Some examples of applications of MCMC methods in latent variable models, mostly mixtures!

Christian P. Robert

Université Paris Dauphine

www.ceremade.dauphine.fr/~xian

Joint works with O. Cappé, A. Doucet, M. Hurn, A. Justel & T. Rydén

1. Mixtures of distributions

Representation of heterogeneous structures

$$x \sim f(x) = \sum_{i=1}^{k} p_i f_i(x)$$

or

$$x \sim f(x) = \sum_{i=1}^{k} p_i f(x|\theta_i)$$

Inference on

- homogeneity classes $x_j \sim f(x|\theta_i)$
- parameters p_i, θ_i
- heterogeneity factor k

Latent variable = indicator

$$\begin{aligned} x|z &\sim f(x|\theta_z) \\ z &\sim \mathcal{M}(1; p_1, \ldots, p_k) \end{aligned}$$

DIFFICULTIES

- Weak identifiability of mixtures

$$P(z_i \neq j, i = 1, \ldots, n) = (1 - p_i)^n$$

and invariance by permutation

- Prohibition of independent improper priors

$$\int \pi(\theta, p) d\theta\, dp = \infty \Rightarrow \int \pi(\theta, p | x_1, \ldots, x_n) d\theta\, dp = \infty$$

- Unbounded likelihood function in most setups

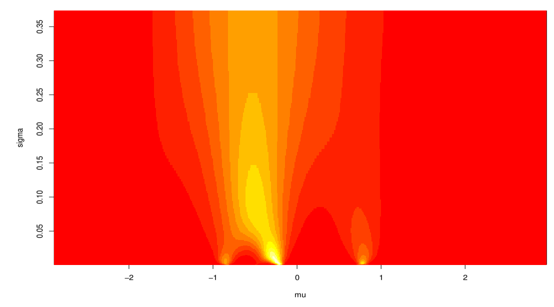
$$\lim_{\sigma_1 \to 0} \mathrm{L}(\theta, \sigma, p | x) = \infty$$

- Multimodal surface of likelihood and posterior distributions

- Combinatorial explosion in n

$$\sum_{i=1}^{k^n} \prod_{j=1}^{k} (p_j/\sigma_j)^{n_{ij}} \exp\{-n_{ij}(\bar{x}_{ij} - \mu_j)^2/2 - s_{ij}/2\}$$

- Difficulties with optimization methods (gradient, EM)



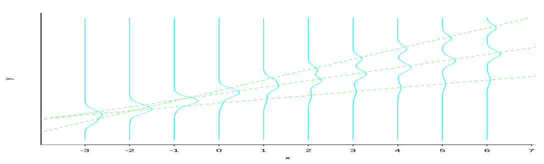
Generalizations

Mixture of linear regressions

$$y_i | x_{i,j} \sim \mathcal{N}(x_i^t \beta_j, \sigma_j^2)$$

$$\downarrow$$

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



Mixtures of logistic regressions

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

1.2. Choice of the prior distribution

Often perceived as a major **difficulty** with the Bayesian approach but

- 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]
- scales the difference between components [What makes a component *one component*?!]

Classes of priors

- possible symmetric empirical Bayes priors

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

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)]

- Reparameterization to compact set for use of uniform priors as, e.g.,

$$\mu_i \longrightarrow \frac{e^{\mu_i}}{1+e^{\mu_i}}, \qquad \sigma_i \longrightarrow \frac{\sigma_i}{1+\sigma_i}$$

- dependent weakly informative priors

$$\begin{aligned} \underline{p} &\sim \mathcal{D}(k, \dots, 1), \\ \theta_i &\sim \mathcal{N}(\theta_{i-1}, \zeta\sigma_{i-1}^2), \\ \sigma_i &\sim (1/2)\mathcal{U}([0, \sigma_{i-1}]) + (1/2)(1/\mathcal{U}([0, 1/\sigma_{i-1}])) \end{aligned}$$

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

- reference priors

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

[Moreno & Liseo (1999)]

1.3. Resolving the identifiability issues

A crucial feature:

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

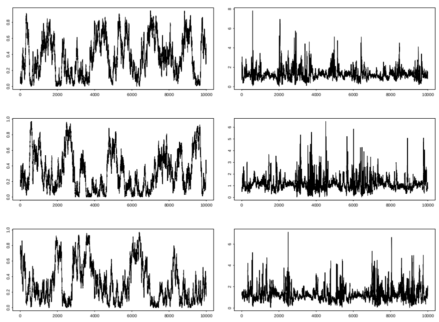
is invariant to relabelling of the components

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

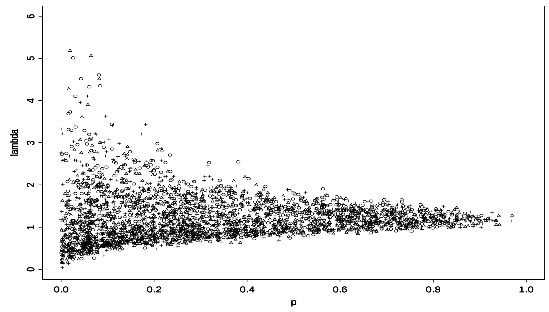
Example — Exponential mixture model

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

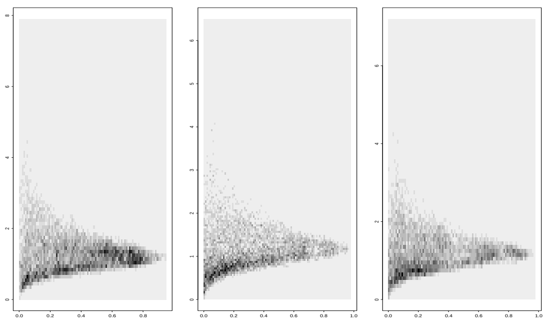
with $p_0 + \ldots + p_k = 1$ and $\lambda_i > 0$ $(0 \le i \le k)$.



