Bayesian Statistics today: a personal view

Christian P. Robert

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

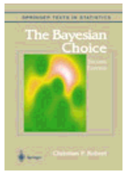
Outline

1. The running example
2. Choice of the prior distribution
3. Computational techniques
4. Model choice

A Reference

Berger, J.O. (2000) Bayesian Analysis: A look at today and thoughts of tomorrow. *JASA* **95**, 1269–1277.

Another:



0. The running example:

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

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

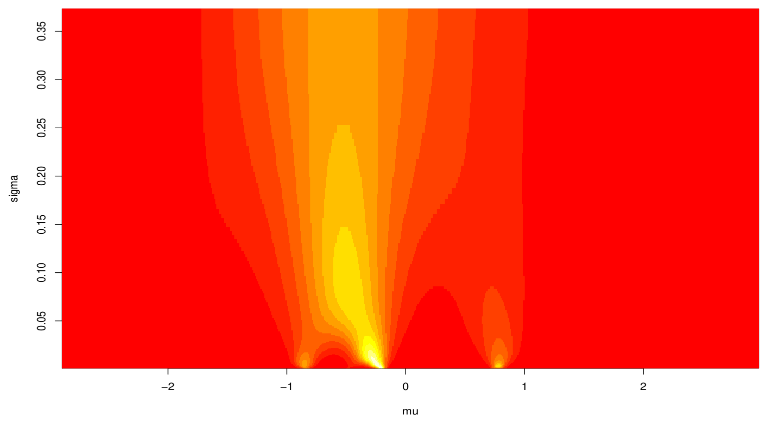
  and invariance by permutation
- Unbounded likelihood function in most setups

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

- 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)



1. Basics

Sampling model

$$(\mathcal{X}, f(x|\theta))$$

Prior modelling

$$\pi(\theta),\ \sigma\text{-finite measure on } \Theta$$

Bayesian paradigm based on posterior distribution

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int \pi(\theta) f(x|\theta) d\theta}$$

if

$$\int \pi(\theta) f(x|\theta) d\theta < \infty \qquad [ae]$$

Interpretation

- reference measure in a likelihood analysis
- penalization factor for the likelihood
- summary of anterior information

Difficulties

(1) **Conceptual**

Unknown fixed parameter θ versus random parameter $\theta \sim \pi$

- formal tool for probability calculus
- coherent updating
- inverse perspective
- implementation of the likelihood principle
- irrelevant in Asymptotia

(2) **Practical**

Choice of the prior distribution π

- low dimension parametric family
- device of imaginary observations
- robust classes
- sensitivity analysis
- **noninformative priors**
- hierarchical modelling
- nonparametric estimation

(3) **Computational**

Resolution of minimization / integration programs

- simple priors
- numerical integration
- analytical approximations
- asymptotic expansions
- Monte Carlo simulation
- **Markov chain Monte Carlo simulation**

2. The 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]

Noninformative priors

- Total lack of prior information
- Requirements of “objectivity”
- Test of prior information
- Conflicting prior opinions
- Theoretic comparisons with frequentist procedures
- Reference

Choice of prior distribution based on the sampling distribution $f(x|\theta)$

- Often improper
- Possible contradiction with the likelihood principle
- Not fit for every aspect of Bayesian inference

(1) **Uniform priors**

Principle of *insufficient reason* (Laplace)

$$\Theta = \{\theta_1, \cdots, \theta_p\} \qquad \pi(\theta_i) = 1/p$$

Extension to continuous spaces

$$\pi(\theta) \propto 1$$

- Lack of reparameterization invariance

$$\psi = e^\theta \qquad \pi_1(\psi) = \frac{1}{\psi} \neq \pi_2(\psi) = 1$$

- Problems of properness

$$x \sim \mathcal{N}(\theta, \sigma^2), \qquad \pi(\theta, \sigma) = 1$$

$$\begin{aligned} & \pi(\theta, \sigma|x) \quad \propto \quad e^{-(x-\theta)^2/2\sigma^2}\sigma^{-1} \\ \Rightarrow \quad & \pi(\sigma|x) \quad \propto \quad 1 \end{aligned}$$

(2) **Invariance priors**

Principle: Agree with the natural symmetries of the problem

- Identify invariance structures as group action

$$\begin{aligned} \mathcal{G} &: \quad x \to g(x) \sim f(g(x)|\bar{g}(\theta)) \\ \bar{\mathcal{G}} &: \quad \theta \to \bar{g}(\theta) \\ \mathcal{G}^* &: \quad L(d, \theta) = L(g^*(d), \bar{g}(\theta)) \end{aligned}$$

- Determine an invariant prior

$$\pi(\bar{g}(A)) = \pi(A)$$

Solution: Right Haar measure

But...

