SSC 2000
Ottawa, June 2000

Bayesian analysis of heterogeneity using mixtures and related models

by Peter Green (University of Bristol,
P.J.Green@bristol.ac.uk
http://www.stats.bris.ac.uk/∼peter).

Joint work with:
Sylvia Richardson (INSERM → Imperial)
Carmen Fernández (Bristol → St. Andrews)
Agostino Nobile (Bristol → Glasgow)
Valerie Viallefont (INSERM → Lancaster)
Laurence Watier, Isabelle Deltour (INSERM)

- finite mixture models, Bayesian formulation
- MCMC moves for variable numbers of components
- illustration for normal mixtures
- *Poisson mixtures*
- *connections with Dirichlet process models*
- *random effects in mixed models*
- *measurement error*
- analysis of factorial experiments
- adaptation to spatial settings, disease mapping

Bayesian analysis of finite mixtures

(with Sylvia Richardson (INSERM → Imperial))

$y_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j)$ independently
$f(\cdot|\theta)$ is a given parametric family
$\{y_i\}$ observed, $\{\theta_j\}$, $\{w_j\}$, k unknown

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

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

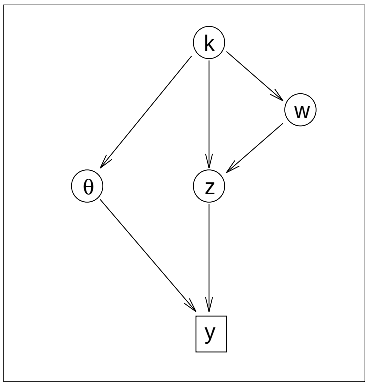
$$y_i|z \sim f(\cdot|\theta_{z_i}) \quad \text{independently for } i = 1, 2, \ldots, n$$

Context 2: 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

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

MCMC methodology

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 ⇒ reversible jump MCMC (PJG, *Biometrika*, 1995)

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.

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

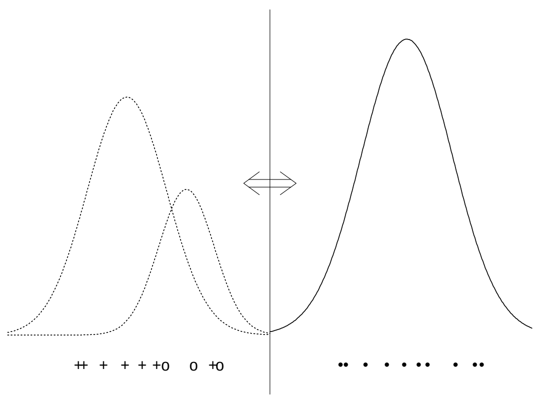
Propose to split a randomly chosen component ($k \to k+1$) or combine two *adjacent* randomly chosen components ($k \to k-1$), *and* reallocate affected observations.

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

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

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

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

What do you get from the Bayesian approach?

- Avoids unsatisfactory and difficult hypothesis testing for number of components
- Exposes multiple explanations
- Averaging over models gives superior predictions
- Natural basis for classification and prediction
- Allows use of real prior information if available

Generalisations

- Other distributions
- Other prior structures
- Skewness
- Structured data

Bayesian analysis of factorial experiments by mixture modelling

(with Agostino Nobile (Bristol $\to$ Glasgow))

Biometrika, 2000.

Approach applies to any factorial setting – here we look at '2 way ANOVA with interaction'.

Sampling model: $y_{ijk} \sim N(\theta_{ij}, \sigma_{ij})$
$i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, r_{ij}$.

Factorial structure: $\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$

Prior set up:

$$
\begin{aligned}
\mu &\sim N(\eta, \sigma^\mu) \\
\alpha_i &\sim \textstyle\sum_t w_t^\alpha N(\mu_t^\alpha, \sigma_t^\alpha) \\
\beta_j &\sim \textstyle\sum_t w_t^\beta N(\mu_t^\beta, \sigma_t^\beta) \\
\gamma_{ij} &\sim \textstyle\sum_t w_t^\gamma N(\mu_t^\gamma, \sigma_t^\gamma)
\end{aligned}
$$

$$\sigma_{ij}^{-1} \sim \Gamma(a, b) \quad \text{with} \quad b \sim \Gamma(q, h)$$

Hyperpriors

We want factor effect components to adapt to the data.

$$\begin{aligned} w^{\alpha} &\sim Dir(1, 1, \ldots, 1) \\ \mu_t^{\alpha} &\sim N(\xi_t^{\alpha}, 1/\tau^{\alpha}) \\ \tau^{\alpha} &\sim \Gamma(a^{\tau\alpha}, b^{\tau\alpha}) \\ (\sigma_t^{\alpha})^{-1} &\sim \Gamma(a_t^{\alpha}, b_t^{\alpha}) \end{aligned}$$

and similarly for β and γ parameters.

