IMS/ENAR, Memphis, March 1997

Statistical problems where the parameter dimension varies: MCMC theory and some applications

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

IMS Special Invited Lecture

Bayesian computation 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

Key ideas

- sample-based computation allows flexible extraction of information from posterior
- reversible jump MCMC samplers used to sample functional representations with varying number of terms
- challenges of extracting and presenting information about complex posterior distributions
- difficulties of specifying weak prior information

Plan

- generic hierarchical model
- reversible jump MCMC methodology
- toy example
- multiple change point analysis
- mixture estimation
- discussion

Bayesian model determination

Three-stage hierarchical model:

model indicator	$k \in \mathcal{K}$ (countable)
parameter vector	$\theta^{(k)} \in \mathcal{R}^{n_k}$ (for example)
data	y

joint = model probability × prior × likelihood

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

– interested in computing the posterior $p(k, \theta^{(k)}|y)$ (whatever we intend to do with it!)

– may be interested in k and $\theta^{(k)}$ **jointly** (across-model inference, e.g. model averaging), or **separately**: $p(k, \theta^{(k)}|y) = p(k|y)p(\theta^{(k)}|k, y)$ (within-model inference).

– abbreviate (k, θ) as x, $p(x|y)$ as $\pi(x)$.

Computation for Bayesian inference

Under the posterior distribution, the components of x are generally *dependent*, so we have to compute with a multivariate distribution, often in a high number of dimensions, with arbitrarily complex patterns of dependence.

Here, "compute with" could mean almost anything; an example would be to calculate a marginal (posterior) density.

- Exact analytic integration is usually not available (and we don't want to be restricted in our model construction to use conjugate priors, etc., to make it possible).
- Asymptotic analytic approximations (e.g. Laplace) are awkward to set up, and can be unreliable.

- Conventional numerical methods require expertise and careful design to set up, and are only efficient in a low number of dimensions.
- Ordinary ("static") simulation is always available in principle, since the posterior distribution can be factorised as

$$p(x|y) = \pi(x_1, x_2, \ldots, x_p)$$

$$= \pi(x_1)\pi(x_2|x_1) \ldots \pi(x_p|x_1, \ldots, x_{p-1})$$

but the univariate distributions on the right hand side are rarely all available for simulation purposes (even after re-ordering).

Example: Weibull/Gamma experiment

Data are a random sample, possibly censored, from Weibull(ρ, κ):

$$f(y|\rho, \kappa) = \kappa^m \rho^{m\kappa} \prod_U y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum y_i^\kappa\right)$$

where m and $\prod_U$ are number of and product over uncensored observations.
Independent Gamma priors on ρ and κ:

$$\pi_0(\rho, \kappa) \propto \rho^{\alpha-1} e^{-\beta\rho} \kappa^{\gamma-1} e^{-\delta\kappa}$$

Posterior:

$$p(\rho, \kappa|y) \propto \kappa^m \rho^{m\kappa} \prod_U y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum y_i^\kappa\right) \rho^{\alpha-1} e^{-\beta\rho} \kappa^{\gamma-1} e^{-\delta\kappa}$$

Markov chain Monte Carlo

Construct a Markov chain with limiting distribution $\pi(x) = p(x|y)$; simulate this to obtain a realisation

$$x^{(1)}, x^{(2)}, \ldots, x^{(N)},$$

and treat this as an approximate sample from $p(x|y)$; for example:

Estimate	$E(x|y)$	by	$\sum_{t=1}^{N} x^{(t)}/N$
and	$P(x \in A|y)$	by	$\#\{t : x^{(t)} \in A\}/N$

Recipes

- Gibbs sampler
- Metropolis(-Hastings) algorithm

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

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

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.

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 $\mathcal{R}^2$: swap x_1 and x_2.

(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 $= 1$.