- Requires invariance to be part of the decision problem
- Missing in most discrete setups (Poisson)

(3) **Jeffrey's prior**

Based on Fisher information

$$I(\theta) = \mathbb{E}_\theta \left[\frac{\partial \ell}{\partial \theta} \frac{\partial \ell'}{\partial \theta} \right]$$

For invariance-under-reparameterization reasons,

$$\pi(\theta) \propto |I(\theta)|^{1/2}$$

- Agrees with most invariant priors
- Parameterization invariant
- Suffers from dimensionality curse

Example

$$x \sim \mathcal{N}_p(\theta, I_p), \qquad \eta = \|\theta\|^2, \qquad \pi(\eta) = \eta^{p/2-1}$$

$$\mathbb{E}^\pi[\eta|x] = \|x\|^2 + p \qquad \text{Bias } 2p$$

(4) **Reference Priors**

Generalizes Jeffreys priors by distinguishing between nuisance and interest parameters

Principle: maximize the information brought by the data

$$\mathbb{E}^n[\int \pi(\theta|x_n) \log(\pi(\theta|x_n)/\pi(\theta)) d\theta]$$

and consider the limit of the π_n

Outcome: most usually, Jeffreys prior

Nuisance parameters:

For $\theta = (\lambda, \omega)$,

$$\pi(\lambda|\omega) = \pi_J(\lambda) \qquad \text{with fixed } \omega$$

and

$$\pi(\omega) = \pi_J(\omega)$$

for the marginal model

$$f(x|\omega) \propto \int f(x|\theta) \pi_J(\lambda|\omega) d\lambda$$

- Depends on ordering
- Problems of definition

Mixtures of distributions

Classes of priors

- Prohibition of independent improper priors

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

- 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)]

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

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

[Chopin (2000)]

- dependent weakly informative priors

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

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

- reference priors

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

[Moreno & Liseo (1999)]

Identifiability issues

A crucial feature:

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

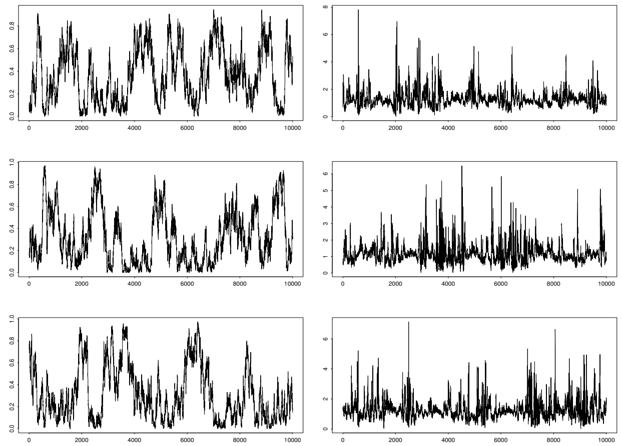
is invariant to relabelling of the components

