Inférence [bayésienne] dans les mélanges de lois

Christian P. Robert

Université Paris Dauphine

www.ceremade.dauphine.fr/∼xian

1. Bayesian inference

Works conditional on data $x_1, \ldots, x_n$

Likelihood $L(\theta|x_1, \ldots, x_n)$ plus prior $\pi(\theta)$ σ-finite measure on Θ lead to *posterior*

$$\begin{aligned} \pi(\theta|x_1, \ldots, x_n) &= \frac{L(\theta|x_1, \ldots, x_n)\pi(\theta)}{\int L(\theta|x_1, \ldots, x_n)\pi(\theta)d\theta} \\ &\propto L(\theta|x_1, \ldots, x_n)\pi(\theta) \end{aligned}$$

Interpretation

- reference measure in a likelihood/maxentropy/fiducial analysis
- penalization factor for the likelihood
- measure of information (cf. Kullback, Fisher)
- objective quantity for classical D.T. optimality concepts (admissibility, invariance)
- summary of previous experiment/anterior information
- subjective beliefs of experimenter/statistician
- philosophical representation of unknown quantities as random
- necessary tool to condition on observables
- artefact whose effect on inference can be evaluated a posteriori

[CPR, The Bayesian Choice (2nd edition), April 2001]

Choice of the prior distribution...

Often [wrongly] perceived as a major **difficulty** with the Bayesian approach

- influence on resulting inference goes decreasing with sample size
- effect of the choice of π can be assessed by [MCMC based] robustness studies
- existence of common knowledge/reference/"noninformative" priors to fall upon
- necessary medium to deal with weakly identifiable structures and heavily multimodal likelihoods [culls extremes & boundaries]

The special and complicated case of mixtures...

Mixture models

$$y_i \sim f(y) = \sum_{j=1}^{k} p_j f(y|\theta_j) \ , \quad i = 1, \cdots, n$$

$$p_i \geq 0 \ \text{ with } \ \sum_{j=1}^{k} p_j = 1$$

$$\text{eg } \ f(y|\theta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{1}{2\sigma_j^2}(y - \mu_j)^2 \right\}$$

Extensions

(i) Mixture of linear regressions

$$y|x \sim \sum_{j=1}^{k} p_j \mathcal{N}(x^t \beta_j, \sigma_j^2)$$

(ii) Mixtures of logistic regressions

$$P(y = 1|x) = \sum_{j=1}^{k} p_j \frac{\exp\{x^t \beta_j\}}{1 + \exp\{x^t \beta_j\}}$$

(iii) Mixtures of Poisson regressions

$$y|x \sim \sum_{j=1}^{k} p_j \, \mathcal{P}(\exp(x^t \beta_j)) \ ,$$

Classes of priors

- possible symmetric empirical Bayes priors

$$\underline{p} \sim \mathcal{D}(\gamma, \ldots, \gamma), \quad \theta_i \sim \mathcal{N}(\hat{\mu}, \hat{\omega}\sigma_i^2), \quad \sigma_i^{-2} \sim \mathcal{G}a(\nu, \hat{\epsilon}\nu)$$

  which can be replaced with hierarchical priors

  [Diebolt & Robert (1990), Richardson & Green (1997)]

- independent improper priors on θ_j's **prohibited**, thus standard "flat" and Jeffreys priors impossible to use (except with the *exclude-empty-component* trick)

  [Diebolt & Robert (1990), Wasserman (1999)]

- dependent weakly informative priors

$$\underline{p} \sim \mathcal{D}(k, \ldots, 1), \quad \theta_i \sim \mathcal{N}(\theta_{i-1}, \zeta\sigma_{i-1}^2), \quad \sigma_i \sim \mathcal{U}([0, \sigma_{i-1}])$$

[Mengersen & Robert (1996), Roeder & Wasserman (1997)]

- reference priors

$$\underline{p} \sim \mathcal{D}(1, \ldots, 1), \quad \theta_i \sim \mathcal{N}(\mu_0, (\sigma_i^2 + \tau_0^2)/2), \quad \sigma_i^2 \sim \mathcal{C}^+(0, \tau_0^2)$$

[Moreno & Liseo (1999)]

- more ad-hoc priors for beta mixtures

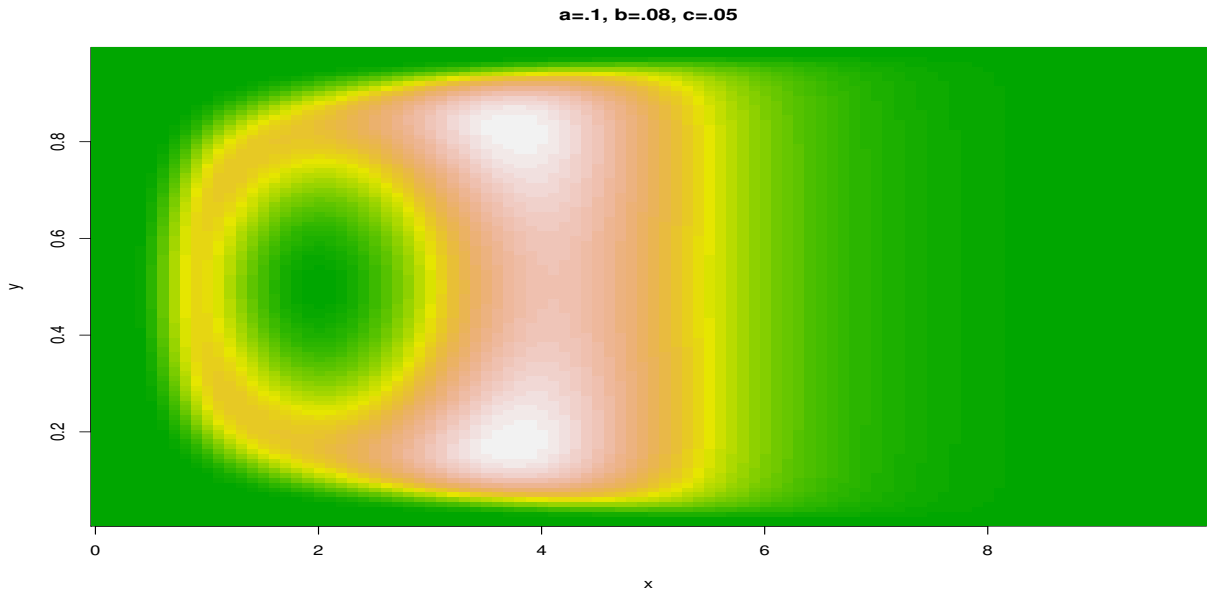
$$p_0 + (1-p_0)\sum_{i=1}^{k} \frac{\omega_i}{\sum_\ell \omega_\ell} \mathcal{B}e(\alpha_i\epsilon_i, \alpha_i(1-\epsilon_i))$$

in goodness-of-fit tests, since

$$p_0 \sim \mathcal{U}([0,1]), \qquad \omega_i \sim \mathcal{B}e(1, i+1),$$

and

$$(\alpha_i, \epsilon_i) \sim [1 - \exp\{-(\alpha_i - 2)^2 - (\epsilon_i - .5)^2\}]$$
$$\exp[-1/(\alpha_i^2 \epsilon_i (1 - \epsilon_i)) - 0.2\alpha_i^2/2]$$



