Bayesian modelling of heterogeneity : algorithms, mixture models and applications

Sylvia Richardson (INSERM, Paris)
richardson@vjf.inserm.fr

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

Introduction

- modelling heterogeneity : two contexts

  - observations drawn from a population consisting of unknown groups
  $\rightarrow$ mixture models

  - construction of priors with flexible shapes reflecting latent heterogeneity
  $\longrightarrow$ semi-parametric density estimation

- MCMC algorithms
  - enable Bayesian inference for "highly structured" models
  $\rightarrow$ conditional independence modelling

  - need to consider state space of variable dimension $\rightarrow$ reversible jump MCMC

- mesurement error models involve
  latent variables with unknown prior distributions
  combining information from several sources
  propagating uncertainty

Plan

- mixture models and priors
- MCMC methodology
- statistical performance
- 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

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

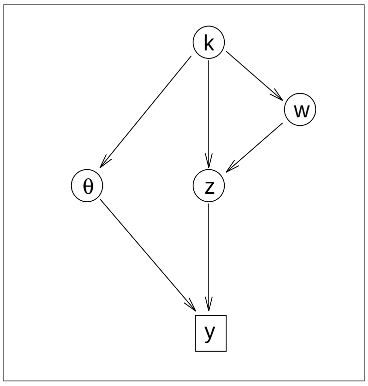
$$p(z_i = j) = w_j \quad \text{independently for } i = 1, 2, \ldots, n$$

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

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.

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

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

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

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

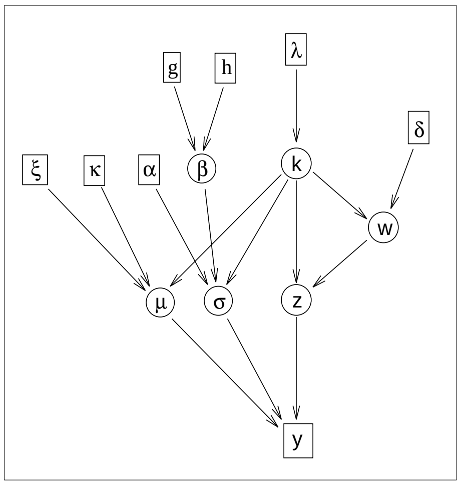
$$\alpha \approx 2 \quad \text{and} \quad \sqrt{(\beta/\alpha)} \approx \text{range}/10$$

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

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 (P.J. Green, *Biometrika*, 1995)

Can be thought of as Metropolis-Hastings for general state space.

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 chose to preserve combined weight, mean and variance.

Combine proposal :

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

Split proposal

- Choose a component j^* at random
- Draw 3 random variables with Beta distributions : u_1, u_2, u_3

$$\begin{aligned} w_{j_1} &= w_{j^*}u_1 \quad , \quad w_{j_2} = w_{j^*}(1-u_1) \\ \mu_{j_1} &= \mu_{j^*} - u_2\sigma_{j^*}\sqrt{\frac{w_{j_2}}{w_{j_1}}} \quad , \quad \mu_{j_2} = \mu_{j^*} + u_2\sigma_{j^*}\sqrt{\frac{w_{j_1}}{w_{j_2}}} \\ \sigma_{j_1}^2 &= u_3(1-u_2^2)\sigma_{j^*}^2\frac{w_{j^*}}{w_{j_1}} \quad , \quad \sigma_{j_2}^2 = (1-u_3)(1-u_2^2)\sigma_{j^*}^2\frac{w_{j^*}}{w_{j_2}} \end{aligned}$$

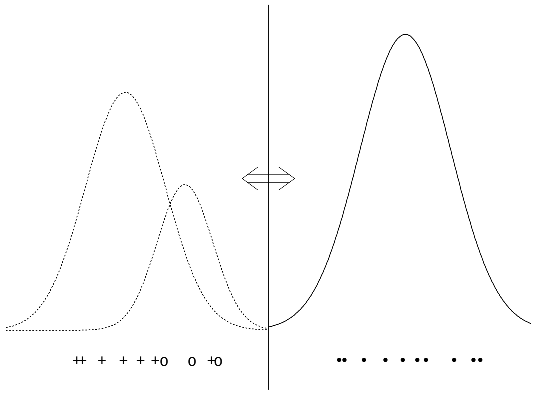
This choice satisfies moment preserving condition.

! Need to check adjacency condition, otherwise reject move.

Propose reallocation between j_1 and j_2 as in standard updating.

Acceptance probability : Metropolis ratio involving a product of likelihood, prior, proposal and a Jacobian.