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

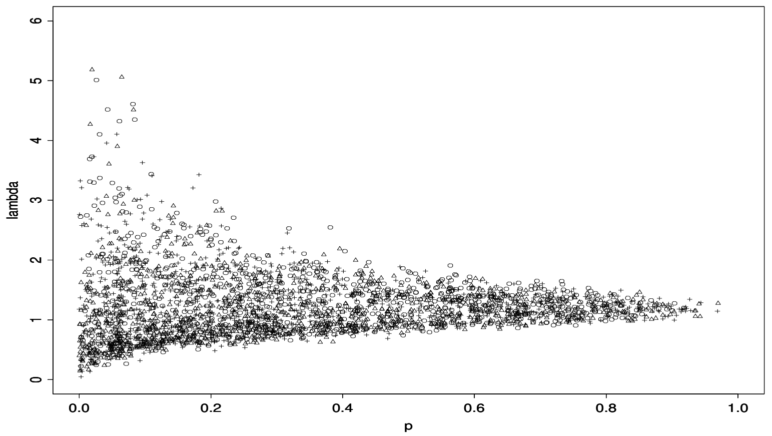
Example — Exponential mixture model

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

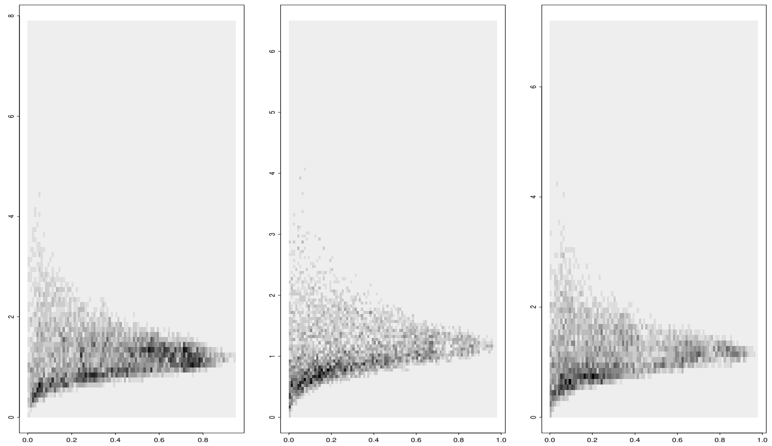
with $p_0 + \ldots + p_k = 1$ and $\lambda_i > 0$ $(0 \le i \le k)$.



Successive values of the Markov chains



Joint distribution



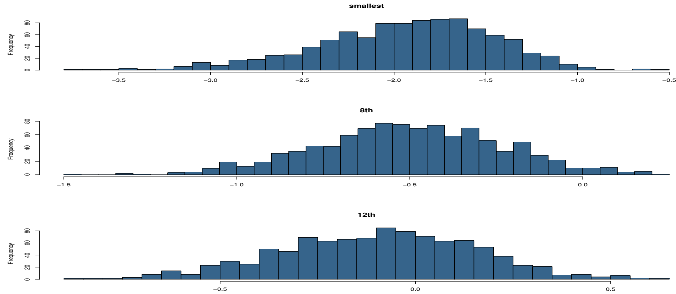
Marginal distributions

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...

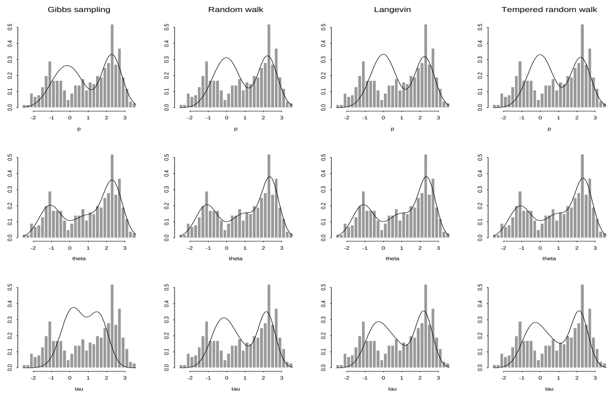
- modifies the prior distribution quite strongly for large #'s of components *[e.g. the marginal on θ_1 or θ_k may be quite different from $\mathcal{N}(0,1)$]*



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

[Diebolt & Robert, 1994]

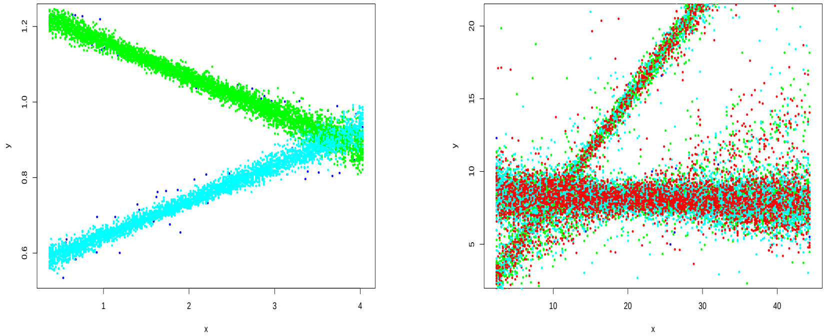
- induces biases *[may force components apart or increase the variance of the first component, for instance]*