Successive values of the Markov chains

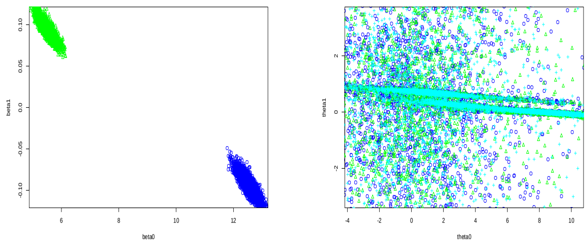


Joint distribution



Marginal distributions

Example — MCMC outputs for two chemical datasets



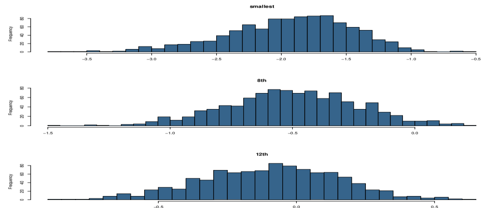
[Hurn, Justel & Robert (2001)]

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...

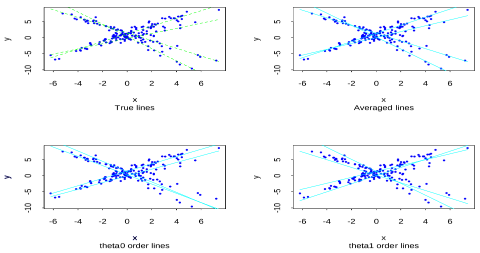
- modifies the prior distribution quite strongly for large #'s of components *[e.g. the marginal on θ_1 or θ_k may be quite different from $\mathcal{N}(0,1)$]*



- hampers MCMC samplers, sometimes to the extent of creating trapping states *[example of the "inverted order"]*

[Diebolt & Robert, 1994]

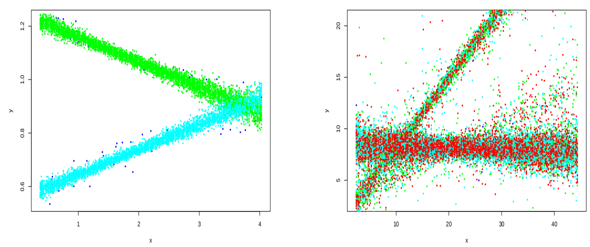
- induces biases *[may force components apart or increase the variance of the first component, for instance]*



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

$$\mathbb{E}[(\mu_i, \sigma_i, p_i)|x_1, \ldots, x_n] = \mathbb{E}[(\mu_j, \sigma_j, p_j)|x_1, \ldots, x_n]$$

Simply forces us to think harder and better!



[Hurn, Justel & Robert (2001)]

1.4. Exploring the whole posterior distribution

Thanks to the MCMC revolution,

[Robert & Casella, 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

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

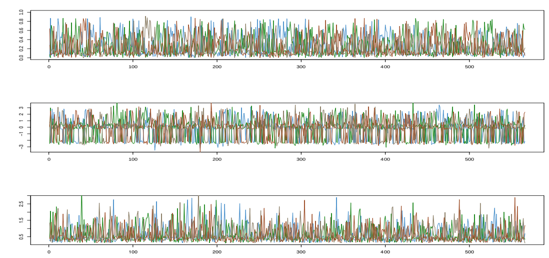
- Gibbs

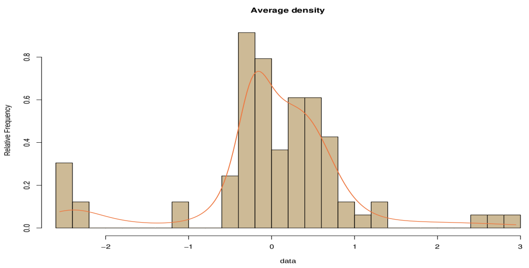
  'Natural' completion by latent indicators

$$z_t \sim \mathcal{M}_k(1; p_1, \ldots, p_k), \qquad x_t | z_t \sim f(x|\theta_{z_t})$$

  - Slow moves sometimes
  - Large increase in dimension, order $O(n)$
  - Good theoretical properties *(Duality principle)*

Example — Galaxy benchmark ($k = 4$)





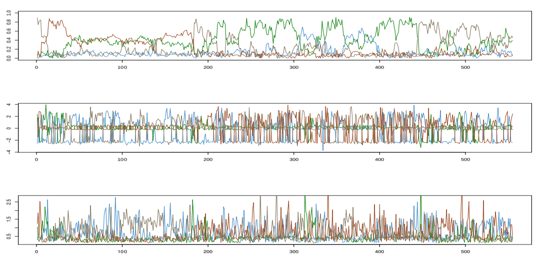
- Random walk Metropolis–Hastings
  - Avoids completion
  - Available (Normal vs. Cauchy vs... moves)
  - Calibrated against acceptance rate
  - Depends on parameterisation