Illustration of split/combine proposal



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

Statistical performance

- posterior distribution of k
- predictive densities
- parameter estimates / choice of labelling
- mixing
- sensitivity
- classification

Data sets

- enzyme data
- (simulated data)

Analyses with default hierarchical random β model (except sensitivity analyses)

Predictive densities

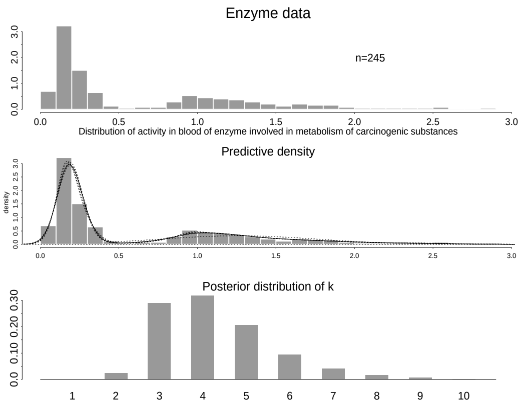
- average across the MCMC run conditional on k

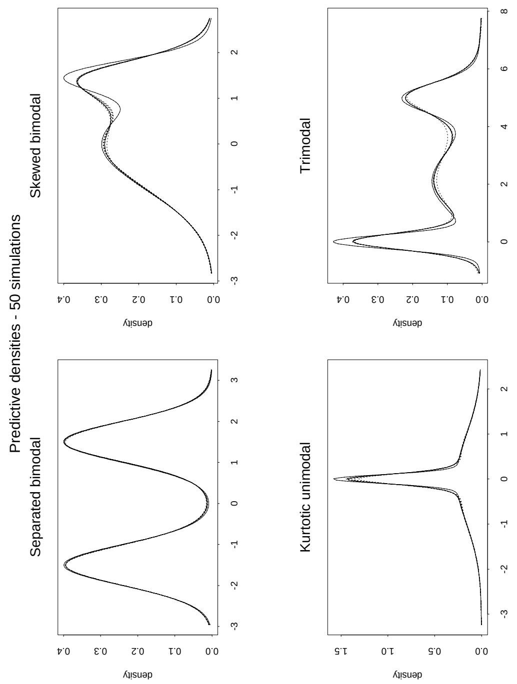
$$\longrightarrow E(f(\cdot|k,w,\theta)|k,y),$$

  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.





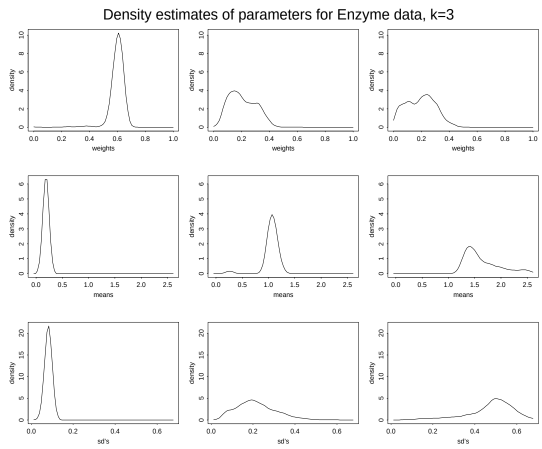
Parameter estimates

Label components at each sweep to summarise posterior (marginal) distribution of parameters

- interpretation is conditional on a choice of labelling
- for overlapping subpopulations: no unambiguous labelling
  $\Rightarrow$ search for *unimodal* densities for all parameters;
  – interpret persisting multimodality in terms of competing explanations

Postprocessing MCMC output

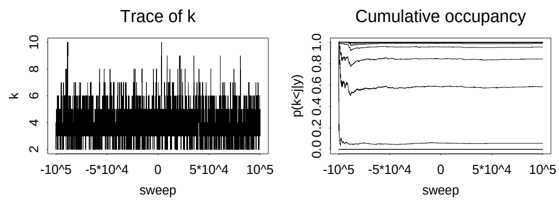
- investigate different choices of labels
- separate out competing explanations



MCMC Performance Jump moves

Effective mixing over k

- % split/combine moves is typically around 10%
- stability of cumulative occupancy fractions after $\simeq 50000$ sweeps



Sensitivity to prior specification

- Prior information on the variances
  - We have found hierarchical model with random β succesfully expresses weak prior information:

  $\Rightarrow$ posterior of k insensitive to hyperparameters

- Prior model for the means

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

  with $\quad \xi = \text{midpoint}$

  - We have found sensitivity to value of κ : can be interpreted as prior more or less permissive of mixtures with very close μ_j