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)]

3. Computational tools

Computational difficulties with the Bayesian approach:

- Computing the posterior distribution

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta)$$

- Resolution of

$$\arg\min \int_\Theta \mathrm{L}(\theta,\delta)\pi(\theta)f(x|\theta)d\theta,$$

- Maximisation of the marginal posterior

$$\arg\max \int_{\Theta_{-1}} \pi(\theta|x)d\theta_{-1}$$

- Computing posterior quantities

$$\delta^\pi(x) = \int_\Theta h(\theta)\ \pi(\theta|x)d\theta = \frac{\int_\Theta h(\theta)\ \pi(\theta)f(x|\theta)d\theta}{\int_\Theta \pi(\theta)f(x|\theta)d\theta}$$

- Resolution (in k) of

$$P(\pi(\theta|x) \geq k) = \alpha$$

Classical approximation methods

- Numerical integration (quadrature, gradient, ...)
  [Bumps into curse of dimen'ty]
- Analytical approximation (Laplace, ...)
  [Difficult assessment]
- Simulation/Monte Carlo (acceptation-reject, importance sampling, ...)
  [Nonstandard densities]

Markov Chain Monte Carlo methods

Idea Given a density distribution $\pi(\theta|x)$, produce a Markov chain $(\theta^{(t)})_t$ with stationary distribution π

- Posterior expectations

$$\mathbb{E}^\pi[h(\theta)|x] \simeq \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)})$$

- Posterior confidence regions

$$C_x^\pi \simeq [\theta^{(\alpha T/2)}, \theta^{(T-\alpha T/2)}]$$

- Posterior mode
- Normalizing constant
- Properness checkup

[Robert & Casella, 1999]

Solutions :

- **Metropolis algorithm**
  1. Find g such that
  $$\sup \pi \subset \sup g$$
  2. Generate $z_t|z_{t-1}$ as
  $$z_t = \begin{cases} y_t \sim g & \text{with probability} \quad \rho \\ z_{t-1} & \qquad\qquad\qquad 1-\rho \end{cases}$$
  with
  $$\rho = \frac{g(z_{t-1}|y_t)}{g(y_t|z_{t-1})}\,\frac{\pi(y_t)}{\pi(z_{t-1})} \wedge 1$$

- **Gibbs sampling (1)**
  1. Break z and π in $(z_1, \cdots, z_p) = z$ and $\pi_1(z_1|z_2, \cdots, z_p), \cdots, \pi_p(z_p|z_1, z_2, \cdots)$
  2. Generate $z^{(t)}|z^{(t-1)}$ as
  $$\begin{aligned} z_1^{(t)} &\sim \pi_1(z_1|z_2^{(t-1)}, \cdots, z_p^{(t-1)}) \\ z_2^{(t)} &\sim \pi_2(z_2|z_1^{(t)}, z_3^{(t-1)}, \cdots) \\ &\cdots \\ z_p^{(t)} &\sim \pi_p(z_p|z_1^{(t)}, z_2^{(t)}, \cdots) \end{aligned}$$

- **Gibbs sampling (2) : demarginalization**
  1. Complete π in g and z in $y = (z, \nu)$ such that
  $$\int g(y)d\nu = \pi(z)$$
  2. Apply Gibbs sampling (1) and use the sub-chain $z^{(t)}$

