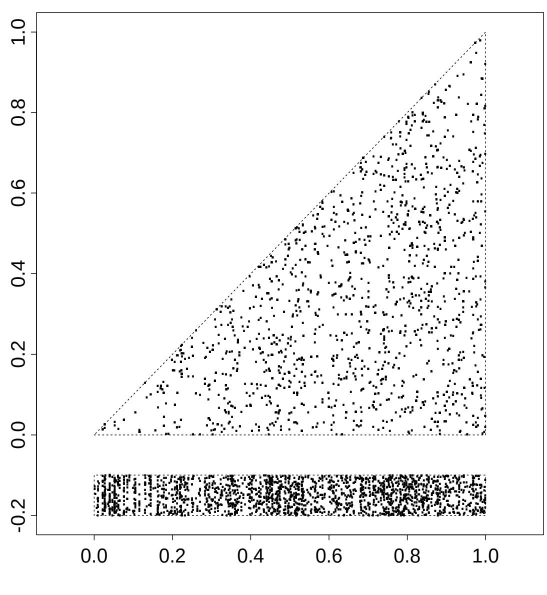
For reverse move, $\alpha = \min\{1, (p_1 r_1)/(2p_2 r_2)\}$.

First 30 steps

$p_1 = 0.4$, $p_2 = 0.6$, $r_1 = 0.7$, $r_2 = 0.4$, $\epsilon = 0.3$.



5000 steps



Some special cases, and related algorithms

- Ordinary Metropolis(-Hastings), with random scan, on a fixed parameter space.
- Birth/death simulation for spatial point processes (c.f. Ripley, 1977).
- Grenander & Miller (1994) jump diffusion processes, with Hastings rejection calculation on diffusion part.
- Carlin & Chib (1995) work on $\bigotimes \mathcal{R}^{n_k}$ instead of $\bigcup \mathcal{R}^{n_k}$.
- Piccioni and Jona-Lasinio embed all $\mathcal{R}^{n_k}$ into a single space.

Multiple change-point analysis

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

- point-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 $x(t)$, $0 < t < L$, 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

Again I will use three moves:

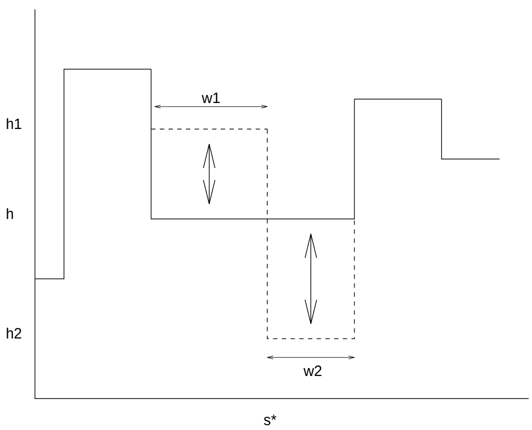
(a) Metropolis change to a randomly chosen step height h_j.

(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_1, h_2)
- death: choose step at random to kill, combine current step heights (h_1, h_2) into one: h
- dimension-balancing: preserve weighted geometric mean: $h_1^{w_1} h_2^{w_2} = h^{w_1+w_2}$.

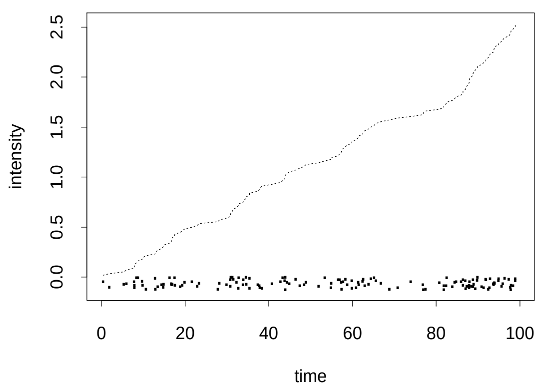
Birth and death of steps



$$h_1^{w_1} h_2^{w_2} = h^{w_1+w_2}$$

Example: cyclones hitting the Bay of Bengal

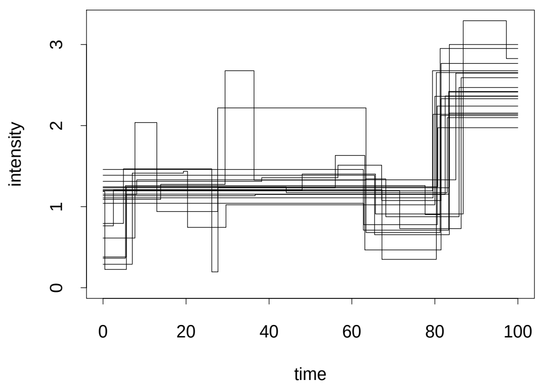
141 cyclones over a period of 100 years.