[Robert & Rousseau, 2001]

2. Resolving the identifiability issues

A crucial feature:

$$\sum_{j=1}^{k} p_j f(y|\theta_j)$$

is invariant to relabelling of the components

Baring some identifiability constraint, under a symmetric prior, posterior is also symmetric in its components/label invariant

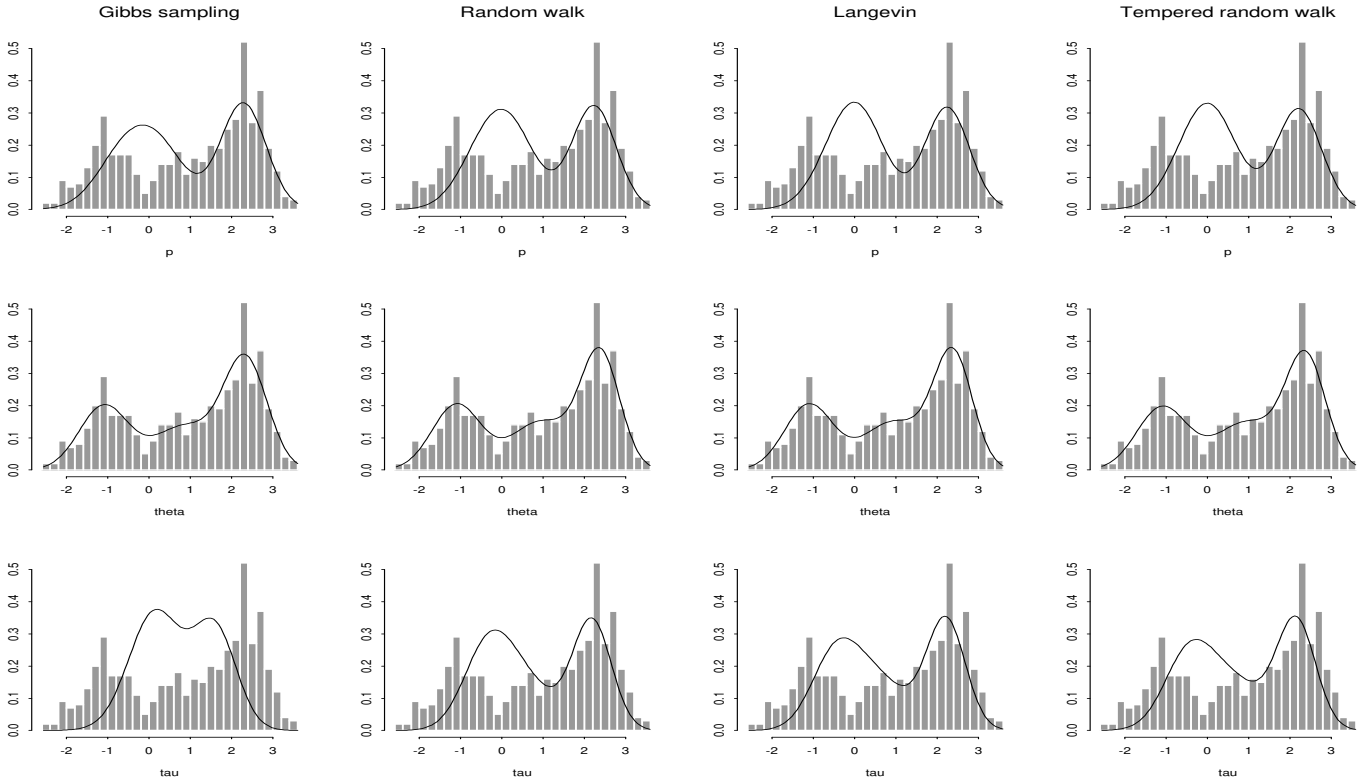
Identifiability can be **imposed** by ordering, like

$$p_1 < p_2 < \cdots < p_k \qquad \text{or} \qquad \mu_1 < \mu_2 < \cdots < \mu_k \qquad \text{or...}$$

but...

- modifies the prior distribution quite strongly for large #'s of components *[e.g. the marginal on θ_1 or θ_k may be quite different from $\mathcal{N}(0,1)$]*
- slows down MCMC samplers, sometimes to the extent of creating trapping states *[example of the "inverted order"]*

- induces biases *[may force components apart or increase the variance of the first component, for instance]*



Not imposing identifiability may result in harder exploration of the posterior distribution **and/or** resulting inference, since, under symmetric prior, all marginals are the same, thus Bayes posterior means are identical for all components

Simply forces us to think harder and better!

Loss functions for mixture estimation...

[Celeux, Hurn & Robert (2000)]

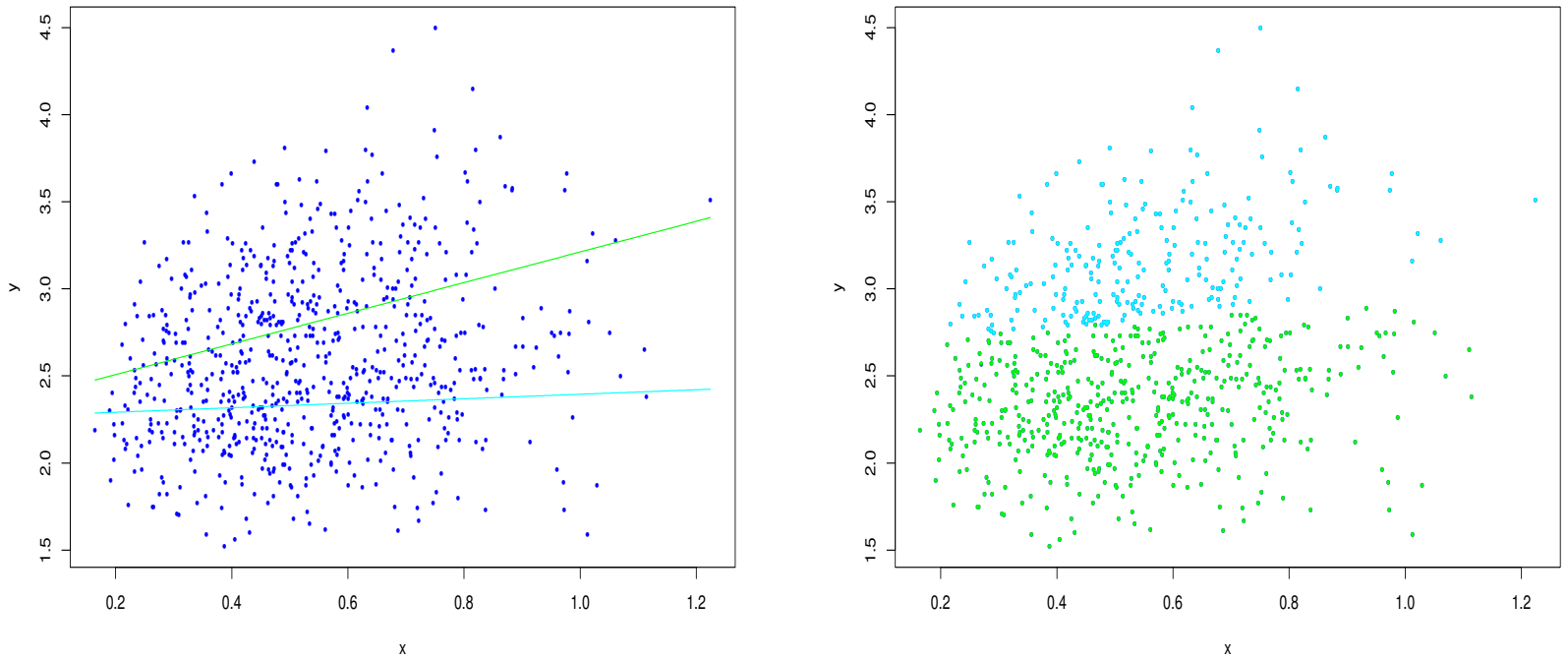
Global loss function that considers distance between predictives