Mixtures of distributions

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

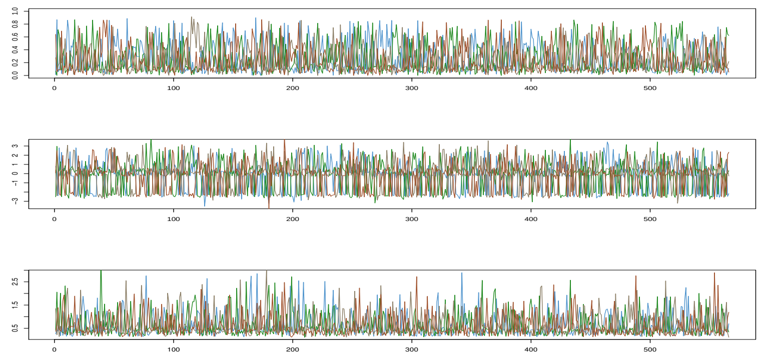
$$\prod_{i=1}^{n}\left(\sum_{j=1}^{k} p_j \frac{e^{-(x_i-\theta_j)^2/2\tau_j^2}}{\tau_j^2}\right)\prod_{j=1}^{k}\frac{e^{-\theta_j^2/2\xi\tau_j^2-\delta/\tau_j^2}}{\tau_j^3}$$

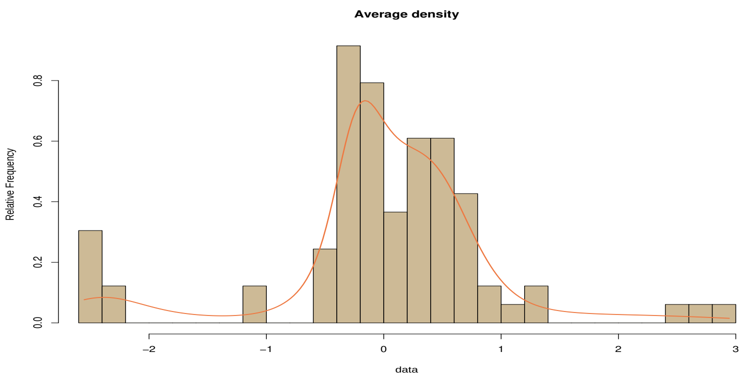
- Gibbs
  'Natural" completion by latent indicators

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

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

Example — Galaxy benchmark ($k = 4$)





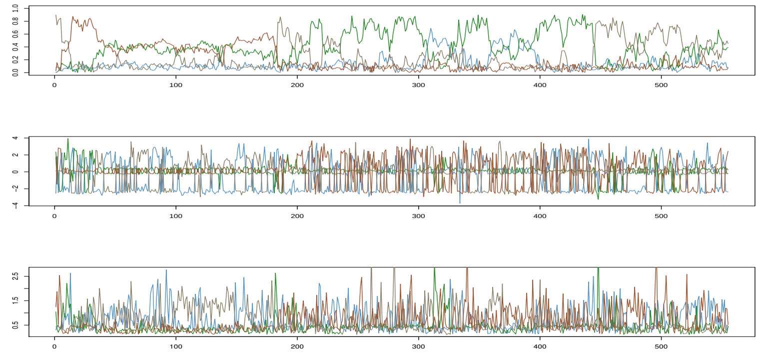
- Random walk Metropolis–Hastings
  - Avoids completion
  - Available (Normal vs. Cauchy vs... moves)
  - Calibrated against acceptance rate
  - Depends on parameterisation

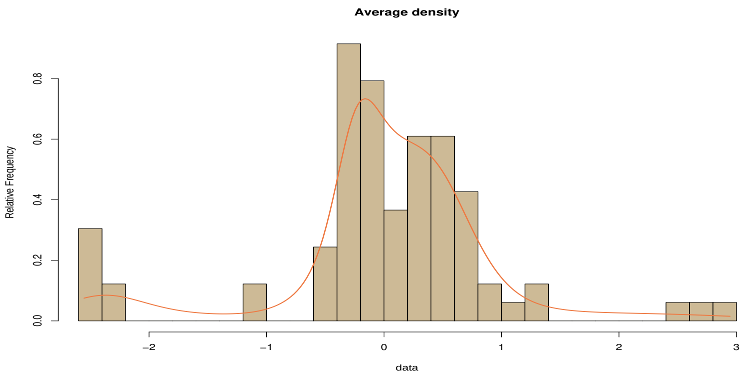
    $\lambda_j \longrightarrow \log \lambda_j \qquad\qquad p_j \longrightarrow \log(p_j/1 - p_k)$