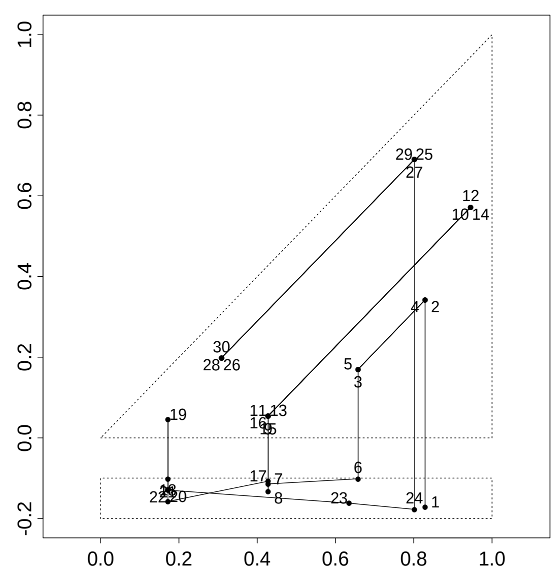
Acceptance decision:

$$\begin{aligned}\alpha &= \min\left\{1, \frac{p(y|x')}{p(y|x)}\frac{p(x')}{p(x)}\frac{g(m,u'|x')}{g(m,u|x)}\left|\frac{\partial(x',u')}{\partial(x,u)}\right|\right\}\\ &= \min\left\{1, \frac{p_2 2I[x_2<x_1]}{p_1}\frac{r_2}{r_1\times 1}|1|\right\}\\ &= \min\left\{1, \frac{2p_2r_2}{p_1r_1}\right\} I[u<x]\end{aligned}$$

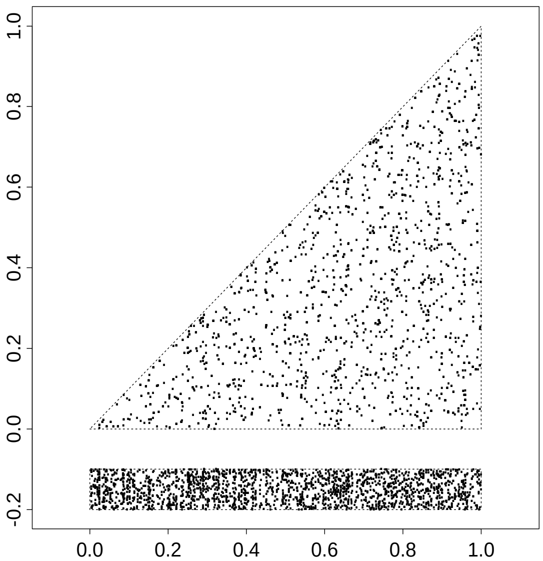
For reverse move, $\alpha = \min\{1, (p_1r_1)/(2p_2r_2)\}$.

First 30 steps

$p_1 = 0.4$, $p_2 = 0.6$, $r_1 = 0.7$, $r_2 = 0.4$.



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:

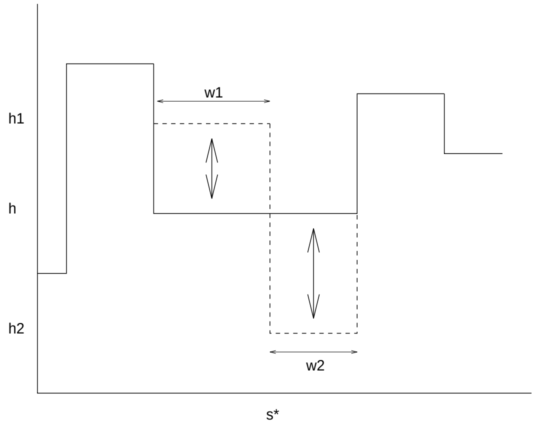
(a) Metropolis change to a randomly chosen step height h_j.

(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: $h_1^{w_1} h_2^{w_2} = h^{w_1+w_2}$.