Choosing hyperparameters

Set $\eta = 0$ and $\sigma^{\mu} = 100 \times \max_{i,j} y_{ij}^2$. Also $\xi^{\alpha} = \xi^{\beta} = \xi^{\gamma} = 0$.

It appears to be completely hopeless to use generic, uninformative priors. We need to control within and between component variability, and this entails a user-specified 'caliper' Δ: two factor levels are regarded as essentially identical if they differ by less than Δ – and this is represented by their coming from the same mixture component.

Choose a_t^{α} and b_t^{α} so that $\mathrm{pr}(|\alpha_i - \alpha_j| \leq \Delta) = 0.95$.

Conversely, we wish there to be small probability of two components' means being closer than Δ, and this gives a method for specifying $a^{\tau\alpha}$ and $b^{\tau\alpha}$.

Finally, choose $a = 3$, $q = 0.2$ and h such that $E(\sigma_{ij}) = E(1/\tau^{\alpha})$.

Some results on survival time dataset

Posterior distribution of partition of poison effects ($\times$ 100 000)

Δ	111	112	121	211	123
1	2 703	75 148	221	5 403	16 525
0.25	1	58 978	0	306	40 715

$$\mathrm{pr}(\alpha_1 \approx \alpha_2 \not\approx \alpha_3) \approx 0.75$$

$$\mathrm{pr}(\alpha_1, \alpha_2, \alpha_3 \text{ all distinct}) \approx 0.17$$

$$\mathrm{pr}(\alpha_1 \approx \alpha_2 \not\approx \alpha_3, \beta_1 \approx \beta_3 \not\approx \beta_2 \approx \beta_4) \approx 0.37$$

Spatial mixtures of Poisson distributions, with application to disease mapping

(with Sylvia Richardson (INSERM $\rightarrow$ Imperial) and Carmen Fernández (Bristol $\rightarrow$ St. Andrews))

Small area disease mapping

In regions indexed $i = 1, 2, \ldots, n$:
y_i = observed count of disease incidence
E_i = expected count based on population size, adjusted for age and sex, etc.

y_i/E_i = standardised mortality (morbidity) ratio (SMR)

Standard assumption: $y_i \sim$ Poisson$(\lambda_i E_i)$
$\Rightarrow$ inference on relative risks $\{\lambda_i\}$

Modelling spatially dependent Poisson data

Some options:

- Direct modelling of dependence at count level – 'auto-Poisson' Markov random field (Besag, 1974) – inflexible, only negative dependence, covariates awkward
- Continuously-distributed (usually Gaussian) MRF for $\log \lambda_i$ – e.g. Besag, York and Mollié (1989) – popular and successful, but some problems in identifiability, specification and interpretation (and over-smoothing?)
- Variations – e.g. Stern and Cressie (1999)
- Hidden Gamma random field model – Wolpert and Ickstadt (1998)

Mixture modelling approach

Basic mixture set-up

$$y_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \quad \text{independently}$$

$$\equiv$$

introduce latent allocation variables $\{z_i\}$ with

$$\begin{aligned} y_i|z &\sim f(\cdot|\theta_{z_i}) \\ p(z_i = j) &= w_j \end{aligned}$$

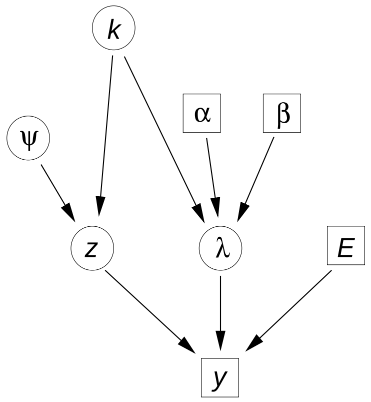
Extension to spatial case

Write relative risk as λ_{z_i} in place of λ_i.

$$y_i|z \sim \text{Poisson}(\lambda_{z_i} E_i)$$

where $\{z_i\}$ is a spatially dependent random field with $z_i \in \{1, 2, \ldots, k\}$

Our model formulations



- $y_i \sim \text{Poisson}(\lambda_{z_i} E_i)$ independently
- $\lambda_j \sim \Gamma(\alpha, \beta)$ independently *and then ordered*
- $k \sim \text{Uniform}(1, 2, \ldots, k_{\max})$
- α, β usually fixed
- various alternatives for $p(z)$, $z \in \{1, 2, \ldots, k\}^n$

Allocation models

In each case, spatial context determined by assumed neighbourhood structure – we say 'adjacent' $\equiv$ 'have common boundary' ($i \sim j$). For rare diseases, more complex dependence not justified.