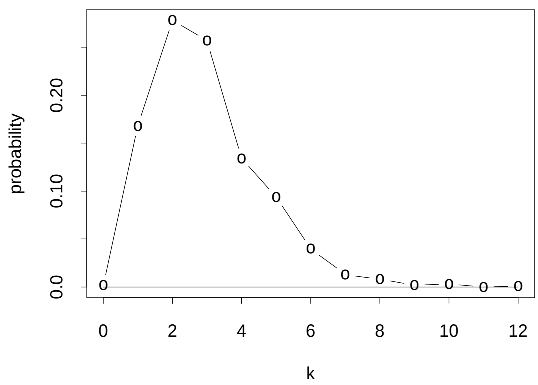
Choices of hyperparameters:

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

Sample of step functions from the posterior:

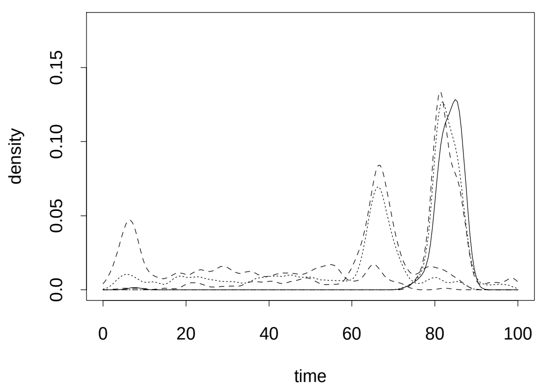


Posterior for the number of change points k

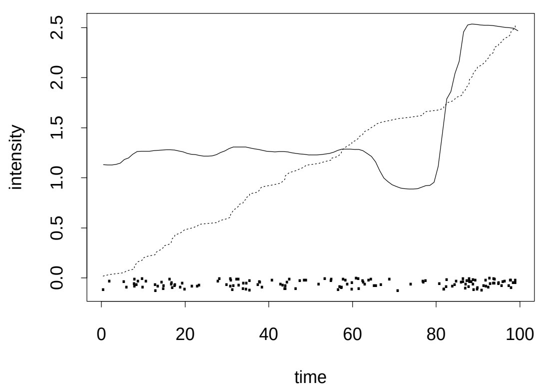


Zero change points is ruled out; $k = 1$ or 2 more probable than under the prior.

Posterior density estimates for change-point positions

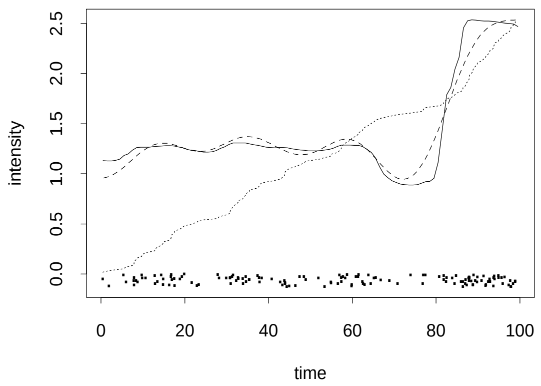


Model-averaged estimate: $E(x(\cdot)|y)$



(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

Finally, here, as further illustration of the flexibility in modelling allowed by the approach, we include another ingredient, that will be justified in many real time-series point process problems: periodicity.

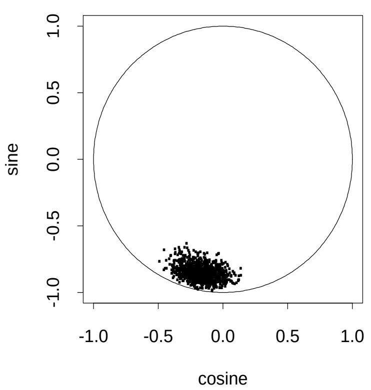
This could be handled in various ways – parametric, nonparametric, with known and unknown period(s) – but the simplest is to take a simple sinusoid, and assume that the data are generated from a Poisson process with instantaneous rate

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

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)



Further reading

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion), *Statistical Science*, **10**, 3–66.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Lauritzen, S. L. (1996) *Graphical models*, Oxford University Press.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society*(B), **60**, 731–792.

Royal Statistical Society discussion meeting on MCMC methods (1993) *Journal of the Royal Statistical Society*, B, **55**, 3–102.