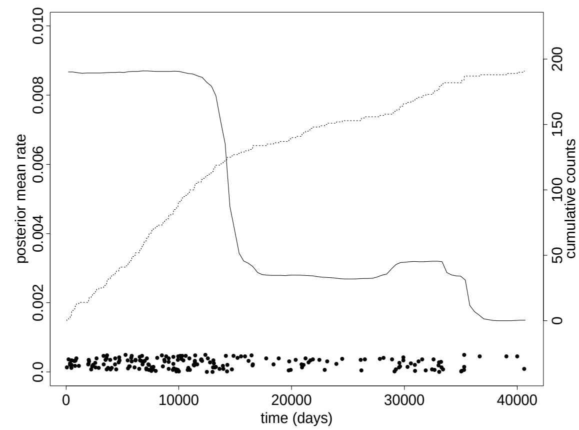
Birth and death of steps



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

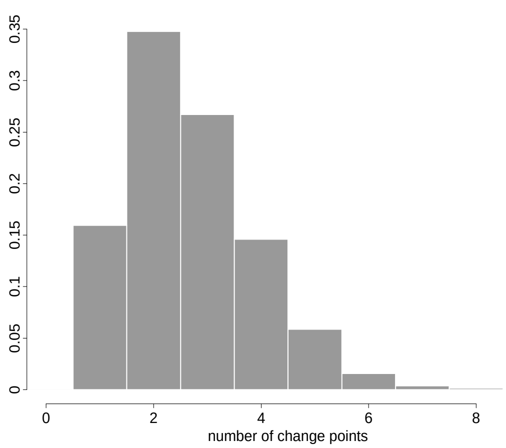
Coal mining disaster data

Times of 191 serious coal mining disasters in the UK, 1851–1962.

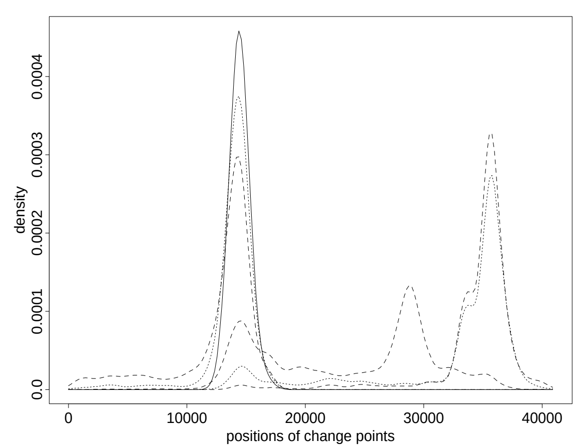


Hyperparameters used: $\lambda = 3$, $\alpha = 1$, $\beta = 200$.

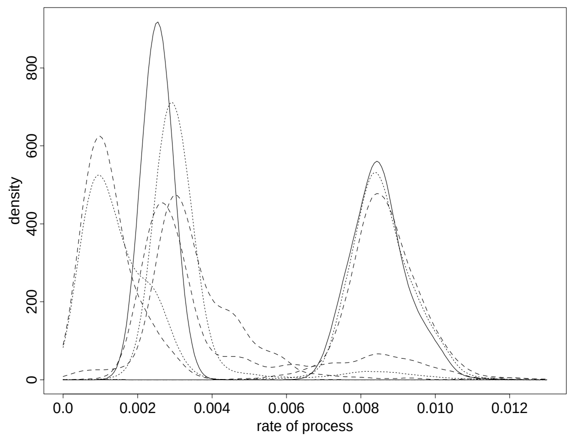
Posterior on k



Posterior density estimates for change-point positions



Posterior density estimates for rates



Mixture estimation

(jointly with Sylvia Richardson, INSERM, Paris)

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

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

Fully Bayesian mixture modelling

- component parameters and number of components modelled jointly
- posterior distributions of parameters and predictive densities, not just point estimates

Bayesian models for mixtures

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

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

Two contexts

Heterogeneous population: Groups $j = 1, 2, \ldots, k$, sizes $\propto w_j$. Observation y_i drawn from unknown group z_i: latent *allocation variable*.