$$L(\xi, \hat{\xi}) = \int_{\mathcal{X}} f_\xi(x) \log \left\{ \frac{f_\xi(x)}{f_{\hat{\xi}}(x)} \right\} dx$$

eliminates the labelling effect

Similar solution for estimating clusters through allocation variables z

$$L(z, \hat{z}) = \sum_{i<j} \left(\mathbb{I}_{[z_i=z_j]}(1 - \mathbb{I}_{[\hat{z}_i=\hat{z}_j]}) + \mathbb{I}_{[\hat{z}_i=\hat{z}_j]}(1 - \mathbb{I}_{[z_i=z_j]}) \right) .$$



3. Exploring the whole posterior distribution

Thanks to the MCMC revolution,

[Cappé & Robert, JASA, 2000,
Robert & Casella, Monte Carlo Statistical Methods, 1999]

many ways of simulating from the posterior distributions

$$\pi(\theta|x_1,\ldots,x_n) \qquad \text{or} \qquad \pi(\theta,z_1,\ldots,z_n|x_1,\ldots,x_n)$$

like

$$\prod_{i=1}^{n}\left(\sum_{j=1}^{k} p_j\lambda_j e^{-\lambda_j x_i}\right)\prod_{j=1}^{k} e^{-\lambda_j}$$

or

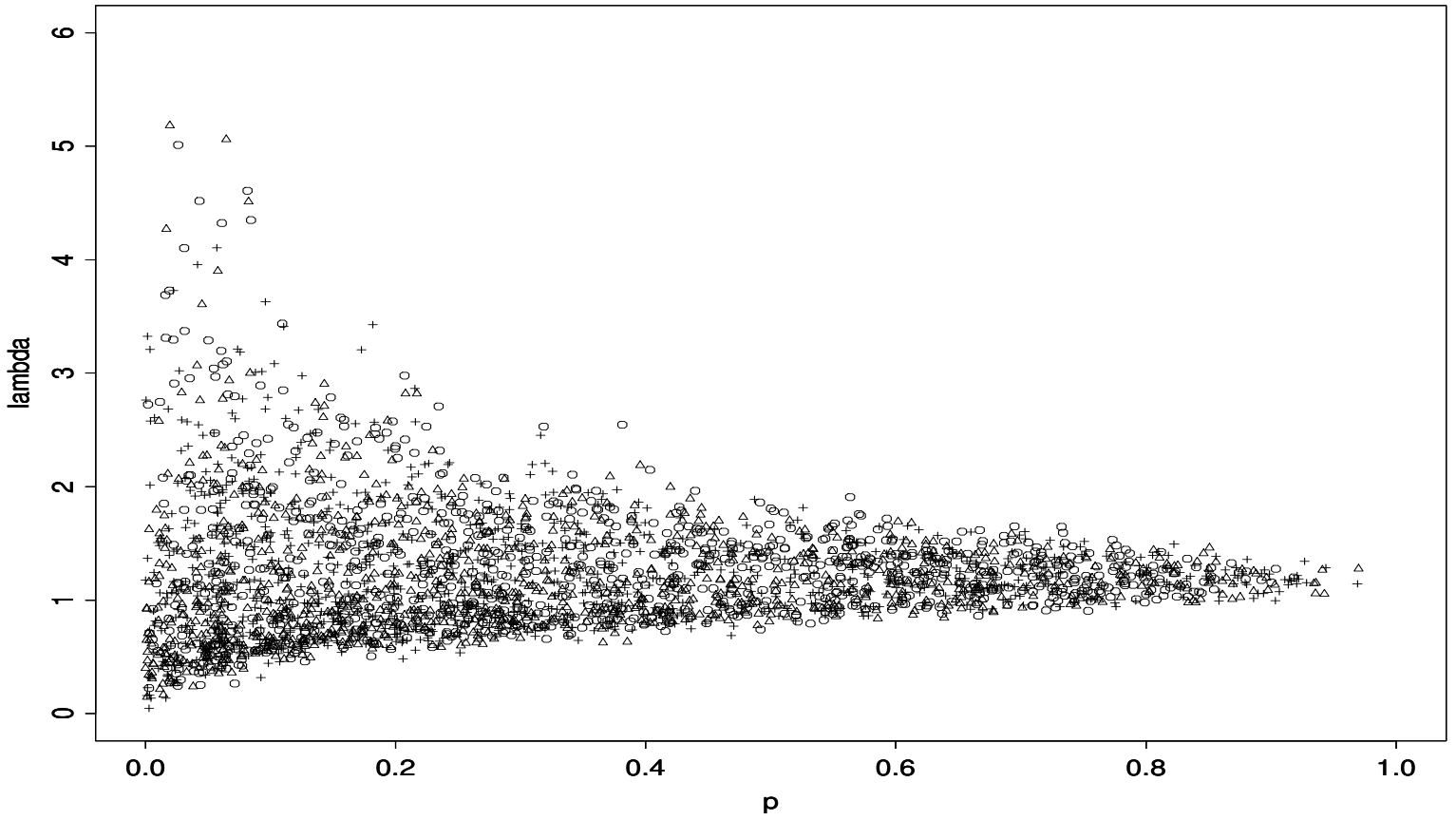
$$\prod_{i=1}^{n}\left(\sum_{j=1}^{k} p_j \frac{e^{-(x_i-\theta_j)^2/2\tau_j^2}}{\tau_j^2}\right)\prod_{j=1}^{k}\frac{e^{-\theta_j^2/2\xi\tau_j^2-\delta/\tau_j^2}}{\tau_j^3}$$

