Hmms with variable dimension structures and extensions

Christian P. Robert

Université Paris Dauphine
www.ceremade.dauphine.fr/∼xian

1 Estimating [not testing] the number of components/states

In the mixture model,

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

what if k is **unknown**?!

1.1 Meaning of the question

- 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]

1.2 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)]

1.2.1 Principles of RJMCMC

– Births and deaths are proposed with probabilities β_k and δ_k, respectively, when having k components.

– Acceptance probability for birth move is $\min(A, 1)$, with

$$A = \text{likelihood ratio} \times \frac{\delta_{k+1}}{\beta_k} \times \frac{(1-w)^{k-1}}{b(w,\phi)}.$$

– Acceptance probability for death move is $\min(A^{-1}, 1)$.

1.2.2 Principles of BDMCMC

– New components are born according to a Poisson process with rate λ_k when having k components.

– Each component (w, ϕ) dies with rate

$$d(w, \phi) = \text{likelihood ratio}^{-1} \times \frac{\lambda_k}{k+1} \times \frac{b(w, \phi)}{(1-w)^{k-1}}.$$

1.2.3 More general moves

Local balance in for Markov jump processes in general:

$$\pi(\theta)q(\theta,\theta') = \pi(\theta')q(\theta',\theta)$$

For birth-death, split-merge moves etc.:

$$\pi(k)\pi(\theta_k|k)L(\theta_k)\times\lambda_k b(u_k)J^{-1} = \pi(k+1)\pi(\theta_{k+1}|k+1)L(\theta_{k+1})\times d(\theta_{k+1},\theta_k)$$

1.2.4 Comparison of mixing properties

– RJMCMC works poorly if

$$A = \text{likelihood ratio} \times \frac{\delta_{k+1}}{\beta_k} \times \frac{(1-w)^{k-1}}{b(w,\phi)}$$

is small.

– If A is small, then

$$\begin{aligned} d(w,\phi) &= \text{likelihood ratio}^{-1} \times \frac{\lambda_k}{k+1} \times \frac{b(w,\phi)}{(1-w)^{k-1}} \\ &\approx N \times \frac{1}{k+1} \times A^{-1} \end{aligned}$$

is large and BDMCMC works poorly.

1.2.5 RJMCMC→BDMCMC

Rescaling time

- In discrete-time RJMCMC, let the time unit be $1/N$, put $\beta_k = \lambda_k/N$ and $\delta_k = 1 - \lambda_k/N$.
- As $N \to \infty$ each birth proposal will be accepted, and having k components births occur according to a Poisson process with rate λ_k.

– As $N \to \infty$, a component (w, ϕ) dies with rate

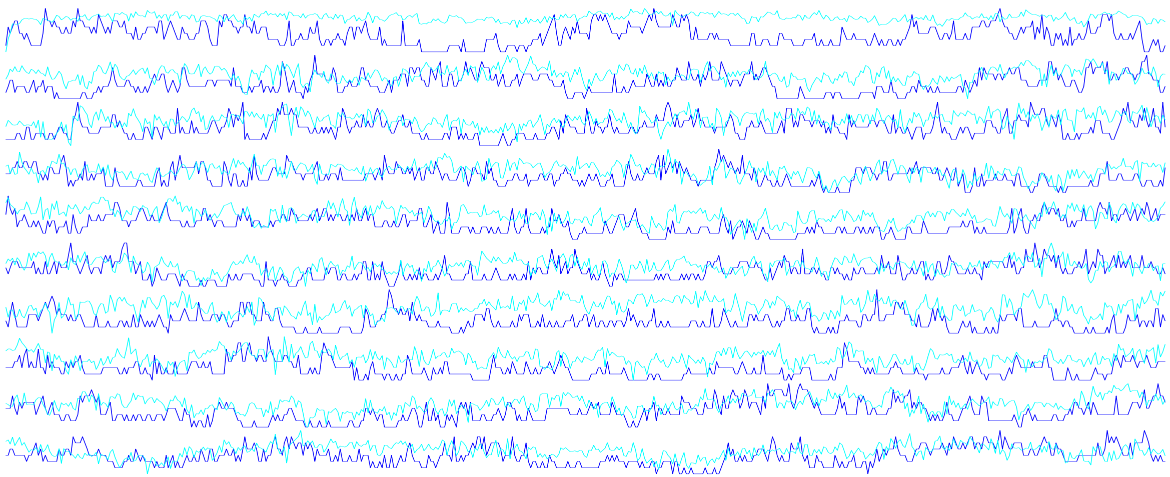
$$\begin{aligned}\lim_{N\to\infty} N\delta_{k+1} \times \frac{1}{k+1} \times \min(A^{-1}, 1) \\ &= \lim_{N\to\infty} N\frac{1}{k+1} \times \text{likelihood ratio}^{-1} \\ &\quad \times \frac{\beta_k}{\delta_{k+1}} \times \frac{b(w,\phi)}{(1-w)^{k-1}} \\ &= \text{likelihood ratio}^{-1} \times \frac{\lambda_k}{k+1} \times \frac{b(w,\phi)}{(1-w)^{k-1}}.\end{aligned}$$

Hence "RJMCMC→BDMCMC". This holds more generally.