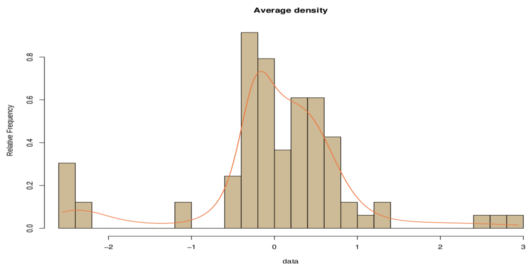
    $\lambda_j \longrightarrow \log \lambda_j \qquad p_j \longrightarrow \log(p_j/1 - p_k)$

    or

    $$\theta_i \longrightarrow \frac{\exp \theta_i}{1 + \exp \theta_i}$$

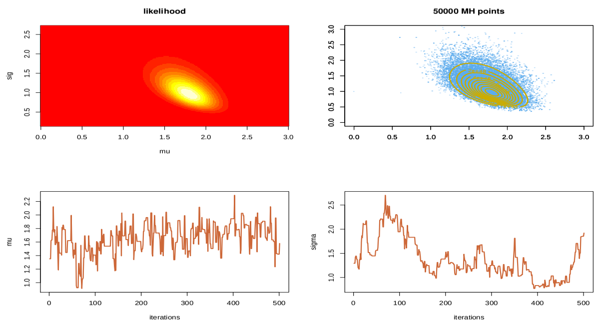
Example — Galaxy benchmark ($k = 4$)





Example — Simulated sample

$$0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(\mu, \sigma^2)$$



1.5. 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**?!

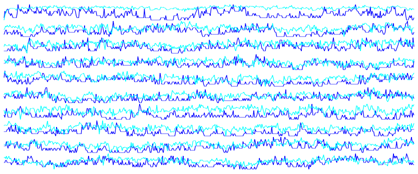
- weak identifiability of mixtures [contiguity of components]
- insolvable “philosophical” problem unless k has a proper intrinsic meaning [and even so...]
- hence testing *per se* is impossible: the data cannot distinguish between k and $k+h$ [unless guided by a firm hand/prior!]
- the choice of a prior 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
  [Carlin & Chib (1995)]
- 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



Saturated models

[Carlin & Chib (1995)]

Range of models $\mathcal{M}_k \quad k = 1, \cdots, K$

Range of priors $\pi_k(\theta_k)$

Range of weights $\rho_j = Pr(\mathcal{M} = \mathcal{M}_j)$

Joint distribution

$$\pi(\underline{\theta}, \mathcal{M} = \mathcal{M}_j | x) \propto f_j(x|\theta_j) \prod_{k=1}^{K} \pi_k^j(\theta_k) \rho_j$$

where π_j^k pseudo-prior, used when $\mathcal{M} = \mathcal{M}_j$ $(k \neq j)$ and $\pi_j^j = \pi_j$

$\boxed{\textbf{Idea:}}$ Use a Gibbs sampler on satured model
with pseudo-priors on irrelevant parameters

$\boxed{\textbf{Validation}}$ Pseudo-priors only used in transitions
of the MCMC algorithm

Reversible jump MCMC

[Green (1995)]

`Facts:`

- No clear dominating measure in variable dimension spaces
- Gibbs sampling does not apply

`Solution:`

- Create fixed dimension moves locally
- Supplements θ_1 from $\mathcal{M}_{k_1}$ and θ_2 from $\mathcal{M}_{k_2}$ by $u_{1\to2}$ and $u_{2\to1}$ resp. so that

$$(\theta_1, u_{1\to2}) \text{ and } (\theta_2, u_{2\to1})$$

are in bijection (one-to-one correspondance):

$$(\theta_2, u_{2\to1}) = T(\theta_1, u_{1\to2})$$

- Use acceptance probability

$$\min\left(\frac{\pi(k_2,\theta_2)\rho_{1\to2}g(u_{1\to2})}{\pi(k_1,\theta_1)\rho_{2\to1}g(u_{1\to2})}\left|\frac{\partial T(\theta_1,u_{1\to2})}{\partial(\theta_1,u_{1\to2})}\right|, 1\right)$$

Example — Mixture of exponential distributions

[Gruet & al. (1999)]

$$\mathcal{M}_k : \sum_{j=1}^{k} p_{jk} \mathcal{E}xp(\lambda_{jk}) \qquad \pi(k)$$

$$\mathcal{M}_{k+1} : \sum_{j=1}^{k+1} p_{j(k+1)} \mathcal{E}xp(\lambda_{jk}) \qquad \pi(k+1)$$

Moves:

`(i) Split`

$$\begin{cases} p_{jk} &= p_{j(k+1)} + p_{(j+1)(k+1)} \\ p_{jk}\lambda_{jk} &= p_{j(k+1)}\lambda_{j(k+1)} + p_{(j+1)(k+1)}\lambda_{(j+1)(k+1)} \end{cases}$$

`(ii) Merge` $(inverse)$

Equivalent:

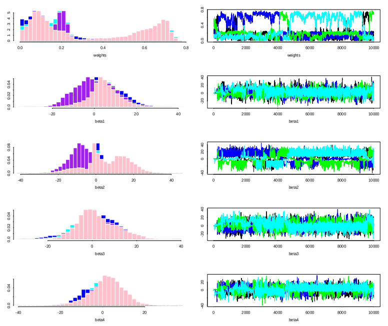
`(i) Split`

$$(T)\begin{cases} u_1, u_2 &\sim \mathcal{U}(0,1) \\ p_{j(k+1)} &= u_1 p_{jk} \\ \lambda_{j(k+1)} &= u_2 \lambda_{jk} \end{cases}$$

$$\frac{\partial T(p,\lambda,u_1,u_2)}{\partial(p,\lambda,u_1,u_2)} = \begin{vmatrix} u_1 & (1-u_1) & 0 & 0 \\ p & -p & 0 & \lambda\frac{1-u_2}{(1-u_1)^2} \\ 0 & 0 & u_2 & \frac{1-u_1 u_2}{1-u_1} \\ 0 & 0 & \lambda & -\frac{\lambda u_1}{1-u_1} \end{vmatrix}$$

$$= p/(1-u_1)$$

Example – An INSEE logistic regression on savings:



2. Switching ARMAs

Model

ARMA structure

$$y_t = \gamma + \sum_{i=1}^{p} \varphi_i (y_{t-i} - \gamma) + \sigma \sum_{j=0}^{q} \theta_j \epsilon_{t-j}$$

plus a latent state Markov chain $s_t \in \{0, 1, \ldots, M\}$:

$$y_t = \gamma_{s_t} + \sum_{i=1}^{p} \varphi_i^{s_t} (y_{t-i} - \gamma_{s_{t-i}}) + \sum_{j=0}^{q} \theta_j^{s_t} \sigma_{s_{t-j}} \epsilon_{t-j} \qquad t \geq 1$$

with

$$P(s_t = i | s_{t-1} = j) = \pi_{ij}$$

[Hamilton (1989)]

Difficulties

1. Iterative definition of the likelihood function due to the MA structure.
2. Non-Markovian process, which can be rescued with artificial states.

   [Billio & Monfort (1989)]
3. No clearcut missing data structure for Bayesian inference besides latent states.
4. Stationarity requirements put constraints on θ and φ parameters.

Special cases

[MA]

$$y_t = \gamma_{s_t} + \sigma_{s_t}\epsilon_t - \theta\sigma_{s_{t-1}}\epsilon_{t-1}$$

[ARMA$_1$]

$$y_t = \gamma_{s_t} + \phi(y_{t-1} - \gamma_{s_{t-1}}) + \sigma_{s_t}\epsilon_t - \theta\sigma_{s_{t-1}}\epsilon_{t-1}$$

[ARMA$_2$]

$$y_t = \mu + \phi_{s_t}(y_{t-1} - \mu) + \sigma\epsilon_t - \theta_{s_t}\sigma\epsilon_{t-1}$$

Prior modelling

Identifiability requirement

$$\sigma_1 < \sigma_0$$

and stationarity constraints

$$|\theta_i| < 1, \quad |\phi_i| < 1$$

Noninformative priors:

$$\begin{aligned} \pi(\xi) &\propto \sigma_0^{-3} e^{-(\gamma_0-\gamma_1)^2\zeta/2\sigma_0^2} \mathbb{I}_{\sigma_0>\sigma_1} \\ \pi(\xi) &\propto \sigma_0^{-3} e^{-(\gamma_0-\gamma_1)^2\zeta/2\sigma_0^2} \mathbb{I}_{\sigma_0>\sigma_1} \\ \pi(\xi) &\propto \sigma^{-1} \end{aligned}$$

Mcmc implementation

1. Generate the missing states (s_t) from

$$f(s_0,\ldots,s_T|\xi,y_0,\ldots,y_T) \propto \prod_{t=1}^{T} \exp\{-(y_t-\gamma_{s_t}+\theta\eta_{t-1})^2/2\sigma_{s_t}^2\}\sigma_{s_t}^{-1}\pi_{s_{t-1}s_t}$$

with

$$\eta_t=\sigma_{s_t}\epsilon_t=y_t-\gamma_{s_t}+\theta\eta_{t-1}, \qquad t\ge 1.$$

2. Generate the parameters ξ from

$$f(\xi|s_0,\ldots,s_T,y_0,\ldots,y_T)$$

Basic idea:

1. Replace at each burdensome step the "true" innovations η_t with innovations $\hat\eta_t$ based on previous values.
2. Use the resulting distribution with these fixed innovations $\hat\eta_t$ as a proposal/instrumental distribution.
3. Validate the method by an Hastings acceptance step.

Example — [MA]

1. *Missing states* — At iteration $m+1$ of the hybrid Gibbs sampler, define

$$\eta_0^{(m+1)}=0, \quad \eta_t^{(m+1)}=y_t-\gamma^{(m)}_{s_t^{(m+1)}}+\theta^{(m)}\eta_{t-1}^{(m+1)}$$

between the generations of $s_t^{(m+1)}$ and of $s_{t+1}^{(m+1)}$.