- *Gibbs*
  - Natural completion by latent indicators
  - Slow moves (not always for exponential mixtures)
  - Large increase in dimension, order $\mathrm{O}(k)$
  - Good theoretical properties *(Duality principle)*
- *Random walk Metropolis–Hastings*
  - Available (Normal vs. Cauchy move)
  - Calibrated against acceptance rate
  - Depends on parameterisation
    $\lambda_j \to \log \lambda_j \qquad p_j \to \log(p_j/1 - p_k)$

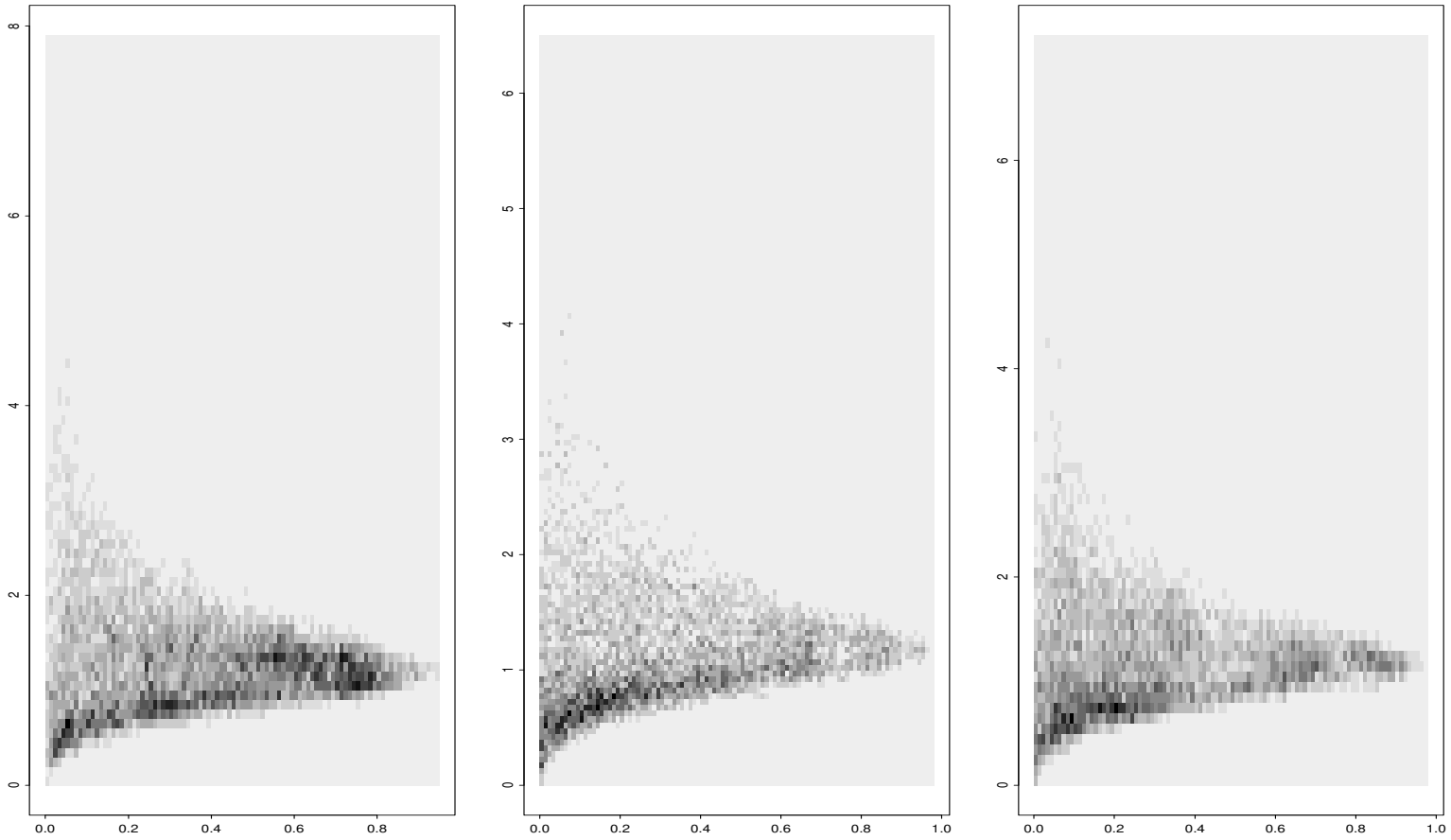
Example — Exponential mixture model

$$\sum_{i=0}^{k} p_i \, \mathcal{E}xp(\lambda_i),$$

with $p_0 + \ldots + p_k = 1$ and $\lambda_i > 0$ $(0 \leq i \leq k)$.



Joint distribution



Marginal distributions

Example — Beta mixture model

$$p_0 + (1-p_0)\sum_{i=1}^{k} \frac{\omega_i}{\sum_\ell \omega_\ell} \mathcal{B}e(\alpha_i\epsilon_i, \alpha_i(1-\epsilon_i))$$

Gibbs sampler

- $p_0 \sim \mathcal{B}e(n_0+1, n-n_0+1)$
- since

$$\pi(\omega_1,\ldots,\omega_k) \propto \prod_{i=1}^{k} \omega_i^{n_i} \left\{\sum_\ell \omega_\ell\right\}^{n-n_0}$$

  use of a $\mathcal{B}e(n_i+1, j+1)$ proposal on each ω_j with Metropolis–Hastings acceptance

[Acceptance 0.2157]