    or

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

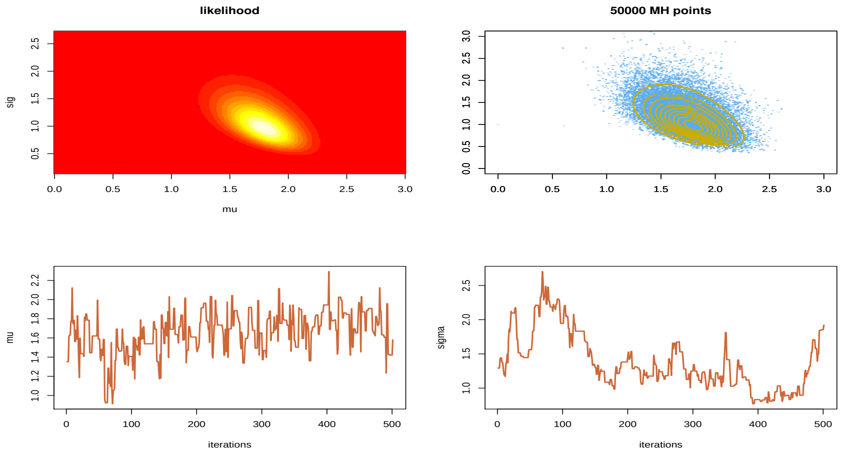
Example — Galaxy benchmark ($k = 4$)





Example — Simulated sample

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



5. Model choice

Choice of possible models/distributions

$$\mathcal{M}_i : x \sim f_i(x|\theta_i), \qquad i \in I$$

- model construction (nonparametrics)
- model checking (goodness of fit)
- model improvement (expansion)
- model prunning (contraction)
- model comparison

Bayesian solution:

- allocate probabilities p_i to each model $\mathcal{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i

Warning : parameters common to several models must be treated as separate entities!

Formally over:

1. Compute

$$p(\mathcal{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}$$

2. Take largest $p(\mathcal{M}_i|x)$ to determine model, or use

$$\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j$$

as predictive

Bayes factors

Models $\mathcal{M}_1$ vs. $\mathcal{M}_2$

$$B_{12} = \frac{Pr(\mathcal{M}_1|x)}{Pr(\mathcal{M}_2|x)} \Big/ \frac{Pr(\mathcal{M}_1)}{Pr(\mathcal{M}_2)} = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2}$$

[Good, 1958 & Jeffreys, 1961]

A battery of difficulties

1. Improper priors not allowed here
2. Vague proper priors not the solution (Jeffreys' paradox)
3. Dependence on priors only vanishes asymptotically
4. Computation of integrals

Recent solutions [methodology]

◇ AIBF $= B_{ji}^N \frac{1}{L} \sum_{\ell} B_{ij}(x_{[\ell]})$

◇ MIBF $= B_{ji}^N \; \text{Med}[B_{ij}(x_{[\ell]})]$

[Berger & Perichi, 1996]

◇ GIBF $= B_{ji}^N \; \exp \frac{1}{L} \sum_{\ell} \log B_{ij}(x_{[\ell]})$

◇ Fractional Bayes factor

$$B_{12}^F = B_{12}(x) \frac{\int L_2^b(\theta_2)\pi_2(\theta_2)d\theta_2}{\int L_1^b(\theta_1)\pi_1(\theta_1)d\theta_1}$$

[O'Hagan, 1995]

Recent solutions [computation]

- **Importance sampling:**

$$\int f(x|\theta)\pi(\theta)d\theta = \int \frac{f(x|\theta)}{g(\theta)} g(\theta)d\theta \approx \frac{1}{T}\sum_{t=1}^{T} f(x|\theta^{(t)}) \frac{\pi(\theta^{(t)})}{cg(\theta^{(t)})}$$

- **Defensive sampling:**

$$g(\theta) = \rho\pi(\theta) + (1-\rho)\pi(\theta|x) \qquad \rho \ll 1$$

[Newton & Raftery, 1994; Hestenberg, 1998]