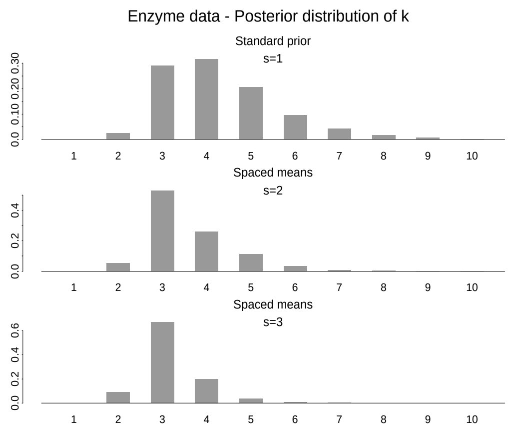
  $\Rightarrow$ separate means by using a dependent prior ?

Joint prior model for the means

For $s \geq 1$, define

$\{\mu_j; j = 1, 2, \ldots, k\}$ to be distributed as the $\{s, 2s; \ldots, ks\}$ order statistics from a sample of size $ks + s - 1$ of $N(\xi, \kappa^{-1})$

$\longrightarrow s > 1$ encourages separation of the means

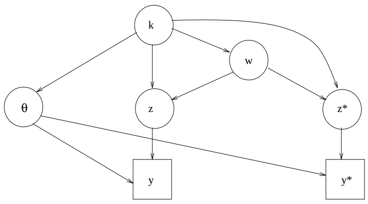


Separation of the means leads to a tighter posterior distribution for k

Classification

need a declared labelling

- within-sample
  compute $\{p(z_i = j|y, k); j = 1, 2, \ldots, k\}$
  Bayesian classification : $\widehat{z_i} = \text{argmax}_j\, p(z_i = j|y, k)$
- predictive : new observation y^*

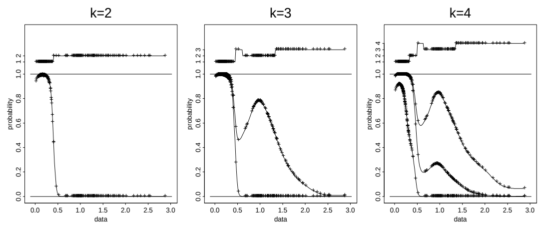


$$\begin{aligned} p(z^* = j|y, y^*, k) &= \int p(z^* = j|y, y^*, k, \theta, w) p(\theta, w|y, y^*, k) d\theta\, dw \\ &\approx \int p(z^* = j|y^*, k, \theta, w) p(\theta, w|y, k) d\theta\, dw \end{aligned}$$

can be computed from MCMC output

Classification of enzyme data set

The large variance of the third component creates a non monotonicity in the classification



Semi-parametric framework

Bayesian modelling of complex data involves often the introduction of latent structures

It is of interest to have flexible models for the distribution of latent structures.

Measurement error problems involve

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

Define local submodels between components

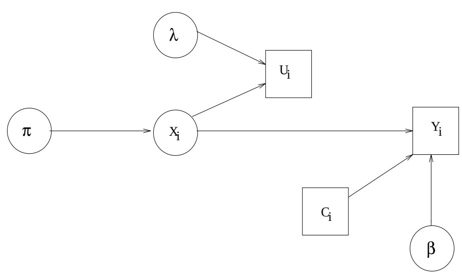
- $[Y|X, \beta]$ regression model
- $[U|X, \lambda]$ measurement model

Distribution for X ?

Typically in epidemiological studies

- Y disease status
- X true risk factor (exposure)
- U surrogate measure of risk factors
- C known covariates

Global structure



3 submodels are linked with conditional independence assumptions ensuring non differentiable measurement error.

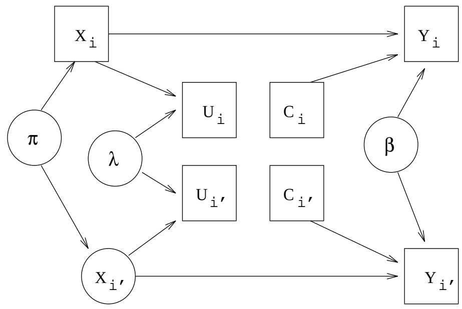
Where does quantitative information on measurement model come from ?

- a priori external information
- study designs

$\longrightarrow$ repeated measures for unbiased measuring instrument

$\longrightarrow$ reference method which can be used to apprehend X on a subgroup where both X and U are recorded, e.g. designs with *validation group*

Designs with a validation group



$\longrightarrow$ Information on λ will be transferred from the validation group to the main study.

$\longrightarrow$ Provide information on relations between X and U in main study.

