ISBA, Istanbul, August 1997

Bayesian mixture estimation

by Peter Green (University of Bristol),
P.J.Green@bristol.ac.uk
and Sylvia Richardson (INSERM, Paris),
richardson@vjf.inserm.fr

Basic mixture model

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

$f(\cdot|\theta)$ 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

Plan

- mixture models and priors
- implementation of reversible jump MCMC methodology for univariate normal mixtures
- connections with Dirichlet Process models
- skewness
- sensitivity issues
- simulation study of statistical performance
- application to measurement error models

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*.

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

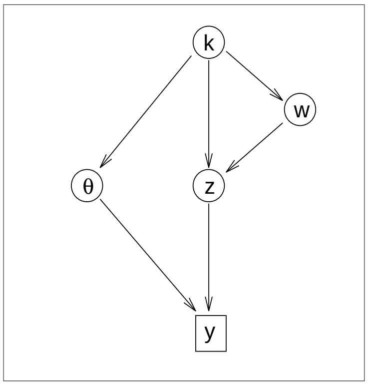
$$y_i|z \sim f(\cdot|\theta_{z_i}) \quad \text{independently for } i = 1, 2, \ldots, n$$

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 $\{\theta_j\}$
⇒ 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

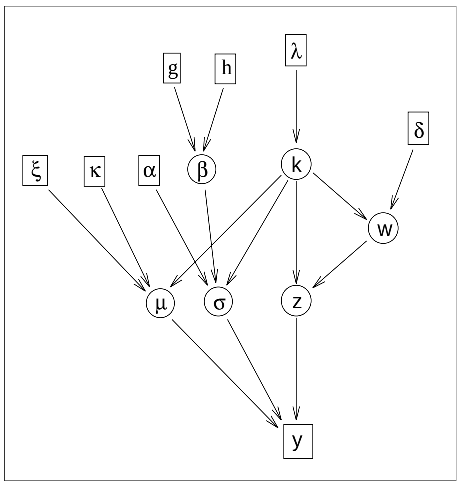
$$\alpha \approx 2 \quad \text{and} \quad \surd(\beta/\alpha) \approx \text{range}/10$$

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

MCMC updating of $(w, \mu, \sigma, z, \beta)$ (by Gibbs sampling for the conjugate priors used here) routine since $\approx$ 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 (PJG, *Biometrika*, 1995)

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

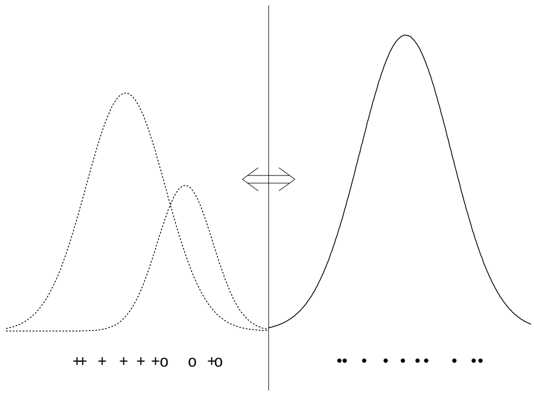
Propose to split a randomly chosen component ($k \to k + 1$) or combine two *adjacent* randomly chosen components ($k \to k - 1$), *and* reallocate affected observations.

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

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



Ordinary Metropolis-Hastings MCMC

Unknowns x, 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. $x = (k, w, \mu, \sigma, z, \beta)$
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

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

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

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

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

Remarks on MCMC methodology

- generic character of moves, exploiting *adjacency*

  – can be adapted to a variety of univariate distributions

- selection of subspace-jumping moves

  – performance comparisons?

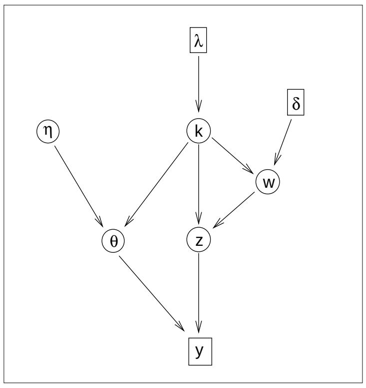
- multivariate extensions: richer algebra of moves?