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

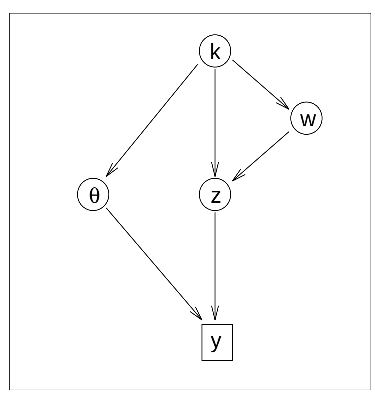
$$y_i|z \sim f(\cdot|\theta_{z_i}) \quad \text{independently for } i = 1, 2, \ldots, n$$

Semi-parametric density estimation: (not prime focus here) use same representation, but $\{z_i\}$ now artificial.

N.B. There may be empty components!

Hierarchical model

$$p(k, \theta, w, z, y) = p(k)p(\theta|k)p(w|k)p(z|w, k)p(y|\theta, z)$$



For flexibility, allow priors for k, θ and w to depend on hyperparameters, drawn from independent hyperpriors.

Univariate normal mixtures

$$\theta_j \to (\mu_j, \sigma_j)$$

with *independent* priors:

$$\mu_j \sim N(\xi, \kappa^{-1}) \quad \text{and} \quad \sigma_j^{-2} \sim \Gamma(\alpha, \beta)$$

Labelling. Model is invariant to relabelling of groups: for identifiability, work with *set*, or choose unique labelling; we generally use

$$\mu_1 < \mu_2 < \cdots < \mu_k$$

Weights.

$$w \sim D(\delta, \delta, \ldots, \delta)$$

Prior on k. Results easily reweighted for any prior, so we typically use

$$k \sim U[1, 2, \ldots, 30]$$

Weak prior information

Possibility of empty components, so

standard *independent* improper non-informative prior distributions for $\{\theta_j\}$
$\Rightarrow$ improper posterior distributions.

We use *weakly informative priors*, which may be range-dependent.

$$\mu_j \sim N(\xi, \kappa^{-1}) \quad \text{independently}$$

with

$$\xi = \text{midpoint} \quad \text{and} \quad \kappa^{-1/2} = \text{range}$$

$$\sigma_j^{-2} \sim \Gamma(\alpha, \beta) \quad \text{independently}$$

with

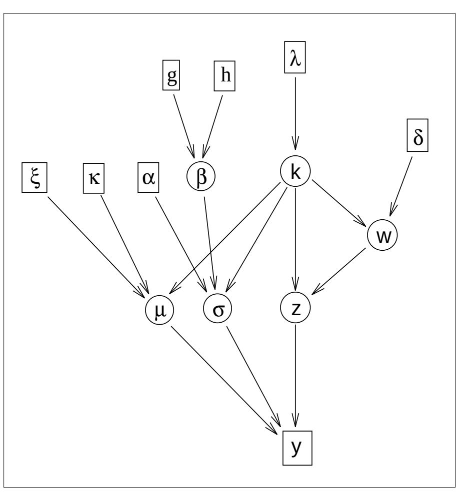
$$\alpha \approx 2 \quad \text{and} \quad \sqrt{(\beta/\alpha)} \approx \text{range}/10$$

We prefer **random β** model: $\sigma_j^{-2} \sim \Gamma(\alpha, \beta)$

$$\beta \sim \Gamma(g, h) \quad \text{with} \quad g < 1 < \alpha$$

expresses belief that σ_j^2 are similar, without being informative about their absolute size. Typically,

$$\alpha = 2, \quad g = 0.2, \quad h = 10/\text{range}^2$$



MCMC methodology for mixtures

MCMC updating of $(w, \mu, \sigma, z, \beta)$ (by Gibbs sampling for the conjugate priors used here) routine since ≈ 1990 (Robert, Diebolt,…).