$\longrightarrow$ Strengthen inference about regression parameters β.

- in many cases it is difficult to specify a parametric model for $[X]$

  'shapeless' distribution of U

  large measurement errors

  small or non existent validation group

- alternatives

  pseudo-likelihood *(Carroll, 1993)* based on plugging an empirical estimate of $[X]$ based on the validation subgroup

  non parametric modelling of $[X]$ *(Roeder, Carroll, Lindsay, JASA 1996)*

  semi-parametric model for $[X]$ via a mixture of gaussian distributions with an unknown number of components

Combining measurement error and mixtures
Joint work with L. Leblond

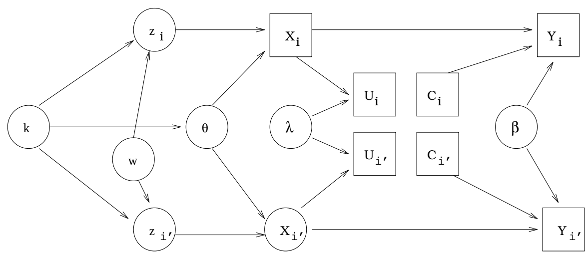
Mixture model for prior distribution of X

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

Measurement error model : normal or lognormal
$U_i \sim N(\alpha_0 + \alpha_1 X_i, \gamma^{-1})$ or $U_i \sim logN(\alpha_0 + \alpha_1 X_i, \gamma^{-1})$

Regression model for binary outcome (disease status): logistic model

$$Y \sim B(p) \quad \text{with } logit\ p = \beta^T(X, C)$$



Performance of mixture priors in measurement error models

Influence of misspecification evaluated by comparing analyses using mixture priors for X versus analyses using a gaussian prior for X.

Simulation set up 50 replications

main study $n = 500$

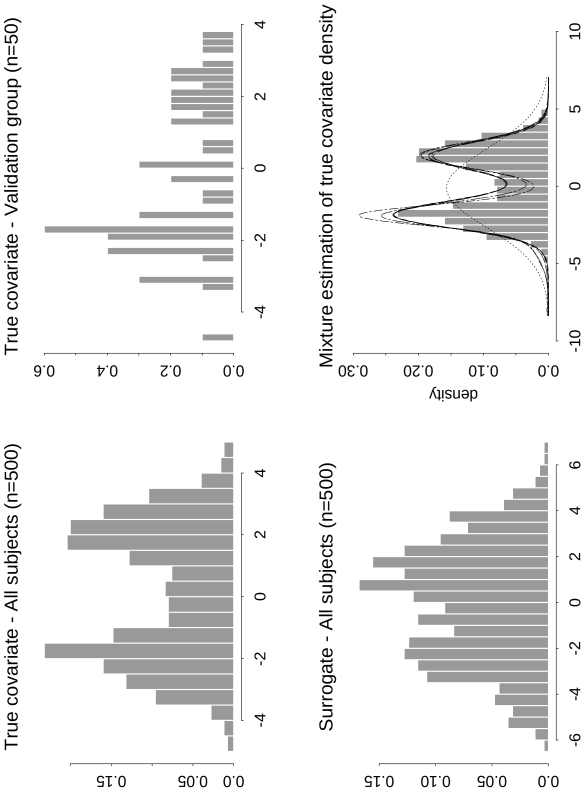
validation subgroup $n = 50$

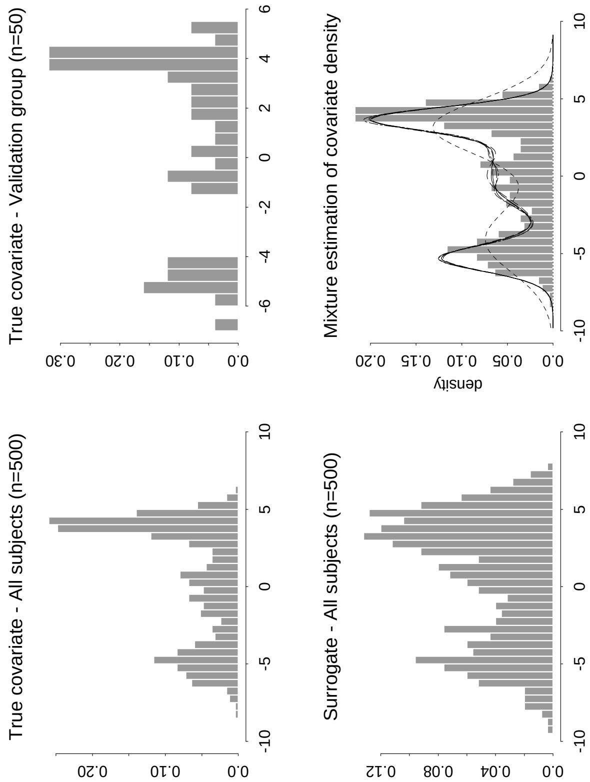
prior of X : $0.5N(-2, 1) + 0.5N(2, (0.75)^{-1})$

measurement model : $U \sim N(X, \theta^{-1})$, $\theta = 0.06$

logistic disease model : $logitP(D|X) = \beta_0 + \beta_1 X$, $\beta_0 = -0.8, \beta_1 = 0.9$