- validation by comparison with

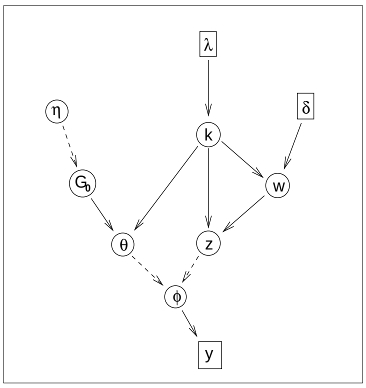
  – analytic calculations on very small data sets
  – prior model in the absence of data

Connections with Dirichlet process models

First, let us go back to the general mixture model, with generic parameter θ and hyperparameter η.



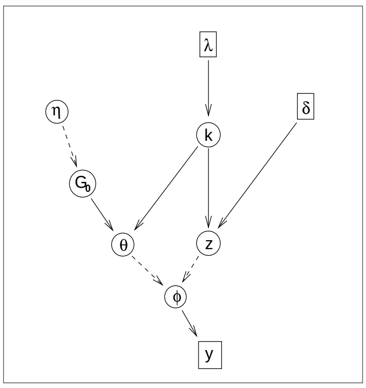
Retain the idea that component parameters θ_j are drawn independently and identically from a prior distribution. Denote this by G_0, and add it explicitly into the graph. Also write $\phi_i = \theta_{z_i}$; $\phi \leftrightarrow (\theta, z)$, and $y_i|\phi \sim f(\cdot|\phi_i)$.



Now integrate out the weights w. We find

$$p(z|k, \delta) = \frac{\Gamma(k\delta)}{(\Gamma(\delta))^k} \frac{\prod_{j=1}^k \Gamma(\delta + n_j)}{\Gamma(k\delta + n)}$$

where $n_j = \#\{i : z_i = j\}$.

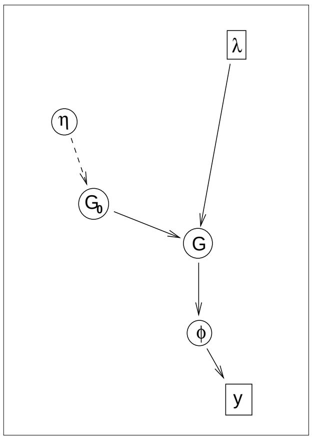


All that is needed is the distribution of ϕ given G_0, λ and δ.
An *alternative* model is that based on the Dirichlet Process:

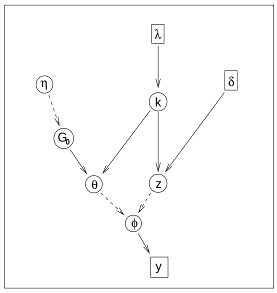
$$\phi_i | G \sim G \quad \text{independently for} \quad i = 1, 2, \ldots, n,$$

where

$$G \sim \mathrm{DP}(\lambda, G_0).$$



But *this* model can be represented exactly in the notation of the previous graph.



We find

$$p(z|k) = \frac{\prod_{j=1}^{k}(n_j - 1)!}{c(n,k)}$$

and

$$p(k) = \frac{\lambda^k c(n,k)}{k!\lambda(\lambda+1)\ldots(\lambda+n-1)}$$

where

$$c(n,k) = \sum_z \prod_j (n_j - 1)!$$

Thus the DP corresponds to taking our Dirichlet/Multinomial distribution for $p(w, z|k)$ and

- conditioning on $n_j > 0 \; \forall \; j$ (no empty components)
- letting $\delta \to 0$ (allocations more uneven);

– note there are no free parameters.

The number of (non-empty) components k has a one-parameter prior distribution, depending also on n.