The novelty here is that when k is altered, the dimension of the whole parameter vector changes: need MCMC moves that can jump between parameter subspaces of different dimensionality
$\Rightarrow$ reversible jump MCMC

Application of reversible jumps to mixtures

We use two dimension-changing moves:

- splitting/combining components
- birth/death of empty components

(the former is essential, the latter is introduced simply to improve mixing in some rather extreme cases)

Split/combine move

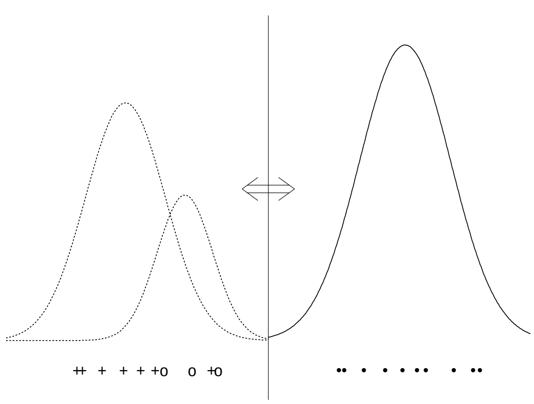
Propose to split a randomly chosen component ($k \to k+1$) or combine two *adjacent* randomly chosen components ($k \to k-1$), *and* reallocate affected observations.

$$(k, w, \mu, \sigma, z) \to (k \pm 1, w', \mu', \sigma', z')$$

Propose a parameter set in the new subspace that is *intuitively* roughly as well supported by the posterior as the old set. We preserve combined weight, mean and variance:

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2} \\ w_{j^*}\mu_{j^*} &= w_{j_1}\mu_{j_1} + w_{j_2}\mu_{j_2} \\ w_{j^*}(\mu_{j^*}^2 + \sigma_{j^*}^2) &= w_{j_1}(\mu_{j_1}^2 + \sigma_{j_1}^2) + w_{j_2}(\mu_{j_2}^2 + \sigma_{j_2}^2) \end{aligned}$$

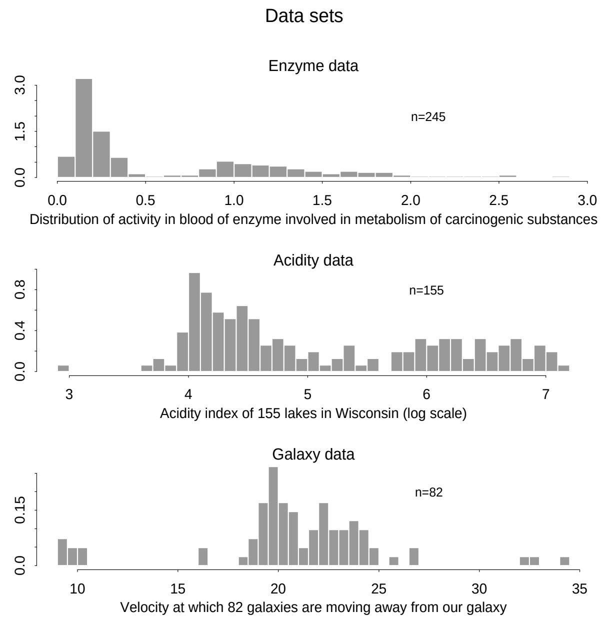
Illustration of split/combine proposal



Acceptance probability. For the split move the probability is $\min(1, A)$, where A is