1.3 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

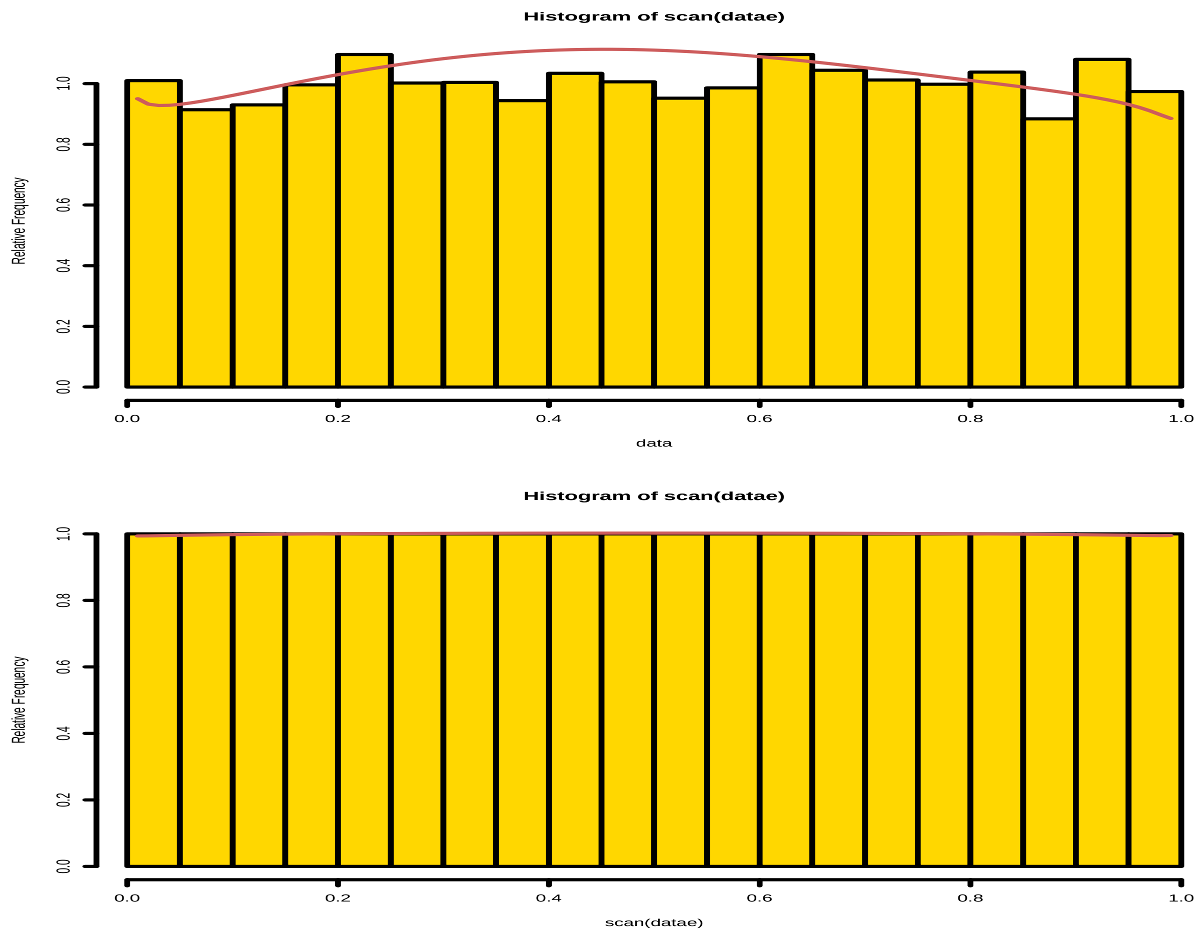
- Inference on k through posterior probabilities and predictive plots of the regression lines

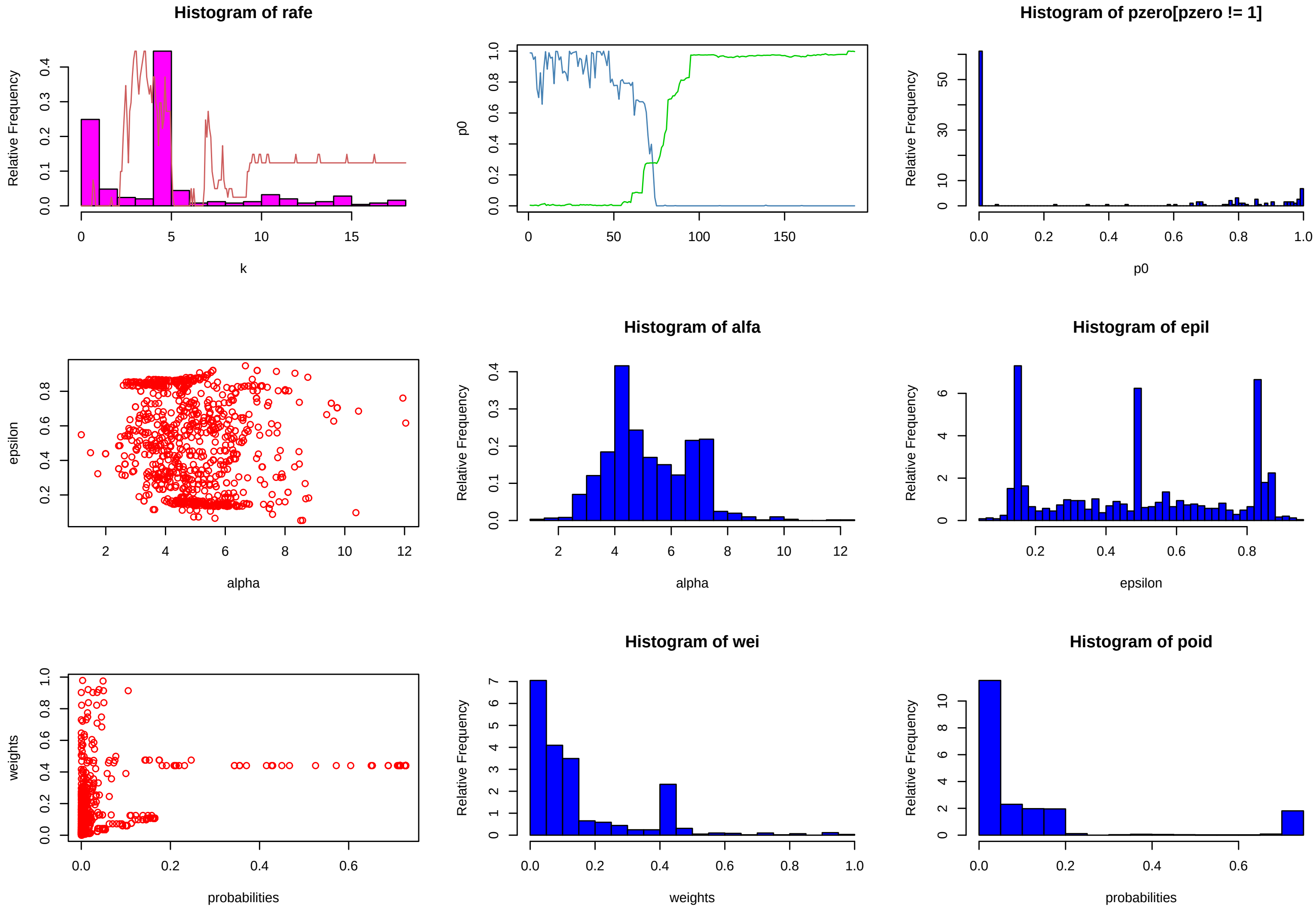


Results on a large uniform sample for the beta mixture:

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

- 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
scan(datae)

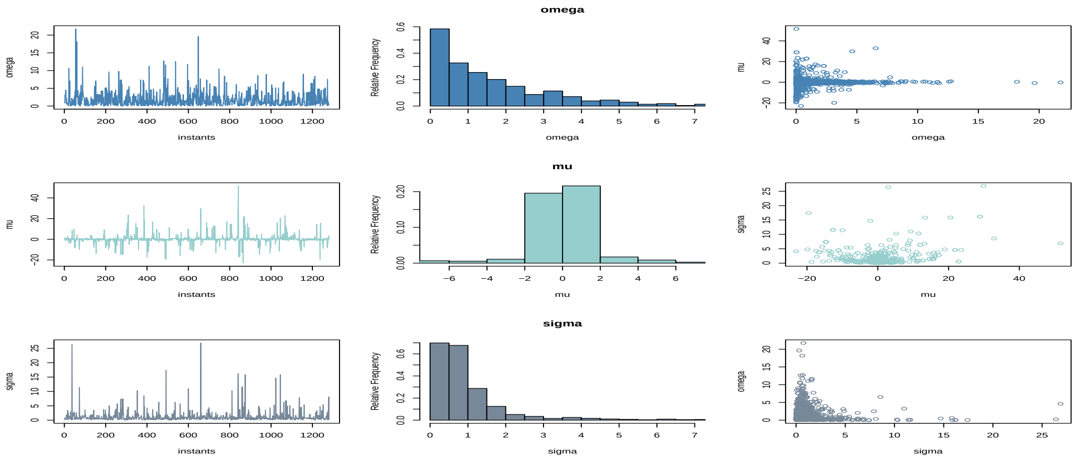

Histogram of rafe
Relative Frequency
k
p0
Histogram of pzero[pzero != 1]
Relative Frequency
p0
epsilon
alpha
Histogram of alfa
Relative Frequency
alpha
Histogram of epil
Relative Frequency
epsilon
weights
probabilities
Histogram of wei
Relative Frequency
weights
Histogram of poid
Relative Frequency
probabilities

2 Extensions to more challenging structures

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

2.1 Hidden Markov models

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



- 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

$$
\begin{aligned}
p_{i.} &\sim \mathcal{D}(n_{i1}+1, \cdots, n_{ik}+1) \\
\mu_i &\sim \mathcal{N}\left(\frac{n_i\sigma_i^{-2}\bar{x}_i + \alpha_{i-1}\mu_{i-1} + \alpha_{i+1}\mu_{i+1}}{n_i\sigma_i^{-2} + \alpha_{i-1} + \alpha_{i+1}}, \right. \\
&\qquad \left. (n_i\sigma_i^{-2} + \alpha_{i-1} + \alpha_{i+1})\right) \\
\sigma_i^2 &\sim \mathcal{IG}\left(\frac{n_i - 1}{2}, \frac{n_i(\bar{x}_i - \mu_i)^2 + s_i^2}{2}\right) \times \mathbb{I}_{\sigma_{i-1} < \sigma_i < \sigma_{i+1}}
\end{aligned}
$$

[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)]

2.1.1 Split-merge moves for HMMs

– Parametrisation:

$$p_{ij} = \omega_{ij} / \sum_{\ell} \omega_{i\ell}, \quad Y_t | X_t = i \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

– Move to split component j_* into j_1 and j_2:

$$\omega_{ij_1} = \omega_{ij_*} \varepsilon_i, \quad \omega_{ij_2} = \omega_{ij_*}(1 - \varepsilon_i), \quad \varepsilon_i \sim \mathcal{U}(0,1);$$

$$\omega_{j_1 j} = \omega_{j_* j} \xi_j, \quad \omega_{j_2 j} = \omega_{j_* j} / \xi_j, \quad \xi_j \sim \log \mathcal{N}(0,1);$$

similar ideas give $\omega_{j_1 j_2}$ etc.;