(a) Generate $s_{t+1}^{(m+1)}$ from

$$p(s_t) \quad\propto\quad \exp\{-(y_t-\gamma_{s_t}^{(m)}+\theta^{(m)}\eta_{t-1}^{(m+1)})^2/2(\sigma_{s_t}^{(m)})^2\} \times(\sigma_{s_t}^{(m)})^{-1}\pi_{s_{t-1}^{(m+1)}s_t}\pi_{s_t s_{t+1}^{(m)}}$$

(b) Accept the new value of (s_t) with probability

$$\omega_s=\frac{\prod_{t\ge1}\pi^{(m)}_{s_t^{(m)}s_{t+1}^{(m+1)}}}{\prod_{t\ge1}\pi^{(m)}_{s_t^{(m+1)}s_{t+1}^{(m)}}} \times\frac{\prod_{t\ge1}\sum_{i=0}^{1}(\sigma_i^{(m)})^{-1}e^{-(y_t-\gamma_i^{(m)}+\theta^{(m)}\eta_{t-1}^{(m+1)})^2/2(\sigma_i^{(m)})^2}}{\prod_{t\ge1}\sum_{i=0}^{1}(\sigma_i^{(m)})^{-1}e^{-(y_t-\gamma_i^{(m)}+\theta^{(m)}\eta_{t-1}^{(m)})^2/2(\sigma_i^{(m)})^2}} \times\frac{\prod_{t\ge1}\pi^{(m)}_{s_{t-1}^{(m+1)}i}\pi^{(m)}_{is_{t+1}^{(m)}}}{\prod_{t\ge1}\pi^{(m)}_{s_{t-1}^{(m)}i}\pi^{(m)}_{is_{t+1}^{(m+1)}}}$$

2. *Locations*— At iteration $m+1$ of the hybrid Gibbs sampler, define $\tilde\eta_0^{(m)}=0$ and

$$\tilde\eta_t^{(m)}=y_t-\gamma^{(m)}_{s_t^{(m+1)}}+\theta^{(m)}\tilde\eta_{t-1}^{(m)}, \quad \tilde y_t^{(m)}=y_t+\theta^{(m)}\tilde\eta_{t-1}^{(m)}.$$

(a) Generate (γ_0,γ_1) from

$$\gamma_0|\gamma_1 \sim \mathcal{N}\left(\frac{n_0^{(m)}\bar y_0^{(m)}+\gamma_1\zeta}{n_0^{(m)}+\zeta},\tau_\gamma^2\frac{(\sigma_0^{(m)})^2}{n_0^{(m)}+\zeta}\right),$$

$$\gamma_1|\gamma_0 \sim \mathcal{N}\left(\frac{n_1^{(m)}\bar y_1^{(m)}+\gamma_0\zeta(\sigma_1^{(m)})^2/(\sigma_0^{(m)})^2}{n_1^{(m)}+\zeta(\sigma_1^{(m)})^2/(\sigma_0^{(m)})^2}, \tau_\gamma^2\frac{\sigma_1^2}{n_1^{(m)}+\zeta(\sigma_1^{(m)})^2/(\sigma_0^{(m)})^2}\right),$$

where

$$n_0^{(m)}=\sum_{t=1}^{T}\mathbb{I}_{\{s_t^{(m+1)}=0\}}, \quad n_0^{(m)}\bar y_0^{(m)}=\sum_{t=1}^{T}\mathbb{I}_{\{s_t^{(m+1)}=0\}}\tilde y_t^{(m)},$$

$$n_1^{(m)}=\sum_{t=1}^{T}\mathbb{I}_{\{s_t^{(m+1)}=1\}}, \quad n_1^{(m)}\bar y_1^{(m)}=\sum_{t=1}^{T}\mathbb{I}_{\{s_t^{(m+1)}=1\}}\tilde y_t^{(m)}.$$

(b) Accept the new value of (γ_0,γ_1) based on the Hastings density ratio.

Note Calibrate τ_γ^2 to achieve acceptance rate in $(.2,.5)$.

3. *MA coefficient*— At iteration $m+1$ of the hybrid Gibbs sampler, define $\hat\eta_t^{(m)} = y_t - \gamma^{(m+1)}_{s_t^{(m+1)}} + \theta^{(m)}\hat\eta^{(m)}_{t-1}$.

   Generate $\tilde\theta$ from $\mathcal{N}(\beta^{(m)}, \tau^{(m)})$, with

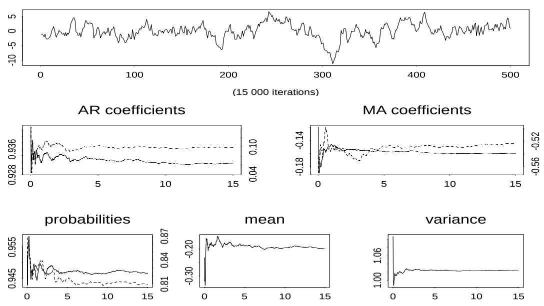
$$\beta^{(m)} = -\frac{\sum_{t=1}^T \left[y_t - \gamma^{(m+1)}_{s_t^{(m+1)}}\right]^2 \hat\eta^{(m)}_{t-1}\left(\sigma^{(m+1)}_{s_t^{(m+1)}}\right)^{-2}}{\sum_{t=1}^T \left(\sigma^{(m+1)}_{s_t^{(m+1)}}\right)^{-2}\left(\hat\eta_t^{(m)}\right)^2},$$

$$\tau^{(m)} = \tau_\theta^2\left[\sum_{t=1}^T \left(\sigma^{(m+1)}_{s_t^{(m+1)}}\right)^{-2}\left(\hat\eta^{(m)}_{t-1}\right)^2\right]^{-1}.$$

and take

$$\hat\theta = \begin{cases} \tilde\theta & \text{if } |\tilde\theta| < 1, \\ 1/\tilde\theta & \text{otherwise.} \end{cases}$$

and accept the new value of θ based on the Hastings density ratio.