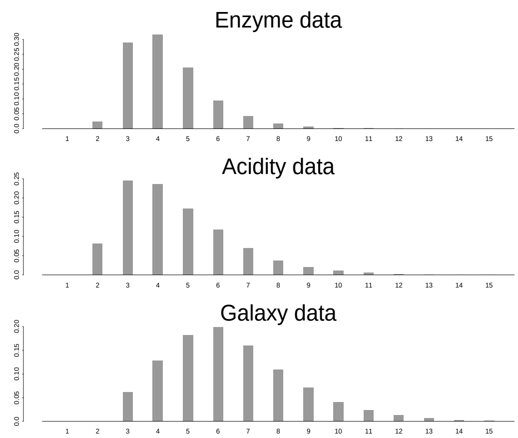
$$\begin{aligned} & (\text{likelihood ratio}) \times \frac{p(k+1)}{p(k)} \times (k+1) \\ \times \quad & \frac{w_{j_1}^{\delta-1+l_1} w_{j_2}^{\delta-1+l_2}}{w_{j^*}^{\delta-1+l_1+l_2} B(\delta, k\delta)} \times \sqrt{\frac{\kappa}{2\pi}} \\ \times \quad & \exp\left[-\frac{1}{2}\kappa\{(\mu_{j_1}-\xi)^2 + (\mu_{j_2}-\xi)^2 - (\mu_{j^*}-\xi)^2\}\right] \\ \times \quad & \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{\sigma_{j_1}^2\sigma_{j_2}^2}{\sigma_{j^*}^2}\right)^{-\alpha-1} \exp\left(-\beta(\sigma_{j_1}^{-2} + \sigma_{j_2}^{-2} - \sigma_{j^*}^{-2})\right) \\ \times \quad & \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times \{g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)\}^{-1} \\ \times \quad & \frac{w_{j^*}|\mu_{j_1} - \mu_{j_2}|\sigma_{j_1}^2\sigma_{j_2}^2}{u_2(1-u_2^2)u_3(1-u_3)\sigma_{j^*}^2} \end{aligned}$$

Performance



Posterior distribution of k

(using uniform prior)

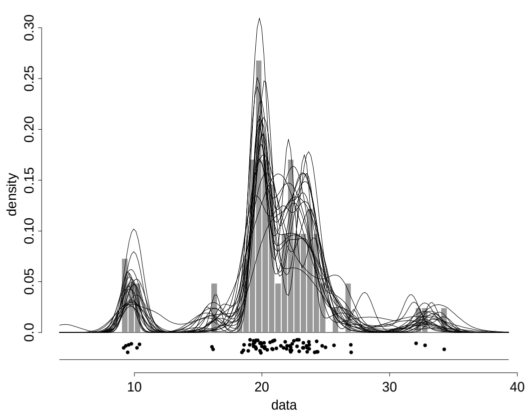


Sample from the posterior distribution of the mixture density

Each sweep:

$$(k, w, \theta) \longrightarrow f(\cdot|k, w, \theta) = \sum_{j=1}^{k} w_j f(\cdot|\theta_j)$$

For Galaxy data:



Predictive densities

- average across the MCMC run conditional on k

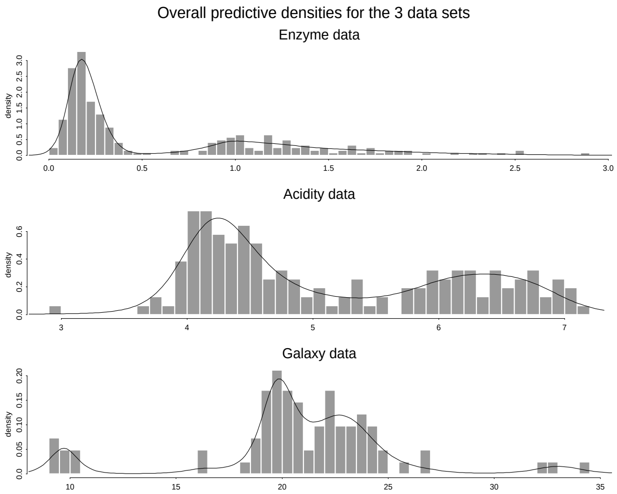
  $$\longrightarrow E(f(\cdot|k, w, \theta)|k, y),$$