- since

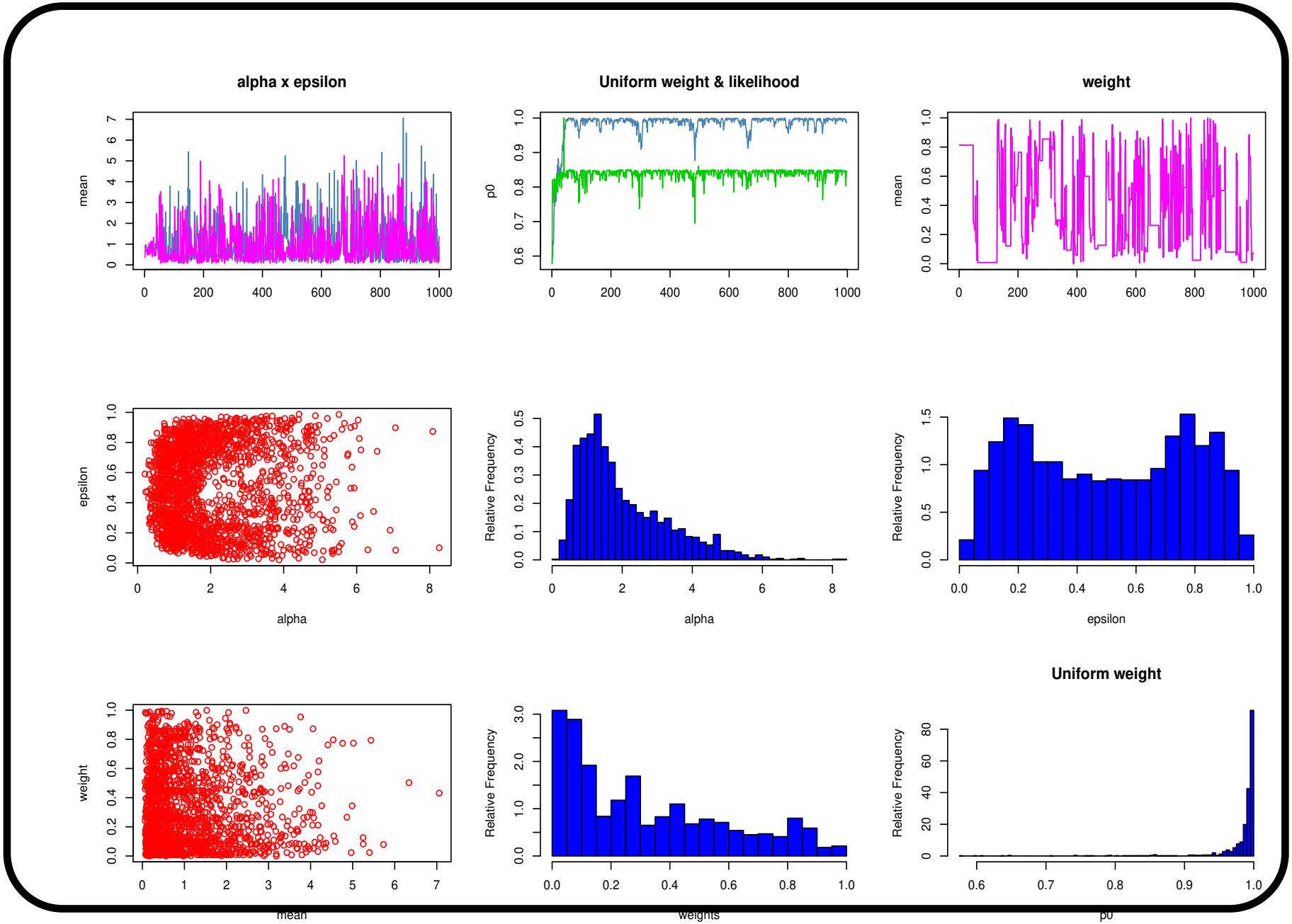
$$\pi(\alpha_i,\epsilon_i) \propto \left\{\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i\epsilon_i)\Gamma(\alpha_i(1-\epsilon_i))}\right\}^{n_i} \prod_{z_t=i} x_t^{\alpha_i\epsilon_i-1}(1-x_t)^{\alpha_i(1-\epsilon_i)-1}\pi(\alpha_i,\epsilon_i)$$

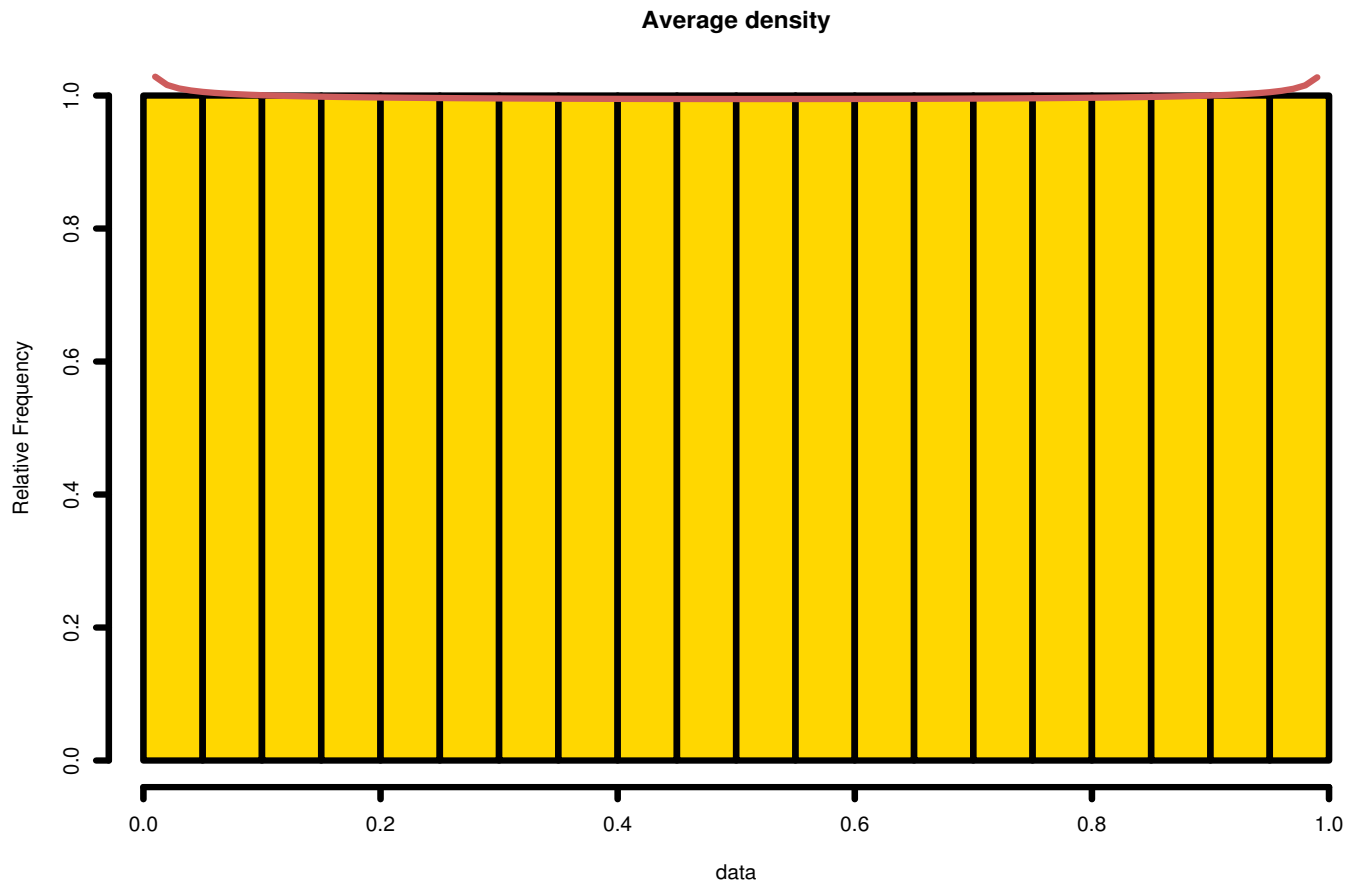
use of a proposal

$$\alpha_i \sim \mathcal{G}a((n_i+1)/2,\hat{\beta}_i) \qquad \log(\epsilon_i/(1-\epsilon_i)) \sim \mathcal{N}(\log(\epsilon_i/(1-\epsilon_i)),,1)$$

and Metropolis–Hastings acceptance step

[Acceptances 0.6502 and 0.3733]


alpha x epsilon
mean
Uniform weight & likelihood
p0
weight
mean
epsilon
alpha
Relative Frequency
alpha
Relative Frequency
epsilon
Uniform weight
weight
mean
Relative Frequency
weights
Relative Frequency
p0


Average density
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
data

Alternatives approaches:

- Importance sampling via partitioning

  [Casella, Robert & Wells (2000)]

- Perfect sampling

  [Casella, Mengersen, Robert & Titterington (2000)]

- Auxilliary variables

  [Robert & Roberts (2001)]

4. Estimating [not testing] the number of components

In

$$y_i \sim f(y) = \sum_{j=1}^{k} p_j f(y|\theta_j) \ , \quad i = 1, \cdots, n$$

what if k is **unknown**?!