In both models, the distinct ϕ_i, that is $\{\theta_j, j = 1, 2, \ldots, k\}$, are drawn i.i.d. from $G_0 = G_0(\cdot|\eta)$.

This comparison is *generic*: common to all DP applications where ϕ is drawn from λ and G_0 in this way, not only to mixture models.

– how "nonparametric" are the DP models?

MCMC samplers for the DP models

Among several MCMC approaches to DP models, as implemented for univariate normal mixtures in particular, it is interesting to compare:

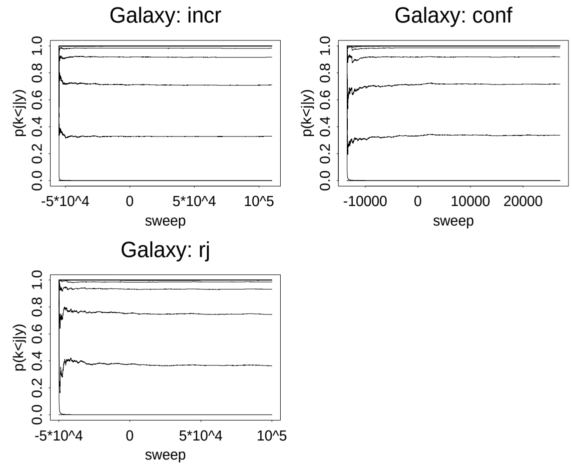
- the "incremental" sampler: e.g., West, Müller and Escobar (1994)
  – update θ for fixed k, z
  – update ϕ one-by-one by Gibbs sampler
- the "configuration" sampler: MacEachern and Müller (1994)
  – update augmented θ for fixed k, z
  – update z for fixed augmented θ
- a reversible jump split/combine sampler
  – update θ for fixed k, z
  – update z for fixed k, θ
  – use split/combine move to change k, with consequences for z and θ

Reversible jump sampler for DP

- needs few modifications to standard model version
- obviates the setup integration needed for incremental method, hence allowing variable hyperparameters

Split/combine approach remains applicable to other DP contexts, with p-dimensional θ.

Some comparisons of MCMC performance



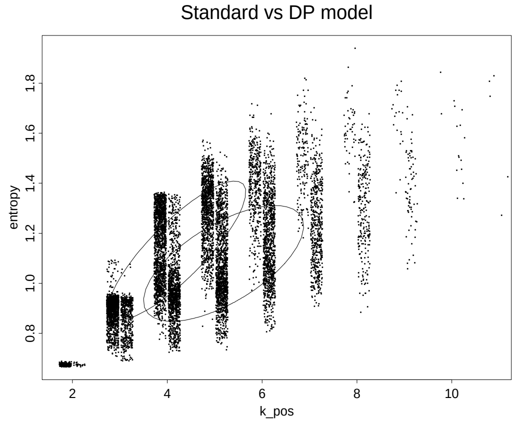
Preliminary statistical comparison

Does the DP model's prior emphasis on unequal allocations persist in the posterior?

Summarise equality of allocation by the *entropy*

$$-\sum_j \frac{n_j}{n} \log \frac{n_j}{n}$$

and look at joint posterior of entropy and k_{pos}, the number of non-empty components



Skew-normal mixture family

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

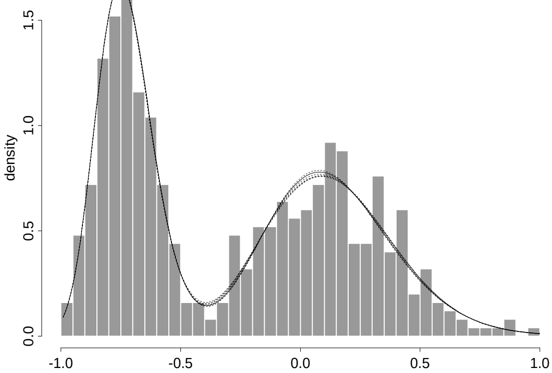
$\phi(\cdot|\mu, \sigma)$ is normal density
$\{\mu_j\}, \{\sigma_j\}, \{w_j\}, k$ unknown
and we observe