  Bayesian predictive density estimate of mixture with k components

- average across values of k

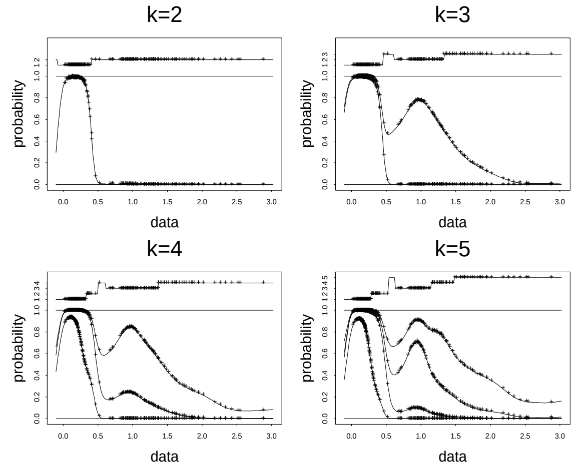
  $$\longrightarrow E(f(\cdot|k, w, \theta)|y)$$

  the ‘overall’ Bayesian predictive density estimate of the distribution of y.



Bayesian classification of Enzyme data

Within-sample and predictive classifications, conditional on k, using same MCMC run:



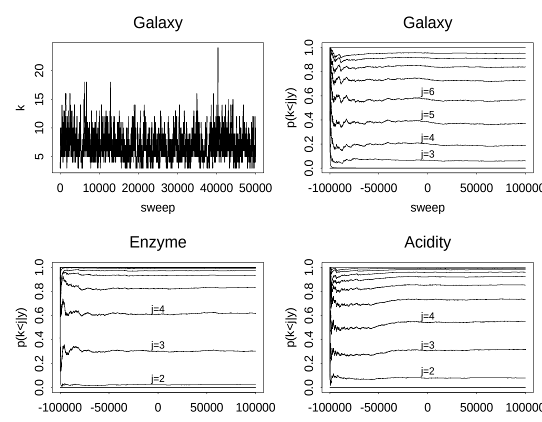
$$p(z_i = j|y, k) \text{ for } j = 1, 2, \ldots, k$$
$$p(z^* = j|y^*, y, k) \text{ for } j = 1, 2, \ldots, k$$

Performance of the MCMC sampler

Effective mixing over k

- % split/combine moves vary between 8% and 14% for the 3 data sets
- stability of cumulative occupancy fractions after $\simeq 50000$ sweeps
- birth/death move for empty components improves mixing, particularly for diffuse posteriors
- prior for k does not affect mixing if it does not unnecessarily restrict range of values of k

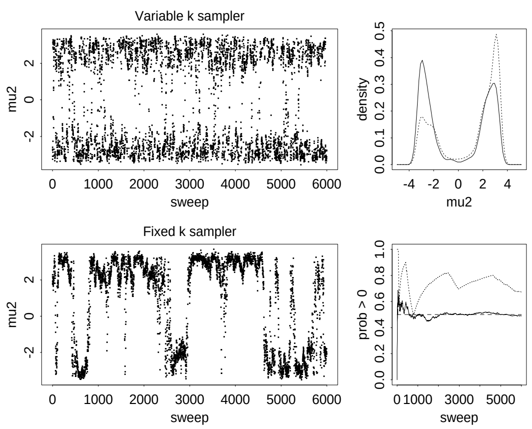
MCMC performance. Jump moves



Effective mixing within k

- no trapping states
- multimodality explored satisfactorily

  $\longrightarrow$ better mixing than a fixed-k sampler in case where posterior has several well separated modes: potential as a method of simulated tempering?