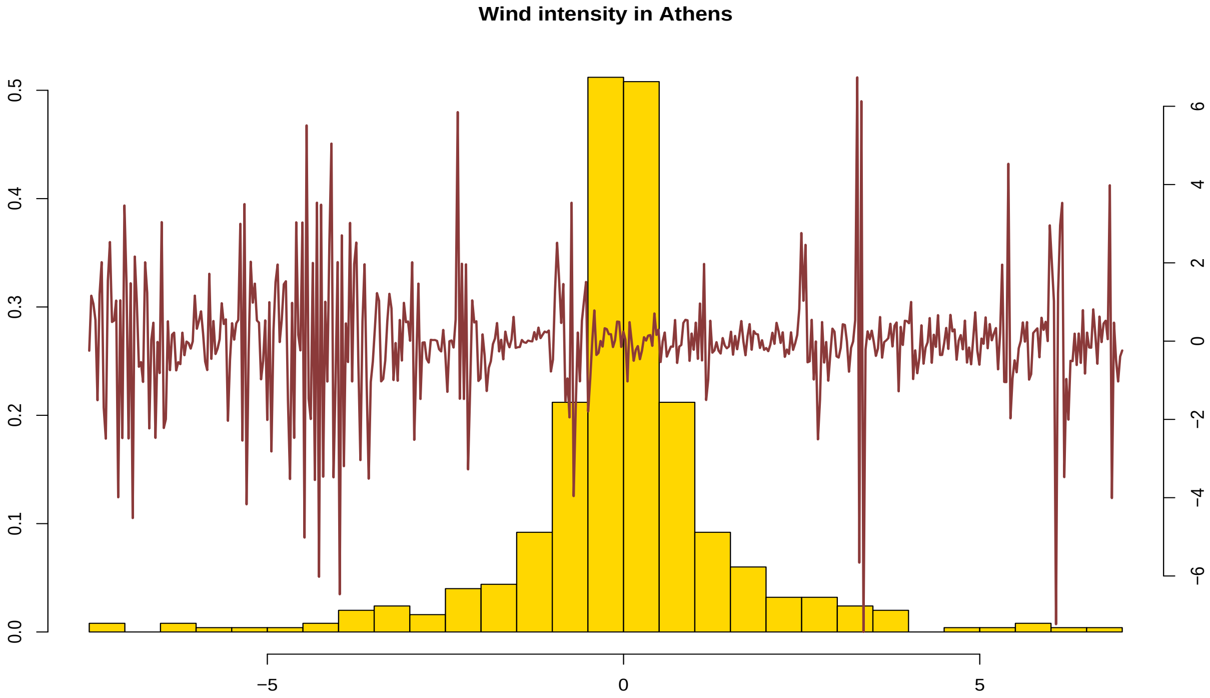
$$\mu_{j_1} = \mu_{j_*} - 3\sigma_{j_*} \varepsilon_\mu, \quad \mu_{j_2} = \mu_{j_*} + 3\sigma_{j_*} \varepsilon_\mu, \quad \varepsilon_\mu \sim \mathcal{N}(0,1);$$

$$\sigma_{j_1}^2 = \sigma_{j_*}^2 \xi_\sigma, \quad \sigma_{j_2}^2 = \sigma_{j_*}^2 / \xi_\sigma, \quad \xi_\sigma \sim \log \mathcal{N}(0,1).$$

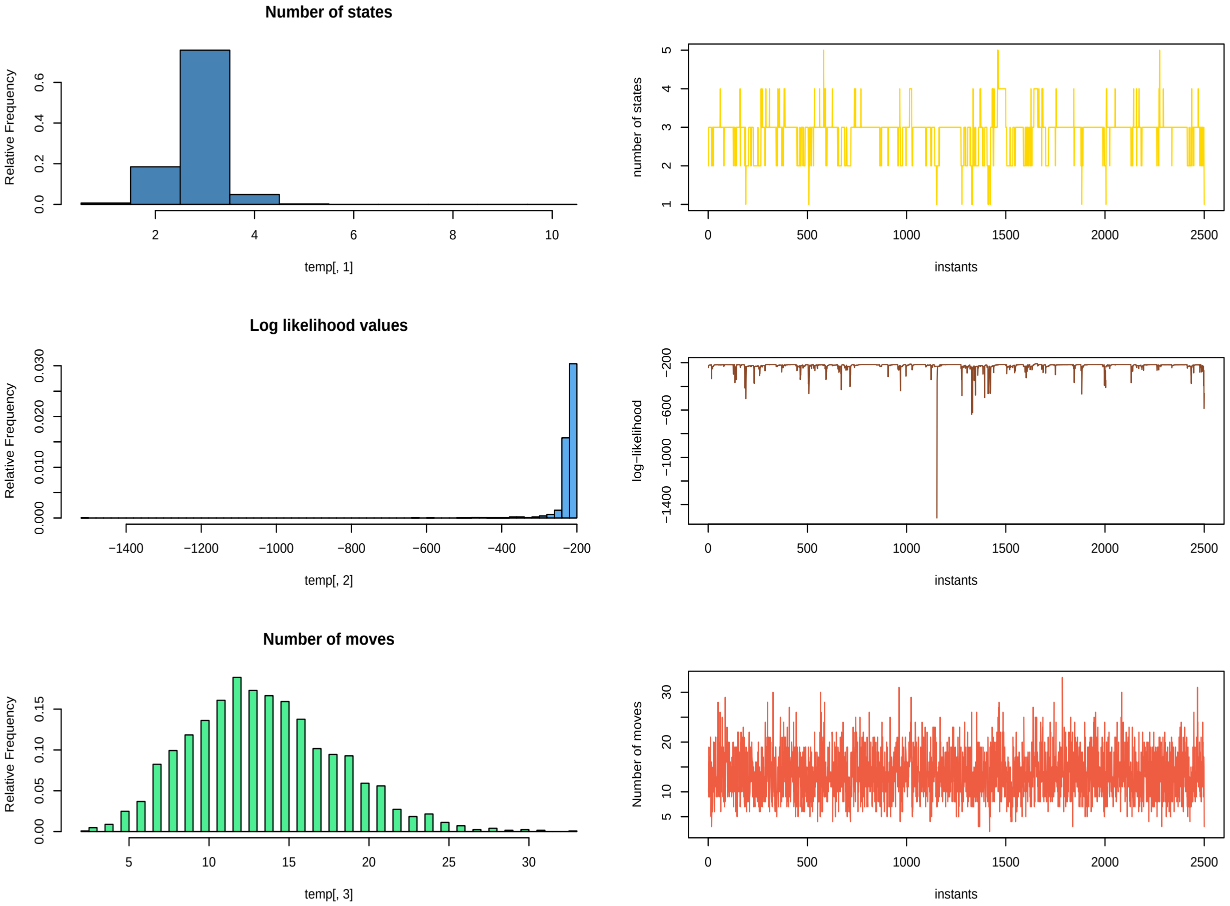
– [Split intensity] $\lambda_{S,k} = k \lambda_B$ [Birth intensity]