- weak identifiability of mixtures
- insolvable "philosophical" problem unless k has a proper intrinsic meaning [and even so...]
- hence testing *per se* is impossible: the data cannot distinguish between k and $k + h$ [unless guided by a firm hand!]
- the choice of a prior on k $\pi(k)$ is thus necessary to translate the degree of details required [equivalence with penalizing factors in likelihood analysis]

Multiplicity of technical solutions:

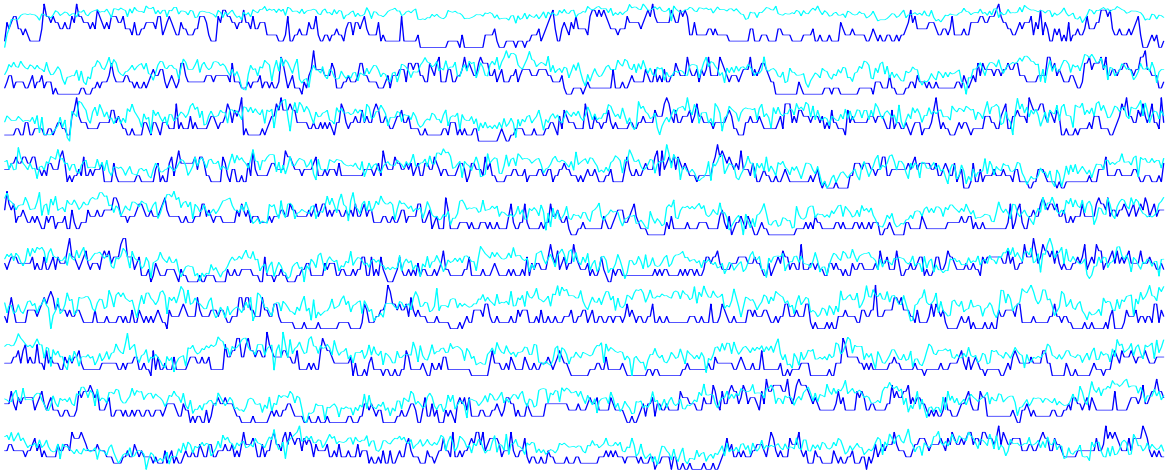
- Saturated models with n mostly empty components
- Reversible jump MCMC techniques for exploration of most models

  [Green (1995); Richardson & Green (1997)]
- Birth and death and other jump processes

  [Preston (1976); Ripley (1977); Stephens (1999,2000)]

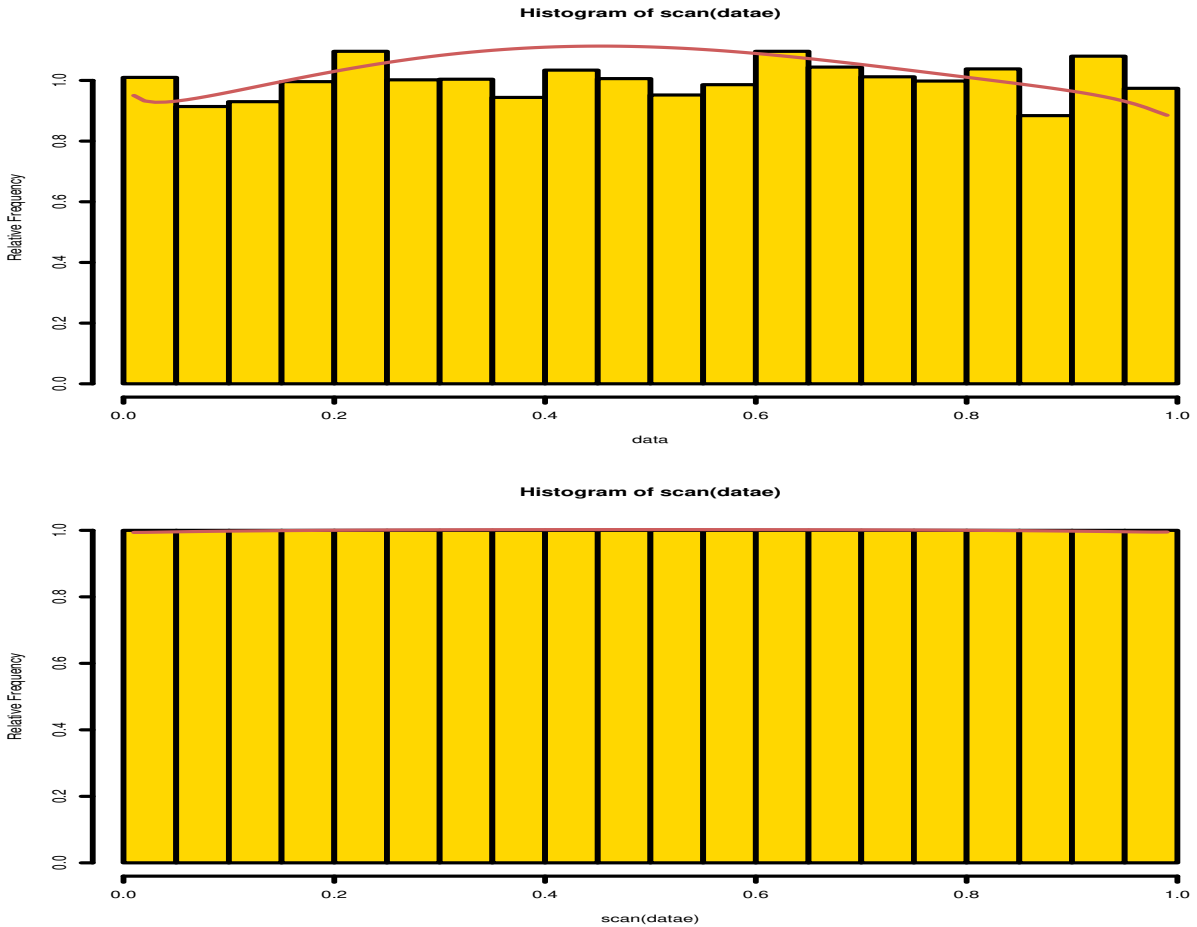
Inference with varying k:

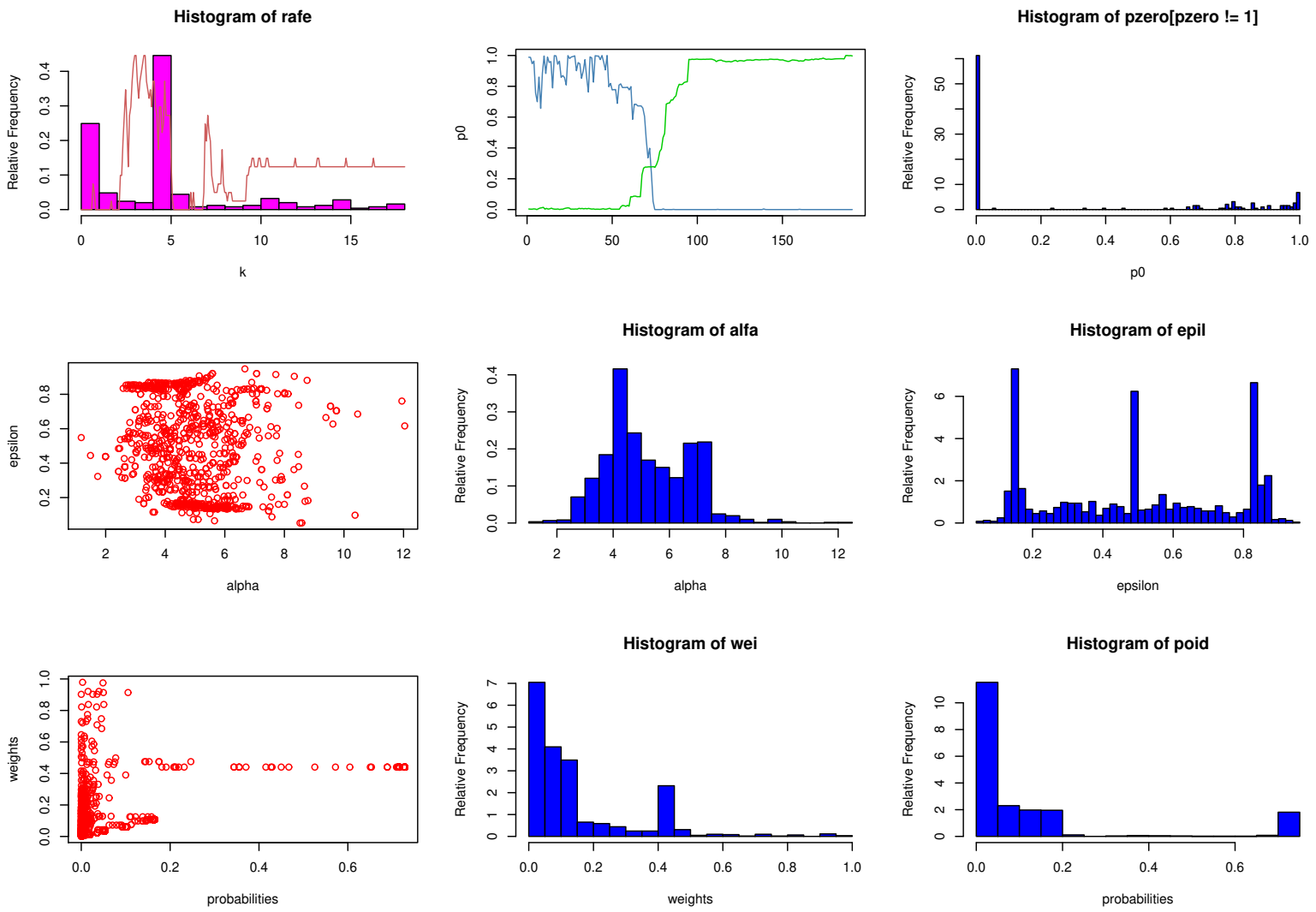
- Little difference from the fixed k setting
- Inference must be conditional on k
- General principle in Bayesian model choice: parameters appearing in different models must be considered as separate entities
- Inference on k through posterior probabilities and predictive plots of the regression lines



Results on a large uniform sample for the beta mixture:

- k never estimated as 0
- p_0 very small
- likelihood widely different from 1
- curve almost flat


Histogram of scan(datae)
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
data
Histogram of scan(datae)
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
scan(datae)

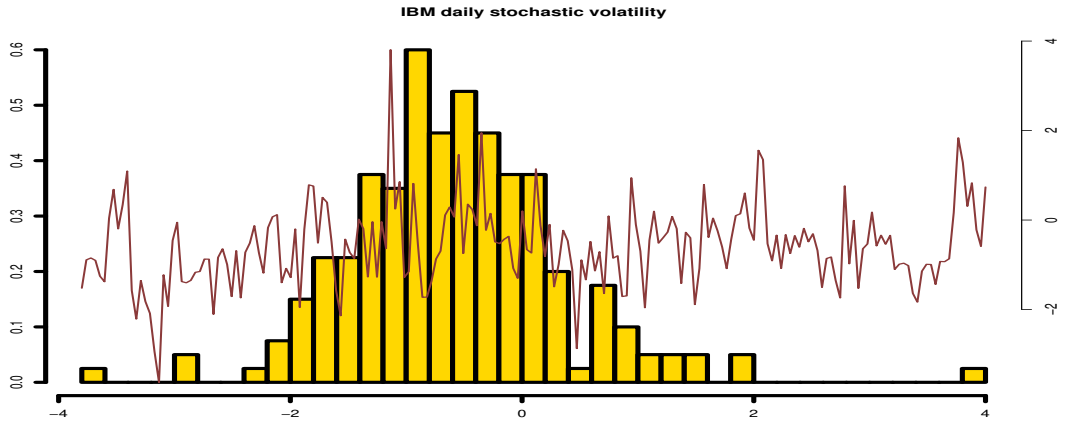

Histogram of rafe
Relative Frequency
k
p0
Histogram of pzero[pzero != 1]
Relative Frequency
p0
epsilon
alpha
Histogram of alfa
Relative Frequency
alpha
Histogram of epil
Relative Frequency
epsilon
weights
probabilities
Histogram of wei
Relative Frequency
weights
Histogram of poid
Relative Frequency
probabilities

5. Exploring more challenging structures

Introduce more advanced models by way of additional latent variables with possible dependence

Example — Hidden Markov models

$$\begin{array}{rcccrcl} z_1 & \sim & \pi & \cdots & z_i & \sim & p_{z_{i-1}z_i} \\ x_1|z_1 & \sim & \mathcal{N}(\mu_{z_1}, \sigma^2_{z_1}) & \cdots & x_i|z_i & \sim & \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i}) \end{array}$$



- Very similar to normal mixture but for additional structure which **improves** estimation
- Still allows for flat priors

$$\pi(\mu, \sigma, P) \propto \frac{1}{\sigma_1^k} \exp\left\{ \frac{-1}{2\sigma^2} \sum (\mu_{i+1} - \mu_i)^2 \right\} \times \mathbb{I}_{\sigma_1 > \cdots > \sigma_k}$$

[Robert & Titterington (1997)]