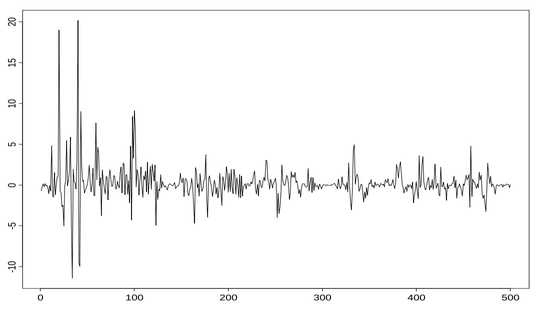


3. Stochastic volatility model

Model ($t = 1, \ldots, T$)

$$\begin{cases} y_t^* &= a + b y^*_{t-1} + c\varepsilon_t^* \\ y_t &= \exp(0.5\, y_t^*)\varepsilon_t \end{cases}$$

with unobserved y_t^*'s.



Simulated stochastic volatility sample

MCMC IMPLEMENTATION

1. Simulate the y_t^*'s conditionaly on the y_t's and the parameters
2. Simulate (a, b, c) for the completed model

Use Metropolis–Hastings schemes with approximations of

$$f(y_t^*|y^*_{t-1}, y^*_{t+1}, y^t, \theta) \propto \exp\left\{-(y_t^* - a - b y^*_{t-1})^2/2c^2 - (y^*_{t+1} - a - b y_t^*)^2/2c^2 - y_t^*/2 - y_t^2 e^{-y_t^*}/2\right\}.$$

[Shephard (1997)] .

- Linearize $(y_t^* + y_t^2 e^{-y_t^*}/2)$ into $(y_t^* - \log(y_t^2))^2/4$ with proposal

$$\mathcal{N}\left(\frac{[a + b y^*_{t-1} + b(y^*_{t+1} - a)]c^{-2} + \log(y_t^2)/2}{(1+b^2)c^{-2} + 1/2}, \left[(1+b^2)c^{-2} + 1/2\right]^{-1}\right),$$

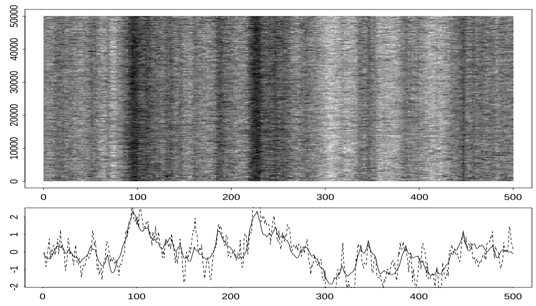
- Use a gamma approximation for $\omega_t = \exp(-y_t^*)$,

$$\omega_t \sim \omega_t^{-1/2} \exp(-y_t^2 \omega_t/2)\, e^{-(-\log(\omega_t))^2(1+b^2)/2c^2} \omega_t^{-[a+b(y_{t-1}^*+y_{t+1}^*-a)]c^{-2}}.$$

as (matching moments or Taylor expansion)

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

[Jacquier, Polson & Rossi (1996)]



Allocation map and average vs. true volatilities

4. Hidden Markov chains

Extension of a mixture model with Markov dependence

$$x_t|z, x_j\ j \neq t \sim \mathcal{N}(\mu_{z_t}, \sigma_{z_t}^2),\ P(z_t = u|z_j\ , j < t) = p_{z_{t-1}u},$$

Noninformative prior distribution on $(\mu, \sigma, \mathbb{P})$ inspired by reparameterization issues: $(j = 1, \ldots, k)$

$$\mu_j = \theta_1 + \tau_1\theta_2 + \ldots + \tau_1 \ldots \tau_{j-1}\theta_j, \quad \sigma_j = \tau_1 \ldots \tau_j$$

and identifiability condition

$$\tau_2 \leq 1, \ldots, \tau_k \leq 1.$$

[Robert & Titterington (1998)]

Resulting prior

$$\pi(\theta_1, \tau_1) = 1/\tau,\ \tau_u \sim \mathcal{U}_{[0,1]},\ \theta_u \sim \mathcal{N}(0, \zeta^2)\ (u = 2, \ldots, k),$$

or

$$\pi(\mu, \sigma) = \sigma_k^2 \sigma_1^{-3} \prod_{u=2}^{k} \sigma_u^{-2} \exp\{-(\mu_u - \mu_{u-1})^2/2\zeta^2\sigma_{u-1}^2\}$$

with $p_{i\cdot} \sim \mathcal{D}(1, \ldots, 1)$.