Skew-normal family

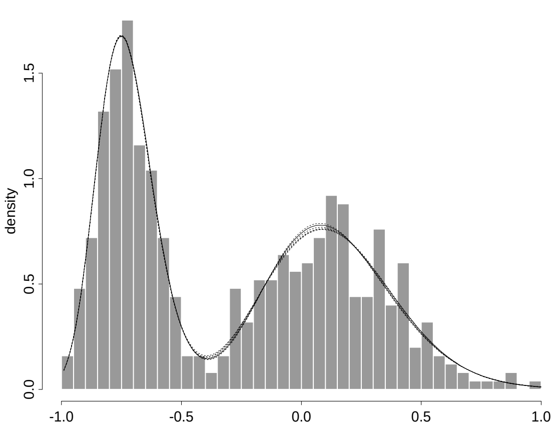
$$v_i \sim \sum_{j=1}^{k} w_j \phi(\cdot|\mu_j, \sigma_j) \quad \text{independently for } i = 1, 2, \ldots, n$$

$\phi(\cdot|\mu,\sigma)$ is normal density
$\{\mu_j\}, \{\sigma_j\}, \{w_j\}, k$ unknown
and we observe

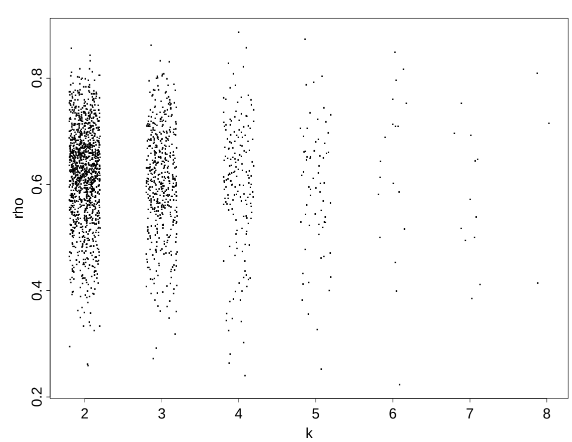
$$y_i = \frac{\exp(\rho v_i) - 1}{\rho}$$

Simulated data, and predictive densities

500 draws from
$0.5 \times N(0.5, 0.04) + 0.5 \times N(-0.5, 0.04)$,
scaled, then skewed by $\rho = 0.6$;
predictive densities for $k = 2$ to 6:



Joint posterior of k and ρ

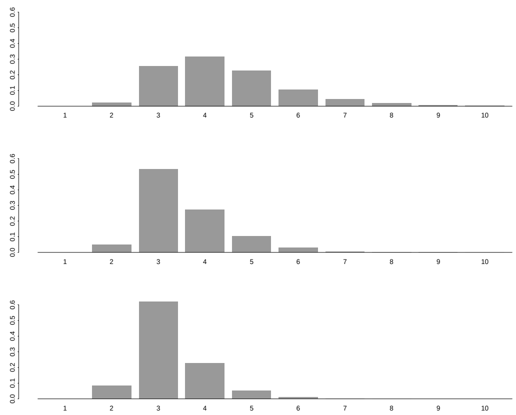


Spacing prior on means

μ_j are drawn *independently* from $N(\xi, \kappa^{-1})$, then ordered by labelling, so no direct control over spacing of means.

Alternative spacing prior draws μ_j as $s^{\text{th}}, 2s^{\text{th}}, \ldots, ks^{\text{th}}$ order statistics from $N(\xi, \kappa^{-1})$.

Effect on posterior for k for Enzyme data:



Extensions

Other univariate distributions

- discrete distributions
- 2 parameter density suitably reparametrised
- adding an overall shape parameter to model mixtures of skewed distributions

Other mixture models

- modelling the weights or allocations (genetic, spatial applications...)
- adapting the algorithm to mixture models based on a Dirichlet process prior: alternative to the incremental sampler

Robust prior modelling in Bayesian analyses

use of mixtures with k unknown as a prior model arises in many contexts of application, *e.g.* measurement error problems, random effects, frailty models.

Discussion

- further evidence for power and flexibility of MCMC in exact computation for complex Bayesian models
- Bayesian approach proves attractive for variable-dimension problems
- multiple explanations for the data are exposed, leaving open some questions about how to present them

Written versions

Biometrika, **82**, 711–732 (1995)

J. Roy. Statist. Soc. B, **59**, (1997) *in press*:
(with Sylvia Richardson, with discussion)

`http://www.stats.bris.ac.uk/~peter`
(preprints, these notes)

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