- **Bridge sampling:**

If

$$\begin{aligned} \pi_1(\theta_1|x) &\propto \tilde\pi_1(\theta_1|x) \\ \pi_2(\theta_2|x) &\propto \tilde\pi_2(\theta_2|x) \end{aligned}$$

then

$$B_{12} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde\pi_1(\theta_i|x)}{\tilde\pi_2(\theta_i)} \qquad \theta_i \sim \pi_2$$

[Chen, Shao & Ibrahim, 2000]

Also

$$B_{12} = \frac{\int \tilde\pi_2(\theta)\alpha(\theta)\pi_1(\theta)d\theta}{\int \tilde\pi_1(\theta)\alpha(\theta)\pi_2(\theta)d\theta} \qquad \forall\ \alpha(\cdot)$$

$$\approx \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\tilde\pi_2(\theta_{1i})\alpha(\theta_{1i})}{\frac{1}{n_2}\sum_{i=1}^{n_2}\tilde\pi_1(\theta_{2i})\alpha(\theta_{2i})} \qquad \theta_{ji}\sim\pi_j(\theta)$$

- **Umbrella sampling:**

$$\begin{aligned}\pi_1(\theta) &= \pi(\theta|\lambda_1) & \pi_2(\theta) &= \pi_1(\theta|\lambda_2)\\ &= \tilde\pi_1(\theta)/c(\lambda_1) & &= \tilde\pi_2(\theta)/c(\lambda_2)\end{aligned}$$

Then

$$\forall\ \pi(\lambda), \qquad \log(c(\lambda_2)/c(\lambda_1)) = \mathbb{E}\left[\frac{\frac{d}{d\lambda}\log\tilde\pi(d\theta)}{\pi(\lambda)}\right]$$

and

$$\log(B_{12}) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\frac{d}{d\lambda}\log\tilde\pi(\theta_i|\lambda_i)}{\pi(\lambda_i)}$$

Variable dimension models

Models $\mathcal{M}_k : x \sim f(x|\theta_k)$ $\qquad \dim(\theta_k) = m_k$

$$k = 1, 2, \cdots, n, \cdots$$

Put a prior distribution on k, $\pi(k)$

Problems

- compute $Pr(\mathcal{M}_k|x) = \pi(k|x)$
- meaning of $\pi(\theta|x)$ *(inference)*
- meaning of $\pi(?|x)$ *(measure 'ory)*
- simulation from $\pi(?|x)$ *(MCMC)*

Mixtures of distributions

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**?!

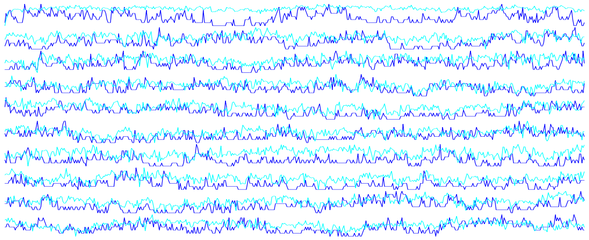
- weak identifiability of mixtures [contiguity of components]
- insolvable "philosophical" problem unless k has a proper intrinsic meaning [and even so...]
- hence testing *per se* is impossible: the data cannot distinguish between k and $k+h$ [unless guided by a firm hand/prior!]
- the choice of a prior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



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 $u_{1\to 2}$ and $u_{2\to 1}$ resp. so that
$$(\theta_1, u_{1\to 2}) \text{ and } (\theta_2, u_{2\to 1})$$
are in bijection (one-to-one correspondance):
$$(\theta_2, u_{2\to 1}) = T(\theta_1, u_{1\to 2})$$

- Use acceptance probability

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

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` $\quad$ *(inverse)*

Equivalent:

`(i) Split`

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

$$
\frac{\partial T(p,\lambda,u_1,u_2)}{\partial(p,\lambda,u_1,u_2)} = \begin{vmatrix} u_1 & (1-u_1) & 0 & 0 \\ p & -p & 0 & \lambda\frac{1-u_2}{(1-u_1)^2} \\ 0 & 0 & u_2 & \frac{1-u_1u_2}{1-u_1} \\ 0 & 0 & \lambda & -\frac{\lambda u_1}{1-u_1} \end{vmatrix}
$$

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