MCMC implementation

Available in this setup for both parameters

$$(p_{u1}, \ldots, p_{uk}) \sim \mathcal{D}(1 + n_{u1}, \ldots, 1 + n_{uk})$$

$$\begin{aligned}
\sigma_1^{-2} &\sim \mathcal{G}a^T\left(\frac{n_1+2}{2}, \frac{n_1(\overline{x}_1 - \mu_1)^2 + s_1^2 + (\mu_2 - \mu_1)^2\zeta^{-2}}{2}\right)\\
\sigma_u^{-2} &\sim \mathcal{G}a^T\left(\frac{n_u-1}{2}, \frac{n_u(\overline{x}_u - \mu_u)^2 + s_u^2 + (\mu_{u+1} - \mu_u)^2\zeta^{-2}}{2}\right)\\
\sigma_k^{-2} &\sim \mathcal{G}a^T\left(\frac{n_k-2}{2}, \frac{n_k(\overline{x}_k - \mu_k)^2 + s_k^2}{2}\right)\\
\mu_1 &\sim \mathcal{N}\left(\frac{n_1\overline{x}_1 + \zeta^{-2}\mu_2}{n_1 + \zeta^{-2}}, \frac{\sigma_1^2}{n_1 + \zeta^{-2}}\right),
\end{aligned}$$

$$\begin{aligned}
\mu_u &\sim \mathcal{N}\left(\frac{\zeta^2 n_u\overline{x}_u + \mu_{u+1} + \mu_{u-1}\frac{\sigma_u^2}{\sigma_{u-1}^2}}{\zeta^2 n_u + 1 + \frac{\sigma_u^2}{\sigma_{u-1}^2}}, \frac{\zeta^2\sigma_u^2}{\zeta^2 n_u + 1 + \frac{\sigma_u^2}{\sigma_{u-1}^2}}\right),\\
\mu_k &\sim \mathcal{N}\left(\frac{n_k\overline{x}_u + \zeta^{-2}\mu_{k-1}\sigma_k^2/\sigma_{k-1}^2}{n_k + \zeta^{-2}\sigma_k^2/\sigma_{k-1}^2}, \frac{\sigma_k^2}{n_k + \zeta^{-2}\sigma_k^2/\sigma_{k-1}^2}\right).
\end{aligned}$$

and latent states $(1 < i < k)$

$$\begin{aligned}
P(z_1 = u|z_2, \ldots, \mathbb{P}) &\propto p_{uz_2} f(x_1|\mu_u, \sigma_u),\\
P(z_i = u|\ldots, z_{i-1}, z_{i+1}, \ldots, \mathbb{P}) &\propto p_{z_{i-1}u} p_{uz_{i+1}} f(x_i|\mu_u, \sigma_u),\\
P(z_k = u|\ldots, z_{k-1}, \ldots, \mathbb{P}) &\propto p_{z_{k-1}u} f(x_k|\mu_u, \sigma_u),
\end{aligned}$$

[Robert & Titterington (1996)]

Alternative implementation for the latent states

Based on the backward formulae

$$\begin{aligned}
P(x_{j+1}, \ldots, x_n \quad | \quad & z_j = i, \mathbb{P})\\
&= \sum_\ell p_{i\ell} P(x_{j+1}, \ldots, x_n|z_{j+1} = \ell, \mathbb{P})\\
&\propto \sum_\ell p_{i\ell} f(x_{j+1}|\mu_\ell, \sigma_\ell) P(x_{j+2}, \ldots, x_n|z_{j+1} = \ell, \mathbb{P}).
\end{aligned}$$

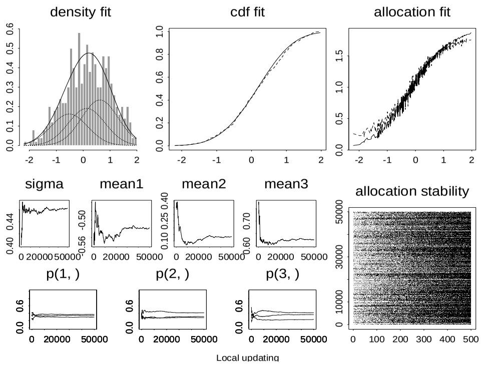
[Baum, Petrie, Soules & Weiss (1970)]

Conditional distribution of z_1

$$\begin{aligned}
P(z_1 = i \quad | \quad & x_1, \ldots, x_n, \mathbb{P})\\
&\propto \pi_i P(x_2, \ldots, x_n|z_1 = i, \mathbb{P}) f(x_1|\mu_i, \sigma_i)\ ,
\end{aligned}$$

then of z_2 conditional on z_1, the observations $x_2, \ldots, x_n$ and the parameters,

$$\begin{aligned}
P(z_j \quad = \quad & i|z_1, \ldots, z_{j-1}, x_1, \ldots, x_n, \mathbb{P})\\
&\propto p_{z_{j-1}i} P(x_{j+1}, \ldots, x_n|z_j = i, \mathbb{P}) f(x_j|\mu_i, \sigma_i)\ .
\end{aligned}$$



4.1 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 $\min(A^{-1}, 1)$.

Other moves in discrete time

- Merge of two adjacent components
- Split of one component into two adjacent components
- Details:
  - move between two models $\mathcal{M}_1$ and $\mathcal{M}_2$
  - bijection by supplementing by auxiliary variables $u_{1\to 2}$ and $u_{2\to 1}$

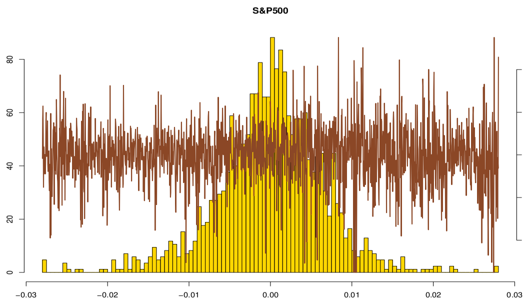
$$(\theta_2, u_{2\to 1}) = \Psi_{1\to 2}(\theta_1, u_{1\to 2}) .$$