– Fixed k moves also used

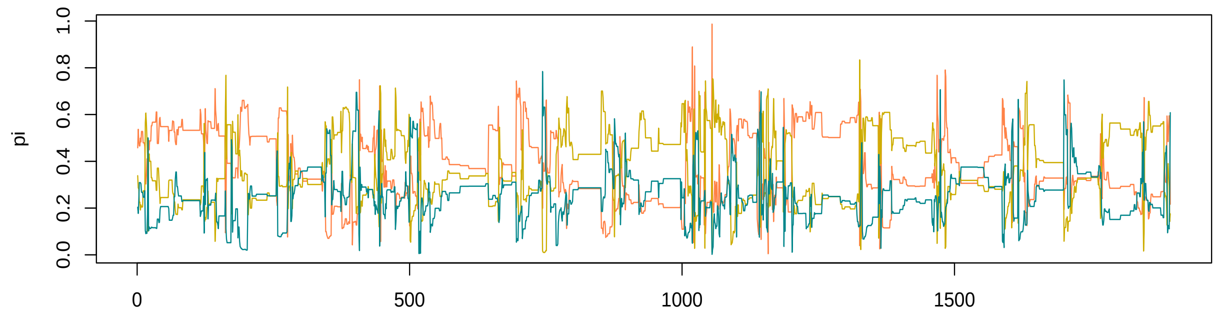
Example :

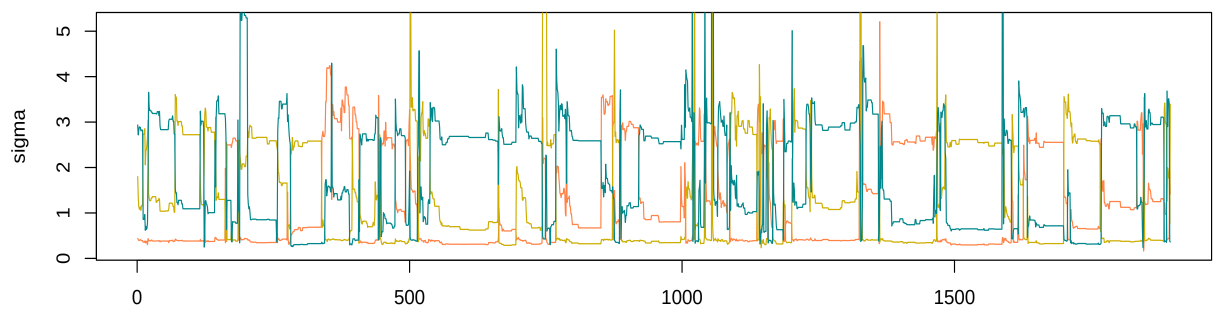


Histogram and rawplot of the dataset

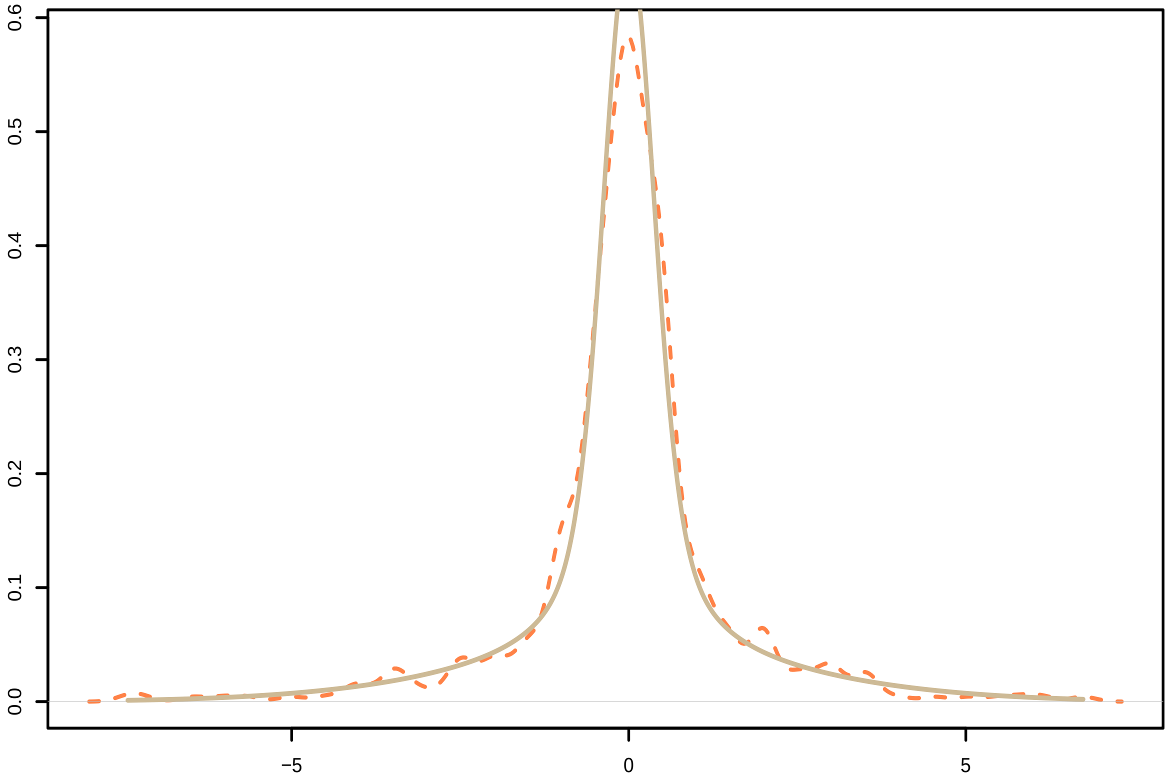


MCMC output on k (histogram and rawplot), number of states, and corresponding likelihood values





MCMC sequence of the parameters of the three components when conditioning on $k = 3$



MCMC evaluation of the marginal density, compared with R **nonparametric** density **estimate.**

2.2 Other latent variable models and hidden structures

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

2.2.1 Hidden semi-Markov models

Example : Ion chanel model

[Hobson, 1999; Carpenter et al., 2001]