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

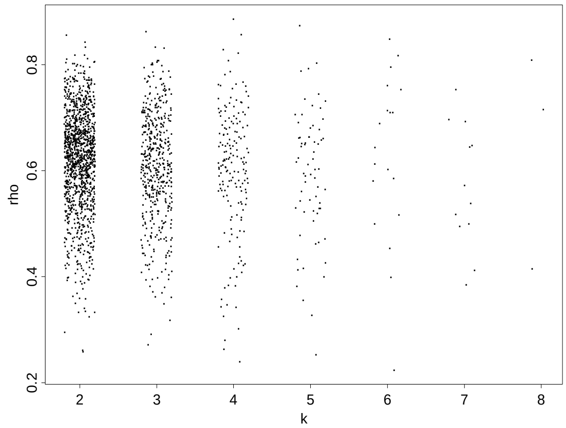
Here ρ is an unknown skewness parameter, common to all components.

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:

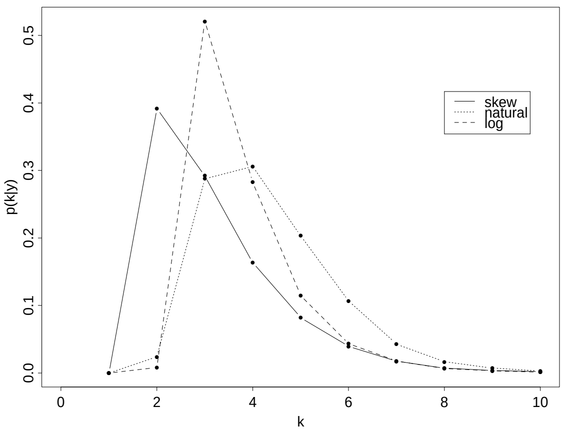


Joint posterior of k and ρ

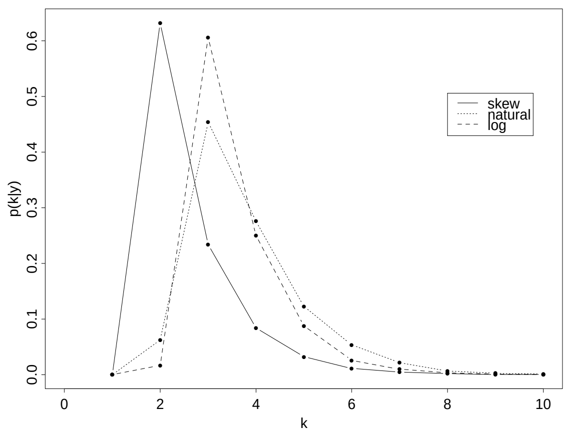


The variable-skew option often yields a stochastically-smaller posterior for k, allowing a more economical and perhaps more readily interpreted description of the data.

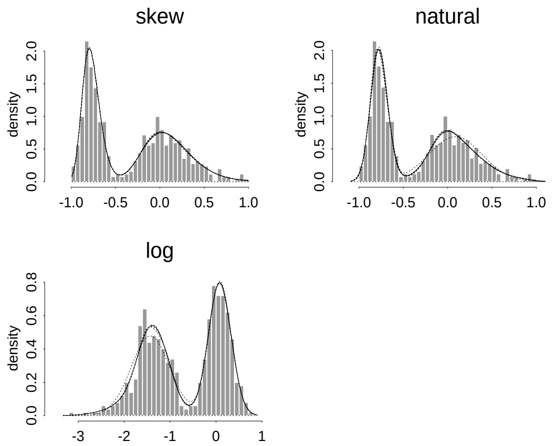
For example, on the Enzyme data



And on simulated data



with also closer coincidence between predictive densities:



Further details

See RSS read paper, with discussion:
Richardson and Green,
J. Roy. Statist. Soc, B, **59**, 731–792.

Email addresses:

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

`richardson@vjf.inserm.fr`

WWW:

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