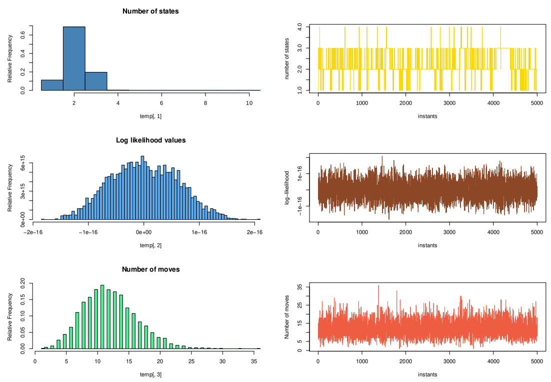
  - acceptance probability

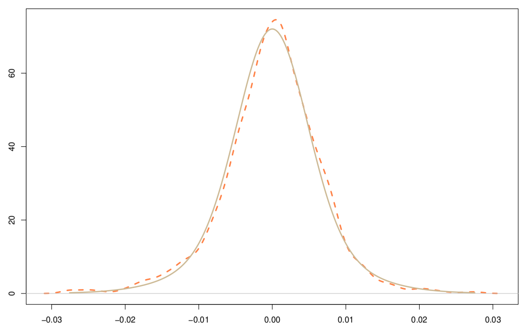
$$\min\left(\frac{\pi_2(\theta_2) g_{2\to 1}(u_{2\to 1})}{\pi_1(\theta_1) g_{1\to 2}(u_{1\to 2})}\left|\frac{\partial \Psi_{1\to 2}(\theta_1, u_{1\to 2})}{\partial(\theta_1, u_{1\to 2})}\right|, 1\right) .$$

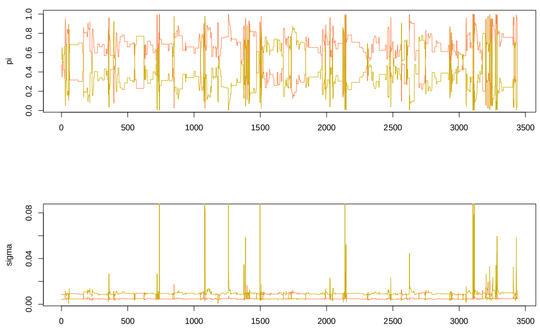
Illustrations

Example – Standard & Poor 500 from the 50's

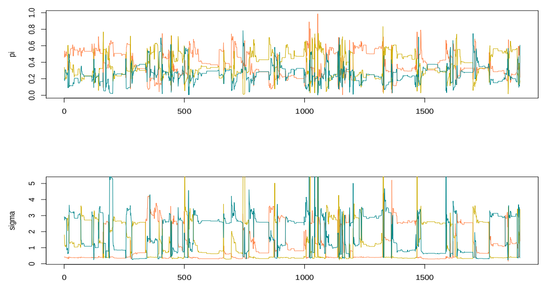


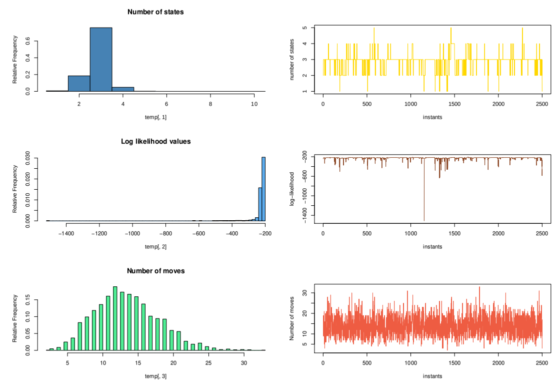


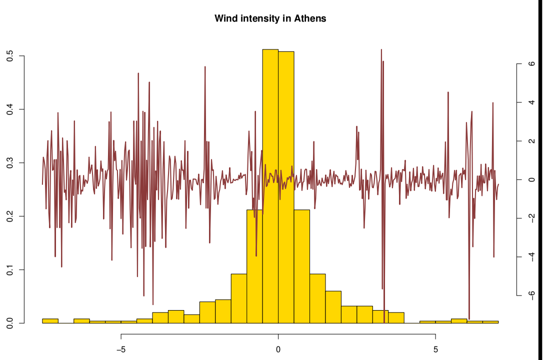


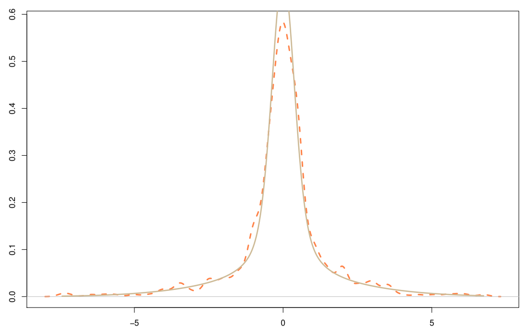


Example – Wind intensity in Athens









4.2. Optimization of the marginal

Role of maximum a posteriori estimation in Bayesian inference

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

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

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

where

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

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

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

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 (SAME)
   [Doucet, Godsill & Robert (2001)]
   Artificially augmented probability model whose marginal distribution is

$$\overline{p}_\gamma(\theta_1 | \mathbf{y}) \propto p^\gamma(\theta_1 | \mathbf{y})$$

via replications of the nuisance parameters:

- Replace θ_2 with γ artificial replications,

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

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

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

- 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