Observables

$$\mathbf{y} = (y_t)_{1 \leq t \leq T}$$

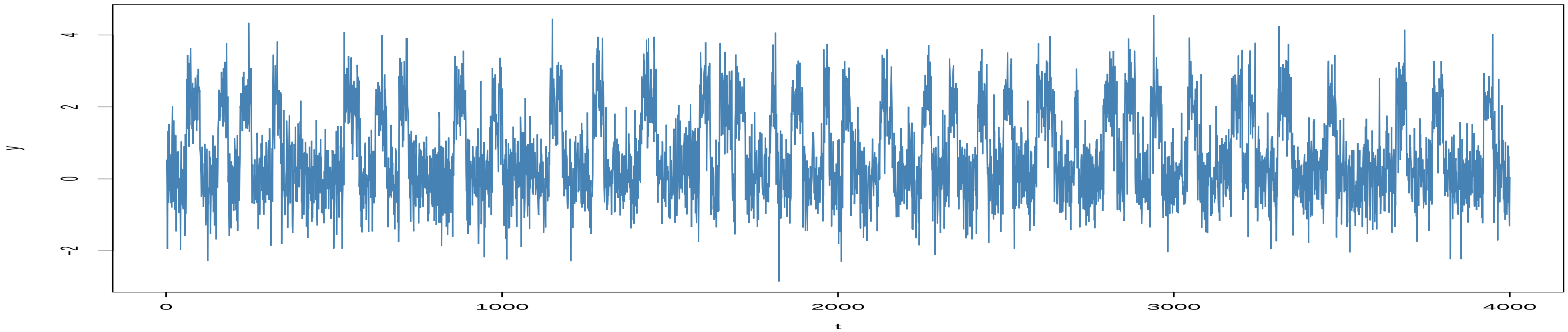
directed by a *hidden* Gamma process $\mathbf{x} = (x_t)_{1 \leq t \leq T}$:

$$y_t | x_t \sim \mathcal{N}(\mu_{x_t}, \sigma^2) \qquad x_t \in \{0, 1\}$$

with durations $(i = 0, 1)$

$$d_j = t_{j+1} - t_j \sim \mathcal{G}a(s_i, \lambda_i)$$

if $x_t = i$ for $t_j \leq t < t_{j+1}$.



Complex likelihood structure with no closed form expression

Prior assumptions

- conjugate normal-gamma prior on the μ's and σ

$$\mathcal{N}(\theta_0, \tau\sigma^2) \times \mathcal{G}(\zeta, \eta)^{-1}$$

- conjugate gamma prior on the λ's

$$\mathcal{G}(\alpha, \beta)$$

- flat prior on the s's on $\{1, \ldots, S\}$

Particle system

Generation of a system of particles

$$(\omega^{(j)}, \mathbf{x}^{(j)})_j \qquad (j = 1, \ldots, J)$$

where

$$\omega = (\mu_0, \mu_1, \sigma, \lambda_0, \lambda_1, s_0, s_1)$$

based on a proposal/instrumental/importance distribution

$$\pi(\omega|\mathbf{y}, \mathbf{x}) \times \pi_H(\mathbf{x}|\mathbf{y}, \omega)$$

where π_H full conditional of a **fitted hidden Markov** model with transition matrix

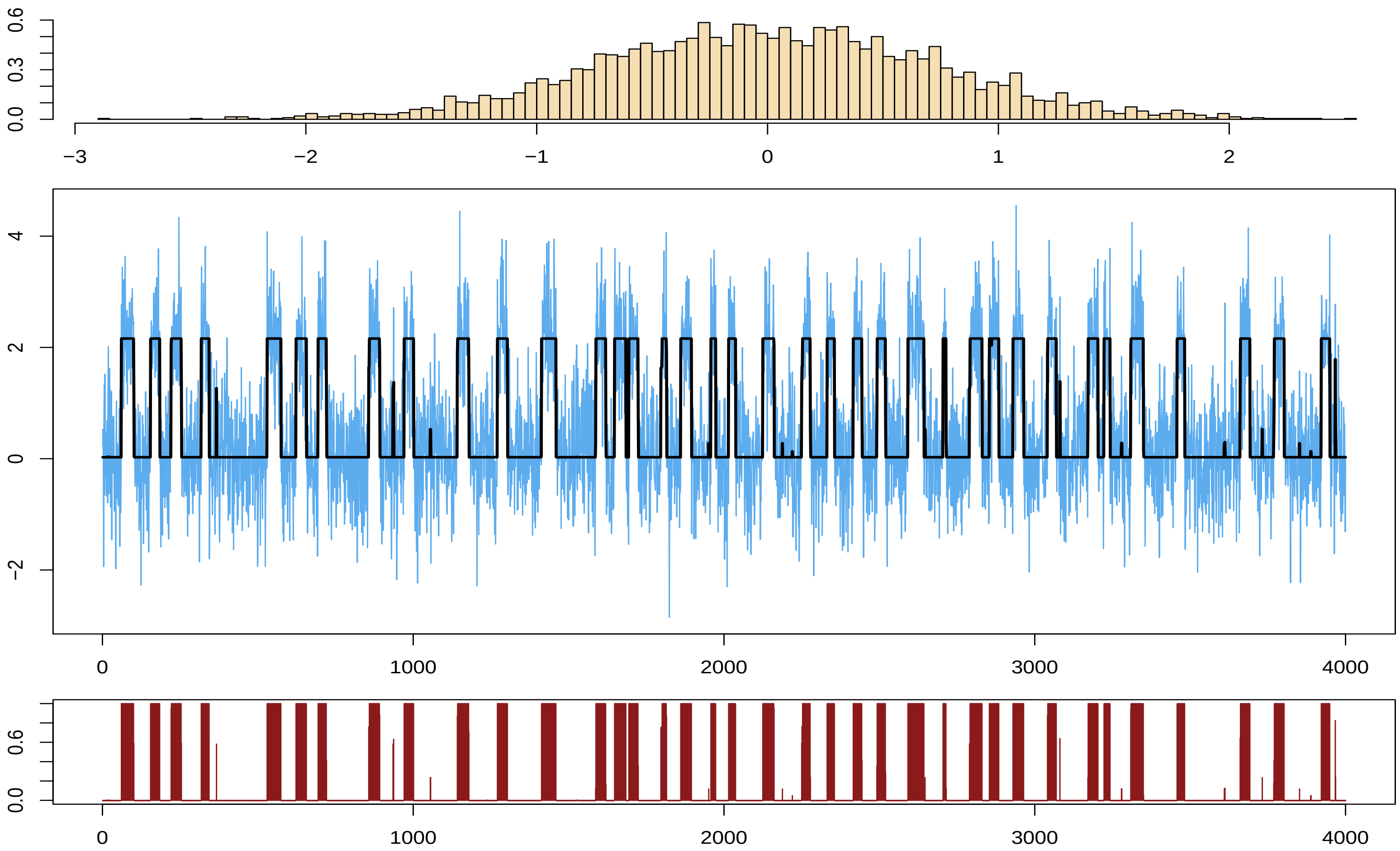
$$\mathbb{P} = \begin{pmatrix} 1 - \frac{\lambda_0}{s_0} & \frac{\lambda_0}{s_0} \\ \frac{\lambda_1}{s_1} & 1 - \frac{\lambda_1}{s_1} \end{pmatrix} ,$$