The formulations we have implemented and explored:

- Potts model: $p(z) = \exp(\psi U(z) - \theta_k(\psi))$ where $U(z) = \#\{i \sim j : z_i = z_j\}$ = number of like-coloured neighbour pairs.
- multinomial allocation – $p(z_i = j) = w_{ij}$ – using either
  - logistic-normal weights: $w_{ij} = \exp(x_{ij}) / \sum_{j'} \exp(x_{ij'})$
  - grouped continuous weights: $w_{ij} = \Psi(x_i - \delta_j) - \Psi(x_i - \delta_{j-1})$

  where (x_{ij}) and (x_i) are Gaussian random fields.

MCMC moves for Potts model formulation

1. allocation variables z updated one-by-one by Gibbs
2. interaction parameter ψ: full conditional $\propto p(\psi)\exp(\psi U(z) - \theta_k(\psi))$ – use Metropolis
3. component parameters λ_j – simultaneous update by Metropolis, with multiplicative perturbation followed by ordering
4. split/merge move to create or remove components

Detail of move (3) – simultaneous update of λ_j

Make simultaneous zero-mean normal perturbations to all $\log \lambda_j$ – the modified λ values are then ordered to give proposed updates λ'_j.

Proposal density is sum of $k!$ terms, but after re-arrangement, the terms in numerator and denominator of Metropolis-Hastings ratio are in proportion - acceptance probability reduces to $\min\{1, R\}$ where $R =$

$$\prod_{j=1}^{k} \left[\left(\frac{\lambda'_j}{\lambda_j} \right)^{\alpha + \sum_{i:z_i=j} y_i} \exp\{-(\lambda'_j - \lambda_j)(\beta + \sum_{i:z_i=j} E_i)\} \right].$$

Detail of move (4) – split/merge

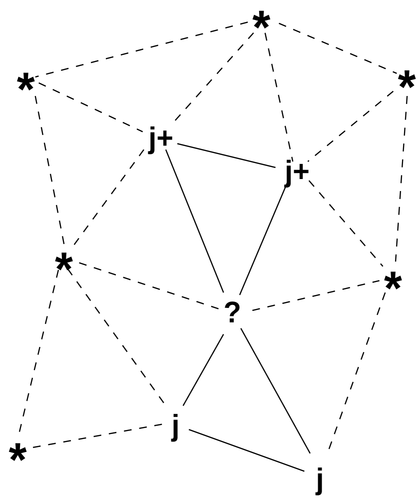
As in independent random sample case, number of components changes only by ± 1, by splitting and merging components.

The proposal simultaneously changes k, amends the vector λ, and reallocates regions.

Steps:

1. choose between split and merge
2. (if split) choose component j at random
3. propose new values $\lambda_{j-} < \lambda_j < \lambda_{j+}$
4. reject if whole λ vector out of order
5. scan through regions allocated to j, re-allocating between $j-$ and $j+$ randomly *but not independently* accumulating allocation probability
6. compute acceptance probability and accept/reject

Illustrating split/merge allocation proposal



Allocation of **?** to **j-** or **j+** according to a Potts model on the $\{$**?, j-, j+**$\}$ sites – in this case it will be **j+** with odds of $\exp(2\psi)$ to 1.

Acceptance probability for this complete split proposal is $\min\{1, R\}$ where

$$
\begin{aligned}
R &= \prod_{i:z'_i=-} e^{-(\lambda_- - \lambda_j)E_i} \left(\frac{\lambda_-}{\lambda_j}\right)^{y_i} \\
&\times \prod_{i:z'_i=+} e^{-(\lambda_+ - \lambda_j)E_i} \left(\frac{\lambda_+}{\lambda_j}\right)^{y_i} \\
&\times \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{\lambda_-\lambda_+}{\lambda_j}\right)^{\alpha-1} e^{-\beta(\lambda_- + \lambda_+ - \lambda_j)} (k+1) \frac{p_{k+1}}{p_k} \\
&\times \exp\{\psi(U(z') - U(z)) + \theta_k(\psi) - \theta_{k+1}(\psi)\} \\
&\times \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times \frac{2c}{u}.
\end{aligned}
$$

Potts model normalising constants

The normalising constant

$$\theta_k(\psi) = \log \left[\sum_{z \in \{1,2,\ldots,k\}^n} \exp(\psi U(z)) \right]$$

is intractable, but accessible by Monte Carlo methods; for example the identity (exponential families!)

$$\theta_k(\psi) = n \log k + \int_0^\psi E(U|\psi', k) \mathrm{d}\psi'$$

can be used.

Interpretation and inference in spatial mixtures

Do we really believe there are k groups of regions with identical relative risks?

- model is being used in a 'semi-parametric' fashion, not to identify clusters
- inference on $\{\lambda_{z_i}\}$ rather robust to details of prior structure – 'borrows strength' between regions in an adaptive way (by Bayesian model averaging)
- avoid over-smoothing of relative risks
- interpret inference on k and z with caution (diagnostic/exploratory)

References and preprints available from

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

P.J.Green@bristol.ac.uk