Analysis	mixture prior	gaussian prior
θ	0.65 (0.10)	0.65 (0.11)
β_0	- 0.76 (0.18)	-0.79 (0.21)
β_1	0.93 (0.12)	1.07 (0.18)
MSE	0.02	0.04





Performance comparison between mixture and non-parametric approaches

Simulation set-up as in Roeder et al (1996)

main study : 90 cases 90 controls

validation subgroup: 30 cases 30 controls

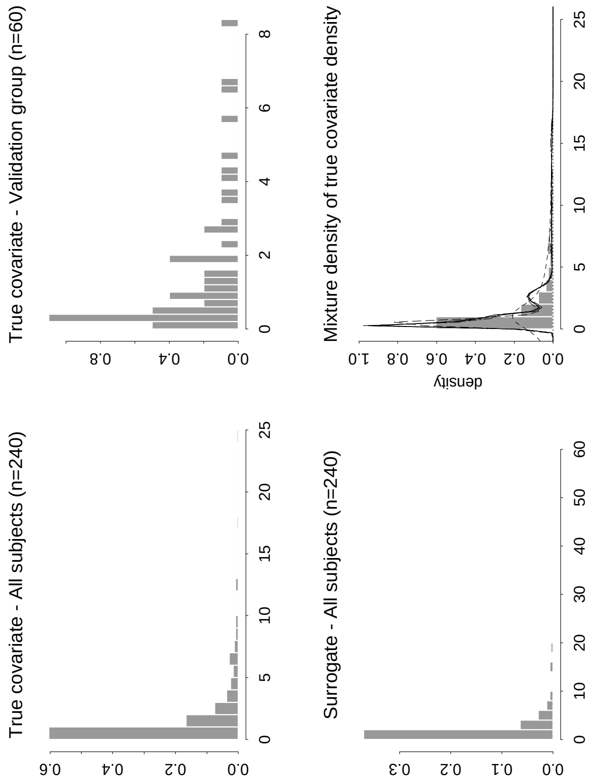
prior distribution of X : $\log N(-0.58, 1.16)$

measurement model : $\log U \sim N(\log X, \theta^{-1})$ for different values of θ

logistic disease model with $\beta_0 = -3.09, \beta_1 = 0.5$

50 replications

Analysis	mixture	P.L.	N.P.	mixture	P.L.	N.P.
true θ		4.0			1.77	
$\overline{\theta}$	3.57 (0.64)	-	-	1.64 (0.27)	-	-
$\overline{\beta_1}$	0.53 (0.14)	0.47	0.47	0.56 (0.16)	0.49	0.49
MSE	0.021	0.034	0.027	0.028	0.067	0.048



Extensions

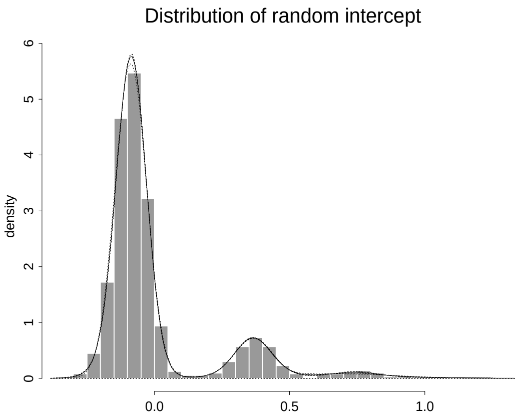
Use of mixture priors with discrete or continuous components and unknown k as distributions of latent variables in different contexts :

- random effects introduced to model correlated data, $e.g$ clustered binary outcomes, spatially dependent counts ...
- frailties in survival data analysis, $e.g$ discrete frailty models, with unknown number of values = alleles following a genetic model, are being developped to analyse familial survival data

  collaborative project with D. Thomas, USC

Analysis of correlated binary outcomes : pregnancy outcomes amongst 1473 women

Logistic mixed model with random intercept



Further details

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

Email addresses :
richardson@vjf.inserm.fr
P.J.Green@bristol.ac.uk

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