by analogy with average sojourn times for both models.

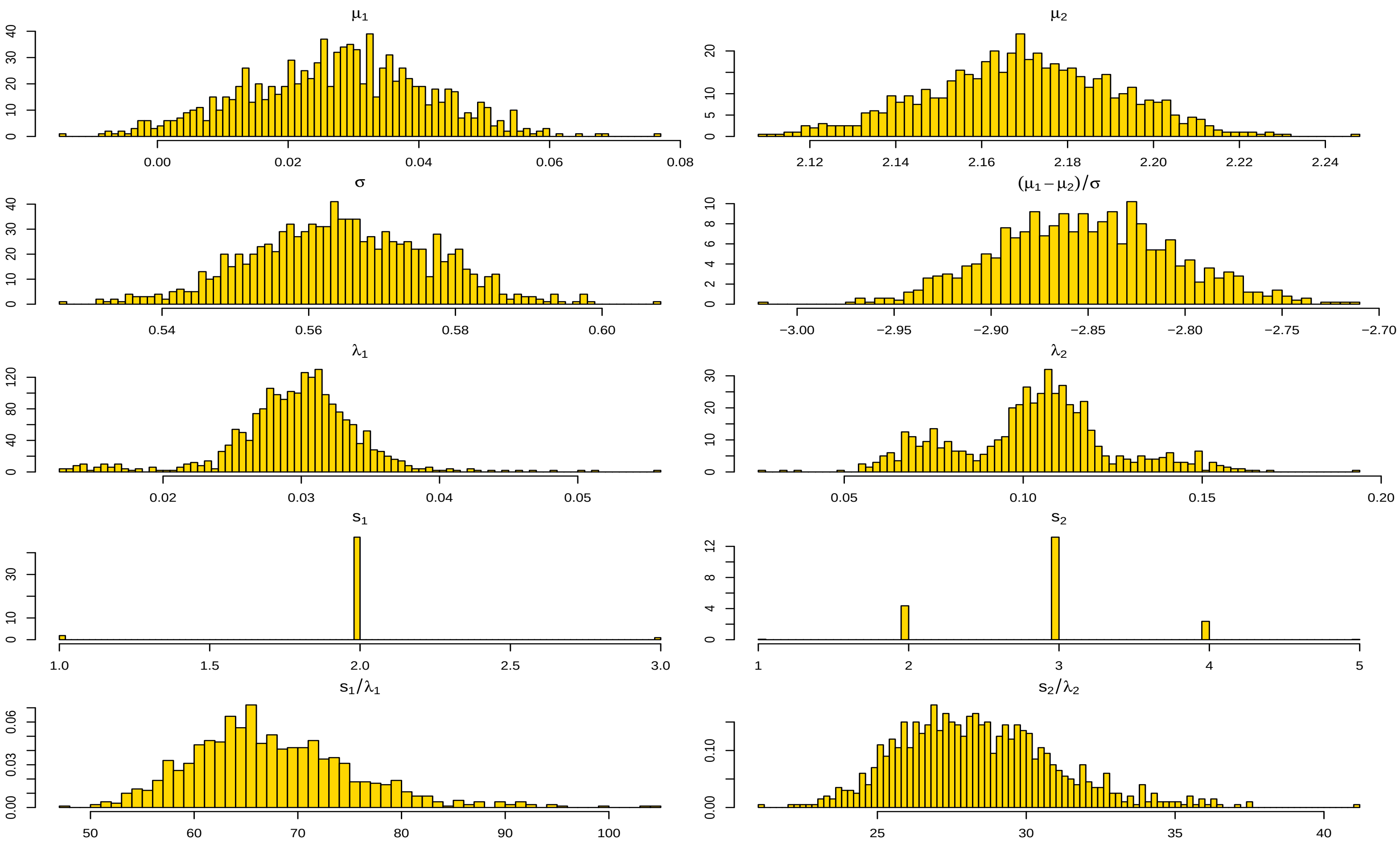
Simulation

Use of
forward–backward formulae,
of conjugate structure for the μ's, λ's and σ and
of finite support for s, distributed as

$$s_i|\mathbf{x} \sim \pi(s_i|\mathbf{x}) \propto \left[\frac{\Delta_i}{(\beta+v_i)^{n_i}}\right]^{s_i} \frac{\Gamma(n_i s_i+\alpha)}{\Gamma(s_i)^{n_i}} \mathbb{I}_{\{1,2,\ldots,S\}}(s_i)$$



Fitted series with residuals (top) and allocation probabilities (bottom)



Fitted series with residuals (top) and allocation probabilities (bottom)

Iterated particle system

Repeated calls to importance sampling with systematic resampling steps to improve fit

- How many steps?
- Which improvement?
- Why bother?!