- Gibbs implementation straightforward
  1. Generate "missing data"

$$p(z_i = j | z_{i-1}, z_{i+1}, \underline{\theta}, \underline{p})$$

2. Generate parameters

$$p_{i\cdot} \sim \mathcal{D}(n_{i1}+1,\cdots,n_{ik}+1)$$

$$\mu_i \sim \mathcal{N}\left(\frac{n_i\sigma_i^{-2}\bar{x}_i+\alpha_{i-1}\mu_{i-1}+\alpha_{i+1}\mu_{i+1}}{n_i\sigma_i^{-2}+\alpha_{i-1}+\alpha_{i+1}}, \frac{}{(n_i\sigma_i^{-2}+\alpha_{i-1}+\alpha_{i+1})}\right)$$

$$\sigma_i^2 \sim \mathcal{IG}\left(\frac{n_i-1}{2}, \frac{n_i(\bar{x}_i-\mu_i)^2+s_i^2}{2}\right) \times \mathbb{I}_{\sigma_{i-1}<\sigma_i<\sigma_{i+1}}$$

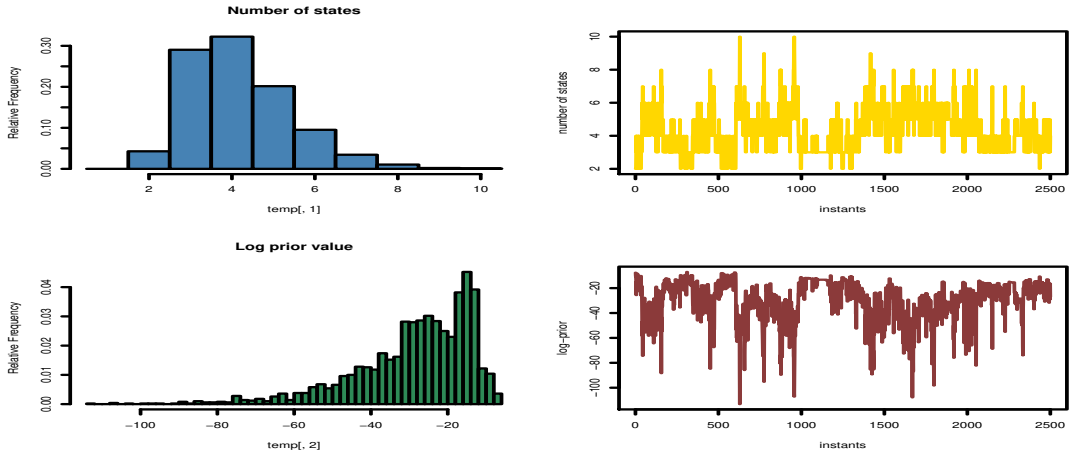
[Celeux, Diebolt & Robert (1993)]

- Non-Gibbsic implementation also possible, without the missing states, thanks to *forward-backward* formulae
- Estimation of k possible via reversible jump

  [Robert, Rydén & Titterington (1999)]

  and other jump process methods

  [Cappé, Robert & Rydén (2001)]



Other latent variable models and hidden structures:

- Hidden semi-Markov models
- Switching ARMA models
- Stochastic volatility and ARCH models

6. Solving optimization problems

Role of maximum a posteriori estimation in Bayesian inference

$$\theta = (\theta_1, \theta_2) \in \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \sim p(\theta)$$

especially *when posterior means are useless* but difficulty with marginal MAP (MMAP) estimates because nuisance parameters must be integrated out

$$\theta_1^{MMAP} = \arg_{\boldsymbol{\Theta}_1} \max p(\theta_1 | \mathbf{y})$$

where

$$p(\theta_1 | \mathbf{y}) = \int_{\boldsymbol{\Theta}_2} p(\theta_1, \theta_2 | \mathbf{y}) \, d\theta_2$$

1. If integration possible in closed-form, use
   Expectation-Maximization (EM) algorithm

   [Dempster et al. (1977)]

   Deterministic algorithm which depends on initialization and is limited to certain classes of models.

   Stochastic variants like Stochastic EM (SEM) or Monte Carlo EM (MCEM)

   [Celeux & Diebolt (1985), Wei & Tanner (1991)]

   Parameter of interest always updated deterministically in the M step

2. "Standard" (and Markov chain) Monte Carlo: draw random samples from the joint posterior distribution

$$p\left(\theta_1, \theta_2 | \, \mathbf{y}\right)$$

or MCMC (approximate, dependent) sample

$$\left\{ \left(\theta_1^{(i)}, \theta_2^{(i)} \right) ; i = 1, \ldots, N \right\}$$

and discard nuisance parameters. More suited to integration than to optimization.

3. Simulated annealing (SA) for maximizing $p(\theta_1|\mathbf{y})$

   Non-homogeneous variant of MCMC for global optimization: invariant distribution at iteration i proportional to

$$p^{\gamma(i)}(\theta_1|\mathbf{y}) ,$$

   $\gamma(i)$ increasing function diverging at infinity.

   Idea: as $\gamma(i)$ goes to infinity, $p^{\gamma(i)}(\theta_1|\mathbf{y})$ concentrates itself upon the set of global modes.

4. State Augmentation for Marginal Estimation

[Doucet, Godsill & Robert (2001)]

Artificially augmented probability model whose marginal distribution is

$$\overline{p}_{\gamma}\left(\theta_1 | \mathbf{y}\right)$$

via replications of the nuisance parameters:

- Replace θ_2 with γ artificial replications,

$$\theta_2\left(1\right), \ldots, \theta_2\left(\gamma\right)$$

- Treat the $\theta_2\left(j\right)$'s as distinct random variables:

$$q_{\gamma}\left(\theta_1, \theta_2\left(1\right), \ldots, \theta_2\left(\gamma\right) | \mathbf{y}\right) \propto \prod_{k=1}^{\gamma} p\left(\theta_1, \theta_2\left(k\right) | \mathbf{y}\right)$$

- Use corresponding marginal for θ_1

$$\begin{aligned} q_\gamma(\theta_1|\mathbf{y}) &= \int q_\gamma(\theta_1, \theta_2(1), \ldots, \theta_2(\gamma)|\mathbf{y})\, d\theta_2(1) \ldots d\theta_2(\gamma) \\ &\propto \int \prod_{k=1}^{\gamma} p(\theta_1, \theta_2(k)|\mathbf{y})\, d\theta_2(1) \ldots d\theta_2(\gamma) \\ &= \overline{p}_\gamma(\theta_1|\mathbf{y}) \end{aligned}$$

- Build a MCMC algorithm in the augmented space, with invariant distribution

$$q_\gamma(\theta_1, \theta_2(1), \ldots, \theta_2(\gamma)|\mathbf{y})$$

- Use simulated subsequence

$$\left\{\theta_1^{(i)}; i \in \mathbb{N}\right\}$$

as drawn from marginal posterior $\overline{p}_\gamma(\theta_1|\mathbf{y})$

Application to the benchmark galaxy dataset

[Roeder (1992)]

82 observations of galaxy velocities from 3 (?) groups

Algorithm	EM	MCEM	SAME
Mean log-posterior	65.47	60.73	66.22
Std dev of log-posterior	2.31	4.48	0.02