Algorithm

Step 0. Generate $(j = 1, \ldots, J)$

1. $\omega^{(j)} \sim \pi(\omega)$
2. $\mathbf{x}_{-}^{(j)} = (x_t^{(j)})_{1 \le t \le T} \sim \pi_H(\mathbf{x}|\mathbf{y}, \omega^{(j)})$

and compute the weights $(j = 1, \ldots, J)$

$$\varrho_j \propto \frac{\pi(\omega^{(j)}, \mathbf{x}_{-}^{(j)}|\mathbf{y})}{\pi(\omega^{(j)})\pi_H(\mathbf{x}_{-}^{(j)}|\mathbf{y}, \omega^{(j)})}$$

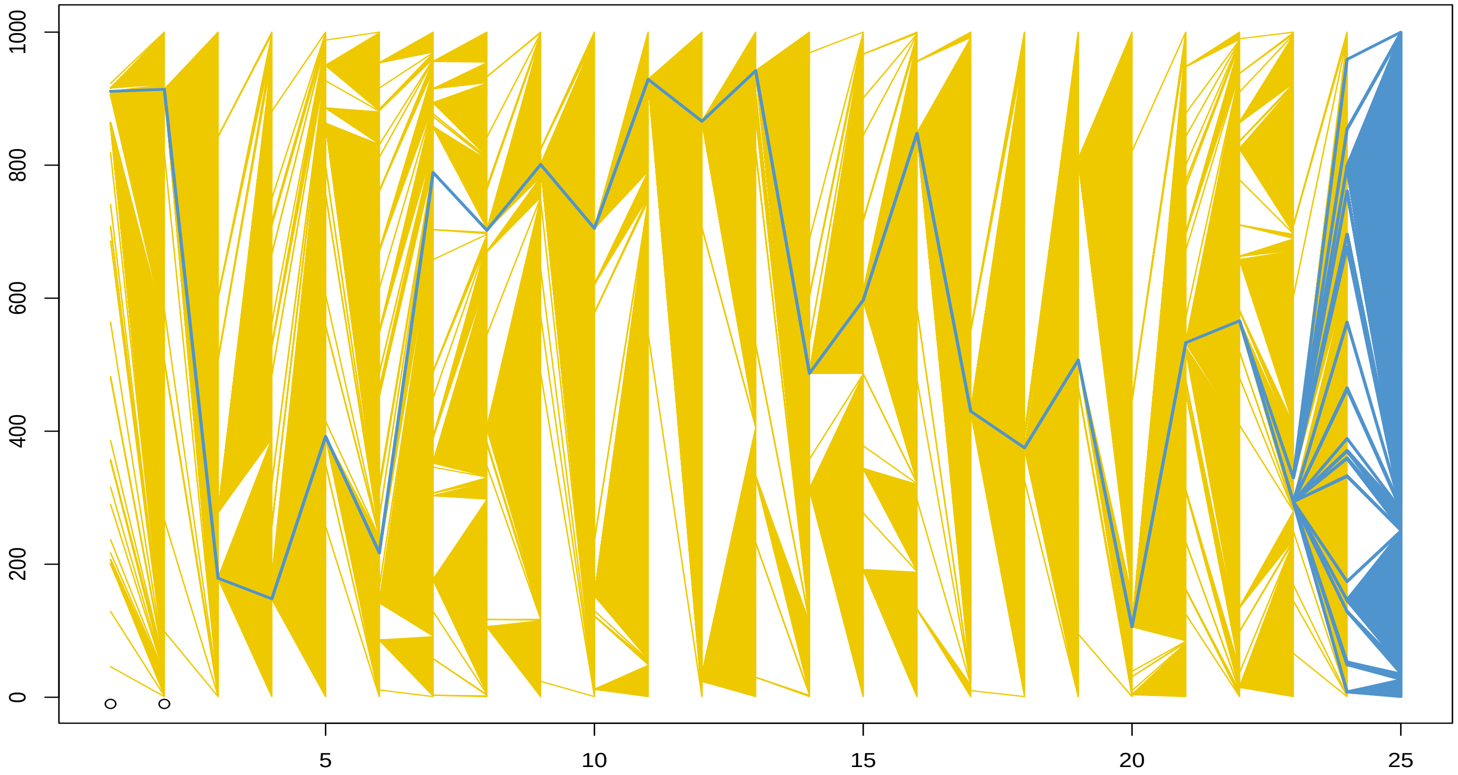
Step i. $(i = 1, \ldots)$ Generate $(j = 1, \ldots, J)$

1. $\omega^{(j)} \sim \pi(\omega | \mathbf{y}, \mathbf{x}_-^{(j)})$
2. $\mathbf{x}_+^{(j)} = (x_t^{(j)})_{1 \le t \le T} \sim \pi_H(\mathbf{x} | \mathbf{y}, \omega^{(j)})$

compute the weights $(j = 1, \ldots, J)$

$$\varrho_j \propto \frac{\pi(\omega^{(j)}, \mathbf{x}_+^{(j)} | \mathbf{y})}{\pi(\omega^{(j)} | \mathbf{y}, \mathbf{x}_-^{(j)}) \pi_H(\mathbf{x}_+^{(j)} | \mathbf{y}, \omega^{(j)})}$$

resample the couples $\omega^{(j)}, \mathbf{x}_+^{(j)}$ from the weights ϱ_j, and take $\mathbf{x}_-^{(j)} = \mathbf{x}_+^{(j)}$ $(j = 1, \ldots, J)$.



History and ancestry of the particle system

3 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

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

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 $p\left(\theta_1 \mid \mathbf{y}\right)$

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

$$p^{\gamma(i)}\left(\theta_1 \mid \mathbf{y}\right),$$

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

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

3.1 State Augmentation for Marginal Estimation

[Doucet, Godsill & Robert (2001)]

Artificially augmented probability model whose marginal distribution is

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

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:

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

- Use corresponding marginal for θ_1

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

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

$$q_\gamma\left(\theta_1, \theta_2\left(1\right), \ldots, \theta_2\left(\gamma\right) \mid \mathbf{y}\right)$$

- Use simulated subsequence

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

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

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